Chapter 6

INFERENCE IN THE MULTIPARAMETER CASE

6.1 INFERENCE FOR GAUSSIAN LINEAR MODELS

Most modern statistical questions involve large data sets, the modeling of whose stochastic structure involves complex models governed by several, often many, real parameters and frequently even more semi- or nonparametric models. In this final chapter of Volume I we develop the analogues of the asymptotic analyses of the behaviors of estimates, tests, and confidence regions in regular one-dimensional parametric models for d-dimensional models $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^d$. We have presented several such models already, for instance, the multinomial (Examples $1.6.7$, $2.3.3$), multiple regression models (Examples $1.1.4$, $1.4.3$, $2.1.1$) and more generally have studied the theory of multiparameter exponential families (Sections $1.6.2$, 2.2, 2.3). However, with the exception of Theorems $5.2.2$ and $5.3.5$, in which we looked at asymptotic theory for the MLE in multiparameter exponential families, we have not considered asymptotic inference, testing, confidence regions, and prediction in such situations. We begin our study with a thorough analysis of the Gaussian linear model with known variance in which exact calculations are possible. We shall show how the exact behavior of likelihood procedures in this model correspond to limiting behavior of such procedures in the unknown variance case and more generally in large samples from regular d-dimensional parametric models and shall illustrate our results with a number of important examples.

This chapter is a lead-in to the more advanced topics of Volume II in which we consider the construction and properties of procedures in non- and semiparametric models. The approaches and techniques developed here will be successfully extended in our discussions of the delta method for function-valued statistics, the properties of nonparametric MLEs, curve estimates, the bootstrap, and efficiency in semiparametric models. There is, however, an important aspect of practical situations that is not touched by the approximation, the fact that d, the number of parameters, and n, the number of observations, are often both large and commensurate or nearly so. The inequalities of Vapnik–Chervonenkis, Talagrand type and the modern empirical process theory needed to deal with such questions will also appear in the later chapters of Volume II.

Notational Convention: In this chapter we will, when there is no ambiguity, let expressions such as $\boldsymbol{\theta}$ refer to both column and row vectors.

6.1.1 The Classical Gaussian Linear Model

Many of the examples considered in the earlier chapters fit the framework in which the ith measurement Y_i among n independent observations has a distribution that depends on known constants $z_{i1}, \dots, z_{ip}$. In the classical Gaussian (normal) linear model this dependence takes the form

$$Y_i = \sum_{j=1}^{p} z_{ij}\beta_j + \epsilon_i,\ i = 1, \dots, n \tag{6.1.1}$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. In vector and matrix notation, we write

$$Y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, \dots, n \tag{6.1.2}$$

and

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{J}) \tag{6.1.3}$$

where $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})^T$, $\mathbf{Z} = (z_{ij})_{n \times p}$, and $\mathbf{J}$ is the $n \times n$ identity matrix.

Here Y_i is called the *response* variable, the z_{ij} are called the *design values*, and $\mathbf{Z}$ is called the *design matrix*.

I n this section we will derive exact statistical procedures under the assumptions of the model $(6.1.3)$. These are among the most commonly used statistical techniques. In Section 6.6 we will investigate the sensitivity of these procedures to the assumptions of the model. It turns out that these techniques are sensible and useful outside the narrow framework of model $(6.1.3)$. Here is Example $1.1.2(4)$ in this framework.

Example 6.1.1. *The One-Sample Location Problem.* We have n independent measurements $Y_1, \dots, Y_n$ from a population with mean $\beta_1 = E(Y)$. The model is

$$Y_i = \beta_1 + \epsilon_i,\ i = 1, \dots, n \tag{6.1.4}$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Here $p = 1$ and $\mathbf{Z}_{n \times 1} = (1, \dots, 1)^T$. □

The regression framework of Examples $1.1.4$ and $2.1.1$ is also of the form $(6.1.3)$:

Example 6.1.2. *Regression.* We consider experiments in which n cases are sampled from a population, and for each case, say the ith case, we have a response Y_i and a set of $p - 1$ covariate measurements denoted by $z_{i2}, \dots, z_{ip}$. We are interested in relating the mean of the response to the covariate values. The normal linear regression model is

$$Y_i = \beta_1 + \sum_{j=2}^{p} z_{ij}\beta_j + \epsilon_i,\ i = 1, \dots, n \tag{6.1.5}$$

where β_1 is called the regression intercept and $\beta_2, \dots, \beta_p$ are called the regression coefficients. If we set $z_{i1} = 1$, $i = 1, \dots, n$, then the notation $(6.1.2)$ and $(6.1.3)$ applies.

We treat the covariate values z_{ij} as fixed (nonrandom). In this case, $(6.1.5)$ is called the *fixed design* normal linear regression model. The random design Gaussian linear regression model is given in Example 1.4.3. We can think of the fixed design model as a conditional version of the random design model with the inference developed for the conditional distribution of Y given a set of observed covariate values. □

Example 6.1.3. *The p-Sample Problem or One-Way Layout.* In Example 1.1.3 and Section 4.9.3 we considered experiments involving the comparisons of two population means when we had available two independent samples, one from each population. Two-sample models apply when the design values represent a qualitative factor taking on only two values. Frequently, we are interested in qualitative factors taking on several values. If we are comparing pollution levels, we want to do so for a variety of locations; we often have more than two competing drugs to compare, and so on.

To fix ideas suppose we are interested in comparing the performance of $p \geq 2$ treatments on a population and that we administer only one treatment to each subject and a sample of n_k subjects get treatment k, $1 \leq k \leq p$, $n_1 + \cdots + n_p = n$. If the control and treatment responses are independent and normally distributed with the same variance σ^2, we arrive at the *one-way layout* or *p-sample* model,

$$Y_{kl} = \beta_k + \epsilon_{kl},\ 1 \leq l \leq n_k,\ 1 \leq k \leq p \tag{6.1.6}$$

where Y_{kl} is the response of the lth subject in the group obtaining the kth treatment, β_k is the mean response to the kth treatment, and the ϵ_{kl} are independent $\mathcal{N}(0, \sigma^2)$ random variables.

To see that this is a linear model we relabel the observations as $Y_1, \dots, Y_n$, where $Y_1, \dots, Y_{n_1}$ correspond to the group receiving the first treatment, $Y_{n_1+1}, \dots, Y_{n_1+n_2}$ to that getting the second, and so on. Then for $1 \leq j \leq p$, if $n_0 = 0$, the design matrix has elements:

$$\begin{aligned} z_{ij} &= 1 \text{ if } \sum_{k=1}^{j-1} n_k + 1 \leq i \leq \sum_{k=1}^{j} n_k \\ &= 0 \text{ otherwise} \end{aligned}$$

and

$$\mathbf{Z} = \begin{pmatrix} \mathbf{I}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_p \end{pmatrix}$$

where $\mathbf{I}_j$ is a column vector of n_j ones and the $\mathbf{0}$ in the "row" whose jth member is $\mathbf{I}_j$ is a column vector of n_j zeros. The model $(6.1.6)$ is an example of what is often called *analysis of variance models*. Generally, this terminology is commonly used when the design values are qualitative.

The model $(6.1.6)$ is often reparametrized by introducing $\alpha = p^{-1} \sum_{k=1}^{p} \beta_k$ and $\delta_k = \beta_k - \alpha$ because then δ_k represents the difference between the kth and average treatment effects, $k = 1, \dots, p$. In terms of the new parameter $\boldsymbol{\beta}^* = (\alpha, \delta_1, \dots, \delta_p)^T$, the linear model is

$$\mathbf{Y} = \mathbf{Z}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{J})$$

where $\mathbf{Z}^*_{n\times(p+1)} = (\mathbf{1}, \mathbf{Z})$ and $\mathbf{1}_{n\times 1}$ is the vector with n ones. Note that $\mathbf{Z}^*$ is of rank p and that $\boldsymbol{\beta}^*$ is not identifiable for $\boldsymbol{\beta}^* \in R^{p+1}$. However, $\boldsymbol{\beta}^*$ is identifiable in the p-dimensional linear subspace $\{\boldsymbol{\beta}^* \in R^{p+1} : \sum_{k=1}^p \delta_k = 0\}$ of R^{p+1} obtained by adding the linear restriction $\sum_{k=1}^p \delta_k = 0$ forced by the definition of the δ_k's. This type of linear model with the number of columns d of the design matrix larger than its rank r, and with the parameters identifiable only once $d - r$ additional linear restrictions have been specified, is common in analysis of variance models. □

Even if $\boldsymbol{\beta}$ is not a parameter (is unidentifiable), the vector of means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ of $\mathbf{Y}$ always is. It is given by

$$\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j \mathbf{c}_j$$

where the $\mathbf{c}_j$ are the columns of the design matrix,

$$\mathbf{c}_j = (z_{1j}, \dots, z_{nj})^T,\ j = 1, \dots, p.$$

The parameter set for $\boldsymbol{\beta}$ is R^p and the parameter set for $\boldsymbol{\mu}$ is

$$\omega = \{\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\beta};\ \boldsymbol{\beta} \in R^p\}.$$

Note that ω is the linear space spanned by the columns $\mathbf{c}_j$, $j = 1, \dots, n$, of the design matrix. Let r denote the number of linearly independent $\mathbf{c}_j$, $j = 1, \dots, p$, then r is the rank of $\mathbf{Z}$ and ω has dimension r. It follows that the parametrization $(\boldsymbol{\beta}, \sigma^2)$ is identifiable if and only if $r = p$ (Problem 6.1.17). We assume that $n \geq r$.

The Canonical Form of the Gaussian Linear Model

The linear model can be analyzed easily using some geometry. Because $\dim \omega = r$, there exists (e.g., by the Gram–Schmidt process) (see Section B.3.2), an orthonormal basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ for R^n such that $\mathbf{v}_1, \dots, \mathbf{v}_r$ span ω. Recall that orthonormal means $\mathbf{v}_i^T\mathbf{v}_j = 0$ for $i \neq j$ and $\mathbf{v}_i^T\mathbf{v}_i = 1$. When $\mathbf{v}_i^T\mathbf{v}_j = 0$, we call $\mathbf{v}_i$ and $\mathbf{v}_j$ *orthogonal.* Note that any $\mathbf{t} \in R^n$ can be written

$$\mathbf{t} = \sum_{i=1}^{n} (\mathbf{v}_i^T\mathbf{t})\mathbf{v}_i \tag{6.1.7}$$

and that

$$\mathbf{t} \in \omega \Leftrightarrow \mathbf{t} = \sum_{i=1}^{r} (\mathbf{v}_i^T\mathbf{t})\mathbf{v}_i \Leftrightarrow \mathbf{v}_i^T\mathbf{t} = 0,\ i = r+1, \dots, n.$$

We now introduce the *canonical variables* and *means*

$$U_i = \mathbf{v}_i^T\mathbf{Y},\ \eta_i = E(U_i) = \mathbf{v}_i^T\boldsymbol{\mu},\ i = 1, \dots, n.$$

Theorem 6.1.1. *The U_i are independent and $U_i \sim \mathcal{N}(\eta_i, \sigma^2)$, $i = 1, \dots, n$, where*

$$\eta_i = 0,\ i = r+1, \dots, n,$$

while $(\eta_1, \dots, \eta_r)^T$ varies freely over R^r.

Proof. Let $\mathbf{A}_{n\times n}$ be the orthogonal matrix with rows $\mathbf{v}_1^T, \dots, \mathbf{v}_n^T$. Then we can write $\mathbf{U} = \mathbf{AY}$, $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\mu}$, and by Theorem B.3.2, $U_1, \dots, U_n$ are independent normal with variance σ^2 and $E(U_i) = \mathbf{v}_i^T\boldsymbol{\mu} = 0$ for $i = r+1, \dots, n$ because $\boldsymbol{\mu} \in \omega$. □

Note that

$$\mathbf{Y} = \mathbf{A}^{-1}\mathbf{U}. \tag{6.1.8}$$

So, observing $\mathbf{U}$ and $\mathbf{Y}$ is the same thing. $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ are equivalently related,

$$\boldsymbol{\mu} = \mathbf{A}^{-1}\boldsymbol{\eta}. \tag{6.1.9}$$

whereas

$$\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{U}) = \sigma^2\mathbf{J}_{n\times n}. \tag{6.1.10}$$

It will be convenient to obtain our statistical procedures for the canonical variables $\mathbf{U}$, which are sufficient for $(\boldsymbol{\mu}, \sigma^2)$ using the parametrization $(\boldsymbol{\eta}, \sigma^2)^T$, and then translate them to procedures for $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, and σ^2 based on $\mathbf{Y}$ using (6.1.8)–(6.1.10). We start by considering the log likelihood $l_{\mathbf{u}}(\eta)$ based on $\mathbf{U}$

$$\begin{aligned} l_{\mathbf{u}}(\eta) &= -\frac{1}{2\sigma^2}\sum_{i=1}^n (u_i - \eta_i)^2 - \frac{n}{2}\log(2\pi\sigma^2) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^n u_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^r \eta_i u_i - \sum_{i=1}^r \frac{\eta_i^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2). \end{aligned} \tag{6.1.11}$$

6.1.2 Estimation

We first consider the σ^2 known case, which is the guide to asymptotic inference in general.

Theorem 6.1.2. *In the canonical form of the Gaussian linear model with σ^2 known*

(i) $\mathbf{T} = (U_1, \dots, U_r)^T$ *is sufficient for* $\boldsymbol{\eta}$.

(ii) $U_1, \dots, U_r$ *is the MLE of* $\eta_1, \dots, \eta_r$.

(iii) U_i *is the UMVU estimate of* η_i, $i = 1, \dots, r$.

(iv) *If* $c_1, \dots, c_r$ *are constants, then the MLE of* $\alpha = \sum_{i=1}^r c_i\eta_i$ *is* $\widehat{\alpha} = \sum_{i=1}^r c_iU_i$. $\widehat{\alpha}$ *is also UMVU for* α.

(v) *The MLE of* $\boldsymbol{\mu}$ *is* $\widehat{\boldsymbol{\mu}} = \sum_{i=1}^r \mathbf{v}_iU_i$ *and* $\widehat{\mu}_i$ *is UMVU for* μ_i, $i = 1, \dots, n$. *Moreover,* $U_i = \mathbf{v}_i^T\widehat{\boldsymbol{\mu}}$, *making* $\widehat{\boldsymbol{\mu}}$ *and* $\mathbf{U}$ *equivalent.*

Proof. (i) By observation, (6.1.11) is an exponential family with sufficient statistic $\mathbf{T}$.

(ii) $U_1, \dots, U_r$ are the MLEs of $\eta_1, \dots, \eta_r$ because, by observation, (6.1.11) is a function of $\eta_1, \dots, \eta_r$ only through $\sum_{i=1}^r (u_i - \eta_i)^2$ and is minimized by setting $\eta_i = u_i$. (We could also apply Theorem 2.3.1.)

(iii) By Theorem 3.4.4 and Example 3.4.6, U_i is UMVU for $E(U_i) = \eta_i$, $i = 1, \dots, r$.

(iv) By the invariance of the MLE (Section 2.2.2), the MLE of $q(\theta) = \sum_{i=1}^r c_i \eta_i$ is $q(\widehat{\theta}) = \sum_{i=1}^r c_i U_i$. If all the c's are zero, $\widehat{\alpha}$ is UMVU. Assume that at least one c is different from zero. By Problem 3.4.10, we can assume without loss of generality that $\sum_{i=1}^r c_i^2 = 1$. By Gram–Schmidt orthogonalization, there exists an orthonormal basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ of R^n with $\mathbf{v}_1 = \mathbf{c} = (c_1, \dots, c_r, 0, \dots, 0)^T \in R^n$. Let $W_i = \mathbf{v}_i^T \mathbf{U}$, $\xi_i = \mathbf{v}_i \boldsymbol{\eta}$, $i = 1, \dots, n$, then $\mathbf{W} \sim \mathcal{N}(\boldsymbol{\xi}, \sigma^2 \mathbf{J})$ by Theorem B.3.2, where $\mathbf{J}$ is the $n \times n$ identity matrix. The distribution of $\mathbf{W}$ is an exponential family, $W_1 = \widehat{\alpha}$ is sufficient for $\xi_1 = \alpha$, and is UMVU for its expectation $E(W_1) = \alpha$.

(v) Follows from (iv). □

Next we consider the case in which σ^2 is unknown and assume $n \geq r + 1$.

Theorem 6.1.3. *In the canonical Gaussian linear model with σ^2 unknown,*

(i) $\widetilde{\mathbf{T}} = \left(U_1, \dots, U_r, \sum_{i=r+1}^n U_i^2\right)^T$ *is sufficient for* $(\eta_1, \dots, \eta_r, \sigma^2)^T$.

(ii) *The MLE of* σ^2 *is* $n^{-1} \sum_{i=r+1}^n U_i^2$.

(iii) $s^2 \equiv (n-r)^{-1} \sum_{i=r+1}^n U_i^2$ *is an unbiased estimator of* σ^2.

(iv) *The conclusions of Theorem* 6.1.2 (ii),...,(v) *are still valid.*

Proof. By (6.1.11), $\left(U_1, \dots, U_r, \sum_{i=1}^n U_i^2\right)^T$ is sufficient. But because $\sum_{i=1}^n U_i^2 = \sum_{i=1}^r U_i^2 + \sum_{i=r+1}^n U_i^2$, this statistic is equivalent to $\widetilde{\mathbf{T}}$ and (i) follows. To show (ii), recall that the maximum of (6.1.11) has $\eta_i = U_i$, $i = 1, \dots, r$. That is, we need to maximize

$$-\frac{1}{2\sigma^2} \sum_{i=r+1}^n U_i^2 - \frac{n}{2}(\log 2\pi\sigma^2)$$

as a function of σ^2. The maximizer is easily seen to be $n^{-1} \sum_{i=r+1}^n U_i^2$ (Problem 6.1.1). (iii) is clear because $EU_i^2 = \sigma^2$, $i \geq r+1$. To show (iv), apply Theorem 3.4.3 and Example 3.4.6 to the canonical exponential family obtained from (6.1.11) by setting $T_j = U_j$, $\theta_j = \eta_j / \sigma^2$, $j = 1, \dots, r$, $T_{r+1} = \sum_{i=1}^n U_i^2$ and $\theta_{r+1} = -1/2\sigma^2$.

Projections

We next express $\widehat{\boldsymbol{\mu}}$ in terms of $\mathbf{Y}$, obtain the MLE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, and give a geometric interpretation of $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\beta}}$, and s^2. To this end, define the norm $|\mathbf{t}|$ of a vector $\mathbf{t} \in R_n$ by $|\mathbf{t}|^2 = \sum_{i=1}^n t_i^2$.

Definition 6.1.1. The *projection* $\mathbf{y}_0 = \pi(\mathbf{y} \mid \omega)$ of a point $\mathbf{y} \in R^n$ on ω is the point

$$\mathbf{y}_0 = \arg\min\{|\mathbf{y} - \mathbf{t}|^2 : \mathbf{t} \in \omega\}.$$

The maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ maximizes

$$\log p(\mathbf{y}, \boldsymbol{\beta}, \sigma) = -\frac{1}{2\sigma^2}|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}|^2 - \frac{n}{2}\log(2\pi\sigma^2)$$

or, equivalently,

$$\widehat{\boldsymbol{\beta}} = \arg\min\{|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}|^2 : \boldsymbol{\beta} \in R^p\}.$$

That is, the MLE of $\boldsymbol{\beta}$ equals the least squares estimate (LSE) of $\boldsymbol{\beta}$ defined in Example 2.1.1 and Section 2.2.1. We have

Theorem 6.1.4. *In the Gaussian linear model*

(i) $\widehat{\boldsymbol{\mu}}$ *is the unique projection of* $\mathbf{Y}$ *on* ω *and is given by*

$$\widehat{\boldsymbol{\mu}} = \mathbf{Z}\widehat{\boldsymbol{\beta}}. \tag{6.1.12}$$

(ii) $\widehat{\boldsymbol{\mu}}$ *is orthogonal to* $\mathbf{Y} - \widehat{\boldsymbol{\mu}}$.

(iii)

$$s^2 = |\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2/(n-r) \tag{6.1.13}$$

(iv) *If* $p = r$, *then* $\boldsymbol{\beta}$ *is identifiable,* $\boldsymbol{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\mu}$, *the MLE = LSE of* $\boldsymbol{\beta}$ *is unique and given by*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\widehat{\boldsymbol{\mu}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}. \tag{6.1.14}$$

(v) If $p = r$, then $\widehat{\beta}_j$ *is the UMVU estimate of* β_j, $j = 1, \dots, p$, *and* $\widehat{\mu}_i$ *is the UMVU estimate of* μ_i, $i = 1, \dots, n$.

Proof. (i) is clear because $\mathbf{Z}\boldsymbol{\beta}$, $\boldsymbol{\beta} \in R^p$, spans ω. (ii) and (iii) are also clear from Theorem 6.1.3 because $\widehat{\boldsymbol{\mu}} = \sum_{i=1}^r \mathbf{v}_i U_i$ and $\mathbf{Y} - \widehat{\boldsymbol{\mu}} = \sum_{j=r+1}^n \mathbf{v}_j U_j$. To show (iv), note that $\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\beta}$ and (6.1.12) implies $\mathbf{Z}^T\boldsymbol{\mu} = \mathbf{Z}^T\mathbf{Z}\boldsymbol{\beta}$ and $\mathbf{Z}^T\widehat{\boldsymbol{\mu}} = \mathbf{Z}^T\mathbf{Z}\widehat{\boldsymbol{\beta}}$ and, because $\mathbf{Z}$ has full rank, $\mathbf{Z}^T\mathbf{Z}$ is nonsingular, and $\boldsymbol{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mu$, $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\widehat{\boldsymbol{\mu}}$. To show $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$, note that the space $\omega^\perp$ of vectors $\mathbf{s}$ orthogonal to ω can be written as

$$\omega^\perp = \{\mathbf{s} \in R^n : \mathbf{s}^T(\mathbf{Z}\boldsymbol{\beta}) = 0 \text{ for all } \boldsymbol{\beta} \in R^p\}.$$

It follows that $\boldsymbol{\beta}^T(\mathbf{Z}^T\mathbf{s}) = 0$ for all $\boldsymbol{\beta} \in R^p$, which implies $\mathbf{Z}^T\mathbf{s} = 0$ for all $\mathbf{s} \in \omega^\perp$. Thus, $\mathbf{Z}^T(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) = 0$ and the second equality in (6.1.14) follows.

$\widehat{\beta}_j$ and $\widehat{\mu}_i$ are UMVU because, by (6.1.9), any linear combination of Y's is also a linear combination of U's, and by Theorems 6.1.2(iv) and 6.1.3(iv), any linear combination of U's is a UMVU estimate of its expectation. □

Note that in Example 2.1.1 we give an alternative derivation of $(6.1.14)$ and the normal equations $(\mathbf{Z}^T\mathbf{Z})\boldsymbol{\beta} = \mathbf{Z}^T\mathbf{Y}$.

The estimate $\widehat{\boldsymbol{\mu}} = \mathbf{Z}\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\mu}$ is called the *fitted value* and $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\boldsymbol{\mu}}$ is called the *residual* from this fit. The goodness of the fit is measured by the *residual sum of squares* (RSS) $|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2 = \sum_{i=1}^n \widehat{\epsilon}_i^2$. Example 2.2.2 illustrates this terminology in the context of Example 6.1.2 with $p = 2$. There the points $\widehat{\mu}_i = \widehat{\beta}_1 + \widehat{\beta}_2 z_i$, $i = 1, \dots, n$ lie on the regression line fitted to the data $\{(z_i, y_i),\ i = 1, \dots, n\}$; moreover, the residuals $\widehat{\epsilon}_i = [y_i - (\widehat{\beta}_1 + \widehat{\beta}_2 z_i)]$ are the vertical distances from the points to the fitted line.

Suppose we are given a value of the covariate $\mathbf{z}$ at which a value Y following the linear model $(6.1.3)$ is to be taken. By Theorem 1.4.1, the best MSPE predictor of Y if $\boldsymbol{\beta}$ is known as well as $\mathbf{z}$ is $E(Y) = \mathbf{z}^T\boldsymbol{\beta}$ and its best (UMVU) estimate not knowing $\boldsymbol{\beta}$ is $\widehat{Y} \equiv \mathbf{z}^T\widehat{\boldsymbol{\beta}}$. Taking $\mathbf{z} = \mathbf{z}_i$, $1 \le i \le n$, we obtain $\widehat{\mu}_i = \mathbf{z}_i^T\widehat{\boldsymbol{\beta}}$, $1 \le i \le n$. In this method of "prediction" of Y_i, it is common to write $\widehat{Y}_i$ for $\widehat{\mu}_i$, the ith component of the fitted value $\widehat{\boldsymbol{\mu}}$. That is, $\widehat{\mathbf{Y}} = \widehat{\boldsymbol{\mu}}$. Note that by $(6.1.12)$ and $(6.1.14)$, when $p = r$,

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

where

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T.$$

The matrix $\mathbf{H}$ is the *projection matrix* mapping R^n into ω, see also Section B.10. In statistics it is also called the *hat matrix* because it "puts the hat on $\mathbf{Y}$." As a projection matrix $\mathbf{H}$ is necessarily symmetric and idempotent,

$$\mathbf{H}^T = \mathbf{H},\ \mathbf{H}^2 = \mathbf{H}.$$

It follows from this and (B.5.3) that if $\mathbf{J} = \mathbf{J}_{n\times n}$ is the identity matrix, then

$$\text{Var}(\widehat{\mathbf{Y}}) = \mathbf{H}(\sigma^2\mathbf{J})\mathbf{H}^T = \sigma^2\mathbf{H}. \tag{6.1.15}$$

Next note that the residuals can be written as

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{J} - \mathbf{H})\mathbf{Y}.$$

The residuals are the projection of $\mathbf{Y}$ on the orthocomplement of ω and

$$\text{Var}(\widehat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{J} - \mathbf{H}). \tag{6.1.16}$$

We can now conclude the following.

Corollary 6.1.1. *In the Gaussian linear model*

(i) *the fitted values* $\widehat{\mathbf{Y}} = \widehat{\boldsymbol{\mu}}$ *and the residual* $\widehat{\boldsymbol{\epsilon}}$ *are independent,*

(ii) $\widehat{\mathbf{Y}} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{H})$,

(iii) $\widehat{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{J} - \mathbf{H}))$, *and*

(iv) *if* $p = r$, $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1})$.

Proof. $(\widehat{\mathbf{Y}}, \widehat{\boldsymbol{\epsilon}})$ is a linear transformation of $\mathbf{U}$ and, hence, joint Gaussian. The independence follows from the identification of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\epsilon}}$ in terms of the U_i in the theorem. $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$ follows from (B.5.3). □

We now return to our examples.

Example 6.1.1. *One Sample (continued).* Here $\mu = \beta_1$ and $\widehat{\mu} = \widehat{\beta}_1 = \bar{Y}$. Moreover, the unbiased estimator s^2 of σ^2 is $\sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$, which we have seen before in Problem 1.3.8 and (3.4.2). □

Example 6.1.2. *Regression (continued).* If the design matrix $\mathbf{Z}$ has rank p, then the MLE = LSE estimate is $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$ as seen before in Example 2.1.1 and Section 2.2.1. We now see that the MLE of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \mathbf{Z}\widehat{\boldsymbol{\beta}}$ and that $\widehat{\beta}_j$ and $\widehat{\mu}_i$ are UMVU for β_j and μ_i respectively, $j = 1, \dots, p$, $i = 1, \dots, n$. The variances of $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\mu}} = \widehat{\mathbf{Y}}$, and $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}}$ are given in Corollary 6.1.1. In the Gaussian case $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{Y}}$, and $\widehat{\boldsymbol{\epsilon}}$ are normally distributed with $\widehat{\mathbf{Y}}$ and $\widehat{\boldsymbol{\epsilon}}$ independent. The error variance $\sigma^2 = \text{Var}(\epsilon_1)$ can be unbiasedly estimated by $s^2 = (n-p)^{-1}|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2$. □

Example 6.1.3. *The One-Way Layout (continued).* In this example the normal equations $(\mathbf{Z}^T\mathbf{Z})\boldsymbol{\beta} = \mathbf{Z}\mathbf{Y}$ become

$$n_k\beta_k = \sum_{l=1}^{n_k} Y_{kl}, \ k = 1, \dots, p.$$

At this point we introduce an important notational convention in statistics. If $\{c_{ijk} \dots\}$ is a multiple-indexed sequence of numbers or variables, then replacement of a subscript by a dot indicates that we are considering the average over that subscript. Thus,

$$Y_{k\cdot} = \frac{1}{n_k}\sum_{l=1}^{n_k} Y_{kl}, \ Y_{\cdot\cdot} = \frac{1}{n}\sum_{k=1}^{p}\sum_{l=1}^{n_k} Y_{kl}$$

where $n = n_1 + \cdots + n_p$ and we can write the least squares estimates as

$$\widehat{\beta}_k = Y_{k\cdot}, \ k = 1, \dots, p.$$

By Theorem 6.1.3, in the Gaussian model, the UMVU estimate of the average effect of all the treatments, $\alpha = \beta_\cdot$, is

$$\widehat{\alpha} = \frac{1}{p}\sum_{k=1}^{p} Y_{k\cdot} \ (\text{not } Y_{\cdot\cdot} \text{ in general})$$

and the UMVU estimate of the incremental effect $\delta_k = \beta_k - \alpha$ of the kth treatment is

$$\widehat{\delta}_k = Y_{k\cdot} - \widehat{\alpha}, \ k = 1, \dots, p.$$

□

Remark 6.1.1. An alternative approach to the MLEs for the normal model and the associated LSEs of this section is an approach based on MLEs for the model in which the errors $\epsilon_1, \dots, \epsilon_n$ in $(6.1.1)$ have the Laplace distribution with density

$$\frac{1}{2\sigma} \exp\left\{-\frac{1}{\sigma}|t|\right\}$$

and the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are least absolute deviation estimates (LADEs) obtained by minimizing the absolute deviation distance $\sum_{i=1}^n |y_i - \mathbf{z}_i^T \boldsymbol{\beta}|$. The LADEs were introduced by Laplace before Gauss and Legendre introduced the LSEs—see Stigler (1986). The LSEs are preferred because of ease of computation and their geometric properties. However, the LADEs are obtained fairly quickly by modern computing methods; see Koenker and D'Orey (1987) and Portnoy and Koenker (1997). For more on LADEs, see Problems 1.4.7 and 2.2.31.

6.1.3 Tests and Confidence Intervals

The most important hypothesis-testing questions in the context of a linear model correspond to restriction of the vector of means $\boldsymbol{\mu}$ to a linear subspace of the space ω, which together with σ^2 specifies the model. For instance, in a study to investigate whether a drug affects the mean of a response such as blood pressure we may consider, in the context of Example 6.1.2, a regression equation of the form

$$\text{mean response} = \beta_1 + \beta_2 z_{i2} + \beta_3 z_{i3}, \qquad (6.1.17)$$

where z_{i2} is the dose level of the drug given the ith patient, z_{i3} is the age of the ith patient, and the matrix $\|z_{ij}\|_{n\times 3}$ with $z_{i1} = 1$ has rank 3. Now we would test $H : \beta_2 = 0$ versus $K : \beta_2 \neq 0$. Thus, under H, $\{\boldsymbol{\mu} : \mu_i = \beta_1 + \beta_3 z_{i3},\ i = 1, \dots, n\}$ is a two-dimensional linear subspace of the full model's three-dimensional linear subspace of R^n given by $(6.1.17)$.

Next consider the p-sample model of Example 1.6.3 with β_k representing the mean response for the kth population. The first inferential question is typically "Are the means equal or not?" Thus we test $H : \beta_1 = \dots = \beta_p = \beta$ for some $\beta \in R$ versus K: "the β's are not all equal." Now, under H, the mean vector is an element of the space $\{\boldsymbol{\mu} : \mu_i = \beta \in R,\ i = 1, \dots, p\}$, which is a one-dimensional subspace of R^n, whereas for the full model $\boldsymbol{\mu}$ is in a p-dimensional subspace of R^n.

In general, we let ω correspond to the full model with dimension r and let ω_0 be a q-dimensional linear subspace over which $\boldsymbol{\mu}$ can range under the null hypothesis H; $0 \leq q < r$.

We first consider the σ^2 known case and consider the likelihood ratio statistic

$$\lambda(\mathbf{y}) = \frac{\sup\{p(\mathbf{y}, \boldsymbol{\mu}) : \boldsymbol{\mu} \in \omega\}}{\sup\{p(\mathbf{y}, \boldsymbol{\mu}) : \boldsymbol{\mu} \in \omega_0\}}$$

for testing $H : \boldsymbol{\mu} \in \omega_0$ versus $K : \boldsymbol{\mu} \in \omega - \omega_0$. Because

$$p(\mathbf{Y}, \boldsymbol{\mu}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}|\mathbf{Y} - \boldsymbol{\mu}|^2\right\} \tag{6.1.18}$$

then, by Theorem 6.1.4,

$$\lambda(\mathbf{Y}) = \exp\left\{-\frac{1}{2\sigma^2}\left(|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2 - |\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0|^2\right)\right\}$$

where $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\mu}}_0$ are the projections of $\mathbf{Y}$ on ω and ω_0, respectively.

But if we let $\mathbf{A}_{n\times n}$ be an orthogonal matrix with rows $\mathbf{v}_1^T, \dots, \mathbf{v}_n^T$ such that $\mathbf{v}_1, \dots, \mathbf{v}_q$ span ω_0 and $\mathbf{v}_1, \dots, \mathbf{v}_r$ span ω and set

$$\mathbf{U} = \mathbf{A}\mathbf{Y},\ \boldsymbol{\eta} = \mathbf{A}\boldsymbol{\mu} \tag{6.1.19}$$

then, by Theorem 6.1.2(v),

$$\lambda(\mathbf{Y}) = \exp\left\{\frac{1}{2\sigma^2}\sum_{i=q+1}^{r} U_i^2\right\} = \exp\left\{\frac{1}{2\sigma^2}|\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0|^2\right\}. \tag{6.1.20}$$

It follows that

$$2\log\lambda(\mathbf{Y}) = \sum_{i=q+1}^{r}(U_i/\sigma)^2.$$

Note that (U_i/σ) has a $\mathcal{N}(\theta_i, 1)$ distribution with $\theta_i = \eta_i/\sigma$. In this case the distribution of $\sum_{i=q+1}^{r}(U_i/\sigma)^2$ is called a *chi-square distribution with $r - q$ degrees of freedom* and *noncentrality parameter* $\theta^2 = |\boldsymbol{\theta}|^2 = \sum_{i=q+1}^{r}\theta_i^2$, where $\boldsymbol{\theta} = (\theta_{q+1}, \dots, \theta_r)^T$ (see Problem B.3.12). We write $\chi^2_{r-q}(\theta^2)$ for this distribution. We have shown the following.

Proposition 6.1.1. *In the Gaussian linear model with σ^2 known, $2\log\lambda(\mathbf{Y})$ has a $\chi^2_{r-q}(\theta^2)$ distribution with*

$$\theta^2 = \sigma^{-2}\sum_{i=q+1}^{r}\eta_i^2 = \sigma^{-2}|\boldsymbol{\mu} - \boldsymbol{\mu}_0|^2 \tag{6.1.21}$$

where $\boldsymbol{\mu}_0$ is the projection of $\boldsymbol{\mu}$ on ω_0. In particular, when H holds, $2\log\lambda(\mathbf{Y}) \sim \chi^2_{r-q}$.

Proof. We only need to establish the second equality in $(6.1.21)$. Write $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\mu}$ where $\mathbf{A}$ is as defined in $(6.1.19)$, then

$$\sum_{i=q+1}^{r}\eta_i^2 = |\boldsymbol{\mu} - \boldsymbol{\mu}_0|^2.$$

□

Next consider the case in which σ^2 is unknown. We know from Problem 6.1.1 that the MLEs of σ^2 for $\boldsymbol{\mu} \in \omega$ and $\boldsymbol{\mu} \in \omega_0$ are

$$\widehat{\sigma}^2 = \frac{1}{n}|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2 \text{ and } \widehat{\sigma}_0^2 = \frac{1}{n}|\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0|^2,$$

respectively. Substituting $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\sigma}^2$, and $\widehat{\sigma}_0^2$ into the likelihood ratio statistic, we obtain

$$\lambda(\mathbf{y}) = \frac{p(\mathbf{y}, \widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2)}{p(\mathbf{y}, \widehat{\boldsymbol{\mu}}_0, \widehat{\sigma}_0^2)} = \left\{ \frac{|\mathbf{y} - \widehat{\boldsymbol{\mu}}_0|^2}{|\mathbf{y} - \widehat{\boldsymbol{\mu}}|^2} \right\}^{\frac{n}{2}}$$

where $p(\mathbf{y}, \boldsymbol{\mu}, \sigma^2)$ denotes the right-hand side of (6.1.18).

The resulting test is intuitive. It consists of rejecting H when the fit, as measured by the residual sum of squares under the model specified by H, is poor compared to the fit under the general model. For the purpose of finding critical values it is more convenient to work with a statistic equivalent to $\lambda(\mathbf{Y})$,

$$T = \frac{n-r}{r-q} \frac{|\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0|^2 - |\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2}{|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2} = \frac{(r-q)^{-1}|\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0|^2}{(n-r)^{-1}|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2}. \tag{6.1.22}$$

Because $T = (n-r)(r-q)^{-1}\{[\lambda(\mathbf{Y})]^{2/n} - 1\}$, T is an increasing function of $\lambda(\mathbf{Y})$ and the two test statistics are equivalent. T is called *the F statistic* for the general linear hypothesis.

We have seen in Proposition 6.1.1 that $\sigma^{-2}|\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0|^2$ have a $\chi^2_{r-q}(\theta^2)$ distribution with $\theta^2 = \sigma^{-2}|\boldsymbol{\mu} - \boldsymbol{\mu}_0|^2$. By the canonical representation (6.1.19), we can write $\sigma^{-2}|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2 = \sum_{i=r+1}^{n}(U_i/\sigma)^2$, which has a χ^2_{n-r} distribution and is independent of $\sigma^{-2}|\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0|^2 = \sum_{i=q+1}^{r}(U_i/\sigma)^2$. Thus, T has the representation

$$T = \frac{(\text{noncentral } \chi^2_{r-q} \text{ variable})/df}{(\text{central } \chi^2_{n-r} \text{ variable})/df}$$

with the numerator and denominator independent. The distribution of such a variable is called the *noncentral $\mathcal{F}$ distribution with noncentrality parameter θ^2 and $r-q$ and $n-r$ degrees of freedom* (see Problem B.3.14). We write $\mathcal{F}_{k,m}(\theta^2)$ for this distribution where $k = r - q$ and $m = n - r$. We have shown the following.

Proposition 6.1.2. *In the Gaussian linear model the F statistic defined by* (6.1.22)*, which is equivalent to the likelihood ratio statistic for $H : \boldsymbol{\mu} \in \omega_0$ for $K : \boldsymbol{\mu} \in \omega - \omega_0$, has the noncentral F distribution $\mathcal{F}_{r-q,n-r}(\theta^2)$ where $\theta^2 = \sigma^{-2}|\boldsymbol{\mu} - \boldsymbol{\mu}_0|^2$. In particular, when H holds, T has the (central) $\mathcal{F}_{r-q,n-r}$ distribution.*

Remark 6.1.2. In Proposition 6.1.1 suppose the assumption "σ^2 is known" is replaced by "σ^2 is the same under H and K and estimated by the MLE $\widehat{\sigma}^2$ for $\boldsymbol{\mu} \in \omega$." In this case, it can be shown (Problem 6.1.5) that if we introduce the *variance equal likelihood ratio statistic*

$$\widetilde{\lambda}(\mathbf{y}) = \frac{\max\{p(\mathbf{y}, \boldsymbol{\mu}, \widehat{\sigma}^2) : \boldsymbol{\mu} \in \omega\}}{\max\{p(\mathbf{y}, \boldsymbol{\mu}, \widehat{\sigma}^2) : \boldsymbol{\mu} \in \omega_0\}} \tag{6.1.23}$$

then $\widetilde{\lambda}(\mathbf{Y})$ equals the likelihood ratio statistic for the σ^2 known case with σ^2 replaced by $\widehat{\sigma}^2$. It follows that

$$2\log\widetilde{\lambda}(\mathbf{Y}) = \frac{r-q}{(n-r)/n}T = \frac{\text{noncentral } \chi^2_{r-q}}{\text{central } \chi^2_{n-r}/n} \tag{6.1.24}$$

where T is the F statistic $(6.1.22)$.

Remark 6.1.3. The canonical representation $(6.1.19)$ made it possible to recognize the identity

$$|\mathbf{Y}-\widehat{\boldsymbol{\mu}}_0|^2 = |\mathbf{Y}-\widehat{\boldsymbol{\mu}}|^2 + |\widehat{\boldsymbol{\mu}}-\widehat{\boldsymbol{\mu}}_0|^2, \tag{6.1.25}$$

which we exploited in the preceding derivations. This is the *Pythagorean identity*. See Figure $6.1.1$ and Section B.10.

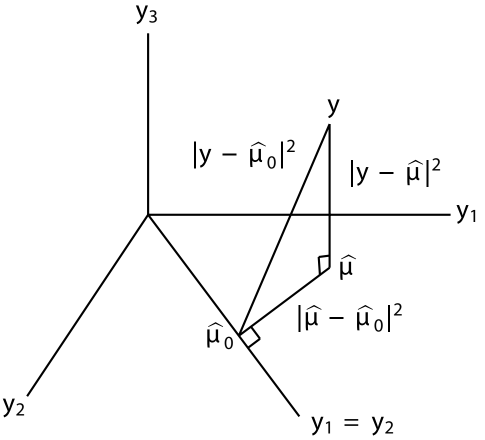

Figure 6.1.1. The projections $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\mu}}_0$ of $\mathbf{Y}$ on ω and ω_0; and the Pythagorean identity.

We next return to our examples.

Example 6.1.1. *One Sample (continued).* We test $H : \beta_1 = \mu_0$ versus $K : \beta \neq \mu_0$. In this case $\omega_0 = \{\mu_0\}$, $q = 0$, $r = 1$ and

$$T = \frac{(\bar{Y}-\mu_0)^2}{(n-1)^{-1}\Sigma(Y_i-\bar{Y})^2},$$

which we recognize as t^2/n, where t is the one-sample Student t statistic of Section $4.9.2$. □

Example 6.1.2. *Regression (continued).* We consider the possibility that a subset of $p-q$ covariates does not affect the mean response. Without loss of generality we ask whether the last $p-q$ covariates in multiple regression have an effect after fitting the first q. To formulate this question, we partition the design matrix $\mathbf{Z}$ by writing it as $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ where $\mathbf{Z}_1$ is $n \times q$ and $\mathbf{Z}_2$ is $n \times (p-q)$, and we partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$ where $\boldsymbol{\beta}_2$ is a $(p-q) \times 1$ vector of main (e.g., treatment) effect coefficients and $\boldsymbol{\beta}_1$ is a $q \times 1$ vector of "nuisance" (e.g., age, economic status) coefficients. Now the linear model can be written as

$$\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \mathbf{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}. \tag{6.1.26}$$

We test $H : \boldsymbol{\beta}_2 = \mathbf{0}$ versus $K : \boldsymbol{\beta}_2 \neq \mathbf{0}$. In this case $\widehat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$ and $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\mathbf{Z}_1^T\mathbf{Y}$ are the MLEs under the full model (6.1.26) and H, respectively. Using (6.1.22) we can write the F statistic version of the likelihood ratio test in the intuitive form

$$F = \frac{(RSS_H - RSS_F)/(df_H - df_F)}{RSS_F/df_F}$$

where $RSS_F = |\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2$ and $RSS_H = |\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0|^2$ are the residual sums of squares under the full model and H, respectively; and $df_F = n-p$ and $df_H = n-q$ are the corresponding degrees of freedom. The F test rejects H if F is large when compared to the αth quantile of the $\mathcal{F}_{p-q,n-p}$ distribution.

Under the alternative F has a noncentral $\mathcal{F}_{p-q,n-p}(\theta^2)$ distribution with noncentrality parameter (Problem 6.1.7)

$$\theta^2 = \sigma^{-2}\boldsymbol{\beta}_2^T\{\mathbf{Z}_2^T\mathbf{Z}_2 - \mathbf{Z}_2^T\mathbf{Z}_1(\mathbf{Z}_1^T\mathbf{Z}_1)^{-1}\mathbf{Z}_1^T\mathbf{Z}_2\}\boldsymbol{\beta}_2. \tag{6.1.27}$$

In the special case that $\mathbf{Z}_1^T\mathbf{Z}_2 = \mathbf{0}$ so the variables in $\mathbf{Z}_1$ are orthogonal to the variables in $\mathbf{Z}_2$, θ^2 simplifies to $\sigma^{-2}\boldsymbol{\beta}_2^T(\mathbf{Z}_2^T\mathbf{Z}_2)\boldsymbol{\beta}_2$, which only depends on the second set of variables and coefficients. However, in general θ^2 depends on the sample correlations between the variables in $\mathbf{Z}_1$ and those in $\mathbf{Z}_2$. This issue is discussed further in Example 6.2.1. □

Example 6.1.3. *The One-Way Layout (continued).* Recall that the least squares estimates of $\beta_1, \dots, \beta_p$ are $Y_{1\cdot}, \dots, Y_{p\cdot}$. As we indicated earlier, we want to test $H : \beta_1 = \cdots = \beta_p$. Under H all the observations have the same mean so that,

$$\widehat{\boldsymbol{\mu}}_0 = (Y_{\cdot\cdot}, \dots, Y_{\cdot\cdot})^T.$$

Thus,

$$|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0|^2 = \sum_{k=1}^{p}\sum_{l=1}^{n_k}(Y_{k\cdot} - Y_{\cdot\cdot})^2 = \sum_{k=1}^{p} n_k(Y_{k\cdot} - Y_{\cdot\cdot})^2.$$

Substituting in (6.1.22) we obtain the F statistic for the hypothesis H in the one-way layout

$$T = \frac{n-p}{p-1}\frac{\sum_{k=1}^{p} n_k(Y_{k\cdot} - Y_{\cdot\cdot})^2}{\sum_{k=1}^{p}\sum_{l=1}^{n_k}(Y_{kl} - Y_{k\cdot})^2}.$$

When H holds, T has a $\mathcal{F}_{p-1,n-p}$ distribution. If the β_i are not all equal, T has a noncentral $\mathcal{F}_{p-1,n-p}$ distribution with noncentrality parameter

$$\delta^2 = \frac{1}{\sigma^2}\sum_{k=1}^{p} n_k(\beta_k - \bar{\beta})^2, \tag{6.1.28}$$

where $\bar{\beta} = n^{-1}\sum_{i=1}^{p} n_i\beta_i$. To derive δ^2, compute $\sigma^{-2}|\boldsymbol{\mu} - \boldsymbol{\mu}_0|^2$ for the vector $\boldsymbol{\mu} = (\beta_1, \dots, \beta_1, \beta_2, \dots, \beta_2, \dots, \beta_p, \dots, \beta_p)^T$ and its projection $\boldsymbol{\mu}_0 = (\bar{\beta}, \dots, \bar{\beta})^T$.

There is an interesting way of looking at the pieces of information summarized by the F statistic. The sum of squares in the numerator,

$$SS_B = \sum_{k=1}^{p} n_k(Y_{k\cdot} - Y_{\cdot\cdot})^2$$

is a measure of variation *between* the p samples $Y_{11}, \dots, Y_{1n_1}, \dots, Y_{p1}, \dots, Y_{pn_p}$. The sum of squares in the denominator,

$$SS_W = \sum_{k=1}^{p}\sum_{l=1}^{n_k}(Y_{kl} - Y_{k\cdot})^2,$$

measures variation *within* the samples. If we define the *total* sum of squares as

$$SS_T = \sum_{k=1}^{p}\sum_{l=1}^{n_k}(Y_{kl} - Y_{\cdot\cdot})^2,$$

which measures the variability of the pooled samples, then by the Pythagorean identity (6.1.25)

$$SS_T = SS_B + SS_W. \tag{6.1.29}$$

Thus, we have a decomposition of the variability of the whole set of data, SS_T, the *total* sum of squares, into two constituent components, SS_B, the *between groups* (or treatment) sum of squares and SS_W, the *within groups* (or *residual*) sum of squares. SS_T/σ^2 is a (noncentral) χ^2 variable with $(n-1)$ degrees of freedom and noncentrality parameter δ^2. Because SS_B/σ^2 and SS_W/σ^2 are independent χ^2 variables with $(p-1)$ and $(n-p)$ degrees of freedom, respectively, we see that the decomposition $(6.1.29)$ can also be viewed stochastically, identifying δ^2 and $(p-1)$ degrees of freedom as "coming" from SS_B/σ^2 and the remaining $(n-p)$ of the $(n-1)$ degrees of freedom of SS_T/σ^2 as "coming" from SS_W/σ^2.

This information as well as $SS_B/(p-1)$ and $SS_W/(n-p)$, the unbiased estimates of δ^2 and σ^2, and the F statistic, which is their ratio, are often summarized in what is known as an *analysis of variance* (ANOVA) table. See Tables 6.1.1 and 6.1.3.

As an illustration, consider the following data$^{(1)}$ giving blood cholesterol levels of men in three different socioeconomic groups labeled I, II, and III with I being the "high" end. We assume the one-way layout is valid. Note that this implies the possibly unrealistic

TABLE 6.1.1. ANOVA table for the one-way layout

	Sum of squares	d.f.	Mean squares	F-value
Between samples	$SS_B = \sum_{k=1}^{p} n_k(Y_{k\cdot} - Y_{\cdot\cdot})^2$	$p-1$	$MS_B = \frac{SS_B}{p-1}$	$\frac{MS_B}{MS_W}$
Within samples	$SS_W = \sum_{k=1}^{p} \sum_{l=1}^{n-k}(Y_{kl} - Y_{k\cdot})^2$	$n-p$	$MS_W = \frac{SS_W}{n-p}$	
Total	$SS_T = \sum_{k=1}^{p} \sum_{l=1}^{n_k}(Y_{kl} - Y_{\cdot\cdot})^2$	$n-1$		

TABLE 6.1.2. Blood cholesterol levels

I	403	311	269	336	259					
II	312	222	302	420	420	386	353	210	286	290
III	403	244	353	235	319	260				

assumption that the variance of the measurement is the same in the three groups (not to speak of normality). But see Section 6.6 for "robustness" to these assumptions.

We want to test whether there is a significant difference among the mean blood cholesterol of the three groups. Here $p = 3$, $n_1 = 5$, $n_2 = 10$, $n_3 = 6$, $n = 21$, and we compute

TABLE 6.1.3. ANOVA table for the cholesterol data

	SS	d.f.	MS	F-value
Between groups	1202.5	2	601.2	0.126
Within groups	$85,750.5$	18	4763.9	
Total	$86,953.0$	20		

From $\mathcal{F}$ tables, we find that the p-value corresponding to the F-value 0.126 is 0.88. Thus, there is no evidence to indicate that mean blood cholesterol is different for the three socioeconomic groups. □

Remark 6.1.4. Decompositions such as $(6.1.29)$ of the response total sum of squares SS_T into a variety of sums of squares measuring variability in the observations corresponding to variation of covariates are referred to as *analysis of variance*. They can be formulated in any linear model including regression models. See Scheffé (1959, pp. 42–45) and Weisberg (1985, p. 48). Originally such decompositions were used to motivate F statistics and to establish the distribution theory of the components via a device known as Cochran's theorem (Graybill, 1961, p. 86). Their principal use now is in the motivation of the convenient summaries of information we call ANOVA tables.

Confidence Intervals and Regions

We next use our distributional results and the method of pivots to find confidence intervals for μ_i, $1 \le i \le n$, β_j, $1 \le j \le p$, and in general, any linear combination

$$\psi = \psi(\boldsymbol{\mu}) = \sum_{i=1}^{n} a_i \mu_i = \mathbf{a}^T \boldsymbol{\mu}$$

of the μ's. If we set $\widehat{\psi} = \sum_{i=1}^{n} a_i \widehat{\mu}_i = \mathbf{a}^T \widehat{\boldsymbol{\mu}}$ and

$$\sigma^2(\widehat{\psi}) = \text{Var}(\widehat{\psi}) = \mathbf{a}^T \ \text{Var}(\widehat{\boldsymbol{\mu}})\mathbf{a} = \sigma^2 \mathbf{a}^T \mathbf{H}\mathbf{a},$$

where H is the hat matrix, then $(\widehat{\psi} - \psi)/\sigma(\widehat{\psi})$ has a $\mathcal{N}(0,1)$ distribution. Moreover,

$$(n-r)s^2/\sigma^2 = |\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2/\sigma^2 = \sum_{i=r+1}^{n} (U_i/\sigma)^2$$

has a χ^2_{n-r} distribution and is independent of $\widehat{\psi}$. Let

$$\widehat{\sigma}(\widehat{\psi}) = (s^2 \mathbf{a}^T \mathbf{H}\mathbf{a})^{\frac{1}{2}}$$

be an estimate of the standard deviation $\sigma(\widehat{\psi})$ of $\widehat{\psi}$. This estimated standard deviation is called *the standard error* of $\widehat{\psi}$. By referring to the definition of the t distribution, we find that the pivot

$$T(\psi) = \frac{(\widehat{\psi} - \psi)/\sigma(\widehat{\psi})}{(s/\sigma)} = (\widehat{\psi} - \psi)/\widehat{\sigma}(\widehat{\psi})$$

has a $\mathcal{T}_{n-r}$ distribution. Let $t_{n-r}\left(1 - \frac{1}{2}\alpha\right)$ denote the $1 - \frac{1}{2}\alpha$ quantile of the $\mathcal{T}_{n-r}$ distribution, then by solving $|T(\psi)| \le t_{n-r}\left(1 - \frac{1}{2}\alpha\right)$ for ψ, we find that

$$\psi = \widehat{\psi} \pm t_{n-r}\left(1 - \tfrac{1}{2}\alpha\right) \widehat{\sigma}(\widehat{\psi})$$

is, in the Gaussian linear model, a $100(1-\alpha)\%$ confidence interval for ψ.

Example 6.1.1. *One Sample (continued).* Consider $\psi = \mu$. We obtain the interval

$$\mu = \bar{Y} \pm t_{n-1}\left(1 - \tfrac{1}{2}\alpha\right) s/\sqrt{n},$$

which is the same as the interval of Example 4.4.1 and Section 4.9.2.

Example 6.1.2. *Regression (continued).* Assume that $p = r$. First consider $\psi = \beta_j$ for some specified regression coefficient β_j. The $100(1-\alpha)\%$ confidence interval for β_j is

$$\beta_j = \widehat{\beta}_j \pm t_{n-p}\left(1 - \tfrac{1}{2}\alpha\right) s\{[(\mathbf{Z}^T\mathbf{Z})^{-1}]_{jj}\}^{\frac{1}{2}}$$

where $[(\mathbf{Z}^T\mathbf{Z})^{-1}]_{jj}$ is the jth diagonal element of $(\mathbf{Z}^T\mathbf{Z})^{-1}$. Computer software computes $(\mathbf{Z}^T\mathbf{Z})^{-1}$ and labels $s\{[(\mathbf{Z}^T\mathbf{Z})^{-1}]_{jj}\}^{\frac{1}{2}}$ as the *standard error* of the (estimated) jth regression coefficient. Next consider $\psi = \mu_i =$ mean response for the ith case, $1 \le i \le n$. The level $(1-\alpha)$ confidence interval is

$$\mu_i = \widehat{\mu}_i \pm t_{n-p}\left(1 - \tfrac{1}{2}\alpha\right) s\sqrt{h_{ii}}$$

where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$. Here $s\sqrt{h_{ii}}$ is called the standard error of the (estimated) mean of the ith case.

Next consider the special case in which $p = 2$ and

$$Y_i = \beta_1 + \beta_2 z_{i2} + \epsilon_i,\ i = 1, \dots, n.$$

If we use the identity

$$\begin{aligned}\sum_{i=1}^{n}(z_{i2} - z_{\cdot 2})(Y_i - \bar{Y}) &= \sum(z_{i2} - z_{\cdot 2})Y_i - \bar{Y}\sum(z_{i2} - z_{\cdot 2}) \\ &= \sum(z_{i2} - z_{\cdot 2})Y_i,\end{aligned}$$

we obtain from Example 2.2.2 that

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{n}(z_{i2} - z_{\cdot 2})Y_i}{\sum_{i=1}^{n}(z_{i2} - z_{\cdot 2})^2}. \tag{6.1.30}$$

Because $\text{Var}(Y_i) = \sigma^2$, we obtain

$$\text{Var}(\widehat{\beta}_2) = \sigma^2 / \sum_{i=1}^{n}(z_{i2} - z_{\cdot 2})^2,$$

and the $100(1-\alpha)\%$ confidence interval for β_2 has the form

$$\beta_2 = \widehat{\beta}_2 \pm t_{n-p}\left(1 - \tfrac{1}{2}\alpha\right) s/\sqrt{\sum(z_{i2} - z_{\cdot 2})^2}.$$

The confidence interval for β_1 is given in Problem 6.1.10.

Similarly, in the $p = 2$ case, it is straightforward (Problem 6.1.10) to compute

$$h_{ii} = \frac{1}{n} + \frac{(z_{i2} - z_{\cdot 2})^2}{\sum_{i=1}^{n}(z_{i2} - z_{\cdot 2})^2}$$

and the confidence interval for the mean response μ_i of the ith case has a simple explicit form. □

Example 6.1.3. *One-Way Layout (continued).* We consider $\psi = \beta_k$, $1 \le k \le p$. Because $\widehat{\beta}_k = Y_{k\cdot} \sim \mathcal{N}(\beta_k, \sigma^2/n_k)$, we find the $100(1-\alpha)\%$ confidence interval

$$\beta_k = \widehat{\beta}_k \pm t_{n-p}\left(1 - \tfrac{1}{2}\alpha\right) s/\sqrt{n_k}$$

where $s^2 = SS_W/(n-p)$. The intervals for $\mu = \beta_\cdot$ and the incremental effect $\delta_k = \beta_k - \mu$ are given in Problem 6.1.11. □

Joint Confidence Regions

We have seen how to find confidence intervals for each individual β_j, $1 \le j \le p$. 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

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

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, $\sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$ 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

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

□

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. Let $\boldsymbol{\Psi} \equiv (\psi_1, \ldots, \psi_p)^T$ where, in the case of minimum contrast estimates based on $\rho(X, \theta)$, $\psi_j = \frac{\partial \rho}{\partial \theta_j}$ is well defined.

A0. We assume that $\bar{\boldsymbol{\theta}}_n$ satisfies

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

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). If $\psi_j(x, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} \log p(x, \boldsymbol{\theta})$, then $\boldsymbol{\theta}(P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ because $\mathcal{Q} \supset \mathcal{P}$. See Section 1.1.2.

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

A3. $\psi_i(\cdot, \boldsymbol{\theta})$, $1 \le i \le p$, 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. $\sup\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(D\boldsymbol{\Psi}(X_i, \mathbf{t}) - D\boldsymbol{\Psi}(X_i, \boldsymbol{\theta}(P)))\right| : |\mathbf{t} - \boldsymbol{\theta}(P)| \le \epsilon_n\right\} \overset{P}{\to} 0$ if $\epsilon_n \to 0$.

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

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

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

where

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

Hence,

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

where

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

and

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

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,

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

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 expression (6.2.6) is called the *Sandwich formula* (Huber(1967)).

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, $\frac{1}{n}\sum_{i=1}^{n} D\boldsymbol{\Psi}(X_i, \boldsymbol{\theta}_n^*)$ is nonsingular.

Note. This result goes beyond Theorem 5.4.2 in making it clear that although the definition of $\bar{\boldsymbol{\theta}}_n$ 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 $\boldsymbol{\theta}(P)$ can be defined uniquely by (6.2.2).

We can again extend the assumptions of Section 5.4.2 to:

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

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

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 $\bar{\boldsymbol{\theta}}_n$ is assumed. Proving consistency usually requires different arguments such as those of Section 5.2. See Section 9.2. It may, however, be shown that with probability tending to 1, a root-finding algorithm starting at a consistent estimate $\boldsymbol{\theta}_n^*$ will find a solution $\bar{\boldsymbol{\theta}}_n$ 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

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

where

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

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 $(6.2.8)$

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

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 $\text{Var}(\mathbf{U}^T, \mathbf{V}^T)^T$ nonsingular

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

Taking inverses of both sides yields

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

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

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

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

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

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

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

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 or M 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 $\mathbf{Z}_1, \dots, \mathbf{Z}_n$, $\widehat{\boldsymbol{\beta}}$ 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

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

and

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

so that by Theorem 6.2.2

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

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, $j = 1, \dots, p$.

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 $\mathbf{Z}_i = \mathbf{z}_i$, $1 \le i \le n$, 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, $(\beta_2 \dots, \beta_p)$ with

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

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

In case (i), we take, without loss of generality, $\alpha = \beta_2 = \cdots = \beta_p = 0$. Let $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^T$, 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, $\widehat{\beta}_1$, in case (ii), 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 $\beta_2, \dots, \beta_p$ 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

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

where $\widehat{\mathbf{Z}}^{(1)}$ is the regression of $(Z_{11}, \dots, Z_{n1})^T$ on the linear space spanned by $(Z_{1j}, \dots, Z_{nj})^T$, $2 \le j \le p$. Similarly,

$$I^{11}(\boldsymbol{\theta}) = \sigma^2/E(Z_{11} - \Pi(Z_{11} \mid Z_{12}, \dots, Z_{1p}))^2 \tag{6.2.27}$$

where $\Pi(Z_{11} \mid Z_{12}, \dots, Z_{1p})$ is the projection of Z_{11} on the linear span of $Z_{12}, \dots, Z_{1p}$ (Problem 6.2.11). Thus, $\Pi(Z_{11} \mid Z_{12}, \dots, Z_{1p}) = \sum_{j=2}^p a_j^* Z_{1j}$ where $(a_2^*, \dots, a_p^*)$ minimizes $E(Z_{11} - \sum_{j=2}^p a_j Z_{1j})^2$ over $(a_2, \dots, a_p) \in R^{p-1}$ (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 $\beta_2, \dots, \beta_p$ when the variables $Z_2, \dots, Z_p$ are in any way correlated with Z_1 and the price is measured by

$$\frac{[E(Z_{11} - \Pi(Z_{11} \mid Z_{12}, \dots, Z_{1p})^2]^{-1}}{E(Z_{11}^2)} = \left(1 - \frac{E(\Pi(Z_{11} \mid Z_{12}, \dots, Z_{1p}))^2}{E(Z_{11}^2)}\right)^{-1}. \tag{6.2.28}$$

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 $(Z_2, \dots, Z_p)$ have no value in predicting Z_1 linearly (see Section 1.4). Correspondingly in the Gaussian linear model (6.1.3) conditional on the $\mathbf{Z}_i$, $i = 1, \dots, n$, $\widehat{\beta}_1$ is undefined if the denominator in (6.2.26) is 0, which corresponds to the case of collinearity and occurs with probability 1 if $E(Z_{11} - \Pi(Z_{11} \mid Z_{12}, \dots, Z_{1p}))^2 = 0$. □

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. with density $\frac{1}{\sigma} f_0\left(\frac{\cdot}{\sigma}\right)$, where f_0 is not necessarily Gaussian, 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 $\psi = -\frac{f_0'}{f_0}$, $\chi(y) = -\left(y\frac{f_0'}{f_0}(y)+1\right)$, $\widehat{\boldsymbol{\beta}}_0 \equiv (\widehat{\beta}_{10}, \dots, \widehat{\beta}_{p0})^T$. 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

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

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

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

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

$$\int \boldsymbol{\Psi}(\mathbf{z}, y, \boldsymbol{\beta}, \sigma)dP = \mathbf{0} \tag{6.2.31}$$

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 $\sigma_0 = c(f_0, \sigma)\sigma$ for some $c(f_0, \sigma)$ 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 $\boldsymbol{\Psi} = (\psi_1, \dots, \psi_p)^T$ 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 $\theta_2, \dots, \theta_p$ 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 $\boldsymbol{\theta}$ 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. Bernstein-von Mises. *If the multivariate versions of* A0–A3, A4(a.s.), A5(a.s.) *and* A6–A8 *hold then, if $\widehat{\theta}$ denotes the MLE,*

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

a.s. under $P_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}$.

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 $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})\prod_{i=1}^n p(X_i, \boldsymbol{\theta})$, 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 $(\theta_3, \dots, \theta_p)$. 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 $Z_2, \dots, Z_p$. 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 $\boldsymbol{\theta}$, $X_1, \dots, X_n$ are i.i.d. $P_{\boldsymbol{\theta}}$, if $\widehat{\boldsymbol{\theta}}$ denotes the MLE for $P_{\boldsymbol{\theta}}$, then the posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$ converges a.s. under P_θ to the $\mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$ 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 cases we can turn to an approximation to the distribution of $\lambda(\mathbf{X})$ based on asymptotic theory, which is usually referred to as *Wilks's theorem* or *approximation*. Other approximations that will be explored in Volume II are based on Monte Carlo and bootstrap simulations. Here is an example in which Wilks's approximation to $\mathcal{L}(\lambda(\mathbf{X}))$ is useful:

Example 6.3.1. Suppose $X_1, X_2, \dots, X_n$ are i.i.d. as X where X has the gamma, $\Gamma(\alpha, \beta)$, distribution with density

$$p(x;\theta) = \beta^\alpha x^{\alpha-1}\exp\{-\beta x\}/\Gamma(\alpha);\ x>0;\ \alpha>0, \beta>0.$$

In Example 2.3.2 we showed that the MLE, $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})$, exists and in Example 2.4.2 we showed how to find $\widehat{\theta}$ as a nonexplicit solution of likelihood equations. Thus, the numerator of $\lambda(\mathbf{x})$ is available as $p(\mathbf{x}, \widehat{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i, \widehat{\theta})$. Suppose we want to test $H : \alpha = 1$ (exponential distribution) versus $K : \alpha \neq 1$. The MLE of β under H is readily seen from (2.3.5) to be $\widehat{\beta}_0 = 1/\bar{\mathbf{x}}$ and $p(\mathbf{x}; 1, \widehat{\beta}_0)$ is the denominator of the likelihood ratio statistic. It remains to find the critical value. This is not available analytically. □

The approximation we shall give is based on the result "$2\log\lambda(\mathbf{X}) \xrightarrow{\mathcal{L}} \chi^2_d$" for degrees of freedom d to be specified later. We next give an example that can be viewed as the limiting situation for which the approximation is exact:

Example 6.3.2. *The Gaussian Linear Model with Known Variance.* Let $Y_1, \dots, Y_n$ be independent with $Y_i \sim \mathcal{N}(\mu_i, \sigma_0^2)$ where σ_0 is known. As in Section 6.1.3 we test whether $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ is a member of a q-dimensional linear subspace of R^n, ω_0, versus the alternative that $\boldsymbol{\mu} \in \omega - \omega_0$ where ω is an r-dimensional linear subspace of R^n and $\omega \supset \omega_0$; and we transform to canonical form by setting

$$\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\mu},\ \mathbf{U} = \mathbf{A}\mathbf{Y}$$

where $\mathbf{A}_{n\times n}$ is an orthogonal matrix with rows $\mathbf{v}_1^T, \dots, \mathbf{v}_n^T$ such that $\mathbf{v}_1, \dots, \mathbf{v}_q$ span ω_0 and $\mathbf{v}_1, \dots, \mathbf{v}_r$ span ω.

Set $\theta_i = \eta_i/\sigma_0$, $i = 1, \dots, r$ and $X_i = U_i/\sigma_0$, $i = 1, \dots, n$. Then $X_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \dots, r$ and $X_i \sim \mathcal{N}(0,1)$, $i = r+1, \dots, n$. Moreover, the hypothesis H is equivalent to $H : \theta_{q+1} = \cdots = \theta_r = 0$. Using Section 6.1.3, we conclude that under H,

$$2\log\lambda(\mathbf{Y}) = \sum_{i=q+1}^{r} X_i^2 \sim \chi^2_{r-q}.$$

Wilks's theorem states that, under regularity conditions, when testing whether a parameter vector is restricted to an open subset of R^q or R^r, $q < r$, the χ^2_{r-q} distribution is an approximation to $\mathcal{L}(2\log\lambda(\mathbf{Y}))$. In this σ^2 known example, Wilks's approximation is exact. □

We illustrate the remarkable fact that χ^2_{r-q} holds as an approximation to the null distribution of $2\log\lambda$ quite generally when the hypothesis is a nice q-dimensional submanifold of an r-dimensional parameter space with the following.

Example 6.3.3. *The Gaussian Linear Model with Unknown Variance.* If Y_i are as in Example 6.3.2 but σ^2 is unknown then $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2)$ ranges over an $r + 1$-dimensional manifold whereas under H, $\boldsymbol{\theta}$ ranges over a $q + 1$-dimensional manifold. In Section 6.1.3, we derived

$$2 \log \lambda(\mathbf{Y}) = n \log \left(1 + \frac{\sum_{i=q+1}^{r} X_i^2}{\sum_{i=r+1}^{n} X_i^2} \right).$$

Apply Example 5.3.7 to $V_n = \sum_{i=q+1}^{r} X_i^2 / n^{-1} \sum_{i=r+1}^{n} X_i^2$ and conclude that $V_n \xrightarrow{\mathcal{L}} \chi^2_{r-q}$. Finally apply Lemma 5.3.2 with $g(t) = \log(1+t)$, $a_n = n$, $c = 0$ and conclude that $2 \log \lambda(\mathbf{Y}) \xrightarrow{\mathcal{L}} \chi^2_{r-q}$ also in the σ^2 unknown case. Note that for $\widetilde{\lambda}(\mathbf{Y})$ defined in Remark 6.1.2, $2 \log \widetilde{\lambda}(\mathbf{Y}) = V_n \xrightarrow{\mathcal{L}} \chi^2_{r-q}$ as well. □

Consider the general i.i.d. case with $X_1, \dots, X_n$ a sample from $p(x, \theta)$, where $x \in \mathcal{X} \subset R^s$, and $\boldsymbol{\theta} \in \Theta \subset R^r$. Write the log likelihood as

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(X_i, \theta).$$

We first consider the simple hypothesis $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Theorem 6.3.1. *Suppose the assumptions of Theorem* 6.2.2 *are satisfied. Then, under* $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$2 \log \lambda(\mathbf{X}) = 2[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)] \xrightarrow{\mathcal{L}} \chi^2_r.$$

Proof. Because $\widehat{\boldsymbol{\theta}}_n$ solves the likelihood equation $D_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}) = 0$, where $D_{\boldsymbol{\theta}}$ is the derivative with respect to $\boldsymbol{\theta}$, an expansion of $l_n(\boldsymbol{\theta})$ about $\widehat{\boldsymbol{\theta}}_n$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ gives

$$2[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)] = n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \mathbf{I}_n(\boldsymbol{\theta}_n^*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{6.3.1}$$

for some $\boldsymbol{\theta}_n^*$ with $|\boldsymbol{\theta}_n^* - \widehat{\boldsymbol{\theta}}_n| \le |\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|$. Here

$$\mathbf{I}_n(\boldsymbol{\theta}) = \left\| -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_k} \frac{\partial}{\partial \theta_j} \log p(X_i, \boldsymbol{\theta}) \right\|_{r \times r}.$$

By Theorem 6.2.2, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0))$, where $I_{r \times r}(\boldsymbol{\theta})$ is the Fisher information matrix.

Because

$$|\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0| \le |\boldsymbol{\theta}_n^* - \widehat{\boldsymbol{\theta}}_n| + |\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| \le 2|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|,$$

we can conclude arguing from A.3 and A.4 that that $\mathbf{I}_n(\boldsymbol{\theta}_n^*) \xrightarrow{P} E\mathbf{I}_n(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)$. Hence,

$$2[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)] \xrightarrow{\mathcal{L}} \mathbf{V}^T I(\boldsymbol{\theta}_0) \mathbf{V}, \mathbf{V} \sim \mathcal{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0)). \tag{6.3.2}$$

The result follows because, by Corollary B.6.2, $\mathbf{V}^T I(\boldsymbol{\theta}_0) \mathbf{V} \sim \chi^2_r$. □

As a consequence of the theorem, the test that rejects $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ when

$$2 \log \lambda(\mathbf{X}) \geq x_r(1-\alpha),$$

where $x_r(1-\alpha)$ is the $1-\alpha$ quantile of the χ^2_r distribution, has approximately level $1-\alpha$, and

$$\{\boldsymbol{\theta}_0 : 2[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)] \leq x_r(1-\alpha)\} \tag{6.3.3}$$

is a confidence region for $\boldsymbol{\theta}$ with approximate coverage probability $1-\alpha$.

Next we turn to the more general hypothesis $H : \boldsymbol{\theta} \in \Theta_0$, where Θ is open and Θ_0 is the set of $\boldsymbol{\theta} \in \Theta$ with $\theta_j = \theta_{0,j}$, $j = q+1, \dots, r$, and $\{\theta_{0,j}\}$ are specified values. Examples 6.3.1 and 6.3.2 illustrate such Θ_0. We set $d = r - q$, $\boldsymbol{\theta}^T = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, $\boldsymbol{\theta}^{(1)} = (\theta_1, \dots, \theta_q)^T$, $\boldsymbol{\theta}^{(2)} = (\theta_{q+1}, \dots, \theta_r)^T$, $\boldsymbol{\theta}_0^{(2)} = (\theta_{0,q+1}, \dots, \theta_{0,r})^T$.

Theorem 6.3.2. *Suppose that the assumptions of Theorem* 6.2.2 *hold for* $p(x, \boldsymbol{\theta})$*,* $\boldsymbol{\theta} \in \Theta$*. Let* $\mathcal{P}_0$ *be the model* $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta_0\}$ *with corresponding parametrization* $\boldsymbol{\theta}^{(1)} = (\theta_1, \dots, \theta_q)$*. Suppose that* $\widehat{\boldsymbol{\theta}}_0^{(1)}$ *is the MLE of* $\boldsymbol{\theta}^{(1)}$ *under* H *and that* $\widehat{\boldsymbol{\theta}}_0^{(1)}$ *satisfies A6 for* $\mathcal{P}_0$*. Let* $\widehat{\boldsymbol{\theta}}_{0,n}^T = (\widehat{\boldsymbol{\theta}}_0^{(1)}, \boldsymbol{\theta}_0^{(2)})$*. Then under* $H : \boldsymbol{\theta} \in \Theta_0$*,*

$$2 \log \lambda(\mathbf{X}) \equiv 2[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\widehat{\boldsymbol{\theta}}_{0,n})] \xrightarrow{\mathcal{L}} \chi^2_{r-q}.$$

Proof. Let $\boldsymbol{\theta}_0 \in \Theta_0$ and write

$$2 \log \lambda(\mathbf{X}) = 2[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)] - 2[l_n(\widehat{\boldsymbol{\theta}}_{0,n}) - l_n(\boldsymbol{\theta}_0)]. \tag{6.3.4}$$

It is easy to see that A0–A6 for $\mathcal{P}$ imply A0–A5 for $\mathcal{P}_0$. By (6.2.10) and (6.3.1) applied to $\widehat{\boldsymbol{\theta}}_n$ and the corresponding argument applied to $\widehat{\boldsymbol{\theta}}_0^{(1)}$, $\widehat{\boldsymbol{\theta}}_{0,n}$ and (6.3.4),

$$2 \log \lambda(\mathbf{X}) = \mathbf{S}^T(\boldsymbol{\theta}_0) I^{-1}(\boldsymbol{\theta}_0) \mathbf{S}(\boldsymbol{\theta}_0) - \mathbf{S}_1^T(\boldsymbol{\theta}_0) I_0^{-1}(\boldsymbol{\theta}_0) \mathbf{S}_1(\boldsymbol{\theta}_0) + o_p(1) \tag{6.3.5}$$

where

$$\mathbf{S}(\boldsymbol{\theta}_0) = n^{-1/2} \sum_{i=1}^n Dl(\mathbf{X}_i, \boldsymbol{\theta})$$

and $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2)^T$ where $\mathbf{S}_1$ is the first q coordinates of $\mathbf{S}$. Furthermore,

$$I_0(\boldsymbol{\theta}_0) = \text{Var}_{\boldsymbol{\theta}_0} \mathbf{S}_1(\boldsymbol{\theta}_0).$$

Make a change of parameter, for given true $\boldsymbol{\theta}_0$ in Θ_0,

$$\boldsymbol{\eta} = M(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

where, dropping the dependence on $\boldsymbol{\theta}_0$,

$$M = PI^{1/2} \tag{6.3.6}$$

and P is an orthogonal matrix such that, if $\boldsymbol{\Delta}_0 \equiv \{\boldsymbol{\theta} - \boldsymbol{\theta}_0 : \boldsymbol{\theta} \in \Theta_0\}$

$$M\boldsymbol{\Delta}_0 = \{\boldsymbol{\eta} : \eta_{q+1} = \cdots = \eta_r = 0, \ \boldsymbol{\eta} \in M\Theta\}.$$

Such P exists by the argument given in Example 6.2.1 because $I^{1/2}\boldsymbol{\Delta}_0$ is the intersection of a q dimensional linear subspace of R^r with $I^{1/2}\{\boldsymbol{\theta} - \boldsymbol{\theta}_0 : \boldsymbol{\theta} \in \Theta\}$. Now write $D_{\boldsymbol{\theta}}$ for differentiation with respect to $\boldsymbol{\theta}$ and $D_{\boldsymbol{\eta}}$ for differentiation with respect to $\boldsymbol{\eta}$. Note that, by definition, λ is invariant under reparametrization

$$\lambda(\mathbf{X}) = \gamma(\mathbf{X}) \tag{6.3.7}$$

where

$$\gamma(\mathbf{X}) = \sup_{\boldsymbol{\eta}}\{p(x, \boldsymbol{\theta}_0 + M^{-1}\boldsymbol{\eta})\} / \sup\{p(\mathbf{x}, \boldsymbol{\theta}_0 + M^{-1}\boldsymbol{\eta}) : \boldsymbol{\theta}_0 + M^{-1}\boldsymbol{\eta} \in \Theta_0\}$$

and from (B.8.13)

$$D_{\boldsymbol{\eta}} l(\mathbf{x}, \boldsymbol{\theta}_0 + M^{-1}\boldsymbol{\eta}) = [M^{-1}]^T D_{\boldsymbol{\theta}} l(\mathbf{x}, \boldsymbol{\theta}). \tag{6.3.8}$$

We deduce from (6.3.6) and (6.3.8) that if

$$\mathbf{T}(\boldsymbol{\eta}) \equiv n^{-1/2} \sum_{i=1}^{n} D_{\boldsymbol{\eta}} l(\mathbf{X}_i, \boldsymbol{\theta}_0 + M^{-1}\boldsymbol{\eta}),$$

then

$$\text{Var}\, \mathbf{T}(\mathbf{0}) = P^T I^{-1/2} I I^{-1/2} P = J. \tag{6.3.9}$$

Moreover, because in terms of $\boldsymbol{\eta}$, H is $\{\boldsymbol{\eta} \in M\Theta : \eta_{q+1} = \cdots = \eta_r = 0\}$, then by applying (6.3.5) to $\gamma(\mathbf{X})$ we obtain,

$$\begin{aligned} 2\log\gamma(\mathbf{X}) &= \mathbf{T}^T(\mathbf{0})\mathbf{T}(\mathbf{0}) - \mathbf{T}_1^T(\mathbf{0})\mathbf{T}_1(\mathbf{0}) + o_p(1) \\ &= \sum_{i=1}^{r} T_i^2(\mathbf{0}) - \sum_{i=1}^{q} T_i^2(\mathbf{0}) + o_p(1) \\ &= \sum_{i=q+1}^{r} T_i^2(\mathbf{0}) + o_p(1), \end{aligned} \tag{6.3.10}$$

which has a limiting χ^2_{r-q} distribution by Slutsky's theorem because $\mathbf{T}(\mathbf{0})$ has a limiting $\mathcal{N}_r(\mathbf{0}, J)$ distribution by (6.3.9). The result follows from (6.3.7). □

Note that this argument is simply an asymptotic version of the one given in Example 6.3.2.

Thus, under the conditions of Theorem 6.3.2, rejecting if $\lambda(\mathbf{X}) \geq x_{r-q}(1-\alpha)$ is an asymptotically level α test of $H : \boldsymbol{\theta} \in \Theta_0$. Of equal importance is that we obtain an asymptotic confidence region for $(\theta_{q+1}, \ldots, \theta_r)$, a piece of $\boldsymbol{\theta}$, with $\theta_1, \ldots, \theta_q$ acting as nuisance parameters. This asymptotic level $1-\alpha$ confidence region is

$$\{(\theta_{q+1}, \ldots, \theta_r) : 2[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\widehat{\theta}_{0,1}, \ldots, \widehat{\theta}_{0,q}, \theta_{q+1}, \ldots, \theta_r)] \leq x_{r-q}(1-\alpha)\} \tag{6.3.11}$$

where $\widehat{\theta}_{0,1}, \dots, \widehat{\theta}_{0,q}$ are the MLEs, themselves depending on $\theta_{q+1}, \dots, \theta_r$, of $\theta_1, \dots, \theta_q$ assuming that $\theta_{q+1}, \dots, \theta_r$ are known.

More complicated linear hypotheses such as $H : \boldsymbol{\theta} - \boldsymbol{\theta}_0 \in \omega_0$ where ω_0 is a linear space of dimension q are also covered. We only need note that if ω_0 is a linear space spanned by an orthogonal basis $\mathbf{v}_1, \dots, \mathbf{v}_q$ and $\mathbf{v}_{q+1}, \dots, \mathbf{v}_r$ are orthogonal to ω_0 and $\mathbf{v}_1, \dots, \mathbf{v}_r$ span R^r then,

$$\omega_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}^T \mathbf{v}_j = 0, \ q+1 \le j \le r\}. \tag{6.3.12}$$

The extension of Theorem 6.3.2 to this situation is easy and given in Problem 6.3.2.

The formulation of Theorem 6.3.2 is still inadequate for most applications. It can be extended as follows.

Suppose H is specified by:

There exist d functions, $g_j : \Theta \to R$, $q+1 \le j \le r$ written as a vector $\mathbf{g}$, such that $D\mathbf{g}(\boldsymbol{\theta})$ exists and is of rank $r-q$ at all $\boldsymbol{\theta} \in \Theta$. Define $H : \boldsymbol{\theta} \in \Theta_0$ with

$$\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}\}. \tag{6.3.13}$$

Evidently, Theorem 6.3.2 falls under this schema with $g_j(\boldsymbol{\theta}) = \theta_j - \theta_{0,j}$, $q+1 \le j \le r$.

Examples such as testing for independence in contingency tables, which require the following general theorem, will appear in the next section.

Theorem 6.3.3. *Suppose the assumptions of Theorem* 6.3.2 *and the previously conditions on* $\mathbf{g}$ *hold. Suppose the MLE* $\widehat{\boldsymbol{\theta}}_{0,n}$ *under* H *is consistent for all* $\boldsymbol{\theta} \in \Theta_0$*. Then, if* $\lambda(\mathbf{X})$ *is the likelihood ratio statistic for* $H : \boldsymbol{\theta} \in \Theta_0$ *given in* (6.3.13), $2 \log \lambda(\mathbf{X}) \xrightarrow{\mathcal{L}} \chi^2_{r-q}$ *under* H.

The proof is sketched in Problems (6.3.2)–(6.3.3). The essential idea is that, if $\boldsymbol{\theta}_0$ is true, $\lambda(\mathbf{X})$ behaves asymptotically like a test for $H : \boldsymbol{\theta} \in \Theta_{00}$ where

$$\Theta_{00} = \{\boldsymbol{\theta} \in \Theta : D\mathbf{g}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \mathbf{0}\} \tag{6.3.14}$$

a hypothesis of the form (6.3.13).

Wilks's theorem depends critically on the fact that not only is Θ open but that if Θ_0 given in (6.3.13) then the set $\{(\theta_1, \dots, \theta_q)^T : \boldsymbol{\theta} \in \Theta\}$ is open in R^q. We need both properties because we need to analyze both the numerator and denominator of $\lambda(\mathbf{X})$. As an example of what can go wrong, let (X_{i1}, X_{i2}) be i.i.d. $\mathcal{N}(\theta_1, \theta_2, J)$, where J is the 2×2 identity matrix and $\Theta_0 = \{\boldsymbol{\theta} : \theta_1 + \theta_2 \le 1\}$. If $\theta_1 + \theta_2 = 1$,

$$\widehat{\boldsymbol{\theta}}_0 = \left(\frac{(X_{\cdot 1} + X_{\cdot 2})}{2} + \frac{1}{2}, \frac{1}{2} - \frac{(X_{\cdot 1} + X_{\cdot 2})}{2} \right)$$

and $2 \log \lambda(\mathbf{X}) \to \chi^2_1$ but if $\theta_1 + \theta_2 < 1$ clearly $2 \log \lambda(\mathbf{X}) = o_p(1)$. Here the dimension of Θ_0 and Θ is the same but the boundary of Θ_0 has lower dimension. More sophisticated examples are given in Problems 6.3.5 and 6.3.6.

6.3.2 Wald's and Rao's Large Sample Tests

The Wald Test

Suppose that the assumptions of Theorem 6.2.2 hold. Then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta})) \text{ as } n \to \infty. \tag{6.3.15}$$

Because $I(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ (Problem 6.3.10), it follows from Proposition B.7.1(a) that

$$I(\widehat{\boldsymbol{\theta}}_n) \xrightarrow{P} I(\boldsymbol{\theta}) \text{ as } n \to \infty. \tag{6.3.16}$$

By Slutsky's theorem B.7.2, (6.3.15) and (6.3.16),

$$n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^T I(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathbf{V}^T I(\boldsymbol{\theta})\mathbf{V},\ \mathbf{V} \sim \mathcal{N}_r(\mathbf{0}, I^{-1}(\boldsymbol{\theta}))$$

where, according to Corollary B.6.2, $\mathbf{V}^T I(\boldsymbol{\theta})\mathbf{V} \sim \chi^2_r$. It follows that the *Wald test* that rejects $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ in favor of $K : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ when

$$W_n(\boldsymbol{\theta}_0) = n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \geq x_r(1-\alpha)$$

has asymptotic level α. More generally $I(\boldsymbol{\theta}_0)$ can be replaced by any consistent estimate of $I(\boldsymbol{\theta}_0)$, in particular $-\frac{1}{n}D^2 l_n(\boldsymbol{\theta}_0)$ or $I(\widehat{\boldsymbol{\theta}}_n)$ or $-\frac{1}{n}D^2 l_n(\widehat{\boldsymbol{\theta}}_n)$. The last Hessian choice is favored because it is usually computed automatically with the MLE. It and $I(\widehat{\boldsymbol{\theta}}_n)$ also have the advantage that the confidence region one generates $\{\boldsymbol{\theta} : W_n(\boldsymbol{\theta}) \leq x_p(1-\alpha)\}$ is an ellipsoid in R^r easily interpretable and computable—see (6.1.31).

For the more general hypothesis $H : \theta \in \Theta_0$ we write the MLE for $\boldsymbol{\theta} \in \Theta$ as $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_n^{(1)}, \widehat{\boldsymbol{\theta}}_n^{(2)})$ where $\widehat{\boldsymbol{\theta}}_n^{(1)} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_q)$ and $\widehat{\boldsymbol{\theta}}_n^{(2)} = (\widehat{\theta}_{q+1}, \ldots, \widehat{\theta}_r)$ and define the *Wald statistic* as

$$W_n(\boldsymbol{\theta}_0^{(2)}) = n(\widehat{\boldsymbol{\theta}}_n^{(2)} - \boldsymbol{\theta}_0^{(2)})^T [I^{22}(\widehat{\boldsymbol{\theta}}_n)]^{-1}(\widehat{\boldsymbol{\theta}}_n^{(2)} - \boldsymbol{\theta}_0^{(2)}) \tag{6.3.17}$$

where $I^{22}(\boldsymbol{\theta})$ is the lower diagonal block of $I^{-1}(\boldsymbol{\theta})$ written as

$$I^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} I^{11}(\boldsymbol{\theta}) & I^{12}(\boldsymbol{\theta}) \\ I^{21}(\boldsymbol{\theta}) & I^{22}(\boldsymbol{\theta}) \end{pmatrix}$$

with diagonal blocks of dimension $q \times q$ and $d \times d$, respectively. More generally, $I^{22}(\widehat{\boldsymbol{\theta}}_n)$ is replaceable by any consistent estimate of $I^{22}(\boldsymbol{\theta})$, for instance, the lower diagonal block of the inverse of $-\frac{1}{n}D^2 l_n(\widehat{\boldsymbol{\theta}}_n)$, the Hessian (Problem 6.3.9).

Theorem 6.3.4. *Under the conditions of Theorem* 6.2.2*, if H is true*,

$$W_n(\boldsymbol{\theta}_0^{(2)}) \xrightarrow{\mathcal{L}} \chi^2_{r-q}. \tag{6.3.18}$$

Proof. $I(\boldsymbol{\theta})$ continuous implies that $I^{-1}(\boldsymbol{\theta})$ is continuous and, hence, I^{22} is continuous. But by Theorem 6.2.2, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^{(2)} - \boldsymbol{\theta}_0^{(2)}) \xrightarrow{\mathcal{L}} \mathcal{N}_d(\mathbf{0}, I^{22}(\boldsymbol{\theta}_0))$ if $\boldsymbol{\theta}_0 \in \Theta_0$ holds. Slutsky's theorem completes the proof. □

The *Wald* test, which rejects iff $W_n(\widehat{\boldsymbol{\theta}}_0^{(2)}) \geq x_{r-q}(1-\alpha)$, is, therefore, asymptotically level α. What is not as evident is that, under H,

$$W_n(\widehat{\boldsymbol{\theta}}_0^{(2)}) = 2\log\lambda(\mathbf{X}) + o_p(1) \tag{6.3.19}$$

where $\lambda(\mathbf{X})$ is the LR statistic for $H : \boldsymbol{\theta} \in \Theta_0$. The argument is sketched in Problem 6.3.9. Thus, the two tests are equivalent asymptotically.

The Wald test leads to the *Wald confidence regions* for $(\theta_{q+1}, \dots, \theta_r)^T$ given by $\{\boldsymbol{\theta}^{(2)} : W_n(\boldsymbol{\theta}^{(2)}) \leq x_{r-q}(1-\alpha)\}$. These regions are ellipsoids in R^d. Although, as (6.3.19) indicates, the Wald and likelihood ratio tests and confidence regions are asymptotically equivalent in the sense that the same conclusions are reached for large n, in practice they can be very different.

The Rao Score Test

For the simple hypothesis $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, Rao's score test is based on the observation that, by the central limit theorem,

$$\sqrt{n}\boldsymbol{\psi}_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}_0)) \tag{6.3.20}$$

where $\boldsymbol{\psi}_n = n^{-1}Dl_n(\boldsymbol{\theta}_0)$ is the likelihood score vector.

It follows from this and Corollary B.6.2 that under H, as $n \to \infty$,

$$R_n(\boldsymbol{\theta}_0) = n\boldsymbol{\psi}_n^T(\boldsymbol{\theta}_0)I^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\psi}_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} \chi_r^2.$$

The test that rejects H when $R_n(\boldsymbol{\theta}_0) \geq x_r(1-\alpha)$ is called the *Rao score test*. This test has the advantage that it can be carried out without computing the MLE, and the convergence $R_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} \chi_r^2$ requires much weaker regularity conditions than does the corresponding convergence for the likelihood ratio and Wald tests.

The extension of the Rao test to $H : \boldsymbol{\theta} \in \Theta_0$ runs as follows. Let

$$\boldsymbol{\Psi}_n(\boldsymbol{\theta}) = n^{-1/2}D_2l_n(\boldsymbol{\theta})$$

where D_1l_n represents the $q \times 1$ gradient with respect to the first q coordinates and D_2l_n the $d \times 1$ gradient with respect to the last d. The Rao test is based on the statistic

$$R_n(\boldsymbol{\theta}_0^{(2)}) \equiv n\boldsymbol{\Psi}_n^T(\widehat{\boldsymbol{\theta}}_{0,n})\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Psi}_n(\widehat{\boldsymbol{\theta}}_{0,n})$$

where $\widehat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$, the asymptotic variance of $\sqrt{n}\boldsymbol{\Psi}_n(\widehat{\boldsymbol{\theta}}_{0,n})$ under H.

It can be shown that (Problem 6.3.8)

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = I_{22}(\boldsymbol{\theta}_0) - I_{21}(\boldsymbol{\theta}_0)I_{11}^{-1}(\boldsymbol{\theta}_0)I_{12}(\boldsymbol{\theta}_0) \tag{6.3.21}$$

where I_{11} is the upper left $q \times q$ block of the $r \times r$ information matrix $I(\boldsymbol{\theta}_0)$, I_{12} is the upper right $q \times d$ block, and so on. Furthermore, (Problem 6.3.9) under A0–A6 and consistency of $\widehat{\boldsymbol{\theta}}_{0,n}$ under H, a consistent estimate of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)$ is

$$n^{-1}[-D_2^2l_n(\widehat{\boldsymbol{\theta}}_{0,n}) + D_{21}l_n(\widehat{\boldsymbol{\theta}}_{0,n})[D_1^2l_n(\widehat{\boldsymbol{\theta}}_{0,n}]^{-1}D_{12}l_n(\widehat{\boldsymbol{\theta}}_{0,n})] \tag{6.3.22}$$

where D_2^2 is the $d \times d$ matrix of second partials of l_n with respect to $\boldsymbol{\theta}^{(1)}$, D_{21} the $d \times d$ matrix of mixed second partials with respect to $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, and so on.

Theorem 6.3.5. *Under $H : \boldsymbol{\theta} \in \Theta_0$ and the conditions A0–A5 of Theorem* 6.2.2 *but with A6 required only for $\mathcal{P}_0$*

$$R_n(\boldsymbol{\theta}_0^{(2)}) \xrightarrow{\mathcal{L}} \chi_d^2.$$

The Rao large sample critical and confidence regions are $\{R_n(\boldsymbol{\theta}_0^{(2)}) \geq x_d(1-\alpha)\}$ and $\{\boldsymbol{\theta}^{(2)} : R_n(\boldsymbol{\theta}^{(2)}) < x_d(1-\alpha)\}$.

The advantage of the Rao test over those of Wald and Wilks is that MLEs need to be computed only under H. On the other hand, it shares the disadvantage of the Wald test that matrices need to be computed and inverted.

Power Behavior of the LR, Rao, and Wald Tests

It is possible as in the one-dimensional case to derive the asymptotic power for these tests for alternatives of the form $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \frac{\boldsymbol{\Delta}}{\sqrt{n}}$ where $\boldsymbol{\theta}_0 \in \Theta_0$. The analysis for $\Theta_0 = \{\theta_0\}$ is relatively easy. For instance, for the Wald test

$$\begin{aligned} & n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T I(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ & = (\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) + \Delta) I(\widehat{\boldsymbol{\theta}}_n)(\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) + \boldsymbol{\Delta}) \xrightarrow{\mathcal{L}} \chi^2_{r-q}(\boldsymbol{\Delta}^T I(\boldsymbol{\theta}_0)\boldsymbol{\Delta}) \end{aligned}$$

where $\chi_m^2(\gamma^2)$ is the noncentral chi square distribution with m degrees of freedom and noncentrality parameter γ^2.

It may be shown that the equivalence (6.3.19) holds under $\boldsymbol{\theta}_n$ and that the power behavior is unaffected and applies to all three tests.

Consistency for fixed alternatives is clear for the Wald test but requires conditions for the likelihood ratio and score tests—see Rao (1973) for more on this.

Summary. We considered the problem of testing $H : \boldsymbol{\theta} \in \Theta_0$ versus $K : \boldsymbol{\theta} \in \Theta - \Theta_0$ where Θ is an open subset of R^r and Θ_0 is the collection of $\boldsymbol{\theta} \in \Theta$ with the last $r - q$ coordinates $\boldsymbol{\theta}^{(2)}$ specified. We established Wilks's theorem, which states that if $\lambda(\mathbf{X})$ is the LR statistic, then, under regularity conditions, $2\log\lambda(\mathbf{X})$ has an asymptotic χ^2_{r-q} distribution under H. We also considered a quadratic form, called the Wald statistic, which measures the distance between the hypothesized value of $\boldsymbol{\theta}^{(2)}$ and its MLE, and showed that this quadratic form has limiting distribution χ^2_{r-q}. Finally, we introduced the Rao score test, which is based on a quadratic form in the gradient of the log likelihood. The asymptotic distribution of this quadratic form is also χ^2_{r-q}.

6.4 LARGE SAMPLE METHODS FOR DISCRETE DATA

In this section we give a number of important applications of the general methods we have developed to inference for discrete data. In particular we shall discuss problems of goodness-of-fit and special cases of log linear and generalized linear models (GLM), treated in more detail in Section 6.5.

6.4.1 Goodness-of-Fit in a Multinomial Model. Pearson's χ^2 Test

As in Examples 1.6.7, 2.2.8, and 2.3.3, consider i.i.d. trials in which $X_i = j$ if the ith trial produces a result in the jth category, $j = 1, \dots, k$. Let $\theta_j = P(X_i = j)$ be the probability of the jth category. Because $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$, we consider the parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{k-1})^T$ and test the hypothesis $H : \theta_j = \theta_{0j}$ for specified θ_{0j}, $j = 1, \dots, k-1$. Thus, we may be testing whether a random number generator used in simulation experiments is producing values according to a given distribution, or we may be testing whether the phenotypes in a genetic experiment follow the frequencies predicted by theory. In Example 2.2.8 we found the MLE $\widehat{\theta}_j = N_j/n$, where $N_j = \sum_{i=1}^n 1\{X_i = j\}$. It follows that the large sample LR rejection region is

$$2 \log \lambda(\mathbf{X}) = 2 \sum_{j=1}^{k} N_j \log(N_j/n\theta_{0j}) \geq x_{k-1}(1-\alpha).$$

To find the Wald test, we need the information matrix $I = \|I_{ij}\|$. For $i, j = 1, \dots, k-1$, we find using (2.2.33) and (3.4.32) that

$$\begin{aligned} I_{ij} &= \left[\frac{1}{\theta_j} + \frac{1}{\theta_k}\right] && \text{if } i = j, \\ &= \frac{1}{\theta_k} && \text{if } i \neq j. \end{aligned}$$

Thus, with $\theta_{0k} = 1 - \sum_{j=1}^{k-1} \theta_{0j}$, the Wald statistic is

$$W_n(\boldsymbol{\theta}_0) = n \sum_{j=1}^{k-1} [\widehat{\theta}_j - \theta_{0j}]^2/\theta_{0j} + n \sum_{j=1}^{k-1} \sum_{i=1}^{k-1} (\widehat{\theta}_i - \theta_{0i})(\widehat{\theta}_j - \theta_{0j})/\theta_{0k}.$$

The second term on the right is

$$n \left[\sum_{j=1}^{k-1} (\widehat{\theta}_j - \theta_{0j})\right]^2 \Bigg/ \theta_{0k} = n(\widehat{\theta}_k - \theta_{0k})^2/\theta_{0k}.$$

Thus,

$$W_n(\boldsymbol{\theta}_0) = \sum_{j=1}^{k} (N_j - n\theta_{0j})^2/n\theta_{0j}.$$

The term on the right is called *Pearson's chi-square* (χ^2) *statistic* and is the statistic that is typically used for this multinomial testing problem. It is easily remembered as

$$\chi^2 = \text{SUM}\,\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \tag{6.4.1}$$

where the sum is over categories and "expected" refers to the expected frequency $E_H(N_j)$. The general form (6.4.1) of Pearson's χ^2 will reappear in other multinomial applications in this section.

To derive the Rao test, note that from Example 2.2.8,

$$\boldsymbol{\psi}_n(\boldsymbol{\theta}) = n^{-1}(\psi_1(\boldsymbol{\theta}), \dots, \psi_{k-1}(\boldsymbol{\theta}))^T,$$

with

$$\psi_j(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_j} l_n(\boldsymbol{\theta}) = \frac{N_j}{\theta_j} - \frac{N_k}{\theta_k},\ j = 1, \dots, k-1.$$

To find I^{-1}, we could invert I or note that by (6.2.11), $I^{-1}(\boldsymbol{\theta}) = \Sigma = \text{Var}(\mathbf{N})$, where $\mathbf{N} = (N_1, \dots, N_{k-1})^T$ and, by A.13.15, $\Sigma = \|\sigma_{ij}\|_{(k-1)\times(k-1)}$ with

$$\sigma_{ii} = \text{Var}(N_i) = n\theta_i(1-\theta_i),\ \sigma_{ij} = -n\theta_i\theta_j,\ i \neq j.$$

Thus, the Rao statistic is

$$\begin{aligned} R_n(\boldsymbol{\theta}_0) &= \left(n \sum_{j=1}^{k-1} \left(\frac{\widehat{\theta}_j}{\theta_{0j}} - \frac{\widehat{\theta}_k}{\theta_{0k}} \right)^2 \theta_{0j} \right) \\ &- \left(n \sum_{j=1}^{k-1} \sum_{i=1}^{k-1} \left(\frac{\widehat{\theta}_i}{\theta_{0i}} - \frac{\widehat{\theta}_k}{\theta_{0k}} \right) \left(\frac{\widehat{\theta}_j}{\theta_{0j}} - \frac{\widehat{\theta}_k}{\theta_{0k}} \right) \theta_{0i}\theta_{0j} \right). \end{aligned} \tag{6.4.2}$$

The second term on the right is

$$-n \left[\sum_{j=1}^{k-1} \left(\frac{\widehat{\theta}_j}{\theta_{0j}} - \frac{\widehat{\theta}_k}{\theta_{0k}} \right) \theta_{0j} \right]^2 = -n \left[\frac{\widehat{\theta}_k}{\theta_{0k}} - 1 \right]^2.$$

To simplify the first term on the right of (6.4.2), we write

$$\frac{\widehat{\theta}_j}{\theta_{0j}} - \frac{\widehat{\theta}_k}{\theta_{0k}} = \{[\theta_{0k}(\widehat{\theta}_j - \theta_{0j})] - [\theta_{0j}(\widehat{\theta}_k - \theta_{0k})]\} \frac{1}{\theta_{0j}\theta_{0k}},$$

and expand the square keeping the square brackets intact. Then, because

$$\sum_{j=1}^{k-1} (\widehat{\theta}_j - \theta_{0j}) = -(\widehat{\theta}_k - \theta_{0k}),$$

the first term on the right of (6.4.2) becomes

$$\begin{aligned}
&= n\left\{\sum_{j=1}^{k-1}\frac{1}{\theta_{0j}}(\widehat{\theta}_j-\theta_{0j})^2+\frac{2}{\theta_{0k}}(\widehat{\theta}_k-\theta_{0k})^2+\frac{1}{\theta_{0k}^2}(1-\theta_{0k})(\widehat{\theta}_k-\theta_{0k})^2\right\}\\
&= n\left\{\sum_{j=1}^{k}\frac{1}{\theta_{0j}}(\widehat{\theta}_j-\theta_{0j})^2+\frac{1}{\theta_{0k}^2}(\widehat{\theta}_k-\theta_{0k})^2\right\}.
\end{aligned}$$

It follows that the Rao statistic equals Pearson's χ^2.

Example 6.4.1. *Testing a Genetic Theory.* In experiments on pea breeding, Mendel observed the different kinds of seeds obtained by crosses from peas with round yellow seeds and peas with wrinkled green seeds. Possible types of progeny were: (1) round yellow; (2) wrinkled yellow; (3) round green; and (4) wrinkled green. If we assume the seeds are produced independently, we can think of each seed as being the outcome of a multinomial trial with possible outcomes numbered 1, 2, 3, 4 as above and associated probabilities of occurrence $\theta_1, \theta_2, \theta_3, \theta_4$. Mendel's theory predicted that $\theta_1 = 9/16$, $\theta_2 = \theta_3 = 3/16$, $\theta_4 = 1/16$, and we want to test whether the distribution of types in the $n = 556$ trials he performed (seeds he observed) is consistent with his theory. Mendel observed $n_1 = 315$, $n_2 = 101$, $n_3 = 108$, $n_4 = 32$. Then, $n\theta_{10} = 312.75$, $n\theta_{20} = n\theta_{30} = 104.25$, $n\theta_{40} = 34.75$, $k = 4$

$$\chi^2 = \frac{(2.25)^2}{312.75}+\frac{(3.25)^2}{104.25}+\frac{(3.75)^2}{104.25}+\frac{(2.75)^2}{34.75} = 0.47,$$

which has a p-value of 0.9 when referred to a χ_3^2 table. There is insufficient evidence to reject Mendel's hypothesis. For comparison $2\log\lambda = 0.48$ in this case. However, this value may be too small! See Note 1.

6.4.2 Goodness-of-Fit to Composite Multinomial Models. Contingency Tables

Suppose $\mathbf{N} = (N_1, \ldots, N_k)^T$ has a multinomial, $\mathcal{M}(n, \boldsymbol{\theta})$, distribution. We will investigate how to test $H : \theta \in \Theta_0$ versus $K : \theta \notin \Theta_0$, where Θ_0 is a composite "smooth" subset of the $(k-1)$-dimensional parameter space

$$\Theta = \{\boldsymbol{\theta} : \theta_i \geq 0,\ 1 \leq i \leq k,\ \sum_{i=1}^{k}\theta_i = 1\}.$$

For example, in the Hardy–Weinberg model (Example 2.1.4),

$$\Theta_0 = \{(\eta^2, 2\eta(1-\eta), (1-\eta)^2),\ 0 \leq \eta \leq 1\},$$

which is a one-dimensional curve in the two-dimensional parameter space Θ. Here testing the adequacy of the Hardy–Weinberg model means testing $H : \boldsymbol{\theta} \in \Theta_0$ versus $K : \boldsymbol{\theta} \in$

Θ_1 where $\Theta_1 = \Theta - \Theta_0$. Other examples, which will be pursued further later in this section, involve restrictions on the θ_i obtained by specifying independence assumptions on classifications of cases into different categories.

We suppose that we can describe Θ_0 parametrically as

$$\Theta_0 = \{(\theta_1(\boldsymbol{\eta}), \dots, \theta_k(\boldsymbol{\eta}) : \boldsymbol{\eta} \in \mathcal{E}\},$$

where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_q)^T$, is a subset of q-dimensional space, and the map $\boldsymbol{\eta} \to (\theta_1(\boldsymbol{\eta}), \dots, \theta_k(\boldsymbol{\eta}))^T$ takes $\mathcal{E}$ into Θ_0. To avoid trivialities we assume $q < k - 1$.

Consider the likelihood ratio test for $H : \theta \in \Theta_0$ versus $K : \theta \notin \Theta_0$. Let $p(n_1, \dots, n_k, \boldsymbol{\theta})$ denote the frequency function of $\mathbf{N}$. Maximizing $p(n_1, \dots, n_k, \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta_0$ is the same as maximizing $p(n_1, \dots, n_k, \boldsymbol{\theta}(\boldsymbol{\eta}))$ for $\boldsymbol{\eta} \in \mathcal{E}$. If a maximizing value, $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \dots, \widehat{\eta}_q)$ exists, the log likelihood ratio is given by

$$\log \lambda(n_1, \dots, n_k) = \sum_{i=1}^{k} n_i[\log(n_i/n) - \log \theta_i(\widehat{\boldsymbol{\eta}})].$$

If $\boldsymbol{\eta} \to \theta(\boldsymbol{\eta})$ is differentiable in each coordinate, $\mathcal{E}$ is open, and $\widehat{\boldsymbol{\eta}}$ exists, then it must solve the likelihood equation for the model, $\{p(\cdot, \boldsymbol{\theta}(\boldsymbol{\eta})) : \boldsymbol{\eta} \in \mathcal{E}\}$. That is, $\widehat{\boldsymbol{\eta}}$ satisfies

$$\frac{\partial}{\partial \eta_j} \log p(n_1, \dots, n_k, \boldsymbol{\theta}(\boldsymbol{\eta})) = 0, \ 1 \le j \le q \tag{6.4.3}$$

or

$$\sum_{i=1}^{k} \frac{n_i}{\theta_i(\boldsymbol{\eta})} \frac{\partial}{\partial \eta_j} \theta_i(\boldsymbol{\eta}) = 0, \ 1 \le j \le q.$$

If $\mathcal{E}$ is not open sometimes the closure of $\mathcal{E}$ will contain a solution of (6.4.3).

To apply the results of Section 6.3 and conclude that, under H, $2 \log \lambda$ approximately has a χ^2_{r-q} distribution for large n, we define $\theta'_j = g_j(\boldsymbol{\theta})$, $j = 1, \dots, r$, where g_j is chosen so that H becomes equivalent to "$(\theta'_1, \dots, \theta'_q)^T$ ranges over an open subset of R^q and $\theta_j = \theta_{0j}$, $j = q + 1, \dots, r$ for specified θ_{0j}." For instance, to test the Hardy–Weinberg model we set $\theta'_1 = \theta_1$, $\theta'_2 = \theta_2 - 2\sqrt{\theta_1}(1 - \sqrt{\theta_1})$ and test $H : \theta'_2 = 0$. Then we can conclude from Theorem 6.3.3 that $2 \log \lambda$ approximately has a χ^2_1 distribution under H.

The Rao statistic is also invariant under reparametrization and, thus, approximately χ^2_{r-q}. Moreover, we obtain the Rao statistic for the composite multinomial hypothesis by replacing θ_{0j} in (6.4.2) by $\theta_j(\widehat{\boldsymbol{\eta}})$. The algebra showing $R_n(\boldsymbol{\theta}_0) = \chi^2$ in Section 6.4.1 now leads to the Rao statistic

$$R_n(\boldsymbol{\theta}(\widehat{\boldsymbol{\eta}})) = \sum_{j=1}^{k} \frac{[N_i - n\theta_j(\widehat{\boldsymbol{\eta}})]^2}{n\theta_j(\widehat{\boldsymbol{\eta}})} = \chi^2$$

where the right-hand side is Pearson's χ^2 as defined in general by (6.4.1).

The Wald statistic is only asymptotically invariant under reparametrization. However, the Wald statistic based on the parametrization $\boldsymbol{\theta}(\boldsymbol{\eta})$ obtained by replacing θ_{0j} by $\theta_j(\widehat{\boldsymbol{\eta}})$, is, by the algebra of Section 6.4.1, also equal to Pearson's χ^2.

Example 6.4.4. *Hardy–Weinberg.* We found in Example 2.2.6 that $\widehat{\eta} = (2n_1 + n_2)/2n$. Thus, H is rejected if $\chi^2 \geq x_1(1-\alpha)$ with

$$\boldsymbol{\theta}(\widehat{\eta}) = \left(\left(\frac{2n_1+n_2}{2n}\right)^2, \frac{(2n_1+n_2)(2n_3+n_2)}{2n^2}, \left(\frac{2n_3+n_2}{2n}\right)^2\right)^T.$$

□

Example 6.4.5. *The Fisher Linkage Model.* A self-crossing of maize heterozygous on two characteristics (starchy versus sugary; green base leaf versus white base leaf) leads to four possible offspring types: (1) sugary-white; (2) sugary-green; (3) starchy-white; (4) starchy-green. If N_i is the number of offspring of type i among a total of n offspring, then $(N_1, \ldots, N_4)$ has a $\mathcal{M}(n, \theta_1, \ldots, \theta_4)$ distribution. A linkage model (Fisher, 1958, p. 301), specifies that

$$\theta_1 = \frac{1}{4}(2+\eta),\ \theta_2 = \theta_3 = \frac{1}{4}(1-\eta),\ \theta_4 = \frac{1}{4}\eta$$

where η is an unknown number between 0 and 1. To test the validity of the linkage model we would take $\Theta_0 = \left\{\left(\frac{1}{4}(2+\eta), \frac{1}{4}(1-\eta), \frac{1}{4}(1-\eta), \frac{1}{4}\eta\right) : 0 \leq \eta \leq 1\right\}$ a "one-dimensional curve" of the three-dimensional parameter space Θ.

The likelihood equation (6.4.3) becomes

$$\frac{n_1}{(2+\eta)} - \frac{(n_2+n_3)}{(1-\eta)} + \frac{n_4}{\eta} = 0, \tag{6.4.4}$$

which reduces to a quadratic equation in $\widehat{\eta}$. The only root of this equation in $[0,1]$ is the desired estimate (see Problem 6.4.1). Because $q=1$, $k=4$, we obtain critical values from the χ_2^2 tables. □

Testing Independence of Classifications in Contingency Tables

Many important characteristics have only two categories. An individual either is or is not inoculated against a disease; is or is not a smoker; is male or female; and so on. We often want to know whether such characteristics are linked or are independent. For instance, do smoking and lung cancer have any relation to each other? Are sex and admission to a university department independent classifications? Let us call the possible categories or states of the first characteristic A and $\bar{A}$ and of the second B and $\bar{B}$. Then a randomly selected individual from the population can be one of four types $AB, A\bar{B}, \bar{A}B, \bar{A}\bar{B}$. Denote the probabilities of these types by $\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}$, respectively. Independent classification then means that the events [being an A] and [being a B] are independent or in terms of the θ_{ij},

$$\theta_{ij} = (\theta_{i1} + \theta_{i2})(\theta_{1j} + \theta_{2j}).$$

To study the relation between the two characteristics we take a random sample of size n from the population. The results are assembled in what is called a 2×2 *contingency table* such as the one shown.

	B	$\bar{B}$
A	N_{11}	N_{12}
$\bar{A}$	N_{21}	N_{22}

The entries in the boxes of the table indicate the number of individuals in the sample who belong to the categories of the appropriate row and column. Thus, for example N_{12} is the number of sampled individuals who fall in category A of the first characteristic and category $\bar{B}$ of the second characteristic. Then, if $\mathbf{N} = (N_{11}, N_{12}, N_{21}, N_{22})^T$, we have $\mathbf{N} \sim \mathcal{M}(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$. We test the hypothesis $H : \boldsymbol{\theta} \in \Theta_0$ versus $K : \theta \notin \Theta_0$, where Θ_0 is a two-dimensional subset of Θ given by

$$\Theta_0 = \{(\eta_1\eta_2, \eta_1(1-\eta_2), \eta_2(1-\eta_1), (1-\eta_1)(1-\eta_2)) : 0 \leq \eta_1 \leq 1,\ 0 \leq \eta_2 \leq 1\}.$$

Here we have relabeled $\theta_{11} + \theta_{12}, \theta_{11} + \theta_{21}$ as η_1, η_2 to indicate that these are parameters, which vary freely.

For $\boldsymbol{\theta} \in \Theta_0$, the likelihood equations $(6.4.3)$ become

$$\begin{aligned} \frac{(n_{11}+n_{12})}{\widehat{\eta}_1} &= \frac{(n_{21}+n_{22})}{(1-\widehat{\eta}_1)} \\ \frac{(n_{11}+n_{21})}{\widehat{\eta}_2} &= \frac{(n_{12}+n_{22})}{(1-\widehat{\eta}_2)} \end{aligned} \tag{6.4.5}$$

whose solutions are

$$\begin{aligned} \widehat{\eta}_1 &= (n_{11}+n_{12})/n \\ \widehat{\eta}_2 &= (n_{11}+n_{21})/n, \end{aligned} \tag{6.4.6}$$

the proportions of individuals of type A and type B, respectively. These solutions are the maximum likelihood estimates. Pearson's statistic is then easily seen to be

$$\chi^2 = n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(N_{ij} - R_iC_j/n)^2}{R_iC_j}, \tag{6.4.7}$$

where $R_i = N_{i1} + N_{i2}$ is the ith row sum, $C_j = N_{1j} + N_{2j}$ is the jth column sum.

By our theory if H is true, because $k = 4$, $q = 2$, χ^2 has approximately a χ_1^2 distribution. This suggests that χ^2 may be written as the square of a single (approximately) standard normal variable. In fact (Problem 6.4.2), the $(N_{ij} - R_iC_j/n)$ are all the same in absolute value and,

$$\chi^2 = Z^2$$

where

$$\begin{aligned} Z &= \left(N_{11} - \frac{R_1C_1}{n}\right)\sqrt{\sum_{i=1}^{2}\sum_{j=1}^{2}\left[\frac{R_iC_j}{n}\right]^{-1}} \\ &= \left(N_{11} - \frac{R_1C_1}{n}\right)\left[\frac{R_1R_2C_1C_2}{n^3}\right]^{-1/2}. \end{aligned}$$

An important alternative form for Z is given by

$$Z = \left(\frac{N_{11}}{C_1} - \frac{N_{12}}{C_2}\right)\sqrt{\frac{C_1C_2n}{R_1R_2}}. \tag{6.4.8}$$

Thus,

$$Z = \sqrt{n}[\widehat{P}(A \mid B) - \widehat{P}(A \mid \bar{B})]\left[\frac{\widehat{P}(B)}{\widehat{P}(A)}\frac{\widehat{P}(\bar{B})}{\widehat{P}(\bar{A})}\right]^{1/2}$$

where $\widehat{P}$ is the empirical distribution and where we use $A, B, \bar{A}, \bar{B}$ to denote the event that a randomly selected individual has characteristic $A, B, \bar{A}, \bar{B}$. Thus, if χ^2 measures deviations from independence, Z indicates what directions these deviations take. Positive values of Z indicate that A and B are positively associated (i.e., that A is more likely to occur in the presence of B than it would in the presence of $\bar{B}$). It may be shown (Problem 6.4.3) that if A and B are independent, that is, $P(A \mid B) = P(A \mid \bar{B})$, then Z is approximately distributed as $\mathcal{N}(0,1)$. Therefore, it is reasonable to use the test that rejects, if and only if,

$$Z \geq z(1-\alpha)$$

as a level α one-sided test of $H : P(A \mid B) = P(A \mid \bar{B})$ (or $P(A \mid B) \leq P(A \mid \bar{B})$) versus $K : P(A \mid B) > P(A \mid \bar{B})$. The χ^2 test is equivalent to rejecting (two-sidedly) if, and only if,

$$|Z| \geq z\left(1 - \frac{\alpha}{2}\right).$$

Next we consider contingency tables for two nonnumerical characteristics having a and b states, respectively, $a, b \geq 2$ (e.g., eye color, hair color). If we take a sample of size n from a population and classify them according to each characteristic we obtain a vector N_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, b$ where N_{ij} is the number of individuals of type i for characteristic 1 and j for characteristic 2. If $\theta_{ij} = P[A$ randomly selected individual is of type i for 1 and j for 2], then

$$\{N_{ij} : 1 \leq i \leq a,\ 1 \leq j \leq b\} \sim \mathcal{M}(n, \theta_{ij} : 1 \leq i \leq a,\ 1 \leq j \leq b).$$

The hypothesis that the characteristics are assigned independently becomes $H : \theta_{ij} = \eta_{i1}\eta_{j2}$ for $1 \leq i \leq a$, $1 \leq j \leq b$ where the η_{i1}, η_{j2} are nonnegative and $\sum_{i=1}^{a} \eta_{i1} = \sum_{j=1}^{b} \eta_{j2} = 1$.

The N_{ij} can be arranged in a $a \times b$ contingency table,

	1	2	$\cdots$	b	
1	N_{11}	N_{12}	$\cdots$	N_{1b}	R_1
$\vdots$	$\vdots$	$\vdots$	$\cdots$ $\cdots$	$\vdots$	$\vdots$
a	N_{a1}	$\cdot$	$\cdots$	N_{ab}	R_a
	C_1	C_2	$\cdots$	C_b	n

with row and column sums as indicated. Maximum likelihood and dimensionality calculations similar to those for the 2×2 table show that Pearson's χ^2 for the hypothesis of independence is given by

$$\chi^2 = n \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\left(N_{ij} - \frac{R_i C_j}{n}\right)^2}{R_i C_j}, \tag{6.4.9}$$

which has approximately a $\chi^2_{(a-1)(b-1)}$ distribution under H. The argument is left to the problems as are some numerical applications.

6.4.3 Logistic Regression for Binary Responses

In Section 6.1 we considered linear models that are appropriate for analyzing continuous responses $\{Y_i\}$ that are, perhaps after a transformation, approximately normally distributed and whose means are modeled as $\mu_i = \sum_{j=1}^{p} z_{ij}\beta_j = \mathbf{z}_i^T \boldsymbol{\beta}$ for known constants $\{z_{ij}\}$ and unknown parameters $\beta_1, \dots, \beta_p$. In this section we will consider Bernoulli responses Y that can only take on the values 0 and 1. Examples are (1) medical trials where at the end of the trial the patient has either recovered $(Y = 1)$ or has not recovered $(Y = 0)$, (2) election polls where a voter either supports a proposition $(Y = 1)$ or does not $(Y = 0)$, or (3) market research where a potential customer either desires a new product $(Y = 1)$ or does not $(Y = 0)$. As is typical, we call $Y = 1$ a "success" and $Y = 0$ a "failure."

We assume that the distribution of the response Y depends on the known covariate vector $\mathbf{z}^T$. In this section we assume that the data are grouped or replicated so that for each fixed i, we observe the number of successes $X_i = \sum_{j=1}^{m_i} Y_{ij}$ where Y_{ij} is the response on the jth of the m_i trials in block i, $1 \le i \le k$. Thus, we observe independent $X_1, \dots, X_k$ with X_i binomial, $\mathcal{B}(m_i, \pi_i)$, where $\pi_i = \pi(\mathbf{z}_i)$ is the probability of success for a case with covariate vector $\mathbf{z}_i$. Next we choose a parametric model for $\pi(\mathbf{z})$ that will generate useful procedures for analyzing experiments with binary responses. Because $\pi(\mathbf{z})$ varies between 0 and 1, a simple linear representation $\mathbf{z}^T\boldsymbol{\beta}$ for $\pi(\cdot)$ over the whole range of $\mathbf{z}$ is impossible. Instead we turn to the *logistic transform* $g(\pi)$, usually called the *logit*, which we introduced in Example 1.6.8 as the canonical parameter

$$\eta = g(\pi) = \log[\pi/(1-\pi)]. \tag{6.4.10}$$

Other transforms, such as the *probit* $g_1(\pi) = \Phi^{-1}(\pi)$ where Φ is the $\mathcal{N}(0,1)$ d.f. and the *log-log transform* $g_2(\pi) = \log[-\log(1-\pi)]$ are also used in practice.

The log likelihood of $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)^T$ based on $\mathbf{X} = (X_1, \dots, X_k)^T$ is

$$\sum_{i=1}^{k} \left[X_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i)\right] + \sum_{i=1}^{k} \log \begin{pmatrix} m_i \\ X_i \end{pmatrix}. \tag{6.4.11}$$

When we use the logit transform $g(\pi)$, we obtain what is called the *logistic linear regression model* where

$$\eta_i = \log[\pi_i/(1-\pi_i)] = \mathbf{z}_i^T \boldsymbol{\beta}.$$

The special case $p = 2$, $\mathbf{z}_i = (z_{i1}, z_{i2})^T = (1, z_{i2})^T$, $1 \le i \le k$, is the *logistic regression* model of Problem 2.3.1. The log likelihood of $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is, if $N = \sum_{i=1}^k m_i$,

$$l_N(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j T_j - \sum_{i=1}^k m_i \log(1 + \exp\{\mathbf{z}_i \boldsymbol{\beta}\}) + \sum_{i=1}^k \log \begin{pmatrix} m_i \\ X_i \end{pmatrix} \tag{6.4.12}$$

where $T_j = \sum_{i=1}^k z_{ij} X_i$, $1 \le j \le p$, and we make the dependence on N explicit. Note that $l_N(\boldsymbol{\beta})$ is the log likelihood of a p-parameter canonical exponential model with parameter vector $\boldsymbol{\beta}$ and sufficient statistic $\mathbf{T} = (T_1, \dots, T_p)^T$. It follows that the MLE of $\boldsymbol{\beta}$ solves $E_{\boldsymbol{\beta}}(T_j) = T_j$, $j = 1, \dots, p$, or $E_{\boldsymbol{\beta}}(\mathbf{Z}^T \mathbf{X}) = \mathbf{Z}^T \mathbf{X}$, where $\mathbf{Z} = \|z_{ij}\|_{k \times p}$ is the design matrix. Thus, Theorem 2.3.1 applies and we can conclude that if $0 < X_i < m_i$ and $\mathbf{Z}$ has rank p, the solution to this equation exists and gives the unique MLE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. The condition is sufficient but not necessary for existence—see Problem 2.3.1. We let $\mu_i = E(X_i) = m_i \pi_i$. Then $E(T_j) = \sum_{i=1}^k z_{ij} \mu_i$, the likelihood equations are just

$$\mathbf{Z}^T(\mathbf{X} - \boldsymbol{\mu}) = 0. \tag{6.4.13}$$

By Theorem 2.3.1 and by Proposition 3.4.4 the Fisher information matrix is

$$I(\boldsymbol{\beta}) = \mathbf{Z}^T \mathbf{W} \mathbf{Z} \tag{6.4.14}$$

where $\mathbf{W} = \text{diag}\{m_i \pi_i (1 - \pi_i)\}_{k \times k}$. The coordinate ascent iterative procedure of Section 2.4.2 can be used to compute the MLE of $\boldsymbol{\beta}$.

Alternatively with a good initial value the Newton–Raphson algorithm can be employed. Although unlike coordinate ascent Newton–Raphson need not converge, we can guarantee convergence with probability tending to 1 as $N \to \infty$ as follows. As the initial estimate use

$$\widehat{\boldsymbol{\beta}}_0 = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{V} \tag{6.4.15}$$

where $\mathbf{V} = (V_1, \dots, V_k)^T$ with

$$V_i = \log \left(\frac{X_i + \frac{1}{2}}{m_i - X_i + \frac{1}{2}} \right), \tag{6.4.16}$$

the *empirical logistic transform*. Because $\boldsymbol{\beta} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \boldsymbol{\eta}$ and $\eta_i = \log[\pi_i (1 - \pi_i)^{-1}]$, $\widehat{\boldsymbol{\beta}}_0$ is a plug-in estimate of β where π_i and $(1 - \pi_i)$ in η_i has been replaced by

$$\pi_i^* = \frac{X_i}{m_i} + \frac{1}{2m_i}, \ (1 - \pi_i)^* = 1 - \frac{X_i}{m_i} + \frac{1}{2m_i}.$$

Here the adjustment $1/2m_i$ is used to avoid $\log 0$ and $\log \infty$. Similarly, in (6.4.14), $\mathbf{W}$ is estimated using

$$\widehat{\mathbf{W}}_0 = \text{diag}\{m_i \pi_i^* (1 - \pi_i)^*\}.$$

Using the δ-method, it follows by Theorem 5.3.3, that if $m \to \infty$, $\pi_i > 0$ for $1 \le i \le k$

$$m_i^{\frac{1}{2}} \left(V_i - \log \left(\frac{\pi_i}{1 - \pi_i} \right) \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, [\pi_i(1 - \pi_i)]^{-1}).$$

Because $\mathbf{Z}$ has rank p, it follows (Problem 6.4.14) that $\widehat{\boldsymbol{\beta}}_0$ is consistent.

To get expressions for the MLEs of $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$, recall from Example 1.6.8 that the inverse of the logit transform g is the logistic distribution function

$$g^{-1}(y) = (1+e^{-y})^{-1}.$$

Thus, the MLE of π_i is $\widehat{\pi}_i = g^{-1}\left(\sum_{j=1}^{p} x_{ij}\widehat{\beta}_j\right)$.

Testing

In analogy with Section 6.1, we let $\omega = \{\boldsymbol{\eta} : \eta_i = \mathbf{z}_i^T\boldsymbol{\beta},\ \boldsymbol{\beta} \in R^p\}$ and let r be the dimension of ω. We want to contrast ω to the case where there are no restrictions on $\boldsymbol{\eta}$; that is, we set $\Omega = R^k$ and consider $\boldsymbol{\eta} \in \Omega$. In this case the likelihood is a product of independent binomial densities, and the MLEs of π_i and μ_i are X_i/m_i and X_i. The LR statistic $2\log\lambda$ for testing $H : \boldsymbol{\eta} \in \omega$ versus $K : \boldsymbol{\eta} \in \Omega - \omega$ is denoted by $D(\mathbf{y}, \widehat{\boldsymbol{\mu}})$, where $\widehat{\boldsymbol{\mu}}$ is the MLE of $\boldsymbol{\mu}$ for $\boldsymbol{\eta} \in \omega$. Thus, from (6.4.11) and (6.4.12)

$$D(\mathbf{X}, \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{k}[X_i \log(X_i/\widehat{\mu}_i) + X_i' \log(X_i'/\widehat{\mu}_i')] \tag{6.4.17}$$

where $X_i' = m_i - X_i$ and $\widehat{\mu}_i' = m_i - \widehat{\mu}_i$. $D(\mathbf{X}, \widehat{\boldsymbol{\mu}})$ measures the distance between the fit $\widehat{\boldsymbol{\mu}}$ based on the model ω and the data $\mathbf{X}$. By the multivariate delta method, Theorem 5.3.4, $D(\mathbf{X}, \widehat{\mu})$ has asymptotically a χ^2_{k-r} distribution for $\boldsymbol{\eta} \in \omega$ as $m_i \to \infty$, $i = 1, \dots, k < \infty$—see Problem 6.4.13.

As in Section 6.1 linear subhypotheses are important. If ω_0 is a q-dimensional linear subspace of ω with $q < r$, then we can form the LR statistic for $H : \boldsymbol{\eta} \in \omega_0$ versus $K : \boldsymbol{\eta} \in \omega - \omega_0$

$$2\log\lambda = 2\sum_{i=1}^{k}\left[X_i \log\left(\frac{\widehat{\mu}_i}{\widehat{\mu}_{0i}}\right) + X_i' \log\left(\frac{\widehat{\mu}_i'}{\widehat{\mu}_{0i}'}\right)\right] \tag{6.4.18}$$

where $\widehat{\boldsymbol{\mu}}_0$ is the MLE of $\boldsymbol{\mu}$ under H and $\widehat{\mu}_{0i}' = m_i - \widehat{\mu}_{0i}$. In the present case, by Problem 6.4.13, $2\log\lambda$ has an asymptotic χ^2_{r-q} distribution as $m_i \to \infty$, $i = 1, \dots, k$. Here is a special case.

Example 6.4.1. *The Binomial One-Way Layout.* Suppose that k treatments are to be tested for their effectiveness by assigning the ith treatment to a sample of m_i patients and recording the number X_i of patients that recover, $i = 1, \dots, k$. The samples are collected independently and we observe $X_1, \dots, X_k$ independent with $X_i \sim \mathcal{B}(\pi_i, m_i)$. For a second example, suppose we want to compare k different locations with respect to the percentage that have a certain attribute such as the intention to vote for or against a certain proposition. We obtain k independent samples, one from each location, and for the ith location count the number X_i among m_i that has the given attribute.

This model corresponds to the one-way layout of Section 6.1, and as in that section, an important hypothesis is that the populations are homogenous. Thus, we test $H: \pi_1 = \pi_2 = \cdots = \pi_k = \pi$, $\pi \in (0,1)$, versus the alternative that the π's are not all equal. Under H the log likelihood in canonical exponential form is

$$\beta T - N \log(1 + \exp\{\beta\}) + \sum_{i=1}^{k} \log \begin{pmatrix} m_i \\ X_i \end{pmatrix}$$

where $T = \sum_{i=1}^k X_i$, $N = \sum_{i=1}^k m_i$, and $\beta = \log[\pi/(1-\pi)]$. It follows from Theorem 2.3.1 that if $0 < T < N$ the MLE exists and is the solution of (6.4.13), where $\boldsymbol{\mu} = (m_1\pi, \dots, m_k\pi)^T$. Using $\mathbf{Z}$ as given in the one-way layout in Example 6.1.3, we find that the MLE of π under H is $\widehat{\pi} = T/N$. The LR statistic is given by (6.4.18) with $\widehat{\mu}_{0i} = m_i\widehat{\pi}$. The Pearson statistic

$$\chi^2 = \sum_{i=1}^{k-1} \frac{(X_i - m_i\widehat{\pi})^2}{m_i\widehat{\pi}(1-\widehat{\pi})}$$

is a Wald statistic and the χ^2 test is equivalent asymptotically to the LR test (Problem 6.4.15).

Summary. We used the large sample testing results of Section 6.3 to find tests for important statistical problems involving discrete data. We found that for testing the hypothesis that a multinomial parameter equals a specified value, the Wald and Rao statistics take a form called "Pearson's χ^2," which equals the sum of standardized squared distances between observed frequencies and expected frequencies under H. When the hypothesis is that the multinomial parameter is in a q-dimensional subset of the $k-1$-dimensional parameter space Θ, the Rao statistic is again of the Pearson χ^2 form. In the special case of testing independence of multinomial frequencies representing classifications in a two-way contingency table, the Pearson statistic is shown to have a simple intuitive form. Finally, we considered logistic regression for binary responses in which the logit transformation of the probability of success is modeled to be a linear function of covariates. We derive the likelihood equations, discuss algorithms for computing MLEs, and give the LR test. In the special case of testing equality of k binomial parameters, we give explicitly the MLEs and χ^2 test.

6.5 GENERALIZED LINEAR MODELS

In Sections 6.1 and 6.4.3 we considered experiments in which the mean μ_i of a response Y_i is expressed as a function of a linear combination

$$\xi_i \equiv \mathbf{z}_i^T \boldsymbol{\beta} = \sum_{j=1}^{p} z_{ij}\beta_j$$

of covariate values. In particular, in the case of a Gaussian response, $\mu_i = \xi_i$. In Section 6.4.3, if $\mu_i = EY_{ij}$, then $\mu_i = \pi_i = g^{-1}(\xi_i)$, where $g^{-1}(y)$ is the logistic distribution function. More generally, McCullagh and Nelder (1983, 1989) synthesized a number of previous generalizations of the linear model, most importantly the log linear model developed by Goodman and Haberman. See Haberman (1974).

The generalized linear model with dispersion depending only on the mean

The data consist of an observation $(\mathbf{Z}, \mathbf{Y})$ where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is $n \times 1$ and $\mathbf{Z}^T_{p \times n} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})^T$ nonrandom and $\mathbf{Y}$ has density $p(\mathbf{y}, \boldsymbol{\eta})$ given by

$$p(\mathbf{y}, \boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^T \mathbf{y} - A(\boldsymbol{\eta})\} h(\mathbf{y}) \tag{6.5.1}$$

where $\boldsymbol{\eta}$ is not in $\mathcal{E}$, the natural parameter space of the n-parameter canonical exponential family $(6.5.1)$, but in a subset of $\mathcal{E}$ obtained by restricting η_i to be of the form

$$\eta_i = h(\mathbf{z}_i, \boldsymbol{\beta}),\ \boldsymbol{\beta} \in \mathcal{B} \subset R^p,\ p \le n$$

where h is a known function. As we know from Corollary $1.6.1$, the mean $\boldsymbol{\mu}$ of $\mathbf{Y}$ is related to $\boldsymbol{\eta}$ via $\boldsymbol{\mu} = \dot{A}(\boldsymbol{\eta})$. Typically, $A(\boldsymbol{\eta}) = \sum_{i=1}^n A_0(\eta_i)$ for some A_0, in which case $\mu_i = A_0'(\eta_i)$. We assume that there is a one-to-one transform $\mathbf{g}(\boldsymbol{\mu})$ of $\boldsymbol{\mu}$, called the *link function*, such that

$$\mathbf{g}(\boldsymbol{\mu}) = \sum_{j=1}^p \beta_j \mathbf{Z}^{(j)} = \mathbf{Z}\boldsymbol{\beta}$$

where $\mathbf{Z}^{(j)} = (z_{1j}, \dots, z_{nj})^T$ is the jth column vector of $\mathbf{Z}$. Note that if $\dot{A}$ is one-one, $\boldsymbol{\mu}$ determines $\boldsymbol{\eta}$ and thereby $\text{Var}(\mathbf{Y}) = \ddot{A}(\boldsymbol{\eta})$. Typically, $\mathbf{g}(\boldsymbol{\mu})$ is of the form $(g(\mu_1), \dots, g(\mu_n))^T$, in which case g is also called the link function.

Canonical links

The most important case corresponds to the link being *canonical*; that is, $\mathbf{g} = \dot{A}^{-1}$ or

$$\boldsymbol{\eta} = \sum_{j=1}^p \beta_j \mathbf{Z}^{(j)} = \mathbf{Z}\boldsymbol{\beta}.$$

In this case, the GLM is the canonical subfamily of the original exponential family generated by $\mathbf{Z}^T\mathbf{Y}$, which is $p \times 1$.

Special cases are:

(i) *The linear model with known variance.*

These Y_i are independent Gaussian with known variance 1 and means $\mu_i = \sum_{j=1}^p Z_{ij}\beta_j$. The model is GLM with canonical $\mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\mu}$, the identity.

(ii) *Log linear models.*

Suppose $(Y_1, \dots, Y_p)^T$ is $\mathcal{M}(n, \theta_1, \dots, \theta_p)$, $\theta_j > 0$, $1 \le j \le p$, $\sum_{j=1}^p \theta = 1$. Then, as seen in Example 1.6.7,

$$\eta_j = \log \theta_j,\ 1 \le j \le p$$

are canonical parameters. If we take $\mathbf{g}(\boldsymbol{\mu}) = (\log \mu_1, \dots, \log \mu_p)^T$, the link is canonical. The models we obtain are called *log linear*—see Haberman (1974) for an extensive treatment. Suppose, for example, that $\mathbf{Y} = \|Y_{ij}\|_{1 \le i \le a}$, $1 \le j \le b$, so that Y_{ij} is the indicator of, say, classification i on characteristic 1 and j on characteristic 2. Then

$$\boldsymbol{\theta} = \|\theta_{ij}\|,\ 1 \le i \le a,\ 1 \le j \le b,$$

and the log linear model corresponding to

$$\log \theta_{ij} = \beta_i + \beta_j,$$

where β_i, β_j are free (unidentifiable parameters), is that of independence

$$\theta_{ij} = \theta_{i+}\theta_{+j}$$

where $\theta_{i+} = \sum_{j=1}^b \theta_{ij}$, $\theta_{+j} = \sum_{i=1}^b \theta_{ij}$.

The log linear label is also attached to models obtained by taking the Y_i independent Bernoulli (θ_i), $0 < \theta_i < 1$ with canonical link $g(\theta) = \log[\theta(1-\theta)]$. This is just the logistic linear model of Section 6.4.3. See Haberman (1974) for a further discussion.

Algorithms

If the link is canonical, by Theorem 2.3.1, if maximum likelihood estimates $\widehat{\boldsymbol{\beta}}$ exist, they necessarily uniquely satisfy the equation

$$\mathbf{Z}^T\mathbf{Y} = \mathbf{Z}^T E_{\widehat{\boldsymbol{\beta}}}\mathbf{Y} = \mathbf{Z}^T \dot{A}(\mathbf{Z}\widehat{\boldsymbol{\beta}})$$

or

$$\mathbf{Z}^T(\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})) = \mathbf{0}. \tag{6.5.2}$$

It's interesting to note that (6.5.2) can be interpreted geometrically in somewhat the same way as in the Gausian linear model—the "residual" vector $\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$ is orthogonal to the column space of $\mathbf{Z}$. But, in general, $\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$ is not a member of that space.

The coordinate ascent algorithm can be used to solve (6.5.2) (or ascertain that no solution exists). With a good starting point $\boldsymbol{\beta}_0$ one can achieve faster convergence with the Newton–Raphson algorithm of Section 2.4. In this case, that procedure is just

$$\widehat{\boldsymbol{\beta}}_{m+1} = \widehat{\boldsymbol{\beta}}_m + (\mathbf{Z}^T\mathbf{W}_m\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{Y} - \widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{\beta}}_m)) \tag{6.5.3}$$

where

$$\mathbf{W}_m = \ddot{A}(\mathbf{Z}\widehat{\boldsymbol{\beta}}_m).$$

In this situation and more generally even for noncanonical links, Newton–Raphson coincides with Fisher's method of scoring described in Problem 6.5.1. If $\widehat{\boldsymbol{\beta}}_0 \overset{P}{\to} \boldsymbol{\beta}_0$, the true value of $\boldsymbol{\beta}$, as $n \to \infty$, then with probability tending to 1 the algorithm converges to the MLE if it exists.

In this context, the algorithm is also called iterated weighted least squares. This name stems from the following interpretation. Let $\widehat{\Delta}_{m+1} \equiv \widehat{\boldsymbol{\beta}}_{m+1} - \widehat{\boldsymbol{\beta}}_m$, which satisfies the equation

$$\widehat{\boldsymbol{\Delta}}_{m+1} = (\mathbf{Z}^T \mathbf{W}_m \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W}_m (\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_m)). \tag{6.5.4}$$

That is, the correction $\widehat{\boldsymbol{\Delta}}_{m+1}$ is given by the weighted least squares formula $(2.2.20)$ when the data are the residuals from the fit at stage m, the variance covariance matrix is $\mathbf{W}_m$ *and* the regression is on the columns of $\mathbf{W}_m \mathbf{Z}$—Problem $6.5.2$.

Testing in GLM

Testing hypotheses in GLM is done via the LR statistic. As in the linear model we can define the biggest possible GLM $\mathcal{M}$ of the form $(6.5.1)$ for which $p = n$. In that case the MLE of μ is $\widehat{\boldsymbol{\mu}}_{\mathcal{M}} = (Y_1, \ldots, Y_n)^T$ (assume that $\mathbf{Y}$ is in the interior of the convex support of $\{y : p(\mathbf{y}, \boldsymbol{\eta}) > 0\}$). Write $\boldsymbol{\eta}(\cdot)$ for $\dot{A}^{-1}$.

We can think of the test statistic

$$2 \log \lambda = 2[l(\mathbf{Y}, \boldsymbol{\eta}(\mathbf{Y})) - l(\mathbf{Y}, \boldsymbol{\eta}(\boldsymbol{\mu}_0)]$$

for the hypothesis that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ within $\mathcal{M}$ as a "measure" of (squared) distance between $\mathbf{Y}$ and $\boldsymbol{\mu}_0$. This quantity, called the *deviance* between $\mathbf{Y}$ and $\boldsymbol{\mu}_0$,

$$D(\mathbf{Y}, \boldsymbol{\mu}_0) = 2\{[\boldsymbol{\eta}^T(\mathbf{Y}) - \boldsymbol{\eta}^T(\boldsymbol{\mu}_0)]\mathbf{Y} - [A(\boldsymbol{\eta}(\mathbf{Y})) - A(\boldsymbol{\eta}(\boldsymbol{\mu}_0))]\} \tag{6.5.5}$$

is always ≥ 0. For the Gaussian linear model with known variance σ_0^2

$$D(\mathbf{y}, \boldsymbol{\mu}_0) = |\mathbf{y} - \boldsymbol{\mu}_0|^2 / \sigma_0^2$$

(Problem $6.5.4$). The LR statistic for $H : \boldsymbol{\mu} \in \omega_0$ is just

$$D(\mathbf{Y}, \widehat{\boldsymbol{\mu}}_0) \equiv \inf\{D(\mathbf{Y}, \boldsymbol{\mu}) : \boldsymbol{\mu} \in \omega_0\}$$

where $\widehat{\boldsymbol{\mu}}_0$ is the MLE of $\boldsymbol{\mu}$ in ω_0. The LR statistic for $H : \boldsymbol{\mu} \in \omega_0$ versus $K : \boldsymbol{\mu} \in \omega_1 - \omega_0$ with $\omega_1 \supset \omega_0$ is

$$D(\mathbf{Y}, \widehat{\boldsymbol{\mu}}_0) - D(\mathbf{Y}, \widehat{\boldsymbol{\mu}}_1)$$

where $\widehat{\boldsymbol{\mu}}_1$ is the MLE under ω_1. We can then formally write an analysis of deviance analogous to the analysis of variance of Section 6.1. If $\omega_0 \subset \omega_1$ we can write

$$D(\mathbf{Y}, \widehat{\boldsymbol{\mu}}_0) = D(\mathbf{Y}, \widehat{\boldsymbol{\mu}}_1) + \Delta(\widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\mu}}_1) \tag{6.5.6}$$

a decomposition of the deviance between $\mathbf{Y}$ and $\widehat{\boldsymbol{\mu}}_0$ as the sum of two nonnegative components, the deviance of $\mathbf{Y}$ to $\widehat{\boldsymbol{\mu}}_1$ and $\Delta(\widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\mu}}_1) \equiv D(Y, \widehat{\boldsymbol{\mu}}_0) - D(Y, \widehat{\boldsymbol{\mu}}_1)$, each of which can be thought of as a squared distance between their arguments. Unfortunately $\Delta \neq D$ generally except in the Gaussian case.

Formally if ω_0 is a GLM of dimension p and ω_1 of dimension q with canonical links, then $\Delta(\widehat{\boldsymbol{\mu}}_0, \widehat{\boldsymbol{\mu}}_1)$ is thought of as being asymptotically χ^2_{p-q}. This can be made precise for stochastic GLMs obtained by conditioning on $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ in the sample $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ from the family with density

$$p(\mathbf{z}, y, \boldsymbol{\beta}) = h(y) q_0(\mathbf{z}) \exp\{(\mathbf{z}^T \boldsymbol{\beta})\mathbf{y} - A_0(\mathbf{z}^T \boldsymbol{\beta})\}. \tag{6.5.7}$$

More details are discussed in what follows.

Asymptotic theory for estimates and tests

If $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ can be viewed as a sample from a population and the link is canonical, the theory of Sections 6.2 and 6.3 applies straightforwardly in view of the general smoothness properties of canonical exponential families. Thus, if we take $\mathbf{Z}_i$ as having marginal density q_0, which we temporarily assume known, then $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ has density

$$p(\mathbf{z}, \mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^n h(y_i) q_0(\mathbf{z}_i) \exp\left\{ \sum_{i=1}^n [(\mathbf{z}_i^T \boldsymbol{\beta}) y_i - A_0(\mathbf{z}_i^T \boldsymbol{\beta})] \right\}. \tag{6.5.8}$$

This is *not* unconditionally an exponential family in view of the $A_0(\mathbf{z}_i^T \boldsymbol{\beta})$ term. However, there are easy conditions under which conditions of Theorems 6.2.2, 6.3.2, and 6.3.3 hold (Problem 6.5.3), so that the MLE $\widehat{\boldsymbol{\beta}}$ is unique, asymptotically exists, is consistent with probability 1, and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I^{-1}(\boldsymbol{\beta})). \tag{6.5.9}$$

What is $I^{-1}(\boldsymbol{\beta})$? The efficient score function $\frac{\partial}{\partial \beta_i} \log p(\mathbf{z}, y, \boldsymbol{\beta})$ is

$$(Y_i - \dot{A}_0(\mathbf{Z}_i^T \boldsymbol{\beta}))\mathbf{Z}_i^T$$

and so,

$$I(\boldsymbol{\beta}) = E(\mathbf{Z}_1^T \mathbf{Z}_1 \ddot{A}_0(\mathbf{Z}_1^T \boldsymbol{\beta})),$$

which, in order to obtain approximate confidence procedures, can be estimated by $\widehat{I} = \widehat{\boldsymbol{\Sigma}} \ddot{A}(\mathbf{Z}\widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\Sigma}}$ is the sample variance matrix of the covariates. For instance, if we assume the covariates in logistic regression with canonical link to be stochastic, we obtain

$$I(\boldsymbol{\beta}) = E(\mathbf{Z}_1^T \mathbf{Z}_1 \pi(\mathbf{Z}_1^T \boldsymbol{\beta}_1)(1 - \pi(\mathbf{Z}_1^T \boldsymbol{\beta})).$$

If we wish to test hypotheses such as $H : \beta_1 = \cdots = \beta_d = 0$, $d < p$, we can calculate

$$2 \log \lambda(\mathbf{Z}_i, Y_i : 1 \le i \le n) = 2 \sum_{i=1}^n [\mathbf{Z}_i^T (\widehat{\boldsymbol{\beta}}_H - \widehat{\boldsymbol{\beta}}) Y_i + A_0(\mathbf{Z}_i^T \widehat{\boldsymbol{\beta}}_H) - A(\mathbf{Z}_i^T \widehat{\boldsymbol{\beta}})] \tag{6.5.10}$$

where $\widehat{\boldsymbol{\beta}}_H$ is the $(p \times 1)$ MLE for the GLM with $\boldsymbol{\beta}^T_{p \times 1} = (0, \dots, 0, \beta_{d+1}, \dots, \beta_p)$, and can conclude that the statistic of $(6.5.10)$ is asymptotically χ^2_d under H. Similar conclusions follow for the Wald and Rao statistics. Note that these tests can be carried out without knowing the density q_0 of $\mathbf{Z}_1$.

These conclusions remain valid for the usual situation in which the $\mathbf{Z}_i$ are not random but their proof depends on asymptotic theory for independent nonidentically distributed variables, which we postpone to Volume II.

The generalized linear model

The GLMs considered so far force the variance of the response to be a function of its mean. An additional "dispersion" parameter can be introduced in some exponential family models by making the function h in $(6.5.1)$ depend on an additional scalar parameter τ. It is customary to write the model as, for $c(\tau) > 0$,

$$p(\mathbf{y}, \boldsymbol{\eta}, \tau) = \exp\{c^{-1}(\tau)(\boldsymbol{\eta}^T\mathbf{y} - A(\boldsymbol{\eta}))\}h(\mathbf{y}, \tau). \tag{6.5.11}$$

Because $\int p(\mathbf{y}, \boldsymbol{\eta}, \tau)d\mathbf{y} = 1$, then

$$A(\boldsymbol{\eta})/c(\tau) = \log \int \exp\{c^{-1}(\tau)\boldsymbol{\eta}^T\mathbf{y}\}h(\mathbf{y}, \tau)d\mathbf{y}. \tag{6.5.12}$$

The left-hand side of $(6.5.12)$ is of product form $A(\boldsymbol{\eta})[1/c(\tau)]$ whereas the right-hand side cannot always be put in this form. However, when it can, then it is easy to see that

$$E(\mathbf{Y}) = \dot{A}(\boldsymbol{\eta}) \tag{6.5.13}$$

$$\text{Var}(\mathbf{Y}) = c(\tau)\ddot{A}(\boldsymbol{\eta}) \tag{6.5.14}$$

so that the variance can be written as the product of a function of the mean and a general dispersion parameter.

Important special cases are the $\mathcal{N}(\mu, \sigma^2)$ and gamma (p, λ) families. For further discussion of this generalization see McCullagh and Nelder (1983, 1989).

General link functions

Links other than the canonical one can be of interest. For instance, if in the binary data regression model of Section 6.4.3, we take $g(\mu) = \Phi^{-1}(\mu)$ so that

$$\pi_i = \Phi(\mathbf{z}_i^T\boldsymbol{\beta})$$

we obtain the so-called *probit* model. Cox (1970) considers the variance stabilizing transformation

$$g(\mu) = \sin^{-1}(\sqrt{\mu}),\ 0 \leq \mu \leq 1,$$

which makes asymptotic analysis equivalent to that in the standard Gaussian linear model. As he points out, the results of analyses with these various transformations over the range $.1 \leq \mu \leq .9$ are rather similar. From the point of analysis for fixed n, noncanonical links can cause numerical problems because the models are now curved rather than canonical exponential families. Existence of MLEs and convergence of algorithm questions all become more difficult and so canonical links tend to be preferred.

Summary. We considered generalized linear models defined as a canonical exponential model where the mean vector of the vector $\mathbf{Y}$ of responses can be written as a function, called the link function, of a linear predictor of the form $\Sigma\beta_j\mathbf{Z}^{(j)}$, where the $\mathbf{Z}^{(j)}$ are observable covariate vectors and $\boldsymbol{\beta}$ is a vector of regression coefficients. We considered the canonical link function that corresponds to the model in which the canonical exponential model parameter equals the linear predictor. We discussed algorithms for computing MLEs of $\widehat{\boldsymbol{\beta}}$. In the random design case, we use the asymptotic results of the previous sections to develop large sample estimation results, confidence procedures, and tests.

6.6 ROBUSTNESS PROPERTIES AND SEMIPARAMETRIC MODELS

As we most recently indicated in Example 6.2.2, the distributional and implicit structural assumptions of parametric models are often suspect. In Example 6.2.2, we studied what procedures would be appropriate if the linearity of the linear model held but the error distribution failed to be Gaussian. We found that if we assume the error distribution f_0, which is symmetric about 0, the resulting MLEs for $\boldsymbol{\beta}$ optimal under f_0 continue to estimate $\boldsymbol{\beta}$ as defined by (6.2.19) even if the true errors are $\mathcal{N}(0,\sigma^2)$ or, in fact, had any distribution symmetric around 0 (Problem 6.2.5). That is, roughly speaking, if we consider the semiparametric model, $\mathcal{P}_1 = \{P : (\mathbf{Z}^T, Y) \sim P$ given by (6.2.19) with ϵ_i i.i.d. with density f for some f symmetric about $0\}$, then the LSE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is, under further mild conditions on P, still a consistent asymptotically normal estimate of $\boldsymbol{\beta}$ and, in fact, so is any estimate solving the equations based on (6.2.31) with f_0 symmetric about 0. There is another semiparametric model $\mathcal{P}_2 = \{P : (\mathbf{Z}^T, Y)^T \sim P$ that satisfies $E_P(Y \mid \mathbf{Z}) = \mathbf{Z}^T\boldsymbol{\beta}$, $E_P\mathbf{Z}^T\mathbf{Z}$ nonsingular, $E_PY^2 < \infty\}$. For this model it turns out that the LSE is optimal in a sense to be discussed in Volume II. Furthermore, if $\mathcal{P}_3$ is the nonparametric model where we assume only that $(\mathbf{Z}, Y)$ has a joint distribution and if we are interested in estimating the best linear predictor $\mu_L(\mathbf{Z})$ of Y given $\mathbf{Z}$, the right thing to do if one assumes the $(\mathbf{Z}_i, Y_i)$ are i.i.d., is "act as if the model were the one given in Example 6.1.2." Of course, for estimating $\mu_L(\mathbf{Z})$ in a submodel of $\mathcal{P}_3$ with

$$Y = \mathbf{Z}^T\boldsymbol{\beta} + \sigma_0(\mathbf{Z})\epsilon \tag{6.6.1}$$

where ϵ is independent of $\mathbf{Z}$ but $\sigma_0(\mathbf{Z})$ is not constant and σ_0 is assumed known, the LSE is not the best estimate of $\boldsymbol{\beta}$. These issues, whose further discussion we postpone to Volume II, have a fixed covariate exact counterpart, the Gauss–Markov theorem, as discussed below. Another even more important set of questions having to do with selection between nested models of different dimension are touched on in Problem 6.6.8 but otherwise also postponed to Volume II.

Robustness in Estimation

We drop the assumption that the errors $\epsilon_1, \ldots, \epsilon_n$ in the linear model $(6.1.3)$ are normal. Instead we assume the *Gauss–Markov linear model* where

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\epsilon}) = \mathbf{0},\ \text{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{J} \tag{6.6.2}$$

where $\mathbf{Z}$ is an $n \times p$ matrix of constants, $\boldsymbol{\beta}$ is $p \times 1$, and $\mathbf{Y}, \boldsymbol{\epsilon}$ are $n \times 1$. The optimality of the estimates $\widehat{\mu}_i$, $\widehat{\beta}_i$ of Section $6.1.1$ and the LSE in general still holds *when they are compared to other linear estimates.*

Theorem 6.6.1. *Suppose the Gauss–Markov linear model* $(6.6.2)$ *holds. Then, for any parameter of the form* $\alpha = \sum_{i=1}^n a_i\mu_i$ *for some constants* $a_1, \ldots, a_n$, *the estimate* $\widehat{\alpha} = \sum_{i=1}^n a_i\widehat{\mu}_i$ *has uniformly minimum variance among all unbiased estimates linear in* Y_1, $\ldots, Y_n$.

Proof. Because $E(\epsilon_i) = 0$, $E(\widehat{\alpha}) = \sum_{i=1}^n a_i\mu_i = \alpha$, and $\widehat{\alpha}$ is unbiased. Moreover, $\text{Cov}(Y_i, Y_j) = \text{Cov}(\epsilon_i, \epsilon_j)$, and by (B.5.6),

$$\text{Var}_{GM}(\widehat{\alpha}) = \sum_{i=1}^n a_i^2 \text{ Var}(\epsilon_i) + 2\sum_{i<j} a_i a_j \text{ Cov}(\epsilon_i, \epsilon_j) = \sigma^2 \sum_{i=1}^n a_i^2$$

where Var_{GM} refers to the variance computed under the Gauss–Markov assumptions. Let $\widetilde{\alpha}$ stand for any estimate linear in $Y_1, \ldots, Y_n$. The preceding computation shows that $\text{Var}_{GM}(\widetilde{\alpha}) = \text{Var}_G(\widetilde{\alpha})$, where $\text{Var}_G(\widetilde{\alpha})$ stands for the variance computed under the Gaussian assumption that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. By Theorems 6.1.2(iv) and 6.1.3(iv), $\text{Var}_G(\widehat{\alpha}) \leq \text{Var}_G(\widetilde{\alpha})$ for all unbiased $\widetilde{\alpha}$. Because $\text{Var}_G = \text{Var}_{GM}$ for all linear estimators, the result follows. □

Note that the preceding result and proof are similar to Theorem $1.4.4$ where it was shown that the optimal linear predictor in the random design case is the same as the optimal predictor in the multivariate normal case. In fact, in Example 6.1.2, our current $\widehat{\mu}$ coincides with the empirical plug-in estimate of the optimal linear predictor $(1.4.14)$. See Problem 6.1.3.

Many of the properties stated for the Gaussian case carry over to the Gauss–Markov case:

Proposition 6.6.1. *If we replace the Gaussian assumptions on the errors* $\epsilon_1, \ldots, \epsilon_n$ *with the Gauss–Markov assumptions, the conclusions* (1) *and* (2) *of Theorem* $6.1.4$ *are still valid; moreover,* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\mu}}$ *are still unbiased and* $\text{Var}(\widehat{\mu}) = \sigma^2\mathbf{H}$, $\text{Var}(\widehat{\epsilon}) = \sigma^2(\mathbf{I} - \mathbf{H})$, *and if* $p = r$, $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}$.

Example 6.6.1. *One Sample (continued).* In Example 6.1.1, $\mu = \beta_1$, and $\widehat{\mu} = \widehat{\beta}_1 = \bar{Y}$. The Gauss–Markov theorem shows that $\bar{Y}$, in addition to being UMVU in the normal case, is UMVU in the class of linear estimates for all models with $EY_i^2 < \infty$. However, for

n large, $\bar{Y}$ has a larger variance (Problem 6.6.5) than the nonlinear estimate $\widehat{Y}$ = sample median when the density of Y is the Laplace density

$$\frac{1}{2\lambda}\exp\{-\lambda|y-\mu|\},\ y \in R,\ \mu \in R,\ \lambda > 0.$$

For this density and all symmetric densities, $\widehat{Y}$ is unbiased (Problem 3.4.12).

Remark 6.6.1. As seen in Example 6.6.1, a major weakness of the Gauss–Markov theorem is that it only applies to linear estimates. Another weakness is that it only applies to the homoscedastic case where Var(Y_i) is the same for all i. Suppose we have a heteroscedastic version of the linear model where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, but Var$(\epsilon_i) = \sigma_i^2$ depends on i. If the σ_i^2 are known, we can use the Gauss–Markov theorem to conclude that the weighted least squares estimates of Section 2.2 are UMVU. However, when the σ_i^2 are unknown, our estimates are not optimal even in the class of linear estimates.

Robustness of Tests

In Section 5.3 we investigated the robustness of the significance levels of t tests for the one- and two-sample problems using asymptotic and Monte Carlo methods. Now we will use asymptotic methods to investigate the robustness of levels more generally. The δ-method implies that if the observations $\mathbf{X}_i$ come from a distribution P, which does not belong to the model, the asymptotic behavior of the LR, Wald and Rao tests depends critically on the asymptotic behavior of the underlying MLEs $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_0$.

From the theory developed in Section 6.2 we know that if $\Psi(\cdot, \boldsymbol{\theta}) = Dl(\cdot, \boldsymbol{\theta})$, if $\widehat{\boldsymbol{\theta}}$ solves (6.2.1), and P is true we expect that $\widehat{\boldsymbol{\theta}} \overset{P}{\rightarrow} \boldsymbol{\theta}(P)$, which is the unique solution of

$$\int \Psi(\mathbf{x}, \boldsymbol{\theta}) dP(\mathbf{x}) = \mathbf{0} \tag{6.6.3}$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}(P)) \overset{\mathcal{L}}{\rightarrow} \mathcal{N}(\mathbf{0}, \Sigma(\Psi, P)) \tag{6.6.4}$$

with Σ given by (6.2.6). Thus, for instance, consider $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and the Wald test statistics $T_W = n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. If $\boldsymbol{\theta}(P) \neq \boldsymbol{\theta}_0$ evidently we have $T_W \overset{P}{\rightarrow} \infty$. But if $\boldsymbol{\theta}(P) = \boldsymbol{\theta}_0$, then $T_W \overset{\mathcal{L}}{\rightarrow} \mathbf{V}^T I(\boldsymbol{\theta}_0)\mathbf{V}$ where $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \Sigma(\Psi, P))$. Because $\Sigma(\Psi, P) \neq I^{-1}(\boldsymbol{\theta}_0)$ in general, the asymptotic distribution of T_W is not χ_r^2. This observation holds for the LR and Rao tests as well—but see Problem 6.6.4. There is an important special case where all is well: the linear model we have discussed in Example 6.2.2.

Example 6.6.2. *The Linear Model with Stochastic Covariates.* Suppose $(\mathbf{Z}_1, Y_1), \ldots,$ $(\mathbf{Z}_n, Y_n)$ are i.i.d. as $(\mathbf{Z}, Y)$ where $\mathbf{Z}$ is a $(p \times 1)$ vector of random covariates and we model the relationship between $\mathbf{Z}$ and Y as

$$Y = \mathbf{Z}^T\boldsymbol{\beta} + \epsilon \tag{6.6.5}$$

with the distribution P of $(\mathbf{Z}, Y)$ such that ϵ and Z are independent, $E_P\epsilon = 0$, $E_P\epsilon^2 < \infty$, and we consider, say, $H : \boldsymbol{\beta} = \mathbf{0}$. Then, when ϵ is $\mathcal{N}(0, \sigma^2)$, (see Examples 6.2.1 and 6.2.2) the LR, Wald, and Rao tests all are equivalent to the F test: Reject if

$$T_n \equiv \widehat{\boldsymbol{\beta}}^T \mathbf{Z}_{(n)}^T \mathbf{Z}_{(n)} \widehat{\boldsymbol{\beta}} / s^2 \geq f_{p,n-p}(1-\alpha) \tag{6.6.6}$$

where $\widehat{\boldsymbol{\beta}}$ is the LSE, $\mathbf{Z}_{(n)} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)^T$ and

$$s^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \frac{1}{n-p} |\mathbf{Y} - \mathbf{Z}_{(n)} \widehat{\boldsymbol{\beta}}|^2.$$

Now, even if ϵ is not Gaussian, it is still true that if $\boldsymbol{\Psi}$ is given by $(6.2.30)$ then $\boldsymbol{\beta}(P)$ specified by $(6.6.3)$ equals $\boldsymbol{\beta}$ in $(6.6.5)$ and $\sigma^2(P) = \text{Var}_P(\epsilon)$. For instance,

$$\int \left(\frac{(y - \mathbf{z}^T \boldsymbol{\beta})^2}{\sigma^2} - 1 \right) dP = 0 \text{ iff } \sigma^2 = \text{Var}_P(\epsilon).$$

Thus, it is still true by Theorem 6.2.1 that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, E^{-1}(\mathbf{Z}\mathbf{Z}^T)\sigma^2(P)) \tag{6.6.7}$$

and

$$s^2 \xrightarrow{P} \sigma^2(P). \tag{6.6.8}$$

Moreover, by the law of large numbers,

$$n^{-1} \mathbf{Z}_{(n)}^T \mathbf{Z}_{(n)} = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i^T \xrightarrow{P} E(\mathbf{Z}\mathbf{Z}^T) \tag{6.6.9}$$

so that the confidence procedures based on approximating Var$(\widehat{\beta})$ by $n^{-1}\mathbf{Z}_{(n)}^T \mathbf{Z}_{(n)} s^2 / \sqrt{n}$ are still asymptotically of correct level. It follows by Slutsky's theorem that, under $H : \boldsymbol{\beta} = \mathbf{0}$,

$$T_n \xrightarrow{\mathcal{L}} \mathbf{W}^T [\text{Var}(\mathbf{W})]^{-1} \mathbf{W}$$

where $\mathbf{W}_{p \times p}$ is Gaussian with mean $\mathbf{0}$ and, hence, the limiting distribution of T_n is χ_p^2. Because $f_{p,n-p}(1-\alpha) \to x_p(1-\alpha)$ by Example $5.3.7$, the test $(6.6.5)$ has asymptotic level α even if the errors are not Gaussian.

This kind of robustness holds for $H : \beta_{q+1} = \beta_{0,q+1}, \dots, \beta_p = \beta_{0,p}$ or more generally $\boldsymbol{\beta} \in \mathcal{L}_0 + \boldsymbol{\beta}_0$, a q-dimensional affine subspace of R^p. It is intimately linked to the fact that even though the parametric model on which the test was based is false,

(i) the set of P satisfying the hypothesis remains the same and

(ii) the (asymptotic) variance of $\widehat{\boldsymbol{\beta}}$ is the same as under the Gaussian model.

If the first of these conditions fails, then it's not clear what $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ means anymore. If it holds but the second fails, the theory goes wrong. We illustrate with two final examples.

Example 6.6.3. *The Two-Sample Scale Problem.* Suppose our model is that $X = (X^{(1)}, X^{(2)})$ where $X^{(1)}, X^{(2)}$ are independent $\mathcal{E}\left(\frac{1}{\theta_1}\right), \mathcal{E}\left(\frac{1}{\theta_2}\right)$ respectively, the lifetimes of paired pieces of equipment. If we take $H : \theta_1 = \theta_2$ the standard Wald test $(6.3.17)$ is

$$\text{“Reject iff } n\frac{(\widehat{\theta}_1 - \widehat{\theta}_2)^2}{\widehat{\sigma}^2} \geq x_1(1-\alpha)\text{”} \tag{6.6.10}$$

where

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^{(j)} \tag{6.6.11}$$

and

$$\widehat{\sigma} = \frac{1}{\sqrt{2}n}\sum_{i=1}^{n}(X_i^{(1)} + X_i^{(2)}). \tag{6.6.12}$$

The conditions of Theorem 6.3.2 clearly hold. However suppose now that $X^{(1)}/\theta_1$ and $X^{(2)}/\theta_2$ are identically distributed but not exponential. Then H is still meaningful, $X^{(1)}$ and $X^{(2)}$ are identically distributed. But the test $(6.6.10)$ does not have asymptotic level α in general. To see this, note that, under H,

$$\sqrt{n}(\widehat{\theta}_1 - \widehat{\theta}_2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\,\text{Var}_P X^{(1)}) \tag{6.6.13}$$

but

$$\widehat{\sigma} \xrightarrow{P} \sqrt{2}E_P(X^{(1)}) \neq \sqrt{2\,\text{Var}_P X^{(1)}}$$

in general. It is possible to construct a test equivalent to the Wald test under the parametric model and valid in general. Simply replace $\widehat{\sigma}^2$ by

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(X_i^{(1)} - \frac{(\bar{X}^{(1)} + \bar{X}^{(2)})}{2}\right)^2 + \left(X_i^{(2)} - \frac{(\bar{X}^{(1)} + \bar{X}^{(2)})}{2}\right)^2\right\}. \tag{6.6.14}$$

□

Example 6.6.4. *The Linear Model with Stochastic Covariates with ϵ and $\mathbf{Z}$ Dependent.* Suppose $E(\epsilon \mid \mathbf{Z}) = 0$ so that the parameters $\boldsymbol{\beta}$ of $(6.6.5)$ are still meaningful but ϵ and $\mathbf{Z}$ are dependent. For simplicity, let the distribution of ϵ given $\mathbf{Z} = \mathbf{z}$ be that of $\sigma(\mathbf{z})\epsilon'$ where ϵ' is independent of $\mathbf{Z}$. That is, we assume variances are heteroscedastic. Suppose without loss of generality that $\text{Var}(\epsilon') = 1$. Then $(6.6.7)$ fails and in fact by Theorem 6.2.1

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \mathcal{N}(\mathbf{0}, \mathbf{Q}) \tag{6.6.15}$$

where

$$\mathbf{Q} = E^{-1}(\mathbf{Z}\mathbf{Z}^T)E(\sigma^2(\mathbf{Z})\mathbf{Z}\mathbf{Z}^T)E^{-1}(\mathbf{Z}\mathbf{Z}^T)$$

(Problem 6.6.6) and, in general, the test (6.6.6) does not have the correct level. Specialize to the case $\mathbf{Z} = (1, I_1 - \lambda_1, \dots, I_d - \lambda_d)^T$ where $(I_1, \dots, I_d)$ has a multinomial $(\lambda_1, \dots, \lambda_d)$ distribution, $0 < \lambda_j < 1$, $1 \le j \le d$. This is the stochastic version of the d-sample model of Example 6.1.3. It is easy to see that our tests of $H : \beta_2 = \cdots = \beta_d = 0$ fail to have correct asymptotic levels in general unless $\sigma^2(\mathbf{Z})$ is constant or $\lambda_1 = \cdots = \lambda_d = 1/d$ (Problem 6.6.6). A solution is to replace $\mathbf{Z}_{(n)}^T \mathbf{Z}_{(n)}/s^2$ in (6.6.6) by $\widehat{\mathbf{Q}}^{-1}$, where $\widehat{\mathbf{Q}}$ is a consistent estimate of $\mathbf{Q}$. For $d = 2$ above this is just the asymptotic solution of the Behrens–Fisher problem discussed in Section 4.9.4, the two-sample problem with unequal sample sizes and variances. □

To summarize: If hypotheses remain meaningful when the model is false then, in general, LR, Wald, and Rao tests need to be modified to continue to be valid asymptotically. The simplest solution at least for Wald tests is to use as an estimate of Var $(\sqrt{n}\widehat{\boldsymbol{\theta}})$ not $I^{-1}(\widehat{\boldsymbol{\theta}})$ or $-n^{-1}\|D^2 l_n(\widehat{\boldsymbol{\theta}})\|^{-1}$ but the "sandwich estimate" based on (6.2.6) (Huber, 1967),

$$\left[\frac{1}{n} D^2 l_n(\widehat{\boldsymbol{\theta}})\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} [Dl_1][Dl_1]^T(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}) \left\{\left[\frac{1}{n} D^2 l_n(\widehat{\boldsymbol{\theta}})\right]^{-1}\right\}^T. \tag{6.6.16}$$

Summary. We considered the behavior of estimates and tests when the model that generated them does not hold. The Gauss–Markov theorem states that the linear estimates that are optimal in the linear model continue to be so if the i.i.d. $\mathcal{N}(0, \sigma^2)$ assumption on the errors is replaced by the assumption that the errors have mean zero, identical variances, and are uncorrelated, provided we restrict the class of estimates to linear functions of $Y_1, \dots, Y_n$. In the linear model with a random design matrix, we showed that the MLEs and tests generated by the model where the errors are i.i.d. $\mathcal{N}(0, \sigma^2)$ are still reasonable when the true error distribution is not Gaussian. In particular, the confidence procedures derived from the normal model of Section 6.1 are still approximately valid as are the LR, Wald, and Rao tests. We also demonstrated that when either the hypothesis H or the variance of the MLE is not preserved when going to the wider model, then the MLE and LR procedures for a specific model will fail asymptotically in the wider model. In this case, the methods need adjustment, and we gave the sandwich estimate as one possible adjustment to the variance of the MLE for the smaller model.

6.7 PROBLEMS AND COMPLEMENTS

Problems for Section 6.1

1. Show that in the canonical exponential model (6.1.11) with both $\boldsymbol{\eta}$ and σ^2 unknown, (i) the MLE does not exist if $n = r$, (ii) if $n \ge r + 1$, then the MLE $(\widehat{\eta}_1, \dots, \widehat{\eta}_r, \widehat{\sigma}^2)^T$ of $(\eta_1, \dots, \eta_r, \sigma^2)^T$ is $\left(U_1, \dots, U_r, n^{-1}\sum_{i=r+1}^{n} U_i^2\right)^T$. In particular, $\widehat{\sigma}^2 = n^{-1}|\mathbf{Y} - \widehat{\boldsymbol{\mu}}|^2$.

2. For the canonical linear Gaussian model with σ^2 unknown, use Theorem 3.4.3 to compute the information lower bound on the variance of an unbiased estimator of σ^2. Compare this bound to the variance of s^2.

Hint: By A.13.22, $\text{Var}(U_i^2) = 2\sigma^4$.

3. Show that $\widehat{\boldsymbol{\mu}}$ of Example 6.1.2 with $p = r$ coincides with the empirical plug-in estimate of $\boldsymbol{\mu}_L = (\mu_{L1}, \dots, \mu_{Ln})^T$, where $\mu_{Li} = \mu_Y + (\mathbf{z}_i^* - \mu_{\mathbf{z}})\boldsymbol{\beta}$, $\mathbf{z}_i^* = (z_{i2}, \dots, z_{ip})^T$, $\mu_Y = \beta_1$; and $\boldsymbol{\beta}$ and $\mu_{\mathbf{z}}$ are as defined in (1.4.14). Here the empirical plug-in estimate is based on i.i.d. $(\mathbf{Z}_1^*, Y_1), \dots, (\mathbf{Z}_n^*, Y_n)$ where $\mathbf{Z}_i^* = (Z_{i2}, \dots, Z_{ip})^T$.

4. Let Y_i denote the response of a subject at time $i, i = 1, \dots, n$. Suppose that Y_i satisfies the following model

$$Y_i = \theta + \epsilon_i,\ i = 1, \dots, n$$

where ϵ_i can be written as $\epsilon_i = ce_{i-1} + e_i$ for given constant c satisfying $0 \le c \le 1$, and the e_i are independent identically distributed with mean zero and variance σ^2, $i = 1, \dots, n$; $e_0 = 0$ (the ϵ_i are called moving average errors, see Problem 2.2.29). Let

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i,\ \ \widehat{\theta} = \sum_{j=1}^n a_j Y_j$$

where

$$a_j = \sum_{i=0}^{n-j}(-c)^i \left(\frac{1-(-c)^{j+1}}{1+c}\right) \Big/ \sum_{i=1}^n \left(\frac{1-(-c)^i}{1+c}\right)^2.$$

(a) Show that $\widehat{\theta}$ is the weighted least squares estimate of θ.

(b) Show that if $e_i \sim \mathcal{N}(0, \sigma^2)$, then $\widehat{\theta}$ is the MLE of θ.

(c) Show that $\bar{Y}$ and $\widehat{\theta}$ are unbiased.

(d) Show that $\text{Var}(\widehat{\theta}) \le \text{Var}(\bar{Y})$.

(e) Show that $\text{Var}(\widehat{\theta}) < \text{Var}(\bar{Y})$ unless $c = 0$.

5. Show that $\widetilde{\lambda}(\mathbf{Y})$ defined in Remark 6.1.2 coincides with the likelihood ratio statistic $\lambda(\mathbf{Y})$ for the σ^2 known case with σ^2 replaced by $\widehat{\sigma}^2$.

6. Consider the model (see Example 1.1.5)

$$Y_i = \theta + e_i,\ i = 1, \dots, n$$

where $e_i = ce_{i-1} + \epsilon_i$, $i = 1, \dots, n$, $\epsilon_0 = 0$, $c \in [0, 1]$ is a known constant and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Find the MLE $\widehat{\theta}$ of θ.

7. Derive the formula (6.1.27) for the noncentrality parameter θ^2 in the regression example.

8. Derive the formula (6.1.29) for the noncentrality parameter δ^2 in the one-way layout.

9. Show that in the regression example with $p = r = 2$, the $100(1-\alpha)\%$ confidence interval for β_1 in the Gaussian linear model is

$$\beta_1 = \widehat{\beta}_1 \pm t_{n-2}\left(1 - \tfrac{1}{2}\alpha\right) s \left[\frac{1}{n} + \frac{z_{\cdot 2}^2}{\sum_{i=1}^n (z_{i2} - z_{\cdot 2})^2}\right].$$

10. Show that if $p = r = 2$ in Example 6.1.2, then the hat matrix $\mathbf{H} = (h_{ij})$ is given by

$$h_{ij} = \frac{1}{n} + \frac{(z_{i2} - z_{\cdot 2})(z_{j2} - z_{\cdot 2})}{\sum (z_{i2} - z_{\cdot 2})^2}.$$

11. Show that for the estimates $\widehat{\alpha}$ and $\widehat{\delta}_k$ in the one-way layout

(a) $\text{Var}(\widehat{\alpha}) = \frac{\sigma^2}{p^2} \sum_{k=1}^{p} \frac{1}{n_k}$, $\text{Var}(\widehat{\delta}_k) = \frac{\sigma^2}{p^2} \left(\frac{(p-1)^2}{n_i} + \sum_{k \neq i} \frac{1}{n_k} \right)$.

(b) If n is fixed and divisible by p, then $\text{Var}(\widehat{\alpha})$ is minimized by choosing $n_i = c = n/p$.

(c) If n is fixed and divisible by $2(p-1)$, then $\text{Var}(\widehat{\delta}_k)$ is minimized by choosing $n_i = n/2, n_2 = \cdots = n_p = n/2(p-1)$.

(d) Give the $100(1-\alpha)\%$ confidence intervals for α and δ_k.

12. Let $Y_1, \ldots, Y_n$ be a sample from a population with mean μ and variance σ^2, where n is even. Consider the three estimates $T_1 = \bar{Y}$, $T_2 = (1/2n) \sum_{i=1}^{\frac{1}{2}n} Y_i + (3/2n) \sum_{i=\frac{1}{2}n+1}^{n} Y_i$, and $T_3 = \frac{1}{2}(\bar{Y} - 2)$.

(a) Why can you conclude that T_1 has a smaller MSE (mean square error) than T_2?

(b) Which estimate has the smallest MSE for estimating $\theta = \frac{1}{2}(\mu - 2)$?

13. In the one-way layout,

(a) Show that level $(1-\alpha)$ confidence intervals for linear functions of the form $\beta_j - \beta_i$ are given by

$$\beta_j - \beta_i = Y_{j\cdot} - Y_{i\cdot} \pm s t_{n-p}\left(1 - \tfrac{1}{2}\alpha\right) \sqrt{\frac{n_i + n_j}{n_i n_j}}$$

and that a level $(1-\alpha)$ confidence interval for σ^2 is given by

$$(n-p)s^2/x_{n-p}\left(1 - \tfrac{1}{2}\alpha\right) \le \sigma^2 \le (n-p)s^2/x_{n-p}\left(\tfrac{1}{2}\alpha\right).$$

(b) Find confidence intervals for $\psi = \frac{1}{2}(\beta_2 + \beta_3) - \beta_1$ and $\sigma_{\psi}^2 = \text{Var}(\widehat{\psi})$ where $\widehat{\psi} = \frac{1}{2}(\widehat{\beta}_2 + \widehat{\beta}_3) - \widehat{\beta}_1$.

14. Assume the linear regression model with $p = r = 2$. We want to predict the value of a future observation Y to be taken at the pont z.

(a) Find a level $(1-\alpha)$ confidence interval for the best MSPE predictor $E(Y) = \beta_1 + \beta_2 z$.

(b) Find a level $(1-\alpha)$ *prediction interval* for Y (i.e., statistics $\underline{t}(Y_1, \ldots, Y_n)$, $\bar{t}(Y_1, \ldots, Y_n)$ such that $P[\underline{t} \le Y \le \bar{t}] = 1 - \alpha$). Note that Y is independent of $Y_1, \ldots, Y_n$.

15. Often a treatment that is beneficial in small doses is harmful in large doses. The following model is useful in such situations. Consider a covariate x, which is the amount or dose of a treatment, and a response variable Y, which is yield or production. Suppose a good fit is obtained by the equation

$$y_i = e^{\beta_1} e^{\beta_2 x_i} x_i^{\beta_3}$$

where y_i is observed yield for dose x_i. Assume the model

$$\log Y_i = \beta_1 + \beta_2 x_i + \beta_3 \log x_i + \epsilon_i,\ i = 1, \dots, n$$

where $\epsilon_1, \dots, \epsilon_n$ are independent $\mathcal{N}(0, \sigma^2)$.

(a) For the following data (from Hald, 1952, p. 653), compute $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ and level 0.95 confidence intervals for $\beta_1, \beta_2, \beta_3$.

(b) Plot (x_i, y_i) and $(x_i, \widehat{y}_i)$ where $\widehat{y}_i = e^{\widehat{\beta}_1} e^{\widehat{\beta}_2 x_i} x_i^{\widehat{\beta}_3}$. Find the value of x that maximizes the estimated yield $\widehat{y} = e^{\widehat{\beta}_1} e^{\widehat{\beta}_2 x} x^{\widehat{\beta}_3}$.

x (nitrogen)	0.09	0.32	0.69	1.51	2.29	3.06	3.39	3.63	3.77
y (yield)	15.1	57.3	103.3	174.6	191.5	193.2	178.7	172.3	167.5

Hint: Do the regression for $\mu_i = \beta_1 + \beta_2 z_{i1} + \beta_2 z_{i1} + \beta_3 z_{i2}$ where $z_{i1} = x_i - \bar{x}$, $z_{i2} = \log x_i - \frac{1}{n}\sum_{i=1}^{n} \log x_i$. You may use $s = 0.0289$.

16. Show that if $\mathbf{C}$ is an $n \times r$ matrix of rank r, $r \le n$, then the $r \times r$ matrix $\mathbf{C}'\mathbf{C}$ is of rank r and, hence, nonsingular.

Hint: Because $\mathbf{C}'$ is of rank r, $\mathbf{x}\mathbf{C}' = \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0}$ for any r-vector $\mathbf{x}$. But $\mathbf{x}\mathbf{C}'\mathbf{C} = \mathbf{0} \Rightarrow \|\mathbf{x}\mathbf{C}'\|^2 = \mathbf{x}\mathbf{C}'\mathbf{C}\mathbf{x}' = 0 \Rightarrow \mathbf{x}\mathbf{C}' = \mathbf{0}$.

17. In the Gaussian linear model with $n \ge p + 1$ show that the parametrization $(\boldsymbol{\beta}, \sigma^2)^T$ is identifiable if and only if $r = p$.

Problems for Section 6.2

1. (a) Check A0,. . . ,A6 when $\boldsymbol{\theta} = (\mu, \sigma^2)$, $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in R,\ \sigma^2 > 0\}$, $\boldsymbol{\psi}(x, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \rho(x, \boldsymbol{\theta})$, $\rho(x, \boldsymbol{\theta}) = -\log p(x, \boldsymbol{\theta})$, and $Q = \mathcal{P}$.

(b) Let $\boldsymbol{\theta}$, $\mathcal{P}$, and $\rho = -\log p$, $p \in \mathcal{P}$ be as in (a) but let Q be the class of distributions with densities of the form

$$(1 - \epsilon)\varphi_{(\mu,\sigma^2)}(x) + \epsilon\varphi_{(\mu,\tau^2)}(x),\ \sigma^2 < \tau^2,\ \epsilon \le \frac{1}{2}$$

where $\varphi_{(\mu,\sigma^2)}$ is the $\mathcal{N}(\mu, \sigma^2)$ density. For fixed ϵ and τ^2, check A1–A4.

2. (a) Show that A0–A4 and A6 hold for $\mathcal{P} = \mathcal{Q}$ where $\mathcal{P} = \{P_{\boldsymbol{\theta}}\}$, $\boldsymbol{\theta} = (\mu, \tau)^T$, with densities of the form

$$\frac{1}{2}\varphi_{(\mu,1)} + \frac{1}{2}\varphi_{(\mu,\tau^2)},\ \tau^2 > 0.$$

(b) Show that the MLE of (μ, τ^2) does not exist so that A5 doesn't hold.

(c) Construct a method of moment estimate $\bar{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta} = (\mu, \tau^2)$ based on the first two moments which are $\sqrt{n}$ consistent.

(d) Deduce from Problem 6.2.10 that the estimate $\widehat{\boldsymbol{\theta}}$ derived as the limit of Newton–Raphson estimates from $\bar{\boldsymbol{\theta}}_n$ is efficient.

3. In Example 6.2.1, show that $([E\mathbf{Z}\mathbf{Z}^T]^{-1})_{(1,1)} \geq [EZ_1^2]^{-1}$ with equality if and only if $EZ_1 Z_i = 0, i > 1$.

4. In Example 6.2.2, show that the assumptions of Theorem 6.2.2 hold if (i) and (ii) hold.

5. In Example 6.2.2, show that $c(f_0, \sigma) = \sigma_0/\sigma$ is 1 if f_0 is normal and is different from 1 if f_0 is logistic. You may assume that $E(\epsilon F_0(\epsilon)) \neq \frac{1}{2}$, where F_0 = logistic *df* and $\epsilon \sim \mathcal{N}(0, 1)$.

6. (a) In Example 6.2.1 show that MLEs of $\boldsymbol{\beta}$, μ, and σ^2 are as given in (6.2.20), (6.2.21).
Hint: $f_X(x) = f_{Y|\mathbf{Z}}(y) f_{\mathbf{Z}}(\mathbf{z})$.

(b) Suppose that the distribution of **Z** is not known so that the model is semiparametric, $X \sim P_{(\theta, H)}$, $\{P_{(\theta,H)} : \theta \in \Theta, H \in \mathcal{H}\}$, θ Euclidean, $\mathcal{H}$ abstract. In some cases it is possible to find $T(X)$ such that the distribution of X given $T(X) = t$ is Q_θ, which doesn't depend on $H \in \mathcal{H}$. The MLE of θ based on (X, t) is then called a *conditional* MLE. Show that if we identify $X = (\mathbf{Z}^{(n)}, Y)$, $T(X) = \mathbf{Z}^{(n)}$, then $(\widehat{\boldsymbol{\beta}}, \widehat{\mu}, \widehat{\sigma}^2)$ are conditional MLEs.
Hint: (a),(b) The MLE minimizes $\frac{1}{\sigma^2}|\mathbf{Y} - \mathbf{Z}^{(n)}\boldsymbol{\beta}|^2$.

7. Fill in the details of the proof of Theorem 6.2.1.

8. Establish (6.2.24) directly as follows:

(a) Show that if $\bar{\mathbf{Z}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i$ then, given $\mathbf{Z}_{(n)}$, $\sqrt{n}(\widehat{\mu} - \mu, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T)^T$ has a multivariate normal distribution with mean **0** and variance,

$$\begin{pmatrix} \sigma^2 & \mathbf{0} \\ \mathbf{0} & n[\mathbf{Z}_{(n)}^T \mathbf{Z}_{(n)}]^{-1} \end{pmatrix},$$

and that $\widehat{\sigma}^2$ is independent of the preceding vector with $n\widehat{\sigma}^2/\sigma^2$ having a χ^2_{r-p} distribution.

(b) Apply the law of large numbers to conclude that

$$n^{-1}\mathbf{Z}_{(n)}^T \mathbf{Z}_{(n)} \xrightarrow{P} E(\mathbf{Z}\mathbf{Z}^T).$$

(c) Apply Slutsky's theorem to conclude that

$$\mathcal{L}(\sqrt{n}[E\mathbf{Z}\mathbf{Z}^T]^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \to \mathcal{N}(0, \sigma^2 J)$$

and, hence, that

(d) $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{Z}_{(n)}^T \mathbf{Z}_{(n)} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = o_p(n^{-1/2})$.

(e) Show that $\widehat{\sigma}^2$ is unconditionally independent of $(\widehat{\mu}, \widehat{\boldsymbol{\beta}})$.

(f) Combine (a)–(e) to establish (6.2.24).

9. Let $Y_1, \dots, Y_n$ real be independent identically distributed

$$Y_i = \mu + \sigma\epsilon_i$$

where $\mu \in R$, $\sigma > 0$ are unknown and ϵ has known density $f > 0$ such that if $\rho(x) \equiv -\log f(x)$ then $\rho'' > 0$ and, hence, ρ is strictly convex. Examples are f Gaussian, and $f(x) = e^{-x}(1 + e^{-x})^{-2}$, (logistic).

(a) Show that if $\sigma = \sigma_0$ is assumed known a unique MLE for μ exists and uniquely solves

$$\sum_{i=1}^{n} \rho'\left(\frac{x_i - \mu}{\sigma_0}\right) = 0.$$

(b) Write $\theta_1 = \frac{1}{\sigma}$, $\theta_2 = \frac{\mu}{\sigma}$. Show that if $\theta_2 = \theta_2^0$ a unique MLE for θ_1 exists and uniquely solves

$$\frac{1}{n}\sum_{i=1}^{n} X_i \rho'(\theta_1 X_i - \theta_2^0) = \frac{1}{\theta_1}.$$

10. Suppose A0–A4 hold and $\boldsymbol{\theta}_n^*$ is $\sqrt{n}$ consistent; that is, $\boldsymbol{\theta}_n^* = \boldsymbol{\theta}_0 + O_p(n^{-1/2})$.

(a) Let $\bar{\boldsymbol{\theta}}_n$ be the first iterate of the Newton–Raphson algorithm for solving (6.2.1) starting at $\boldsymbol{\theta}_n^*$,

$$\bar{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_n^* - \left[\frac{1}{n}\sum_{i=1}^{n} D\boldsymbol{\psi}(X_i, \boldsymbol{\theta}_n^*)\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(X_i, \boldsymbol{\theta}_n^*).$$

Show that $\bar{\boldsymbol{\theta}}_n$ satisfies (6.2.3).

Hint:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(X_i, \boldsymbol{\theta}_n^*) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(X_i, \boldsymbol{\theta}_0) - \left(\frac{1}{n}\sum_{i=1}^{n} D\boldsymbol{\psi}(X_i, \boldsymbol{\theta}_n^*) + o_p(1)\right)(\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0).$$

(b) Show that under A0–A4 there exists $\epsilon > 0$ such that with probability tending to 1, $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(X_i, \boldsymbol{\theta})$ has a unique $\mathbf{0}$ in $S(\boldsymbol{\theta}_0, \epsilon)$, the ϵ ball about $\boldsymbol{\theta}_0$.

Hint: You may use a uniform version of the inverse function theorem: If $g_n : R^d \to R^d$ are such that:

(i) $\sup\{|D\mathbf{g}_n(\boldsymbol{\theta}) - D\mathbf{g}(\boldsymbol{\theta})| : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \le \epsilon\} \to 0$,

(ii) $\mathbf{g}_n(\boldsymbol{\theta}_0) \to \mathbf{g}(\boldsymbol{\theta}_0)$,

(iii) $D\mathbf{g}(\boldsymbol{\theta}_0)$ is nonsingular,

(iv) $D\mathbf{g}(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$, then, for n sufficiently large, there exists a $\delta > 0$, $\epsilon > 0$ such that $\mathbf{g}_n$ are $1 - 1$ on $\delta(\boldsymbol{\theta}_0, \delta)$ and their image contains a ball $S(\mathbf{g}(\boldsymbol{\theta}_0), \delta)$.

(c) Conclude that with probability tending to 1, iteration of the Newton–Raphson algorithm starting at $\boldsymbol{\theta}_n^*$ converges to the unique root $\bar{\boldsymbol{\theta}}_n$ described in (b) and that $\bar{\boldsymbol{\theta}}_n$ satisfies (6.2.3).

Hint: You may use the fact that if the initial value of Newton–Raphson is close enough to a unique solution, then it converges to that solution.

11. Establish (6.2.26) and (6.2.27).

Hint: Write

$$\sum_{i=1}^{n}(Y_i - \mathbf{Z}_i^T\boldsymbol{\beta})^2 = \sum_{i=1}^{n}\left(Y_i - (Z_{i1} - \widehat{Z}_i^{(1)})\beta_1 - \sum_{j=2}^{p}(\beta_j + c_j\beta_1)Z_{ij}\right)^2$$

where $\widehat{Z}_i^{(1)} = \sum_{j=2}^{p} c_j Z_{ij}$ and the c_j do not depend on $\boldsymbol{\beta}$. Thus, minimizing $\sum_{i=1}^{n}(Y_i - \mathbf{Z}_i^T\boldsymbol{\beta})^2$ over all $\boldsymbol{\beta}$ is the same as minimizing

$$\sum_{i=1}^{n}\left(Y_i - (Z_{i1} - \widehat{Z}_i^{(1)})\beta_1 - \sum_{j=2}^{p}\gamma_j Z_{ij}\right)^2,$$

where $\gamma_j = \beta_j + c_j\beta_1$. Differentiate with respect to β_1. Similarly compute the information matrix when the model is written as

$$Y_i = \beta_1(Z_{i1} - \Pi(Z_{i1} \mid Z_{i2}, \dots, Z_{ip})) + \sum_{j=2}^{p}\gamma_j Z_{ij} + \epsilon_i$$

where $\beta_1, \gamma_2, \dots, \gamma_p$ range freely and ϵ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Problems for Section 6.3

1. Suppose responses $Y_1, \dots, Y_n$ are independent Poisson variables with $Y_i \sim \mathcal{P}(\lambda_i)$, and

$$\log\lambda_i = \theta_1 + \theta_2 z_i, \; 0 < z_1 < \cdots < z_n$$

for given covariate values $z_1, \dots, z_n$. Find the asymptotic likelihood ratio, Wald, and Rao tests for testing $H : \theta_2 = 0$ versus $K : \theta_2 \neq 0$.

2. Suppose that ω_0 is given by (6.3.12) and the assumptions of Theorem (6.3.2) hold for $p(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. Reparametrize $\mathcal{P}$ by $\boldsymbol{\eta}(\boldsymbol{\theta}) = \sum_{j=1}^{q}\eta_j(\boldsymbol{\theta})\mathbf{v}_j$ where $\eta_j(\boldsymbol{\theta}) \equiv \boldsymbol{\theta}^T\mathbf{v}_j$, $\{\mathbf{v}_j\}$ are orthonormal, $q(\cdot, \boldsymbol{\eta}) \equiv p(\cdot, \boldsymbol{\theta})$ for $\boldsymbol{\eta} \in \Xi$ and $\Xi \equiv \{\boldsymbol{\eta}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Show that if $\Xi_0 = \{\boldsymbol{\eta} \in \Xi : \eta_j = 0, q + 1 < j \leq r\}$ then $\lambda(\mathbf{X})$ for the original testing problem is given by

$$\lambda(\mathbf{X}) = \sup\{q(\mathbf{X} \cdot \boldsymbol{\eta}) : \boldsymbol{\eta} \in \Xi\}/\sup\{q(\mathbf{X}, \boldsymbol{\eta}) : \boldsymbol{\eta} \in \Xi_0\}$$

and, hence, that Theorem 6.3.2 holds for Θ_0 as given in (6.3.12).

3. Suppose that $\boldsymbol{\theta}_0 \in \Theta_0$ and the conditions of Theorem 6.3.3 hold. There exists an open ball about $\boldsymbol{\theta}_0$, $S(\boldsymbol{\theta}_0) \subset \Theta$ and a map

$$\boldsymbol{\eta} : R^r \to R^r,$$

which is continuously differentiable such that

(i) $\eta_j(\boldsymbol{\theta}) = g_j(\boldsymbol{\theta})$ on $S(\boldsymbol{\theta}_0)$, $q+1 \le j \le r$ and, hence,

$$S(\boldsymbol{\theta}_0) \cap \Theta_0 = \{\boldsymbol{\theta} \in S(\boldsymbol{\theta}_0) : \eta_j(\boldsymbol{\theta}) = 0,\ q+1 \le j \le r\}.$$

(ii) $\boldsymbol{\eta}$ is $1{-}1$ on $S(\boldsymbol{\theta}_0)$ and $D\boldsymbol{\eta}(\boldsymbol{\theta})$ is a nonsingular $r \times r$ matrix for all $\boldsymbol{\theta} \in S(\boldsymbol{\theta}_0)$.

(Adjoin to $g_{q+1}, \dots, g_r$, $\mathbf{a}_1^T\boldsymbol{\theta}, \dots, \mathbf{a}_q^T\boldsymbol{\theta}$ where $\mathbf{a}_1, \dots, \mathbf{a}_q$ are orthogonal to the linear span of $\left\|\frac{\partial g_j}{\partial \theta_i}(\boldsymbol{\theta}_0)\right\|_{p\times d}$.)

Show that if we reparametrize $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in S(\boldsymbol{\theta}_0)\}$ by $q(\cdot, \boldsymbol{\eta}(\boldsymbol{\theta})) \equiv p(\cdot, \boldsymbol{\theta})$ where $q(\cdot, \boldsymbol{\eta})$ is uniquely defined on $\Xi = \{\boldsymbol{\eta}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in S(\boldsymbol{\theta}_0)\}$ then, $q(\cdot, \boldsymbol{\eta})$ and $\widehat{\boldsymbol{\eta}}_n \equiv \boldsymbol{\eta}(\widehat{\theta}_n)$ and, $\widehat{\boldsymbol{\eta}}_{0,n} \equiv \boldsymbol{\eta}(\widehat{\theta}_{0,n})$ satisfy the conditions of Problem 6.3.2. Deduce that Theorem 6.3.3 is valid.

4. *Testing Simple versus Simple.* Let $\Theta = \{\theta_0, \theta_1\}$, $X_i, \dots, X_n$ i.i.d. with density $p(\cdot, \theta)$. Consider testing $H : \theta = \theta_0$ versus $K : \theta = \theta_1$. Assume that $P_{\theta_1} \ne P_{\theta_0}$, and that for some $\delta \ge 0$,

(i) $E_{\theta_0}\left|\log \frac{p(X_1,\theta_1)}{p(X_1,\theta_0)}\right|^\delta < \infty$,

(ii) $E_{\theta_1}\left|\log \frac{p(X_1,\theta_1)}{p(X_1,\theta_0)}\right|^\delta < \infty$.

(a) Let $\lambda(X_1, \dots, X_n)$ be the likelihood ratio statistic. Show that under H, even if $\delta = 0$, $2\log\lambda(X_1, \dots, X_n) \xrightarrow{P} 0$.

(b) If $\delta = 2$ show that asymptotically the critical value of the most powerful (Neyman–Pearson) test with $T_n = \sum_{i=1}^n (l(X_i, \theta_1) - l(X_i, \theta_0))$ is $-nK(\theta_0, \theta_1) + z_{1-\alpha}\sqrt{n}\sigma(\theta_0, \theta_1)$ where $K(\theta_0, \theta_1)$ is the Kullback–Leibler divergence

$$K(\theta_0, \theta_1) = E_{\theta_0} \log \frac{p(X_1, \theta_0)}{p(X_1, \theta_1)}$$

and

$$\sigma^2(\theta_0, \theta_1) = \mathrm{Var}_{\theta_0}\left(\log \frac{p(X_1, \theta_1)}{p(X_1, \theta_0)}\right).$$

5. Let (X_i, Y_i), $1 \le i \le n$, be i.i.d. with X_i and Y_i independent, $\mathcal{N}(\theta_1, 1)$, $\mathcal{N}(\theta_2, 1)$, respectively. Suppose $\theta_j \ge 0$, $j = 1, 2$. Consider testing $H : \theta_1 = \theta_2 = 0$ versus $K : \theta_1 > 0$ or $\theta_2 > 0$.

(a) Show that whatever be n, under H, $2\log\lambda(X_i, Y_i : 1 \leq i \leq n)$ is distributed as a mixture of point mass at 0, χ_1^2 and χ_2^2 with probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, respectively.

Hint: By sufficiency reduce to $n = 1$. Then

$$2\log\lambda(X_1, Y_1) = X_1^2 1(X_1 > 0) + Y_1^2 1(Y_1 > 0).$$

(b) Suppose X_i, Y_i are as above with the same hypothesis but $\Theta = \{(\theta_1, \theta_2) : 0 \leq \theta_2 \leq c\theta_1, \theta_1 \geq 0\}$. Show that $2\log\lambda(X_i, Y_i : 1 \leq i \leq n)$ has a null distribution, which is a mixture of point mass at 0, χ_1^2 and χ_2^2 but with probabilities $\frac{1}{2} - \frac{\Delta}{2\pi}$, $\frac{1}{2}$ and $\frac{\Delta}{2\pi}$ where $\sin\Delta = \frac{c}{\sqrt{1+c^2}}$, $0 \leq \Delta \leq \frac{\pi}{2}$.

(c) Let (X_1, Y_1) have an $\mathcal{N}_2(\theta_1, \theta_2, \sigma_{10}^2, \sigma_{20}^2, \rho_0)$ distribution and (X_i, Y_i), $1 \leq i \leq n$, be i.i.d. Let $\theta_1, \theta_2 \geq 0$ and H be as above. Exhibit the null distribution of $2\log\lambda(X_i, Y_i : 1 \leq i \leq n)$.

Hint: Consider $\sigma_{10}^2 = \sigma_{20}^2 = 1$ and $Z_1 = X_1$, $Z_2 = \frac{\rho_0 X_1 - Y_1}{\sqrt{1-\rho_0^2}}$.

6. In the model of Problem 5(a) compute the MLE $(\widehat{\theta}_1, \widehat{\theta}_2)$ under the model and show that

(a) If $\theta_1 > 0, \theta_2 > 0$,

$$\mathcal{L}(\sqrt{n}(\widehat{\theta}_1 - \theta_1, \widehat{\theta}_2 - \theta_2)) \to \mathcal{N}(0, 0, 1, 1, 0).$$

(b) If $\theta_1 = \theta_2 = 0$

$$\mathcal{L}(\sqrt{n}(\widehat{\theta}_1, \widehat{\theta}_2) \to \mathcal{L}(|U|, |V|)$$

where $U \sim \mathcal{N}(0, 1)$ with probability $\frac{1}{2}$ and 0 with probability $\frac{1}{2}$ and V is independent of U with the same distribution.

(c) Obtain the limit distribution of $\sqrt{n}(\widehat{\theta}_1 - \theta_1, \widehat{\theta}_2 - \theta_2)$ if $\theta_1 = 0, \theta_2 > 0$.

(d) Relate the result of (b) to the result of Problem 4(a).

Note: The results of Problems 4 and 5 apply generally to models obeying A0–A6 when we restrict the parameter space to a cone (Robertson, Wright, and Dykstra, 1988). Such restrictions are natural if, for instance, we test the efficacy of a treatment on the basis of two correlated responses per individual.

7. Show that $(6.3.19)$ holds.

Hint:

(i) Show that $I(\widehat{\boldsymbol{\theta}}_n)$ can be replaced by $I(\boldsymbol{\theta})$.

(ii) Show that $W_n(\boldsymbol{\theta}_0^{(2)})$ is invariant under affine reparametrizations $\boldsymbol{\eta} = \mathbf{a} + B\boldsymbol{\theta}$ where B is nonsingular.

(iii) Reparametrize as in Theorem 6.3.2 and compute $W_n(\boldsymbol{\theta}_0^{(2)})$ showing that its leading term is the same as that obtained in the proof of Theorem 6.3.2 for $2\log\lambda(\mathbf{X})$.

8. Show that under A0–A5 and A6 for $\widehat{\boldsymbol{\theta}}_n^{(1)}$

$$\sqrt{n}\boldsymbol{\Psi}_n(\widehat{\boldsymbol{\theta}}_{n,0}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is given by (6.3.21).

Hint: Write

$$\boldsymbol{\Psi}_n(\widehat{\boldsymbol{\theta}}_{n0}) = \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0) + \frac{1}{n} D_{21} l_n(\widehat{\boldsymbol{\theta}}_n^{*})(\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}))$$

and apply Theorem 6.2.2 to $\widehat{\boldsymbol{\theta}}_n^{(1)}$.

9. Under conditions A0–A6 for (a) and A0–A6 with A6 for $\widehat{\boldsymbol{\theta}}_n^{(1)}$ for (b) establish that

(a) $\left[-\frac{1}{n} D^2 l_n(\widehat{\boldsymbol{\theta}}_n)\right]^{-1}$ is a consistent estimate of $I^{-1}(\boldsymbol{\theta}_0)$.

(b) (6.3.22) is a consistent estimate of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)$.

Hint: Argue as in Problem 5.3.10.

10. Show that under A2, A3, A6 $\boldsymbol{\theta} \to I(\boldsymbol{\theta})$ is continuous.

Problems for Section 6.4

1. Exhibit the two solutions of (6.4.4) explicitly and find the one that corresponds to the maximizer of the likelihood.

2. **(a)** Show that for any 2×2 contingency table the table obtained by subtracting (estimated) expectations from each entry has all rows and columns summing to zero, hence, is of the form

Δ	$-\Delta$
$-\Delta$	Δ

(b) Deduce that $\chi^2 = Z^2$ where Z is given by (6.4.8)

(c) Derive the alternative form (6.4.8) for Z.

3. In the 2×2 contingency table model let $X_i = 1$ or 0 according as the ith individual sampled is an A or $\bar{A}$ and $Y_i = 1$ or 0 according as the ith individual sampled is a B or $\bar{B}$.

(a) Show that the correlation of X_1 and Y_1 is

$$\rho = \frac{P(A \cap B) - P(A)P(B)}{\sqrt{P(A)(1 - P(A))P(B)(1 - P(B))}}.$$

(b) Show that the sample correlation coefficient r studied in Example 5.3.6 is related to Z of (6.4.8) by $Z = \sqrt{n}r$.

(c) Conclude that if A and B are independent, $0 < P(A) < 1$, $0 < P(B) < 1$, then Z has a limiting $\mathcal{N}(0, 1)$ distribution.

4. **(a)** Let $(N_{11}, N_{12}, N_{21}, N_{22}) \sim \mathcal{M}(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ as in the contingency table. Let $R_i = N_{i1} + N_{i2}$, $C_i = N_{1i} + N_{2i}$. Show that given $R_1 = r_1$, $R_2 = r_2 = n - r_1$, N_{11} and N_{21} are independent $\mathcal{B}(r_1, \theta_{11}/(\theta_{11} + \theta_{12}))$, $\mathcal{B}(r_2, \theta_{21}/(\theta_{21} + \theta_{22}))$.

(b) Show that $\theta_{11}/(\theta_{11} + \theta_{12}) = \theta_{21}/(\theta_{21} + \theta_{22})$ iff R_1 and C_1 are independent.

(c) Show that under independence the conditional distribution of N_{ii} given $R_i = r_i$, $C_i = c_i$, $i = 1, 2$ is $\mathcal{H}(c_i, n, r_i)$ (the hypergeometric distribution).

5. *Fisher's Exact Test*

From the result of Problem 6.2.4 deduce that if $j(\alpha)$ (depending on r_1, c_1, n) can be chosen so that

$$P[\mathcal{H}(c_1, n, r_1) \geq j(\alpha)] \leq \alpha, \ P[\mathcal{H}(c_1, n, r_1) \geq j(\alpha) - 1] \geq \alpha$$

then the test that rejects (conditionally on $R_1 = r_1$, $C_1 = c_1$) if $N_{11} \geq j(\alpha)$ is exact level α. This is known as Fisher's exact test. It may be shown (see Volume II) that the (approximate) tests based on Z and Fisher's test are asymptotically equivalent in the sense of $(5.4.54)$.

6. Let N_{ij} be the entries of an $a \times b$ contingency table with associated probabilities θ_{ij} and let $\eta_{i1} = \sum_{j=1}^{b} \theta_{ij}$, $\eta_{j2} = \sum_{i=1}^{a} \theta_{ij}$. Consider the hypothesis $H : \theta_{ij} = \eta_{i1}\eta_{j2}$ for all i, j.

(a) Show that the maximum likelihood estimates of η_{i1}, η_{j2} are given by

$$\widehat{\eta}_{i1} = \frac{R_i}{n}, \ \widehat{\eta}_{j2} = \frac{C_j}{n}$$

where $R_i = \sum_j N_{ij}$, $C_j = \sum_i N_{ij}$.

(b) Deduce that Pearson's χ^2 is given by $(6.4.9)$ and has approximately a $\chi^2_{(a-1)(b-1)}$ distribution under H.

Hint: (a) Consider the likelihood as a function of η_{i1}, $i = 1, \ldots, a-1, \eta_{j2}$, $j = 1, \ldots, b-1$ only.

7. Suppose in Problem 6.4.6 that H is true.

(a) Show that then

$$\begin{aligned} P[N_{ij} &= n_{ij};\, i = 1, \ldots, a,\ j = 1, \ldots, b \mid R_i = r_i,\ C_j = c_j] \\ &= \frac{\begin{pmatrix} c_1 \\ n_{11}, \ldots, n_{a1} \end{pmatrix} \begin{pmatrix} c_2 \\ n_{12}, \ldots, n_{a2} \end{pmatrix} \cdots \begin{pmatrix} c_a \\ n_{a1}, \ldots, n_{ab} \end{pmatrix}}{\begin{pmatrix} n \\ r_1, \ldots, r_a \end{pmatrix}} \end{aligned}$$

where $\begin{pmatrix} A \\ B, C, D, \ldots \end{pmatrix} = \frac{A!}{B!C!D!\ldots}$ are the multinomial coefficients.

(b) How would you, in principle, use this result to construct a test of H similar to the χ^2 test with probability of type I error independent of η_{i1}, η_{j2}?

8. The following table gives the number of applicants to the graduate program of a small department of the University of California, classified by sex and admission status. Would you accept or reject the hypothesis of independence at the 0.05 level

(a) using the χ^2 test with approximate critical value?

(b) using Fisher's exact test of Problem $6.4.5$?

	Admit	Deny
Men	19	12
Women	5	0

Hint: (b) It is easier to work with N_{22}. Argue that the Fisher test is equivalent to rejecting H if $N_{22} \geq q_2 + n - (r_1 + c_1)$ or $N_{22} \leq q_1 + n - (r_1 + c_1)$, and that under H, N_{22} is conditionally distributed $\mathcal{H}(r_2, n, c_2)$.

9. (a) If A, B, C are three events, consider the assertions,

(i) $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$ $(A, B$ independent given $C)$

(ii) $P(A \cap B \mid \bar{C}) = P(A \mid \bar{C})P(B \mid \bar{C})$ $(A, B$ independent given $\bar{C})$

(iii) $P(A \cap B) = P(A)P(B)$ $(A, B$ independent$)$

($\bar{C}$ is the complement of C.) Show that (i) and (ii) imply (iii), if A and C are independent or B and C are independent.

(b) Construct an experiment and three events for which (i) and (ii) hold, but (iii) does not.

(c) The following 2×2 tables classify applicants for graduate study in different departments of the university according to admission status and sex. Test in both cases whether the events [being a man] and [being admitted] are independent. Then combine the two tables into one, and perform the same test on the resulting table. Give p-values for the three cases.

	Admit	Deny	
Men	235	35	270
Women	38	7	45
	273	42	

$n = 315$

	Admit	Deny	
Men	122	93	215
Women	103	69	172
	225	162	

$n = 387$

(d) Relate your results to the phenomenon discussed in (a), (b).

10. Establish $(6.4.14)$.

11. Suppose that we know that $\beta_1 = 0$ in the logistic model, $\eta_i = \beta_1 + \beta_2 z_i$, z_i not all equal, and that we wish to test $H : \beta_2 \leq \beta_2^0$ versus $K : \beta_2 > \beta_2^0$.

Show that, for suitable α, there is a UMP level α test, which rejects, if and only if, $\sum_{i=1}^{p} z_i N_i \geq k$, where $P_{\beta_2^0}[\sum_{i=1}^{p} z_i N_i \geq k] = \alpha$.

12. Suppose the z_i in Problem 6.4.11 are obtained as realization of i.i.d. Z_i and $m_i \equiv m$ so that (Z_i, X_i) are i.i.d. with $(X_i \mid Z_i) \sim \mathcal{B}(m, \pi(\beta_2 Z_i))$.

(a) Compute the Rao test for $H : \beta_2 \leq \beta_2^0$ and show that it agrees with the test of Problem 6.4.11.

(b) Suppose that β_1 is unknown. Compute the Rao test statistic for $H : \beta_2 \leq \beta_2^0$ in this case.

(c) By conditioning on $\sum_{i=1}^k X_i$ and using the approach of Problem 6.4.5 construct an exact test (level independent of β_1).

13. Show that if $\omega_0 \subset \omega_1$ are nested logistic regression models of dimension $q < r \leq k$ and $m_1, \dots, m_k \to \infty$ and $H : \eta \in \omega_0$ is true then the law of the statistic of $(6.4.18)$ tends to χ^2_{r-q}.

Hint: $(X_i - \mu_i)/\sqrt{m_i \pi_i (1 - \pi_i)}$, $1 \leq i \leq k$ are independent, asymptotically $\mathcal{N}(0, 1)$. Use this to imitate the argument of Theorem 6.3.3, which is valid for the i.i.d. case.

14. Show that, in the logistic regression model, if the design matrix has rank p, then $\widehat{\beta}_0$ as defined by $(6.4.15)$ is consistent.

15. In the binomial one-way layout show that the LR test is asymptotically equivalent to Pearson's χ^2 test in the sense that $2 \log \lambda - \chi^2 \xrightarrow{P} 0$ under H.

16. Let $X_1, \dots, X_k$ be independent $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$ where either $\sigma^2 = \sigma_0^2$ (known) and $\theta_1, \dots, \theta_k$ vary freely, or $\theta_i = \theta_{i0}$ (known) $i = 1, \dots, k$ and σ^2 is unknown.

Show that the likelihood ratio test of $H : \theta_1 = \theta_{10}, \dots, \theta_k = \theta_{k0}$, $\sigma^2 = \sigma_0^2$ is of the form: Reject if $(1/\sigma_0^2) \sum_{i=1}^k (X_i - \theta_{i0})^2 \geq k_2$ or $\leq k_1$. This is an approximation (for large k, n) and simplification of a model under which $(N_1, \dots, N_k) \sim \mathcal{M}(n, \theta_{10}, \dots, \theta_{k0})$ under H, but under K may be either multinomial with $\theta \neq \theta_0$ *or* have $E_\theta(N_i) = n\theta_{i0}$, but $\text{Var}_\theta(N_i) < n\theta_{i0}(1 - \theta_{i0})$("Cooked data").

Problems for Section 6.5

1. *Fisher's Method of Scoring*

The following algorithm for solving likelihood equations was proposed by Fisher—see Rao (1973), for example. Given an initial value $\widehat{\boldsymbol{\theta}}_0$ define iterates

$$\widehat{\boldsymbol{\theta}}_{m+1} = \widehat{\boldsymbol{\theta}}_m + I^{-1}(\widehat{\boldsymbol{\theta}}_m) Dl(\widehat{\boldsymbol{\theta}}_m).$$

Show that for GLM this method coincides with the Newton–Raphson method of Section 2.4.

2. Verify that $(6.5.4)$ is as claimed formula $(2.2.20)$ for the regression described after $(6.5.4)$.

3. Suppose that $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ have density as in $(6.5.8)$ and,

(a) $P[\mathbf{Z}_1 \in \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(k)}\}] = 1$

(b) The linear span of $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(k)}\}$ is R^p.

(c) $P[\mathbf{Z}_1 = \mathbf{z}^{(j)}] > 0$ for all j. Show that the conditions A0–A6 hold for $P = P_{\boldsymbol{\beta}_0} \in \mathcal{P}$ (where q_0 is assumed known).

Hint: Show that if the convex support of the conditional distribution of Y_1 given $\mathbf{Z}_1 = \mathbf{z}^{(j)}$ contains an open interval about μ_j for $j = 1, \dots, k$, then the convex support of the conditional distribution of $\sum_{j=1}^k \lambda_j Y_j \mathbf{z}^{(j)}$ given $\mathbf{Z}_j = \mathbf{z}^{(j)}$, $j = 1, \dots, k$, contains an open ball about $\sum_{j=1}^k \lambda_j \mu_j \mathbf{z}^{(j)}$ in R^p.

4. Show that for the Gaussian linear model with known variance σ_0^2, the deviance is $D(\mathbf{y}, \boldsymbol{\mu}_0) = |\mathbf{y} - \boldsymbol{\mu}_0|^2/\sigma_0^2$.

5. Let $Y_1, \dots, Y_n$ be independent responses and suppose the distribution of Y_i depends on a covariate vector $\mathbf{z}_i$. Assume that there exist functions $h(y,\tau)$, $b(\theta)$, $g(\mu)$ and $c(\tau)$ such that the model for Y_i can be written as

$$p(y, \theta_i) = h(y, \tau) \exp\left\{\frac{\theta_i y - b(\theta_i)}{c(\tau)}\right\}$$

where τ is known, $g(\mu_i) = \mathbf{z}_i^T \boldsymbol{\beta}$, and b' and g are monotone. Set $\xi = g(\mu)$ and $v(\mu) = \text{Var}(Y)/c(\tau) = b''(\theta)$.

(a) Show that the likelihood equations are

$$\sum_{i=1}^n \frac{d\mu_i}{d\xi_i} \frac{(y_i - \mu_i) z_{ij}}{v(\mu_i)} = 0, \ j = 1, \dots, p.$$

Hint: By the chain rule

$$\frac{\partial}{\partial \beta_j} l(y, \theta) = \frac{\partial l}{\partial \theta} \frac{d\theta}{d\mu} \frac{d\mu}{d\xi} \frac{\partial \xi}{\partial \beta_j}.$$

(b) Show that the Fisher information is $\mathbf{Z}_D^T \mathbf{W} \mathbf{Z}_D$ where $\mathbf{Z}_D = \|z_{ij}\|$ is the design matrix and $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$, $w_i = w(\mu_i) = 1/v(\mu_i)(d\xi_i/d\mu_i)^2$.

(c) Suppose $(\mathbf{Z}_1, Y_1), \dots, (\mathbf{Z}_n, Y_n)$ are i.i.d. as $(\mathbf{Z}, Y)$ and that given $\mathbf{Z} = \mathbf{z}$, Y follow the model $p(y, \theta(\mathbf{z}))$ where $\theta(\mathbf{z})$ solves $b'(\theta) = g^{-1}(\mathbf{z}^T \boldsymbol{\beta})$. Show that, under appropriate conditions,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, w(\mathbf{Z}^T \boldsymbol{\beta}) \mathbf{Z}\mathbf{Z}^T).$$

(d) *Gaussian GLM.* Suppose $Y_i \sim \mathcal{N}(\mu_i, \sigma_0^2)$. Give θ, τ, $h(y,\tau)$, $b(\theta)$, $c(\tau)$, and $v(\mu)$. Show that when g is the canonical link, $g = (b')^{-1}$, the result of (c) coincides with (6.5.9).

(e) Suppose that Y_i has the Poisson, $\mathcal{P}(\mu_i)$, distribution. Give θ, τ, $h(y,\tau)$, $b(\theta)$, $c(\tau)$, and $v(\mu)$. In the random design case, give the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Find the canonical link function and show that when g is the canonical link, your result coincides with (6.5.9).

Problems for Section 6.6

1. Consider the linear model of Example 6.6.2 and the hypothesis

$$\beta_{q+1} = \beta_{0,q+1}, \dots, \beta_p = \beta_{0,p}$$

under the sole assumption that $E\epsilon = 0$, $0 < \text{Var}\,\epsilon < \infty$. Show that the LR, Wald, and Rao tests are still asymptotically equivalent in the sense that if $2\log\lambda_n$, W_n, and R_n are the corresponding test statistics, then under H,

$$\begin{aligned} 2\log\lambda_n &= W_n + o_p(1) \\ R_n &= W_n + o_p(1). \end{aligned}$$

Note: $2\log\lambda_n$, W_n and R_n are computed under the assumption of the *Gaussian* linear model with σ^2 known.

Hint: Retrace the arguments given for the asymptotic equivalence of these statistics under parametric models and note that the only essential property used is that the MLEs under the model satisfy an appropriate estimating equation. Apply Theorem 6.2.1.

2. Show that the standard Wald test for the problem of Example 6.6.3 is as given in (6.6.10).

3. Show that $\widetilde{\sigma}^2$ given in (6.6.14) is a consistent estimate of 2 $\text{Var}_P X^{(1)}$ in Example 6.6.3 and, hence, replacing $\widehat{\sigma}^2$ by $\widetilde{\sigma}^2$ in (6.6.10) creates a valid level α test.

4. Consider the Rao test for $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ for the model $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ and assume that A0–A6 hold. Suppose that the true P does not belong to $\mathcal{P}$ but if $\boldsymbol{\theta}(P)$ is defined by (6.6.3) then $\boldsymbol{\theta}(P) = \boldsymbol{\theta}_0$. Show that, if $\text{Var}_P Dl(X, \boldsymbol{\theta}_0)$ is estimated by $I(\boldsymbol{\theta}_0)$, then the Rao test does not in general have the correct asymptotic level, but that if the estimate $\frac{1}{n}\sum_{i=1}^n [Dl][Dl]^T(X_i, \boldsymbol{\theta}_0)$ is used, then it does.

5. Suppose $X_1, \dots, X_n$ are i.i.d. P. By Problem 5.4.1, if P has a positive density f at $\nu(P)$, the unique median of P, then the sample median $\widehat{X}$ satisfies

$$\sqrt{n}(\widehat{X} - \nu(P)) \to \mathcal{N}(0, \sigma^2(P))$$

where $\sigma^2(P) = 1/4f(\nu(p))$.

(a) Show that if f is symmetric about μ, then $\nu(P) = \mu$.

(b) Show that if f is $\mathcal{N}(\mu, \sigma^2)$, then $\sigma^2(P) > \sigma^2 = \text{Var}_P(X_1)$, the information bound and asymptotic variance of $\sqrt{n}(\bar{X} - \mu)$, but if $f_\mu(x) = \frac{1}{2}\exp -|x - \mu|$, then $\sigma^2(P) < \sigma^2$, in fact, $\sigma^2(P)/\sigma^2 = 2/\pi$.

6. Establish (6.6.15) by verifying the condition of Theorem 6.2.1 under this model and verifying the formula given.

7. In the binary data regression model of Section 6.4.3, let $\pi = s(\mathbf{z}_i^T\boldsymbol{\beta})$ where $s(t)$ is the continuous distribution function of a random variable symmetric about 0; that is,

$$s(t) = 1 - s(-t),\ t \in R. \tag{6.7.1}$$

(a) Show that π can be written in this form for both the probit and logit models.

(b) Suppose that $\mathbf{z}_i$ are realizations of i.i.d. $\mathbf{Z}_i$, that $\mathbf{Z}_1$ is bounded with probability 1 and let $\widehat{\beta}_L(\mathbf{X}^{(n)})$, where $\mathbf{X}^{(n)} = \{(Y_i, \mathbf{Z}_i) : 1 \le i \le n\}$, be the MLE for the logit model. Show that if the correct model has π_i given by s as above and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, then $\widehat{\beta}_L$ is not a consistent estimate of $\boldsymbol{\beta}_0$ unless $s(t)$ is the logistic distribution. But if β_L is defined as the solution of $E\mathbf{Z}_1 s(\mathbf{Z}_1^T\boldsymbol{\beta}_0) = \mathbf{Q}(\boldsymbol{\beta})$ where $\mathbf{Q}(\boldsymbol{\beta}) = E(\mathbf{Z}_1^T \dot{A}(\mathbf{Z}_1^T\boldsymbol{\beta}))$ is $p \times 1$, then

$$\sqrt{n}(\widehat{\beta}_L - \boldsymbol{\beta}_L)$$

has a limiting normal distribution with mean 0 and variance

$$\dot{\mathbf{Q}}^{-1}(\boldsymbol{\beta})\text{Var}(\mathbf{Z}_1(Y_1 - \dot{A}(\mathbf{Z}_1^T\boldsymbol{\beta})))[\dot{Q}^{-1}(\boldsymbol{\beta})]$$

where $\dot{\mathbf{Q}}(\boldsymbol{\beta}) = E(\mathbf{Z}_1^T \ddot{A}(\mathbf{Z}_1^T\boldsymbol{\beta}_0)\mathbf{Z}_1)$ is $p \times p$ and necessarily nonsingular.

Hint: Apply Theorem 6.2.1.

8. (Model Selection) Consider the classical Gaussian linear model (6.1.1) $Y_i = \mu_i + \epsilon_i$, $i = 1, \dots, n$, where ϵ_i are i.i.d. Gaussian with mean zero $\mu_i = \mathbf{z}_i^T\boldsymbol{\beta}$ and variance σ^2, $\mathbf{z}_i$ are d-dimensional vectors of covariate (factor) values. Suppose that the covariates are ranked in order of importance and that we entertain the possibility that the last $d - p$ don't matter for prediction, that is, the model with $\beta_{p+1} = \cdots = \beta_d = 0$ may produce better predictors.

Let $\widehat{\beta}(p)$ be the LSE using only the first p covariates and $\widehat{Y}_i^{(p)}$ the corresponding fitted value.

A natural goal to entertain is to predict future values $Y_1^*, \dots, Y_n^*$ at $\mathbf{z}_1, \dots, \mathbf{z}_n$ and evaluate the performance of $\widehat{Y}_1^{(p)}, \dots, \widehat{Y}_n^{(p)}$ as predictors of $Y_1^*, \dots, Y_n^*$ and, hence, of the model with $\beta_{d+1} = \cdots = \beta_p = 0$, by the (average) expected prediction error

$$EPE(p) = n^{-1}E\sum_{i=1}^{n}(Y_i^* - \widehat{Y}_i^{(p)})^2.$$

Here $Y_1^*, \dots, Y_n^*$ are independent of $Y_1, \dots, Y_n$ and Y_i^* is distributed as Y_i, $i = 1, \dots, n$. Let $RSS(p) = \sum(Y_i - \widehat{Y}_i^{(p)})^2$ be the residual sum of squares. Suppose that σ^2 is known.

(a) Show that $EPE(p) = \sigma^2\left(1 + \frac{p}{n}\right) + \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \mu_i^{(p)})^2$ where $\mu_i^{(p)} = \mathbf{z}_i^T\boldsymbol{\beta}^{(p)}$ and $\boldsymbol{\beta}^{(p)} = (\beta_1, \dots, \beta_p, 0, \dots, 0)^T$.

(b) Show that

$$E\left(\frac{1}{n}RSS(p)\right) = \sigma^2\left(1 - \frac{p}{n}\right) + \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \mu_i^{(p)})^2.$$

(c) Show that $\widehat{EPE}(p) \equiv \frac{1}{n}RSS(p) + \frac{2p}{n}\sigma^2$ is an unbiased estimate of $EPE(p)$.

(d) Show that (a),(b) and (c) continue to hold if we assume the Gauss–Markov linear model (6.6.2).

Model selection consists in selecting $p = \widehat{p}$ to minimize $\widehat{EPE}(p)$ and then using $\widehat{\mathbf{Y}}(\widehat{p})$ as a predictor (Mallows, 1973, for instance).

(e) Suppose $d = 2$ and $\mu_i = \beta_1 z_{i1} + \beta_2 z_{i2}$. Evaluate $EPE(p)$ for (i) $p = 1$ and (ii) $p = 2$. Give values of β_1, β_2 and $\{z_{i1}, z_{i2}\}$ such that the EPE in case (i) is smaller than in case (ii) and vice versa. Use $\sigma^2 = 1$ and $n = 10$.

Hint: (a), (b) and (c). Let $\widehat{\mu}_i^{(p)} = \widehat{Y}_i^{(p)}$ and note that

$$\begin{aligned} EPE(p) &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} E(\widehat{\mu}_i^{(p)} - \mu_i^{(p)})^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(\mu_i^{(p)} - \mu_i)^2 \\ \frac{1}{n}RSS(p) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_i)^2 - \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_i^{(p)}) - \mu_i)^2. \end{aligned}$$

Derive the result for the canonical model.

(d) The result depends only on the mean and covariance structure of the $Y_i, Y_i^*, \widehat{\mu}_i^{(p)}$, $i = 1, \dots, n$. See the proof of Theorems 1.4.4 and 6.6.1.

6.8 NOTES

Note for Section 6.1

(1) From the L. A. Heart Study after Dixon and Massey (1969).

Note for Section 6.2

(1) See Problem 3.5.9 for a discussion of densities with heavy tails.

Note for Section 6.4

(1) R. A. Fisher pointed out that the agreement of this and other data of Mendel's with his hypotheses is *too* good. To guard against such situations he argued that the test should be used in a two-tailed fashion and that we should reject H both for large and for *small* values of χ^2. Of course, this makes no sense for the model we discussed in this section, but it is reasonable, if we consider alternatives to H, which are not multinomial. For instance, we might envision the possibility that an overzealous assistant of Mendel "cooked" the data. LR test statistics for enlarged models of this type do indeed reject H for data corresponding to small values of χ^2 as well as large ones (Problem 6.4.16). The moral of the story is that the practicing statisticians should be on their guard! For more on this theme see Section 6.6.

6.9 REFERENCES

COX, D. R., *The Analysis of Binary Data* London: Methuen, 1970.

DIXON, W. AND F. MASSEY, *Introduction to Statistical Analysis*, 3rd ed. New York: McGraw–Hill, 1969.

FISHER, R. A., *Statistical Methods for Research Workers*, 13th ed. New York: Hafner, 1958.

GRAYBILL, F., *An Introduction to Linear Statistical Models*, Vol. I New York: McGraw–Hill, 1961.

HABERMAN, S., *The Analysis of Frequency Data* Chicago: University of Chicago Press, 1974.

HALD, A., *Statistical Theory with Engineering Applications* New York: Wiley, 1952.

HUBER, P. J., “The behavior of the maximum likelihood estimator under nonstandard conditions,” *Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1, Univ. of California Press*, 221–233 (1967).

KASS, R., J. KADANE AND L. TIERNEY, “Approximate marginal densities of nonlinear functions,” *Biometrika, 76*, 425–433 (1989).

KOENKER, R. AND V. D’OREY, “Computing regression quantiles,” *J. Roy. Statist. Soc. Ser. C, 36*, 383–393 (1987).

LAPLACE, P.-S., “Sur quelques points du système du monde,” *Memoires de l’Académie des Sciences de Paris* (Reprinted in *Oevres Complétes, 11*, 475–558. Gauthier–Villars, Paris) (1789).

MALLOWS, C., “Some comments on C_p,” *Technometrics, 15*, 661–675 (1973).

McCULLAGH, P. AND J. A. NELDER, *Generalized Linear Models* London: Chapman and Hall, New York, 1983; second edition, 1989.

PORTNOY, S. AND R. KOENKER, “The Gaussian Hare and the Laplacian Tortoise: Computability of squared-error versus absolute-error estimators,” *Statistical Science, 12*, 279–300 (1997).

RAO, C. R., *Linear Statistical Inference and Its Applications*, 2nd ed. New York: J. Wiley & Sons, 1973.

ROBERTSON, T., F. T. WRIGHT, AND R. L. DYKSTRA, *Order Restricted Statistical Inference* New York: Wiley, 1988.

SCHEFFÉ, H., *The Analysis of Variance* New York: Wiley, 1959.

SCHERVISCH, M., *Theory of Statistics* New York: Springer, 1995.

STIGLER, S., *The History of Statistics: The Measurement of Uncertainty Before 1900* Cambridge, MA: Harvard University Press, 1986.

WEISBERG, S., *Applied Linear Regression*, 2nd ed. New York: Wiley, 1985.

This page intentionally left blank