Chapter 2
Frequentist Inference

2.1 Introduction

Inference from data can take many forms, but primary inferential aims will often be *point estimation*, to provide a "best guess" of an unknown parameter, and *interval estimation*, to produce ranges for unknown parameters that are supported by the data. Under the frequentist approach, parameters and hypotheses are viewed as unknown but fixed (nonrandom) quantities, and consequently there is no possibility of making probability statements about these unknowns.[1] As the name suggests, the frequentist approach is characterized by a frequency view of probability, and the behavior of inferential procedures is evaluated under hypothetical repeated sampling of the data.

Frequentist procedures are not typically universally applicable to all models/ sample sizes and often require "fixes." For example, a number of variants of likelihood have been developed for use in particular situations (Sect. 2.4.2). In contrast, the Bayesian approach, described in Chap. 3, is completely prescriptive, though there are significant practical hurdles to overcome (such as likelihood and prior specification) in pursuing that prescription. In addition, in situations in which frequentist procedures encounter difficulties, Bayesian approaches typically require very careful prior specification to avoid posterior distributions that exhibit anomalous behavior.

The outline of this chapter is as follows. We begin our discussion in Sect. 2.2 with an overview of criteria by which frequentist procedures may be evaluated. In Sect. 2.3 we present a general development of *estimating functions* which provide a unifying framework for defining and establishing the properties of commonly used frequentist procedures. Two important classes of estimating functions are then

[1] *Random effects* models provide one example in which parameters are viewed as random from a frequentist perspective and are regarded as arising from a population of such effects. Frequentist inference for such models is described in Part III of this book.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series
in Statistics, DOI 10.1007/978-1-4419-0925-1_2,

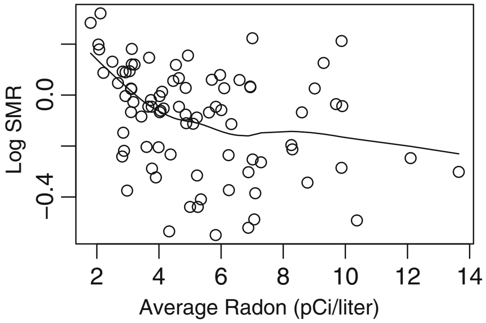

Fig. 2.1 Exploratory plot of log SMR for lung cancer versus average residential radon, with a local smoother superimposed, for 85 counties in Minnesota

introduced: those arising from the specification of a likelihood function, in Sect. 2.4, and those from a quasi-likelihood function, in Sect. 2.5. A recurring theme is the assessment of frequentist procedures under model misspecification. In Sect. 2.6 we discuss the *sandwich estimation* technique which provides estimation of the standard error of estimators in more general circumstances than were assumed in deriving the estimator. Section 2.7 introduces the bootstrap, which is a simulation-based method for making inference with reduced assumptions. Section 2.8 discusses the choice of an estimating function. Hypothesis testing is considered in Sect. 2.9, and the chapter ends with concluding remarks in Sect. 2.10. To provide some numerical relief to the mostly methodological development of this chapter, we provide one running example.

Example: Lung Cancer and Radon

We consider the data introduced in Sect. 1.3.3 and examine the association between counts of lung cancer incidence, Y_i, and the average residential radon, x_i, in county i with $i = 1, \ldots, 85$, indexing the counties within which radon measurements were available (in two counties no radon data were reported). We examine the association using the loglinear model

$$\log \text{E}[\text{SMR}_i \mid x_i] = \beta_0 + \beta_1 x_i. \tag{2.1}$$

where $\text{SMR}_i = Y_i/E_i$ (with E_i the expected count) is the standardized mortality ratio in county i (Sect. 1.3.3) and is a summary measure that controls for the differing age and gender populations across counties. We take as our parameter of interest $\exp(\beta_1)$ which is the multiplicative change in risk associated with a 1 pCi/l increase in radon. In the epidemiological literature this parameter is referred to as the *relative risk*; here it corresponds to the risk ratio for two areas whose radon exposures x differ by one unit.

To first order, $\text{E}[\log \text{SMR} \mid x] \approx \log \text{E}[\text{SMR} \mid x]$, and so if (2.1) is an appropriate model, a plot of $\log \text{SMR}_i$ versus x_i should display an approximately linear trend; Fig. 2.1 shows this plot with a local smoother superimposed and indicates a negative association. This example is illustrative, and so distracting issues, such as the effect of additional covariates (including smoking, the major confounder) and residual spatial dependence in the counts, will be conveniently ignored.

2.2 Frequentist Criteria

In this section we describe frequentist criteria by which competing estimators may be compared and discuss conditions under which optimal estimators exist under these criteria. Under the frequentist approach to inference, the fundamental outlook is that statistical procedures are assessed with respect to their performance under hypothetical, repeated sampling of the data, under fixed values of the parameters. In this section, for simplicity, we consider the estimation of a *univariate* parameter θ and let $\boldsymbol{Y} = [Y_1, \dots, Y_n]^{\mathrm{T}}$ represent a vector of n random variables and $\boldsymbol{y} = [y_1, \dots, y_n]^{\mathrm{T}}$ a realization. Often inference will be summarized via a $100(1-\alpha)\%$ *confidence interval* for θ, which is an interval $[\, a(\boldsymbol{Y}), b(\boldsymbol{Y})\,]$ such that

$$\Pr\left\{\theta \in [\, a(\boldsymbol{Y}), b(\boldsymbol{Y})\,]\right\} = 1 - \alpha, \tag{2.2}$$

for all θ, where the probability statement is with respect to the distribution of $\boldsymbol{Y}$ and $1-\alpha$ is known as the *coverage* probability. For interpretation it is crucial to recognize that the random quantities in (2.2) are the endpoints of the interval $[\, a(\boldsymbol{Y}), b(\boldsymbol{Y})\,]$, so that we are not assigning a probability statement to θ. The correct interpretation of a confidence interval is that, under hypothetical repeated sampling, a proportion $1-\alpha$ of the intervals created will contain the true value θ. We emphasize that we cannot say that the specific interval $[\, a(\boldsymbol{y}), b(\boldsymbol{y})\,]$ contains θ with probability $1-\alpha$.

Ideally, we would like to determine the shortest possible confidence interval for a given α. The search for such intervals is closely linked to the determination of optimal point estimators of θ. The point *estimator* $\widehat{\theta}(\boldsymbol{Y})$ of θ represents a random variable, with an associated *sampling* distribution, while the point *estimate* $\widehat{\theta}(\boldsymbol{y})$ is a specific value. In any given situation a host of potential estimators are available, and we require criteria by which to judge competing choices. Heuristically speaking, a good estimator will have a sampling distribution that is concentrated "close" to the true value θ, where "close" depends on the distance measure that we apply to the distribution of $\widehat{\theta}(\boldsymbol{Y})$.

One natural measure of closeness is the *mean squared error* (MSE) of $\widehat{\theta}(\boldsymbol{Y})$ which arises from a quadratic loss function for estimation and is defined as

$$\begin{aligned}\text{MSE}\left[\widehat{\theta}(\boldsymbol{Y})\right] &= \mathrm{E}_{\boldsymbol{Y}|\theta}\left[\left(\widehat{\theta}(\boldsymbol{Y}) - \theta\right)^2\right] \\ &= \text{var}_{\boldsymbol{Y}|\theta}\left[\widehat{\theta}(\boldsymbol{Y})\right] + \text{bias}\left[\widehat{\theta}(\boldsymbol{Y})\right]^2\end{aligned}$$

where the *bias* of the estimator is

$$\text{bias}\left[\widehat{\theta}(\boldsymbol{Y})\right] = \text{E}_{\boldsymbol{Y}|\theta}\left[\widehat{\theta}(\boldsymbol{Y})\right] - \theta.$$

This notation stresses that all expectations are with respect to the sampling distribution of the estimator, given the true value of the parameter; this is a crucial aspect but the notation is cumbersome and so will be suppressed. Finding estimators with minimum MSE for all values of θ is not possible. For example, $\widehat{\theta}(\boldsymbol{Y}) = 3$ has zero MSE for $\theta = 3$ (and so is optimal for this θ!) but is, in general, a disastrous estimator.

An elegant theory, which is briefly summarized in Appendix G, has been developed to characterize uniformly minimum-variance *unbiased* estimators (UMVUEs). The theory depends first on writing down a full probability model for the data, $p(\boldsymbol{y} \mid \theta)$. We assume conditional independence so that $p(\boldsymbol{y} \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta)$. The Cramér–Rao lower bound for any unbiased estimator $\widehat{\phi}$ of a scalar function of interest $\phi = \phi(\theta)$ is

$$\text{var}(\widehat{\phi}) \geq -\frac{[\phi'(\theta)]^2}{\text{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right]}, \tag{2.3}$$

where $l(\theta) = \sum_{i=1}^{n} \log p(y_i \mid \theta)$ is the log of the joint distribution of the data, viewed as a function of θ. If $T(\boldsymbol{Y})$ is a sufficient statistic of dimension 1, then, under suitable regularity conditions, there is a unique function $\phi(\theta)$ for which a UMVUE exists and its variance attains the Cramér–Rao lower bound. Further, a UMVUE only exists when the data are independently sampled from a one-parameter exponential family. Specifically, suppose that $p(y_i \mid \theta)$ is of one-parameter exponential family form, so that its distribution may be written, for suitably defined functions, as

$$p(y \mid \theta) = \exp\left[\theta T(y) - b(\theta) + c(y)\right]. \tag{2.4}$$

In this situation, there is a unique function of θ for which a UMVUE exists. Unfortunately, this theory only covers a narrow range of circumstances. There are methods available for constructing estimators with the minimal attainable variance in additional situations but even this wider class of models does not come close to covering the range of models that we would like to consider for practical application. UMVUEs are also not always sensible; see Exercise 2.2.

As discussed in Sect. 1.2, model formulation should begin with a model that we would like to fit, before proceeding to examine its mathematical properties. As we will see, exponential family models can provide robust inference, in the sense of performing well even if certain aspects of the assumed model are wrong, but to only consider these models is unnecessarily restrictive.

We now discuss how estimators may be compared in general circumstances *asymptotically*, that is, as $n \to \infty$. There are two hypothetical situations that are being considered here. The first is the repeated sampling aspect for fixed n, and the second is allowing $n \to \infty$. The asymptotic properties of frequentist procedures may be used in two respects. The first is to justify particular procedures, and the second is to carry out inference, for example, to construct confidence intervals. We might question the relevance of asymptotic criteria, since in any practical situation n is finite, and an inconsistent or asymptotically inefficient estimator may have better finite sample properties (a reduced MSE for instance) than a consistent alternative. On the other hand, for many commonly used models, asymptotic inference is often accurate for relatively small sample sizes (as we will see in later chapters).

While unbiasedness of estimators, per se, is of debatable value, a fundamentally important frequentist criterion for assessing an estimator is *consistency*. *Weak consistency* states that as $n \to \infty$, $\widehat{\theta}_n \to_p \theta$ (Appendix F), that is,

$$\Pr(\mid \widehat{\theta}_n - \theta \mid > \epsilon) \to 0 \quad \text{as} \quad n \to \infty \quad \text{for any} \quad \epsilon > 0.$$

Intuitively, the distribution of a consistent estimator concentrates more and more around the true value as the sample size increases. In all but pathological cases, a consistent estimator is asymptotically unbiased, though the contrary is not true. For example, consider the model with $\mathrm{E}[Y_i \mid \theta] = \theta$, $i = 1, \ldots, n$, and the estimator $\widehat{\theta} = Y_1$, this estimator is unbiased but inconsistent.

When assessing an estimator, once consistency has been established, asymptotic normality of the estimator is then typically sought, and interest focuses on the variance of the estimator. In particular, the *asymptotic relative efficiency*, or more simply the *efficiency*, allows an estimator $\widetilde{\theta}_n$ to be compared to the estimator with the smallest variance $\widehat{\theta}_n$ via

$$\frac{\text{var}(\widetilde{\theta}_n)}{\text{var}(\widehat{\theta}_n)}.$$

The $100(1-\alpha)\%$ asymptotic confidence interval associated with an estimator $\widehat{\theta}_n$ is

$$\widehat{\theta}_n \pm z_{1-\alpha/2} \times \sqrt{\text{var}(\widehat{\theta}_n)} \tag{2.5}$$

where $Z \sim \mathrm{N}(0,1)$ and $\Pr(Z < z_{1-\alpha/2}) = 1 - \alpha/2$. If $\widehat{\theta}_n$ is asymptotically efficient, then interval (2.5) is (asymptotically) the shortest available. Maximum likelihood estimation (Sect. 2.4) provides a method for finding efficient estimators.

A difficulty with the interpretation of frequentist inferential summaries is that all probability statements refer to hypothetical data replications and to the estimator, and not to the *estimate* from a specific realization of data. This can lead to intervals with poor properties. Exercise 2.1 describes an instance in which the confidence coverage is correct on average, but for some realizations of the data, the interval has 100% coverage.

We summarize this section and provide a road map to the remainder of the chapter. A fundamental, desirable criterion is to produce confidence intervals that are the shortest possible. Only in stylized situations may estimators with minimum variance be found in non-asymptotic situations. Asymptotically, the picture is rosier, however. In the next section we describe a general class of estimators and give results concerning consistency and asymptotic normality. Subsequently, we show that maximum likelihood estimators attain the smallest asymptotic variance (subject to regularity conditions) *if* the model is correctly specified. We then consider quasi-likelihood, sandwich estimation, and the bootstrap, each of which is designed to reduce the reliance of inference on a full probability model specification.

2.3 Estimating Functions

In the last section we saw that optimal estimators can be found when a full probability model is assumed. The need to specify a full probability model for the data is undesirable. While a practical context may suggest a mean model and perhaps an appropriate mean–variance relationship, it is rare to have faith in a choice for the *distribution* of the data. In this section we give a framework within which the asymptotic properties of a broad range of estimation recipes may be evaluated.

Let $\boldsymbol{Y} = [Y_1, \dots, Y_n]$ represent n observations from a distribution indexed by a p-dimensional parameter $\boldsymbol{\theta}$, with $\text{cov}(Y_i, Y_j \mid \boldsymbol{\theta}) = 0$, $i \neq j$. In the following we will not rigorously derive asymptotic results and only informally discuss regularity conditions under which the results hold. The models discussed subsequently will, unless otherwise stated, obey the necessary conditions.

In the following, for ease of presentation, we assume that Y_i, $i = 1, \dots, n$, are independent and identically distributed (iid).[2] An *estimating function* is a function,

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i), \tag{2.6}$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$\text{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \boldsymbol{0} \tag{2.7}$$

for all $\boldsymbol{\theta}$. The estimating function $\boldsymbol{G}_n(\boldsymbol{\theta})$ is a random variable because it is a function of $\boldsymbol{Y}$. The corresponding *estimating equation* that defines the estimator $\widehat{\boldsymbol{\theta}}_n$ has the form

$$\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\widehat{\boldsymbol{\theta}}_n, Y_i) = \boldsymbol{0}. \tag{2.8}$$

For inference the asymptotic properties of the estimating function are derived (which is why we index the estimating function by n), and these are transferred to the resultant estimator. The estimator $\widehat{\boldsymbol{\theta}}_n$ that solves (2.8) will often be unavailable in closed form and so deriving its distribution from that of the estimating function

[2] In a regression setting we have *independently distributed* observations only, because the distribution of the outcome changes as a function of covariates.

is an ingenious step, because the estimating function may be constructed to be a simple (e.g., linear) function of the data. The estimating function defined in (2.6) is a sum of random variables, which provides the opportunity to evaluate its asymptotic properties via a central limit theorem since the first two moments will often be straightforward to calculate. The art of constructing estimating functions is to make them dependent on distribution-free quantities, for example, the first two moments of the data; robustness of inference to misspecification of higher moments often follows.

We now state an important result that will be used repeatedly in the context of frequentist inference.

Result 2.1. Suppose that $\widehat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \boldsymbol{0},$$

that is, $\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$. Then $\widehat{\boldsymbol{\theta}}_n \to_p \boldsymbol{\theta}$ (consistency) and

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to_d \mathrm{N}_p\left[\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{\mathrm{T}})^{-1}\right] \tag{2.9}$$

(asymptotic normality), where

$$\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{G}(\boldsymbol{\theta}, Y)\right]$$

and

$$\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\mathrm{T}}] = \mathrm{var}\left[\boldsymbol{G}(\boldsymbol{\theta}, Y)\right].$$

Outline Derivation

We refer the interested reader to van der Vaart (1998, Sect. 5.2) for a proof of consistency and present an outline derivation of asymptotic normality, based on van der Vaart (1998, Sect. 5.3). For simplicity we assume that θ is univariate.

We expand $G_n(\theta)$ in a Taylor series around the true value θ:

$$0 = G_n(\widehat{\theta}_n) = G_n(\theta) + (\widehat{\theta}_n - \theta)\ \frac{dG_n}{d\theta}\bigg|_{\theta} + \frac{1}{2}(\widehat{\theta}_n - \theta)^2\ \frac{d^2 G_n}{d\theta^2}\bigg|_{\widetilde{\theta}}, \tag{2.10}$$

where $\widetilde{\theta}$ is a point between $\widehat{\theta}_n$ and θ. We rewrite (2.10) as

$$\sqrt{n}\,(\widehat{\theta}_n - \theta) = \frac{-\sqrt{n}\,G_n(\theta)}{\frac{dG_n}{d\theta}\Big|_{\theta} + \frac{1}{2}(\widehat{\theta}_n - \theta)\ \frac{d^2G_n}{d\theta^2}\Big|_{\widetilde{\theta}}} \tag{2.11}$$

and determine the asymptotic distribution of the right-hand side, beginning with the distribution of $G_n(\theta)$. To apply a central limit theorem, note that $\text{E}[G_n(\theta)] = 0$ and

$$n \times \text{var}\left[G_n(\theta)\right] = \text{var}\left[G(\theta, Y)\right] = \text{E}\left[G(\theta, Y)^2\right] = B$$

(which we assume is finite). Consequently, by the central limit theorem (Appendix G),

$$\sqrt{n}\, G_n(\theta) \;\to_d\; \text{N}\left[0, B(\theta)\right]. \tag{2.12}$$

We now transfer the properties of the estimating function to the estimator $\widehat{\theta}_n$ via (2.11). The first term of the denominator of (2.11),

$$\left.\frac{dG_n}{d\theta}\right|_{\theta} = \frac{1}{n}\sum_{i=1}^{n} \left.\frac{d}{d\theta}G(\theta, Y_i)\right|_{\theta},$$

is an average and so converges to its expectation, provided this expectation exists, by the weak law of large numbers (Appendix G)

$$\left.\frac{dG_n}{d\theta}\right|_{\theta} \to_p \text{E}\left[\frac{d}{d\theta}G(\theta, Y)\right] = A(\theta).$$

Due to consistency, $\widehat{\theta}_n \to_p \theta$, and the second term in the denominator of (2.11) includes the average

$$\left.\frac{d^2G_n}{d\theta^2}\right|_{\widetilde{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\theta^2}G(\theta, Y_i),$$

which, by the law of large numbers, tends to its expectation, that is,

$$\left.\frac{d^2G_n}{d\theta^2}\right|_{\widetilde{\theta}} \to_p \text{E}\left[\frac{d^2}{d\theta^2}G(\theta, Y)\right],$$

provided this average exists. Hence, the second term in the denominator of (2.11) converges in probability to zero and so, by Slutsky's theorem (Appendix G)

$$\sqrt{n}\,(\widehat{\theta}_n - \theta) \to_d \text{N}\left(0, \frac{B}{A^2}\right),$$

as required, where we have suppressed the dependence of $A(\theta)$ and $B(\theta)$ on θ.

□

In practice, $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta})$ and $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\theta})$ are replaced by $\boldsymbol{A}_n(\widehat{\boldsymbol{\theta}}_n)$ and $\boldsymbol{B}_n(\widehat{\boldsymbol{\theta}}_n)$, respectively, with asymptotic normality continuing to hold due to Slutsky's theorem.

In the sections that follow we describe a number of approaches for constructing and using estimating functions. These approaches differ in the number of assumptions that are required for both specifying the estimating function and making inference. At one extreme, in a fully *model-based* approach, a full probability distribution is specified for the data and is used to both specify the estimating function and to evaluate the expectations required in the calculation of $\boldsymbol{A}$ and $\boldsymbol{B}$. At the other extreme, minimal assumptions are made on the data to construct the estimating function, and the expectations required to evaluate var$(\widehat{\boldsymbol{\theta}}_n)$ are calculated empirically from the observed data (see Sect. 2.6).

In the independent but not identically distributed case

$$\left[\boldsymbol{A}_n^{-1}\boldsymbol{B}_n(\boldsymbol{A}_n^{\mathrm{T}})^{-1}\right]^{-1/2}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}) \to_d \mathrm{N}_p(\mathbf{0},\mathrm{I}_p), \tag{2.13}$$

where

$$\boldsymbol{A}_n = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{G}_n(\boldsymbol{\theta})\right]$$
$$\boldsymbol{B}_n = \mathrm{E}\left[\boldsymbol{G}_n(\boldsymbol{\theta})\boldsymbol{G}_n(\boldsymbol{\theta})^{\mathrm{T}}\right] = \mathrm{var}\left[\boldsymbol{G}_n(\boldsymbol{\theta})\right].$$

The previous independent *and* identically distributed situation is a special case, with $\boldsymbol{A}_n = n\boldsymbol{A}$ and $\boldsymbol{B}_n = n\boldsymbol{B}$, in which case (2.13) simplifies to (2.9).

The *sandwich* form of the variance of $\widehat{\boldsymbol{\theta}}_n$ in (2.9) and (2.13)—the covariance of the estimating function, flanked by the expectation of the inverse of the Jacobian matrix of the transformation from the estimating function to the parameter—is one that will appear repeatedly.

Estimators derived from an estimating function are invariant in the sense that if we are interested in a function, $\phi = g(\boldsymbol{\theta})$, then the estimator is $\widehat{\phi}_n = g(\widehat{\boldsymbol{\theta}}_n)$. The delta method (Appendix G) allows the transfer of inference from the parameters of the model to quantities of interest. Specifically, suppose

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}) \to_d \mathrm{N}_p\left[\mathbf{0},\boldsymbol{V}(\boldsymbol{\theta})\right].$$

Then, by the delta method,

$$\sqrt{n}\,\left[g(\widehat{\boldsymbol{\theta}}_n)-g(\boldsymbol{\theta})\right] \to_d \mathrm{N}\left[0,g'(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})g'(\boldsymbol{\theta})^{\mathrm{T}}\right],$$

where $g'(\boldsymbol{\theta})$ is the $1\times p$ vector of derivatives of $g(\cdot)$ with respect to elements of $\boldsymbol{\theta}$. For example, for $p=2$

$$\mathrm{var}\left[g(\boldsymbol{\theta})\right] = V_{11}\left(\left.\frac{\partial g}{\partial\theta_1}\right|_{\boldsymbol{\theta}}\right)^2 + 2V_{12}\left(\left.\frac{\partial g}{\partial\theta_1}\right|_{\boldsymbol{\theta}}\right)\left(\left.\frac{\partial g}{\partial\theta_2}\right|_{\boldsymbol{\theta}}\right) + V_{22}\left(\left.\frac{\partial g}{\partial\theta_2}\right|_{\boldsymbol{\theta}}\right)^2,$$

where V_{jk} denotes the (j,k)th element of $\boldsymbol{V}$, $j,k=1,2$. Again in practice, $\widehat{\boldsymbol{\theta}}_n$ replaces $\boldsymbol{\theta}$ in var$\left[g(\boldsymbol{\theta})\right]$. The accuracy of the asymptotic distribution depends on the parameterization adopted. A rule of thumb is to obtain the asymptotic distribution for a reparameterized parameter defined on the real line; one may then transform back to the parameter of interest, to construct confidence intervals, for example.

The implementation of a frequentist approach usually requires a maximization or root-finding algorithm, but most statistical software packages now contain reliable routines for such endeavors in the majority of situations encountered in practice; hence, we will rarely discuss computational details (in contrast to the Bayesian approach for which computation is typically more challenging).

2.4 Likelihood

For reasons that will become evident, likelihood provides a popular approach to statistical inference and our coverage reflects this. Let $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ be a full probability model for the observed data given a p dimensional vector of parameters, $\boldsymbol{\theta}$. The probability model for the full data is based upon the context and all relevant accumulated knowledge. The level of belief in this model will clearly be context specific, and in many situations, there will be insufficient information available to confidently specify all components of the model. Depending on the confidence in the likelihood, which in turn depends on the sample size (since large n allows more reliable examination of the assumptions of the model), the likelihood may be effectively viewed as approximately "correct," in which case inference proceeds as if the true model were known. Alternatively the likelihood may be seen as an initial working model from which an estimating function is derived; the properties of the subsequent estimator may then be determined under a more general model.

Definition. Viewing $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ gives the *likelihood function*, denoted $L(\boldsymbol{\theta})$.

A key point is that $L(\boldsymbol{\theta})$ is *not* a probability distribution in $\boldsymbol{\theta}$, hence the name likelihood.[3]

2.4.1 Maximum Likelihood Estimation

The value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$ and hence gives the highest probability (density) to the observed data, denoted $\widehat{\boldsymbol{\theta}}$, is known as the maximum likelihood estimator (MLE).

In Part II of this book, we consider models that are appropriate when the data are conditionally independent given $\boldsymbol{\theta}$ so that

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}).$$

[3] We use the label "likelihood" in this section, but strictly speaking we are considering *frequentist* likelihood, since we will evaluate the frequentist properties of an estimator derived from the likelihood. This contrasts with a pure likelihood view, as described in Royall (1997), in which properties are derived from the likelihood function alone, without resorting to frequentist arguments.

For the remainder of this chapter, we assume such conditional independence holds. For both computation and analysis, it is convenient to consider the *log-likelihood* function

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(Y_i \mid \boldsymbol{\theta})$$

and the *score* function

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \right]^{\mathrm{T}} \\ &= [S_1(\boldsymbol{\theta}), \dots, S_p(\boldsymbol{\theta})]^{\mathrm{T}},\end{aligned}$$

which is the $p \times 1$ vector of derivatives of the log-likelihood. As we now illustrate, the score satisfies the requirements of an estimating function.

Definition. Fisher's expected information in a sample of size n is the $p \times p$ matrix

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} l(\boldsymbol{\theta}) \right] = -\mathrm{E}\left[\frac{\partial \boldsymbol{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \right].$$

Result. Under suitable regularity conditions,

$$\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathrm{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}} \right] = \mathbf{0}, \tag{2.14}$$

and

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial \boldsymbol{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \right] = \mathrm{E}\left[\boldsymbol{S}(\boldsymbol{\theta}) \boldsymbol{S}(\boldsymbol{\theta})^{\mathrm{T}} \right]. \tag{2.15}$$

Proof. For simplicity we give a prove for the situation in which θ is univariate, and the observations are independent and identically distributed. Under these circumstances

$$I_n(\theta) = n I_1(\theta),$$

where

$$I_1(\theta) = -\mathrm{E}\left[\frac{d^2}{d\theta^2} \log p(Y \mid \theta) \right].$$

The expectation of the score is

$$\mathrm{E}[S(\theta)] = \sum_{i=1}^{n} \mathrm{E}\left[\frac{d}{d\theta} \log p(Y_i \mid \theta) \right] = n\mathrm{E}\left[\frac{d}{d\theta} \log p(Y \mid \boldsymbol{\theta}) \right]$$

and, under regularity conditions that allow the interchange of differentiation and integration,

$$\begin{aligned}\mathrm{E}\left[\frac{d}{d\theta}\log p(Y \mid \theta)\right] &= \int \left(\frac{d}{d\theta}\log p(y \mid \theta)\right) p(y \mid \theta)dy \\ &= \int \frac{d}{d\theta}p(y \mid \theta)\frac{p(y \mid \theta)}{p(y \mid \theta)}dy = \frac{d}{d\theta}\int p(y \mid \theta)dy = 0,\end{aligned} \tag{2.16}$$

which proves (2.14).

From (2.16),

$$\begin{aligned}0 &= \frac{d}{d\theta}\left[\int \left(\frac{d}{d\theta}\log p(y \mid \theta)\right)\ p(y \mid \theta)dy\right] \\ &= \int \frac{d}{d\theta}\left(\frac{d}{d\theta}\log p(y \mid \theta)p(y \mid \theta)\right)dy \\ &= \int \left(\frac{d^2}{d\theta^2}\log p(y \mid \theta)\right)\ p(y \mid \theta)dy + \int \left(\frac{d}{d\theta}\log p(y \mid \theta)\right)\left(\frac{d}{d\theta}p(y \mid \theta)\right)dy \\ &= \int \left(\frac{d^2}{d\theta^2}\log p(y \mid \theta)\right)\ p(y \mid \theta)dy + \int \left(\frac{d}{d\theta}\log p(y \mid \theta)\right)^2\ p(y \mid \theta)dy \\ &= \mathrm{E}\left[\frac{d^2}{d\theta^2}\log p(Y \mid \theta)\right] + \mathrm{E}\left[\left(\frac{d}{d\theta}\log p(Y \mid \theta)\right)^2\right],\end{aligned}$$

which proves (2.15). □

Viewing the score as an estimating function,

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{S}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta}\log p(Y_i \mid \theta),$$

shows that the MLE satisfies $\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. We have already seen that

$$\mathrm{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \frac{1}{n}\mathrm{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0},$$

and to apply Result 2.1 of Sect. 2.3, we require

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{G}(\boldsymbol{\theta}, Y)\right] = \mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log p(Y \mid \boldsymbol{\theta})\right]$$

and

$$\boldsymbol{B}(\boldsymbol{\theta}) = \mathrm{E}\left[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\mathrm{T}}\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta})\right)^{\mathrm{T}}\right].$$

Equation (2.15) shows that

$$\boldsymbol{I}_1(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta})$$

and, from (2.12)

$$n^{-1/2}\boldsymbol{S}(\boldsymbol{\theta}) \to_d \mathrm{N}\left[\mathbf{0}, \boldsymbol{I}_1(\boldsymbol{\theta})\right]. \tag{2.17}$$

From Result 2.1, the asymptotic distribution of the MLE is therefore

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \;\to_d\; \mathrm{N}_p\left[\mathbf{0}, \boldsymbol{I}_1(\boldsymbol{\theta})^{-1}\right]. \tag{2.18}$$

For independent, but not necessarily identically distributed, random variables $Y_1, \ldots, Y_n$,

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = -\boldsymbol{A}_n(\boldsymbol{\theta}) = \boldsymbol{B}_n(\boldsymbol{\theta}),$$

and

$$\boldsymbol{I}_n(\boldsymbol{\theta})^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \;\to_d\; \mathrm{N}_p(\mathbf{0}, \mathrm{I}_p), \tag{2.19}$$

The information is scaling the statistic and should be growing with n for the asymptotic distribution to be appropriate. Intuitively, the *curvature* of the log-likelihood, as measured by the second derivative, determines the variability of the estimator; the greater the curvature, the smaller the variance of the estimator. The distribution of $\widehat{\boldsymbol{\theta}}_n$ is sometimes written as

$$\widehat{\boldsymbol{\theta}}_n \;\to_d\; \mathrm{N}_p\left[\boldsymbol{\theta}, \boldsymbol{I}_n(\boldsymbol{\theta})^{-1}\right],$$

but this is a little sloppy since the limiting distribution should be independent of n. The variance of the score-based estimating function has the property that $\boldsymbol{A} = \boldsymbol{A}^{\mathrm{T}}$ because the matrix of second derivatives is symmetric, that is,

$$\frac{\partial^2 l}{\partial\theta_j \partial\theta_k} = \frac{\partial^2 l}{\partial\theta_k \partial\theta_j}$$

for $j, k = 1, \ldots, p$.

If there is a unique maximum, then the MLE is consistent and asymptotically normal. The Cramér–Rao bound was given in (2.3). In the present terminology, for any unbiased estimator, $\widetilde{\boldsymbol{\theta}}$, the bound is $\mathrm{var}(\widetilde{\boldsymbol{\theta}}) \geq \boldsymbol{I}_n(\boldsymbol{\theta})^{-1}$ so that the MLE is asymptotically efficient. Asymptotic efficiency under correct model specification is a primary motivation for the widespread use of MLEs.

For inference via (2.18), we may also replace the *expected* information by the *observed* information,

$$\boldsymbol{I}_n^{\star} = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} l(\boldsymbol{\theta}).$$

Asymptotically, their use is equivalent since $\boldsymbol{I}_n^{\star} \to_p \boldsymbol{I}_n$ as $n \to \infty$ by the weak law of large numbers (Appendix G).

The regularity conditions required to derive the asymptotic distribution of the MLE include identifiability so that each element of the parameter space $\boldsymbol{\theta}$ should correspond to a different model $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, otherwise there would be no unique value of $\boldsymbol{\theta}$ to which $\widehat{\boldsymbol{\theta}}$ would converge. We require the interchange of differentiation and integration, and so the range of the data cannot depend on an unknown parameter. Additionally, the true parameter value must lie in the interior of the parameter space, and the Taylor series expansion that was used to determine the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ requires a well-behaved derivative and so the amount of information must increase with sample size. One situation in which one must be wary is when the number of parameters increases with sample size—this number cannot increase too quickly—see Exercise 2.6 for a model in which this condition is violated.

In Sect. 2.4.3, we examine the effects on inference based on the MLE of model misspecification and, in Sects. 2.6 and 2.7, describe methods for determining properties of the estimator that do not depend on correct specification of the full probability model.

Example: Binomial Likelihood

For a single observation from a binomial distribution, $Y \mid p \sim \text{Binomial}(n, p)$, the log-likelihood is

$$l(p) = Y \log p + (n - Y)\log(1 - p),$$

where we omit the term $\log \binom{n}{Y}$ because it is constant with respect to p. The score is

$$S(p) = \frac{dl}{dp} = \frac{Y}{p} - \frac{n - Y}{1 - p},$$

and setting $S(\widehat{p}) = 0$ gives $\widehat{p} = Y/n$. In addition

$$\frac{d^2 l}{dp^2} = -\frac{Y}{p^2} - \frac{n - Y}{(1 - p)^2},$$

and

$$I(p) = -\mathrm{E}\left[\frac{d^2 l}{dp^2}\right] = \frac{n}{p(1 - p)}.$$

We therefore see that the amount of information in the data for p is greater if p is closer to 0 or 1. This is intuitively reasonable since the variance of Y is $np(1 - p)$ and so there is less variability in the data (and hence less uncertainty) if p is close to 0 or 1. The asymptotic distribution of the MLE is

$$\sqrt{n}(\widehat{p} - p) \ \to_d \ \mathrm{N}\left[p, p(1 - p)\right],$$

so that an asymptotic 95% confidence interval for p is

$$\left[\widehat{p} - 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right].$$

Unfortunately, the endpoints of this interval are not guaranteed to lie in (0,1). To rectify this shortcoming, we may parameterize in terms of the logit of p, $\theta = \log[p/(1-p)]$. We could derive the asymptotic distribution using the delta method, but instead we reparameterize the model to give

$$l(\theta) = Y\theta - n\log\left[1 + \exp(\theta)\right],$$

and, proceeding as in the previous parameterization,

$$\widehat{\theta} = \log\left(\frac{Y}{n-Y}\right)$$

and

$$I(\theta) = \frac{n[1+\exp(\theta)]^2}{\exp(\theta)}$$

to give

$$\sqrt{n}(\widehat{\theta} - \theta) \ \to_d \ \mathrm{N}\left(\theta, \frac{\exp(\theta)}{[1+\exp(\theta)]^2}\right).$$

An asymptotic 95% confidence interval for p follows from transforming the endpoints of the interval for θ:

$$\left[\frac{\exp\left(\widehat{\theta} - 1.96 \times \sqrt{\mathrm{var}(\widehat{\theta})/n}\right)}{1 + \exp\left(\widehat{\theta} - 1.96 \times \sqrt{\mathrm{var}(\widehat{\theta})/n}\right)}, \frac{\exp\left(\widehat{\theta} + 1.96 \times \sqrt{\mathrm{var}(\widehat{\theta})/n}\right)}{1 + \exp\left(\widehat{\theta} + 1.96 \times \sqrt{\mathrm{var}(\widehat{\theta})/n}\right)}\right].$$

The endpoints will be contained in (0,1), though $\widehat{\theta}$ is undefined if $Y = 0$ or $Y = n$.

Example: Lung Cancer and Radon

Consider the model

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}(\mu_i),$$

with $\mu_i = E_i \exp(\boldsymbol{x}_i\boldsymbol{\beta})$, $\boldsymbol{x}_i = [1, x_i]$, $i = 1, \dots, n$, and $\boldsymbol{\beta} = [\beta_0, \beta_1]^{\mathrm{T}}$. The probability distribution of $\boldsymbol{y}$ is

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}) = \exp\left(\sum_{i=1}^{n} y_i \log\mu_i - \sum_{i=1}^{n}\mu_i - \sum_{i=1}^{n}\log y_i!\right)$$

to give log-likelihood

$$l(\boldsymbol{\beta}) = \boldsymbol{\beta}^{\mathrm{T}} \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} Y_i - \sum_{i=1}^{n} E_i \exp(\boldsymbol{x}_i \boldsymbol{\beta})$$

and 2 × 1 score vector (estimating function)

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} \left[Y_i - E_i \exp(\boldsymbol{x}_i \boldsymbol{\beta}) \right] \\ &= \boldsymbol{x}^{\mathrm{T}} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}) \right], \end{aligned} \tag{2.20}$$

where $\boldsymbol{x} = [\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_n^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^{\mathrm{T}}$, and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^{\mathrm{T}}$. The equation $\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$ does not, in general, have a closed-form solution, but, pathological datasets aside, numerical solution is straightforward. Asymptotic inference is based on

$$\boldsymbol{I}_n(\widehat{\boldsymbol{\beta}}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_2(\boldsymbol{0}, \mathrm{I}_2),$$

where the information matrix is

$$\boldsymbol{I}_n(\widehat{\boldsymbol{\beta}}_n) = \mathrm{var}(\boldsymbol{S}) = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} \mathrm{var}(Y_i) \boldsymbol{x}_i = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V} \boldsymbol{x},$$

with $\boldsymbol{V}$ the diagonal matrix with elements $\mathrm{var}(Y_i) = E_i \exp(\boldsymbol{x}_i \boldsymbol{\beta})$, $i = 1, \ldots, n$. In this case, the expected and observed information coincide. In practice, the information is estimated by replacing $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}_n$. An important observation is that if the mean is correctly specified the score, (2.20) is a consistent estimator of zero, and $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}$. In particular, if the data do not conform to $\mathrm{var}(Y_i) = \mu_i$, we still have a consistent estimator, but the standard errors will be incorrect.

For the lung cancer data, we have $n = 85$, and the MLE is $\widehat{\boldsymbol{\beta}} = [0.17, -0.036]^{\mathrm{T}}$ with

$$\boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1} = \begin{bmatrix} 0.027^2 & -0.95 \times 0.027 \times 0.0054 \\ -0.95 \times 0.027 \times 0.0054 & 0.0054^2 \end{bmatrix}.$$

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are 0.027 and 0.0054, respectively, and an asymptotic 95% confidence interval for β_1 is $[-0.047, -0.026]$. Leaning on asymptotic normality is appropriate with the large sample size here. A useful inferential summary is an asymptotic 95% confidence interval for the area-level relative risk associated with a one-unit increase in residential radon, which is

$$\exp(-0.036 \pm 1.96 \times 0.0054) = [0.954, 0.975].$$

This interval suggests that the decrease in lung cancer incidence associated with a one-unit increase in residential radon is between 2.5% and 4.6%, though we stress that this is an ecological (area-level) analysis, and we would not transfer inference from the level of the area to the level of the individuals within the areas (as discussed in Sect. 1.3.3).

Example: Weibull Model

The Weibull distribution is useful for the modeling of survival and reliability data and is of the form

$$p(y \mid \boldsymbol{\theta}) = \theta_1 \theta_2^{\theta_1} y^{\theta_1 - 1} \exp\left[-(\theta_2 y)^{\theta_1}\right],$$

where $y > 0$, $\boldsymbol{\theta} = [\theta_1, \theta_2]^{\mathrm{T}}$ and $\theta_1, \theta_2 > 0$. The mean and variance of the Weibull distribution are

$$\begin{aligned} \mathrm{E}[Y \mid \boldsymbol{\theta}] &= \varGamma(1/\theta_1 + 1)/\theta_2 \\ \mathrm{var}(Y \mid \boldsymbol{\theta}) &= [\varGamma(2/\theta_1 + 1) - \varGamma(1/\theta_1 + 1)^2]/\theta_2^2, \end{aligned}$$

where

$$\varGamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$$

is the gamma function. Therefore, the first two moments are not simple functions of θ_1 and θ_2. With independent and identically distributed observations Y_i, $i = 1, \ldots, n$, from a Weibull distribution the log-likelihood is

$$l(\boldsymbol{\theta}) = n \log \theta_1 + n\theta_1 \log \theta_2 + (\theta_1 - 1) \sum_{i=1}^n \log Y_i - \theta_2^{\theta_1} \sum_{i=1}^n Y_i^{\theta_1},$$

with score equations

$$\begin{aligned} S_1(\boldsymbol{\theta}) &= \frac{\partial l}{\partial \theta_1} = \frac{n}{\theta_1} + n \log \theta_2 + \sum_{i=1}^n \log Y_i - \theta_2^{\theta_1} \sum_{i}^{n} Y_i^{\theta_1} \log(\theta_2 Y_i) \\ S_2(\boldsymbol{\theta}) &= \frac{\partial l}{\partial \theta_2} = \frac{n\theta_1}{\theta_2} - \theta_1 \theta_2^{\theta_1 - 1} \sum_{i=1}^n Y_i^{\theta_1}, \end{aligned}$$

which have no closed-form solution and are not a function of a sufficient statistic of dimension less than n. Hence, consistency of $\widehat{\boldsymbol{\theta}}_n$, where $\boldsymbol{S}(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$, cannot be determined from consideration of the first moment (or even the first two moments) of the data only, unlike the Poisson example. In particular, consistency under model misspecification cannot easily be determined.

2.4.2 Variants on Likelihood

Estimation via the likelihood, as defined by $L(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})$, is not always universally applied. In some situations, such as when regularity conditions are violated, alternative versions are required to provide procedures that produce estimators with desirable properties. In other situations, alternative likelihoods provide estimators with better small sample properties, perhaps because nuisance parameters are dealt with more efficiently. Unfortunately, the construction of these likelihoods is not prescriptive and can require a great deal of ingenuity. We describe conditional, marginal, and profile likelihoods.

Conditional Likelihood

Suppose $\boldsymbol{\lambda}$ represent parameters of interest, with $\boldsymbol{\phi}$ being nuisance parameters. Suppose the distribution for $\boldsymbol{y}$ can be factorized as

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) \propto p(\boldsymbol{t}_1 \mid \boldsymbol{t}_2, \boldsymbol{\lambda}) p(\boldsymbol{t}_2 \mid \boldsymbol{\lambda}, \boldsymbol{\phi}), \tag{2.21}$$

where $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ are statistics, that is, functions of $\boldsymbol{y}$. Then inference for $\boldsymbol{\lambda}$ may be based on the *conditional likelihood*

$$L_c(\boldsymbol{\lambda}) = p(\boldsymbol{t}_1 \mid \boldsymbol{t}_2, \boldsymbol{\lambda}). \tag{2.22}$$

The conditional likelihood has similar properties to a regular likelihood. Conditional likelihoods may be used in situations in which we wish to eliminate nuisance parameters. The conditioning statistic, $\boldsymbol{t}_2$, is not ancillary (Appendix F), so that it does depend on $\boldsymbol{\lambda}$, and so some information may be lost in the act of conditioning, but the benefits of elimination are assumed to outweigh this loss. Conditional likelihoods will be used in Sect. 7.7 in the context of Fisher's exact test and individually matched case-control studies (in which the number of parameters increases with sample size) and in Sects. 9.5 and 9.13.4 to eliminate random effects in mixed effects models.

Marginal Likelihood

Let $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, $\boldsymbol{A}$ be a minimal sufficient statistic where $\boldsymbol{A}$ is ancillary (Appendix F), and suppose we have the factorization

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) &\propto p(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{a} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) \\ &= p(\boldsymbol{a}) p(\boldsymbol{s}_1 \mid \boldsymbol{a}, \boldsymbol{\lambda}) p(\boldsymbol{s}_2 \mid \boldsymbol{s}_1, \boldsymbol{a}, \boldsymbol{\lambda}, \boldsymbol{\phi}) \end{aligned}$$

where $\boldsymbol{\lambda}$ are parameters of interest and $\boldsymbol{\phi}$ are the remaining (nuisance) parameters. In contrast to conditional likelihood, marginal likelihoods are based on *averaging* over parts of the data to obtain $p(\boldsymbol{s}_1 \mid \boldsymbol{a}, \boldsymbol{\lambda})$, though operationally marginal likelihoods are often derived without the need for explicit averaging.

Inference for $\boldsymbol{\lambda}$ may be based on the *marginal* likelihood

$$L_m(\boldsymbol{\lambda}) = p(\boldsymbol{s}_1 \mid \boldsymbol{a}, \boldsymbol{\lambda})$$

and is desirable if inference is simplified or if problems with standard likelihood methods are to be avoided.

These advantages may outweigh the loss of efficiency in ignoring the term $p(\boldsymbol{s}_2 \mid \boldsymbol{s}_1, \boldsymbol{a}, \boldsymbol{\lambda}, \boldsymbol{\phi})$. If there is no ancillary statistic, then the marginal likelihood is

$$L_m(\boldsymbol{\lambda}) = p(\boldsymbol{s}_1 \mid \boldsymbol{\lambda}).$$

The marginal likelihood has similar properties to a regular likelihood. We will make use of marginal likelihoods for variance component estimation in mixed effects models in Sect. 8.5.3.

Example: Normal Linear Model

Assume $\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n)$ where $\boldsymbol{x}$ is the $n \times (k+1)$ design matrix and $\dim(\boldsymbol{\beta}) = k+1$. Suppose the parameter of interest is $\lambda = \sigma^2$, with remaining parameters $\boldsymbol{\phi} = \boldsymbol{\beta}$. The MLE for σ^2 is

$$\widetilde{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}) = \frac{\mathrm{RSS}}{n}$$

with $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$. It is well known that $\widetilde{\sigma}^2$ has finite sample bias, because the estimation of $\boldsymbol{\beta}$ is not taken into account. The minimal sufficient statistics are $s_1 = S^2 = \mathrm{RSS}/(n-k-1)$ and $\boldsymbol{s}_2 = \widehat{\boldsymbol{\beta}}$. We write the probability density for $\boldsymbol{y}$ in terms of s_1 and $\boldsymbol{s}_2$:

$$\begin{aligned}
p(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\beta}) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})\right] \\
&\propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(n-k-1)s^2\right] \exp\left[-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] \\
&= p(s_1 \mid \sigma^2) p(\boldsymbol{s}_2 \mid \boldsymbol{\beta}, \sigma^2)
\end{aligned}$$

where going between the first and second line is straightforward if we recognize that

$$\begin{aligned}
(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}} + \boldsymbol{x}\widehat{\boldsymbol{\beta}} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}} + \boldsymbol{x}\widehat{\boldsymbol{\beta}} - \boldsymbol{x}\boldsymbol{\beta}) \\
&= (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \qquad (2.23)
\end{aligned}$$

with the cross term disappearing because of independence between $\widehat{\boldsymbol{\beta}}$ and the vector of residuals $\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}$. Consequently, the marginal likelihood is

$$L_m(\sigma^2) = p(s^2 \mid \sigma^2).$$

Since the data are normal

$$\frac{(n-k-1)s^2}{\sigma^2} \sim \chi^2_{n-k-1} = \text{Ga}\left(\frac{n-k-1}{2}, \frac{1}{2}\right),$$

and so

$$p(s^2 \mid \sigma^2) = \left(\frac{n-k-1}{2\sigma^2}\right)^{(n-k-1)/2} \frac{\left(s^2\right)^{(n-k-1)/2-1}}{\Gamma\left(\frac{n-k-1}{2}\right)} \exp\left[-\frac{(n-k-1)s^2}{2\sigma^2}\right],$$

to give

$$l_m = \log L_m = -(n-k-1)\log\sigma - \frac{(n-k-1)s^2}{2\sigma^2},$$

and marginal likelihood estimator $\widehat{\sigma}^2 = s^2$, the usual unbiased estimator.

Profile Likelihood

Profile likelihood provides a method of examining the behavior of a subset of the parameters. If $\boldsymbol{\theta} = [\boldsymbol{\lambda}, \boldsymbol{\phi}]$, where $\boldsymbol{\lambda}$ again represents a vector of parameters of interest and $\boldsymbol{\phi}$ the remaining parameters, then the profile likelihood $L_p(\boldsymbol{\lambda})$ for $\boldsymbol{\lambda}$ is defined as

$$L_p(\boldsymbol{\lambda}) = \max_{\boldsymbol{\phi}} L(\boldsymbol{\lambda}, \boldsymbol{\phi}). \tag{2.24}$$

If $\widetilde{\boldsymbol{\lambda}}$ denotes the maximum of $L_p(\boldsymbol{\lambda})$ and $\widehat{\boldsymbol{\theta}} = \left[\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\phi}}\right]$ is the MLE, then $\widetilde{\boldsymbol{\lambda}} = \widehat{\boldsymbol{\lambda}}$. Profile likelihoods will be encountered in Sect. 8.5, in the context of the estimation of variance components in linear mixed effects models.

2.4.3 Model Misspecification

In the following, we begin by assuming independent observations. We have seen that if the assumed model is correct then the MLE, $\widehat{\boldsymbol{\theta}}$, has asymptotic distribution

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \to_d \text{N}_p\left[\boldsymbol{0}, \boldsymbol{I}_1(\boldsymbol{\theta})^{-1}\right].$$

In this section we examine the effects of model misspecification. We first determine exactly what quantity the MLE is estimating under misspecification and then examine the asymptotic distribution of the MLE. Let $p(y \mid \boldsymbol{\theta})$ and $p_{\mathrm{T}}(y)$ denote the *assumed* and *true* densities, respectively.

The average of the log-likelihood is such that

$$\frac{1}{n}\sum_{i=1}^{n} \log p(Y_i \mid \boldsymbol{\theta}) \rightarrow_{a.s.} \mathrm{E}_{\mathrm{T}}[\log p(Y \mid \boldsymbol{\theta})], \tag{2.25}$$

by the strong law of large numbers. Hence, asymptotically the MLE maximizes the expectation of the assumed log-likelihood under the true model and $\widehat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}_{\mathrm{T}}$. We now investigate what $\boldsymbol{\theta}_{\mathrm{T}}$ represents when we have assumed an incorrect model. We write

$$\begin{aligned}\mathrm{E}_{\mathrm{T}}[\log p(Y \mid \boldsymbol{\theta})] &= \mathrm{E}_{\mathrm{T}}\left[\log p_{\mathrm{T}}(Y) - \log p_{\mathrm{T}}(Y) + \log p(Y \mid \boldsymbol{\theta})\right] \\ &= \mathrm{E}_{\mathrm{T}}[\log p_{\mathrm{T}}(Y)] - \mathrm{KL}(p_{\mathrm{T}}, p), \end{aligned} \tag{2.26}$$

where

$$\mathrm{KL}(f, g) = \int \log \frac{f(y)}{g(y)} f(y)\, dy \geq 0,$$

is the Kullback–Leibler measure of the “distance” between the densities f and g (the measure is not symmetric so is not a conventional distance measure). The first term of (2.26) does not depend on $\boldsymbol{\theta}$, and so the MLE minimizes $\mathrm{KL}(p_{\mathrm{T}}, p)$, and is therefore that value of $\boldsymbol{\theta}$ which makes the assumed model closest, in a Kullback–Leibler sense, to the true model.

We let $\boldsymbol{S}_n(\boldsymbol{\theta})$ denote the score under the assumed model and state the following result, along with a heuristic derivation.

Result. Suppose $\widehat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation $\boldsymbol{S}_n(\boldsymbol{\theta}) = \boldsymbol{0}$, that is, $\boldsymbol{S}_n(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$. Then

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\mathrm{T}}) \rightarrow_d \mathrm{N}_p\left[\boldsymbol{0}, \boldsymbol{J}^{-1}\boldsymbol{K}(\boldsymbol{J}^{\mathrm{T}})^{-1}\right] \tag{2.27}$$

where

$$\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\theta}_{\mathrm{T}}) = \mathrm{E}_{\mathrm{T}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} \log p(Y \mid \boldsymbol{\theta}_{\mathrm{T}})\right],$$

and

$$\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{\theta}_{\mathrm{T}}) = \mathrm{E}_{\mathrm{T}}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta}_{\mathrm{T}})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}} \log p(Y \mid \boldsymbol{\theta}_{\mathrm{T}})\right)^{\mathrm{T}}\right].$$

Outline Derivation

The derivation closely follows that of Result 2.1, and for simplicity we again assume θ is one-dimensional. We first obtain the expectation and variance of

$$\frac{1}{n} S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta} \log p(y_i \mid \theta),$$

in order to derive the asymptotic distribution of $S_n(\theta)$. Subsequently, we obtain the distribution of $\widehat{\theta}_n$.

Recall that θ_{T} is that value which minimizes the Kullback–Leibler distance, that is,

$$\begin{aligned} 0 = \frac{d}{d\theta} \text{KL}(\theta)\bigg|_{\theta_{\text{T}}} &= \left[\frac{d}{d\theta} \int \log \frac{p_{\text{T}}(y)}{p(y \mid \theta)} p_{\text{T}}(y) dy \right]\bigg|_{\theta_{\text{T}}} \\ &= \left[\int \frac{d}{d\theta} \log p_{\text{T}}(y) p_{\text{T}}(y) dy - \int \frac{d}{d\theta} \log p(y \mid \theta) p_{\text{T}}(y) dy \right]\bigg|_{\theta_{\text{T}}} \\ &= 0 - \left[\int \left(\frac{d}{d\theta} \log p(y \mid \theta) \right) p_{\text{T}}(y) dy \right]\bigg|_{\theta_{\text{T}}}, \end{aligned}$$

and so $\text{E}_{\text{T}}[S(\theta_{\text{T}})] = 0$ (and we have assumed that we can interchange the order of differentiation and integration).

For the second moment,

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{d}{d\theta} \log p(y_i \mid \theta) \right)^2 \to_p \text{E}_{\text{T}} \left[\left(\frac{d}{d\theta} \log p(Y \mid \theta_{\text{T}}) \right)^2 \right] = K,$$

which we assume exists. Hence, by the central limit theorem

$$\frac{1}{n} S(\theta_{\text{T}}) \ \to_d \ \text{N}(0, K).$$

Expanding $S_n(\theta)$ in a Taylor series around θ_{T}:

$$0 = \frac{1}{n} S_n(\widehat{\theta}_n) = \frac{1}{n} S_n(\theta_{\text{T}}) + (\widehat{\theta}_n - \theta_{\text{T}}) \frac{1}{n} \left. \frac{dS_n}{d\theta} \right|_{\theta_{\text{T}}} + \frac{1}{2} (\widehat{\theta}_n - \theta_{\text{T}})^2 \frac{1}{n} \left. \frac{d^2 S_n}{d\theta^2} \right|_{\widetilde{\theta}},$$

where $\widetilde{\theta}$ is between $\widehat{\theta}_n$ and θ_{T} and

$$\frac{1}{n} \left. \frac{dS_n(\theta)}{d\theta} \right|_{\theta_{\text{T}}} = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log p(y \mid \theta) \bigg|_{\theta_{\text{T}}} \to_p \text{E}_{\text{T}} \left[\frac{d^2}{d\theta^2} \log p(Y \mid \theta_{\text{T}}) \right] = J.$$

Following the outline derivation of Result 2.1 gives

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_{\mathrm{T}}) \;\to_d\; \mathrm{N}\left(0, \frac{K}{J^2}\right),$$

as required.

Example: Exponential Assumed Model, Gamma True Model

Suppose that the assumed model is exponential with mean θ but that the true model is gamma $\mathrm{Ga}(\alpha, \beta)$. Minimizing the Kullback–Leibler distance with respect to θ corresponds to maximizing (2.25), that is

$$\mathrm{E}_{\mathrm{T}}\left[-\log\theta - \frac{Y}{\theta}\right] = \log\theta - \frac{\alpha/\beta}{\theta},$$

so that $\theta_{\mathrm{T}} = \alpha/\beta$ is the quantity that is being estimated by the MLE. Hence, the closest exponential distribution to the gamma distribution, in a Kullback–Leibler sense, is the one that possesses the same mean.

2.5 Quasi-likelihood

2.5.1 Maximum Quasi-likelihood Estimation

In this section we describe an estimating function that is based upon the mean and variance of the data only. Specifically, we assume that the first two moments are of the form

$$\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{\beta}] = \boldsymbol{\mu}(\boldsymbol{\beta})$$
$$\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{\beta}) = \alpha \boldsymbol{V}\left[\boldsymbol{\mu}(\boldsymbol{\beta})\right]$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = [\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta})]^{\mathrm{T}}$ represents the regression function, $\boldsymbol{V}$ is a diagonal matrix (so the observations are assumed uncorrelated), with

$$\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V\left[\mu_i(\boldsymbol{\beta})\right],$$

and $\alpha > 0$ is a scalar that does not depend upon $\boldsymbol{\beta}$. We assume $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_k]^{\mathrm{T}}$ so that the dimension of $\boldsymbol{\beta}$ is $k + 1$. The aim is to obtain the asymptotic properties of an estimator of $\boldsymbol{\beta}$ based on these first two moments only. The specification of the mean function in a parametric regression setting is unavoidable, and efficiency will clearly depend on the form of the variance model.

To motivate an estimating function, consider the sum of squares

$$(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})/\alpha, \tag{2.28}$$

where $\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{V}=\boldsymbol{V}(\boldsymbol{\beta})$. To minimize this sum of squares, there are two ways to proceed. Perhaps the more obvious route is to acknowledge that both $\boldsymbol{\mu}$ and $\boldsymbol{V}$ are functions of $\boldsymbol{\beta}$ and differentiate with respect to $\boldsymbol{\beta}$ to give

$$-2\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})/\alpha+(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\frac{\partial\boldsymbol{V}^{-1}}{\partial\boldsymbol{\beta}}(\boldsymbol{Y}-\boldsymbol{\mu})/\alpha, \tag{2.29}$$

where $\boldsymbol{D}$ is the $n\times p$ matrix of derivatives with elements $\partial\mu_i/\partial\beta_j$, $i=1,\dots,n$, $j=1,\dots,p$. Unfortunately, (2.29) is not ideal as an estimating function because it does not necessarily have expectation zero when we only assume $\mathrm{E}[Y\mid\beta]=\mu$, because of the presence of the second term. If the expectation of the estimating function is not zero, then an inconsistent estimator of $\boldsymbol{\beta}$ results.

Alternatively, we may temporarily forget that $\boldsymbol{V}$ is a function of $\boldsymbol{\beta}$ when we differentiate (2.28) and solve the estimating equation

$$\boldsymbol{D}(\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}(\widehat{\boldsymbol{\beta}})^{-1}\left[\boldsymbol{Y}-\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right]/\alpha=\boldsymbol{0}.$$

As shorthand we write this estimating function as

$$\boldsymbol{U}(\boldsymbol{\beta})=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}\right)/\alpha. \tag{2.30}$$

This estimating function is linear in the data and so its properties are straightforward to evaluate. In particular,

1. $\mathrm{E}[U(\boldsymbol{\beta})]=\boldsymbol{0}$, assuming $\mathrm{E}[Y\mid\boldsymbol{\beta}]=\mu(\boldsymbol{\beta})$.
2. $\mathrm{var}\left[\boldsymbol{U}(\boldsymbol{\beta})\right]=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha$, assuming $\mathrm{var}(Y\mid\boldsymbol{\beta})=V$.
3. $-\mathrm{E}\left[\frac{\partial U}{\partial\beta}\right]=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha=\mathrm{var}\left[\boldsymbol{U}(\boldsymbol{\beta})\right]$, assuming $\mathrm{E}[Y\mid\boldsymbol{\beta}]=\mu(\boldsymbol{\beta})$.

The similarity of these properties with those of the score function (Sect. 2.4.1) is apparent and has led to (2.30) being referred to as a *quasi-score* function. Let $\widehat{\boldsymbol{\beta}}_n$ represent the root of (2.30), that is, $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}_n)=\boldsymbol{0}$. We can apply Result 2.1 directly to obtain the asymptotic distribution of the maximum quasi-likelihood estimator (MQLE) as

$$(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{1/2}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\ \to_d\ \mathrm{N}_{k+1}(\boldsymbol{0},\alpha\mathrm{I}_{k+1}),$$

where we have assumed that α is known. Using (B.4) in Appendix B

$$\mathrm{E}[(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})]/\alpha=n,$$

and so if μ were known, an unbiased estimator of α would be

$$\widehat{\alpha}_n=(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})/n.$$

A degree of freedom corrected (but not in general unbiased) estimate is given by the Pearson statistic divided by its degrees of freedom:

$$\widehat{\alpha}_n = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{2.31}$$

where $\widehat{\mu}_i = \widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$. This estimator of the scale parameter is consistent so long as the assumed variance model is correct. The asymptotic distribution that is used in practice is therefore

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}}/\widehat{\alpha}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}) \ \to_d \ \mathrm{N}_{k+1}(\mathbf{0},\mathrm{I}_{k+1}).$$

The inclusion of an estimate for α is justified by applying Slutsky's theorem (Appendix G) to $\widehat{\alpha}_n \times \boldsymbol{U}(\widehat{\boldsymbol{\beta}}_n)$. As usual in such asymptotic calculations, the uncertainty in $\widehat{\alpha}_n$ is not reflected in the variance for $\widehat{\boldsymbol{\beta}}_n$. This development reveals a mixing of inferential approaches with $\widehat{\boldsymbol{\beta}}_n$ a MQLE and $\widehat{\alpha}_n$ a method of moments estimator. A justification for the latter estimator is that it is likely to be consistent in a wider range of circumstances than a likelihood-based estimator. A crucial observation is that if the mean function is correctly specified, the estimator $\widehat{\boldsymbol{\beta}}_n$ is consistent also. Asymptotically appropriate standard errors result if the mean–variance relationship is correctly specified. McCullagh (1983) and Godambe and Heyde (1987) discuss the close links between consistency, the quasi-score function (2.30), and membership of the exponential family; see also Chap. 6.

As an aside, in the above, the mean model does not need to be "correct" since we are simply estimating a specified form of association, and estimation will be performed regardless of whether this model is appropriate. Of course, the usefulness of inference does depend on an appropriate mean model.

As a function of μ, we have the quasi-score

$$\frac{Y-\mu}{\alpha V(\mu)}, \tag{2.32}$$

and integration of this quantity gives

$$l(\mu,\alpha) = \int_y^{\mu}\frac{y-t}{\alpha V(t)}dt,$$

which, if it exists, behaves like a log-likelihood. As an example, for the model $\mathrm{E}[Y]=\mu$ and $\mathrm{var}(Y)=\alpha\mu$

$$l(\mu,\alpha) = \int_y^{\mu}\frac{y-t}{\alpha t}dt = \frac{1}{\alpha}[y\log\mu-\mu+c],$$

where $c=-y\log y - y$ and $y\log\mu-\mu$ is the log-likelihood of a Poisson random variable. Table 2.1 lists some distributions that correspond to particular choices of variance function.

Table 2.1 Variance functions and quasi log-likelihoods

Variance $V(\mu)$	Quasi log likelihood	Distribution
1	$-\frac{1}{\alpha}\left[\frac{1}{2}(y-\mu)^2\right]$	$\text{N}(\mu,\alpha)$
μ	$\frac{1}{\alpha}\left(y\log\mu-\mu\right)$	$\text{Poisson}(\mu)$
μ^2	$\frac{1}{\alpha}\left(-\frac{y}{\mu}-\log\mu\right)$	$\text{Ga}(1/\alpha,\mu/\alpha)$
$n\mu(1-\mu)$	$\frac{1}{\alpha}\left[y\log\left(\frac{\mu}{1-\mu}\right)+n\log(1-\mu)\right]$	$\text{Binomial}(n,\mu)$
$\mu+\mu^2/b$	$\frac{1}{\alpha}\left[y\log\left(\frac{\mu}{b+\mu}\right)+b\log\left(\frac{b}{b+\mu}\right)\right]$	$\text{NegBin}(\mu,b)$, b known
$\mu^2(1-\mu)^2$	$\frac{1}{\alpha}\left[(2y-1)\log\left(\frac{\mu}{1-\mu}\right)-\frac{y}{\mu}-\frac{1-y}{1-\mu}\right]$	No distribution

In all cases $\text{E}[Y]=\mu$. The parameterizations of the distributional forms are as in Appendix D. For the Poisson, binomial, and negative binomial distributions, these are the forms that the quasi-score corresponds to when $\alpha=1$

The word "quasi" refers to the fact that the score may or not correspond to a probability function. For example, in Table 2.1, the variance function $\mu^2(1-\mu)^2$ does not correspond to a probability distribution. In most cases, there is an implied distributional kernel, but the addition of the variance multiplier α often produces a mean–variance relationship that is not present in the implied distribution.

We emphasize that the first two moments do not uniquely define a distribution. For example, the negative binomial distribution may be derived as the marginal distribution of

$$Y \mid \mu,\theta \sim \text{Poisson}(\mu\theta) \tag{2.33}$$

$$\theta \sim \text{Ga}(b,b) \tag{2.34}$$

so that $\text{E}[Y]=\mu$ and

$$\text{var}(Y)=\text{E}[\text{var}(Y\mid\theta)]+\text{var}(\text{E}[Y\mid\theta])=\mu+\frac{\mu^2}{b}. \tag{2.35}$$

These latter two moments are also recovered if we replace the gamma distribution with a lognormal distribution. Specifically, assume the model

$$\begin{aligned} Y \mid \theta^\star &\sim \text{Poisson}(\theta^\star)\\ \theta^\star &\sim \text{LogNorm}(\eta,\sigma^2) \end{aligned}$$

and let $\mu=\text{E}[\theta]=\exp(\eta+\sigma^2/2)$. Then,

$$\text{var}(\theta^\star)=\text{E}[\theta^\star]^2\left[\exp(\sigma^2)-1\right]=\mu^2\left[\exp(\sigma^2)-1\right].$$

Under this model, $\text{E}[Y]=\mu$ and

$$\text{var}(Y)=\text{E}[\text{var}(Y\mid\theta^\star)]+\text{var}[\text{E}(Y\mid\theta^\star)]=\mu+\mu^2\left[\exp(\sigma^2)-1\right]$$

which, on writing $b^\star = [\exp(\sigma^2) - 1]^{-1}$, gives the same form of quadratic variance function, (2.35), as with the gamma model.

If the estimating function (2.30) corresponds to the score function for a particular probability distribution, then the subsequent estimator corresponds to the MLE (because α does not influence the estimation of $\boldsymbol{\beta}$), though the variance of the estimator will usually differ. A great advantage of the use of quasi-likelihood is its computational simplicity.

A prediction interval for an observable, Y, is not possible with quasi-likelihood since there is no probabilistic mechanism with which to reflect the stochastic component of the prediction.

Example: Lung Cancer and Radon

We return to the lung cancer example and now assume the quasi-likelihood model

$$\mathrm{E}[Y_i \mid \boldsymbol{\beta}] = E_i \exp(\boldsymbol{x}_i\boldsymbol{\beta}), \quad \mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha \mathrm{E}[Y_i \mid \boldsymbol{\beta}].$$

Fitting this model yields identical point estimates to the MLEs and $\widehat{\alpha} = 2.81$ so that the quasi-likelihood standard errors are $\sqrt{\widehat{\alpha}} = 1.68$ times larger than the Poisson model-based standard errors. The variance–covariance matrix is

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}})^{-1}\widehat{\alpha} = \begin{bmatrix} 0.045^2 & -0.95 \times 0.045 \times 0.0090 \\ -0.95 \times 0.045 \times 0.0090 & 0.0090^2 \end{bmatrix}.$$

An asymptotic 95% confidence interval for the relative risk associated with a one-unit increase in radon is $[0.947, 0.982]$ which is $\sqrt{\widehat{\alpha}} = 1.68$ wider than the Poisson interval evaluated previously.

2.5.2 A More Complex Mean–Variance Model

For comparison, we now describe a more general model than considered under the quasi-likelihood approach. Suppose we specify the first two moments of the data as

$$\mathrm{E}[Y_i \mid \boldsymbol{\beta}] = \mu_i(\boldsymbol{\beta}) \tag{2.36}$$

$$\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = V_i(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{2.37}$$

where $\boldsymbol{\alpha}$ is an $r \times 1$ vector of parameters that appear only in the variance model. Let $\widehat{\boldsymbol{\alpha}}_n$ be a consistent estimator of $\boldsymbol{\alpha}$. We state without proof the following result. The estimator $\widehat{\boldsymbol{\beta}}_n$ that satisfies the estimating equation

$$\boldsymbol{G}(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\alpha}}_n) = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)\boldsymbol{V}^{-1}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)\left[\boldsymbol{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_n)\right] \tag{2.38}$$

has asymptotic distribution

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \to_d \ \mathrm{N}_{k+1}(\mathbf{0}, \mathrm{I}_{k+1})$$

where $\widehat{\boldsymbol{D}} = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$.

The difference between this model and that in the quasi-likelihood approach is that $\boldsymbol{V}$ may now depend on additional variance–covariance parameters $\boldsymbol{\alpha}$ in a more complex way. Under quasi-likelihood it is assumed that $\text{var}(Y_i) = \alpha V_i(\mu_i)$, so that the estimating function does not depend on α. Consequently, $\widehat{\boldsymbol{\beta}}$ also does not depend on α, though the standard errors are proportional to $\sqrt{\alpha}$. This is a motivating factor in the development of quasi-likelihood, since standard software may be used for implementation and, perhaps more importantly, consistency of $\boldsymbol{\beta}$ is guaranteed if the mean model is correctly specified.

The form of the mean–variance relationship given by (2.36) and (2.36) suggests an iterative scheme for estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Set $t = 0$ and let $\widehat{\boldsymbol{\alpha}}^{(0)}$ be an initial estimate for $\boldsymbol{\alpha}$. Now iterate between

1. Solve $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(t)}) = \mathbf{0}$ to give $\widehat{\boldsymbol{\beta}}^{(t+1)}$,
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(t+1)}$ with $\widehat{\mu}_i = \mu_i\left(\widehat{\boldsymbol{\beta}}^{(t+1)}\right)$. Set $t \to t + 1$ and return to 1.

The model given by (2.36) and (2.36) is more flexible than that provided by quasi-likelihood but requires the correct specification of mean and variance for a consistent estimator of $\boldsymbol{\beta}$.

Example: Lung Cancer and Radon

As an example of the mean–variance model discussed in the previous section, we fit a negative binomial model to the lung cancer data. This model is motivated via the random effects formulation given by (2.33) and (2.34) with loglinear model $\mu_i = \mu_i(\boldsymbol{\beta}) = E_i \exp(\beta_0 + \beta_1 x_i)$, $i = 1, \ldots, n$. In the lung cancer context, the random effects are area-specific perturbations from the mean μ_i. The introduction of the random effects may be seen as a device for inducing overdispersion. Integrating over θ_i, we obtain the negative binomial distribution

$$\Pr(y_i \mid \boldsymbol{\beta}, b) = \frac{\Gamma(y_i + b)}{\Gamma(b) y_i!} \frac{\mu_i^{y_i} b^b}{(\mu_i + b)^{y_i+b}},$$

for $y_i = 0, 1, 2, \ldots$, with

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \text{var}(Y_i \mid \boldsymbol{\beta}, b) &= \mu_i(\boldsymbol{\beta})\left[1 + \frac{\mu_i(\boldsymbol{\beta})}{b}\right], \end{aligned} \tag{2.39}$$

so that smaller values of b correspond to greater degrees of overdispersion and as $b \to \infty$ we recover the Poisson model. For consistency with later chapters we use b rather than α for the parameter occurring in the variance model. Care is required with the negative binomial distribution since a number of different parameterizations are available; see Exercise 2.4. The log-likelihood is

$$l(\boldsymbol{\beta}, b) = \sum_{i=1}^{n} \log \frac{\Gamma(y_i + b)}{\Gamma(b) y_i!} + y_i \log \mu_i + b \log b - (y_i + b) \log(\mu_i + b) \tag{2.40}$$

giving the score function for $\boldsymbol{\beta}$ as

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} \frac{y_i - \mu_i}{\mu_i + \mu_i^2 / b} \\ &= \sum_{i=1}^{n} \boldsymbol{D}(\boldsymbol{\beta})_i^{\mathrm{T}} \boldsymbol{V}_i^{-1}(b) \left[Y_i - \mu_i(\boldsymbol{\beta}) \right] \end{aligned}$$

which corresponds to (2.38). Hence, for fixed b, we can solve this estimating equation to obtain an estimator $\widehat{\boldsymbol{\beta}}$. Usually we will also wish to estimate b (as opposed to assuming a fixed value). One possibility is maximum likelihood though a quick glance at (2.40) reveals that no closed-form estimator will be available and numerical maximization will be required (which is not a great impediment). We describe an alternative method of moments estimator which may be more robust.

For the *quadratic* variance model (2.39), the variance is

$$\text{var}(Y_i \mid \boldsymbol{\beta}, b) = \mathrm{E}[(Y_i - \mu_i)^2] = \mu_i (1 + \mu_i / b),$$

so that

$$b^{-1} = \mathrm{E} \left[\frac{(Y_i - \mu_i)^2 - \mu_i}{\mu_i^2} \right],$$

for $i = 1, \ldots, n$, leading to the method of moments estimator

$$\widehat{b} = \left[\frac{1}{n - k - 1} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2 - \widehat{\mu}_i}{\widehat{\mu}_i^2} \right]^{-1}, \tag{2.41}$$

with $k = 1$ in the lung cancer example. If we have a consistent estimator $\widehat{b}$ (which follows if the quadratic variance model is correct) and the mean correctly specified, then valid inference follows from

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}} \widehat{\boldsymbol{V}}(\widehat{b})^{-1} \widehat{\boldsymbol{D}})^{1/2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}_2(\mathbf{0}, \mathrm{I}_2).$$

We fit this model to the lung cancer data. The estimates (standard errors) are $\widehat{\beta}_0 = 0.090\ (0.047)$ and $\widehat{\beta}_1 = -0.030\ (0.0085)$. The latter point estimate differs a little

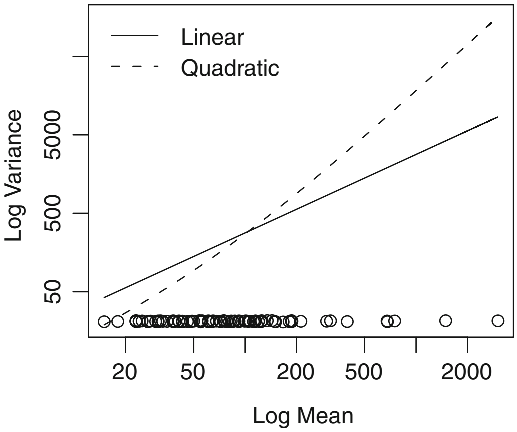

Fig. 2.2 Linear and quadratic variance functions for the lung cancer data

from the MLE (and MQLE) of -0.036, reflecting the different variance weighting in the estimating function. The moment-based estimator was $\widehat{b} = 57.8$ (the MLE is 61.3 and so close to this value). An asymptotic 95% confidence interval for the relative risk $\exp(\beta_1)$ is [0.955,0.987], so that the upper limit is closer to unity than the intervals we have seen previously.

In terms of the first two moments, the difference between quasi-likelihood and the negative binomial model is that the variances are, respectively, linear and quadratic functions of the mean. In Fig. 2.2, we plot the estimated linear and quadratic variance functions over the range of the mean for these data. To produce a clearer plot, the log of the variance is plotted against the log of the mean, and the log of the observed counts, y_i, $i = 1, \dots, 85$, is added to the plot (with a small amount of jitter). Over the majority of the data, the two variance functions are similar, but for large values of the mean in particular, the variance functions are considerably different which leads to the differences in inference, since large observations are being weighted very differently by the two variance functions. Based on this plot, we might expect even greater differences. However, closer examination of the data reveals that the x's associated with the large y values are all in the midrange, and consequently, these points are not influential.

Examination of the residuals gives some indication that the quadratic mean–variance model is more appropriate for these data (see Sect. 6.9). It is typically very difficult to distinguish between the two models, unless there are sufficient points across a large spread of mean values.

2.6 Sandwich Estimation

A general method of avoiding stringent modeling conditions when the variance of an estimator is calculated is provided by *sandwich estimation*. Recall from Sect. 2.3 the estimating function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}(\boldsymbol{\theta}, Y_i).$$

Based on independent and identically distributed observations, we have the sandwich form for the variance

$$\text{var}(\widehat{\boldsymbol{\theta}}_n) = \frac{\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{\mathrm{T}})^{-1}}{n} \tag{2.42}$$

where

$$\boldsymbol{A} = \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{G}(\boldsymbol{\theta}, Y)\right]$$

and

$$\boldsymbol{B} = \mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\mathrm{T}}].$$

For (2.42) to be asymptotically appropriate, the expectations need to be evaluated under the true model (as discussed in Sect. 2.4.3).

So far we have used an assumed model to calculate the expectations. An alternative is to evaluate $\boldsymbol{A}$ and $\boldsymbol{B}$ *empirically* via

$$\widehat{\boldsymbol{A}}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{G}(\widehat{\boldsymbol{\theta}}, Y_i),$$

and

$$\widehat{\boldsymbol{B}}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}(\widehat{\boldsymbol{\theta}}, Y_i)\boldsymbol{G}(\widehat{\boldsymbol{\theta}}, Y_i)^{\mathrm{T}}.$$

By the weak law of large numbers, $\widehat{\boldsymbol{A}}_n \to_p \boldsymbol{A}$ and $\widehat{\boldsymbol{B}}_n \to_p \boldsymbol{B}$, and

$$\text{var}(\widehat{\boldsymbol{\theta}}_n) = \frac{\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{A}}^{\mathrm{T}})^{-1}}{n} \tag{2.43}$$

is a consistent estimator of the variance. The great advantage of sandwich estimation is that it provides a consistent estimator of the variance in very broad situations. An important assumption is that the observations are uncorrelated (this will be relaxed in Part III of the book when generalized estimating equations are described).

We now consider the situation in which the estimating function arises from the score and suppose we have independent and identically distributed data. In this situation

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}l_i(\boldsymbol{\theta}),$$

with $l_i(\boldsymbol{\theta}) = \log p(Y_i \mid \boldsymbol{\theta})$, to give

$$\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}l(\boldsymbol{\theta})\right]$$

and

$$\boldsymbol{B} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta})\right)^{\mathrm{T}}\right]$$

where $l(\boldsymbol{\theta}) = \log p(Y \mid \boldsymbol{\theta})$. Then, *under the model*,

$$\boldsymbol{I}_1(\boldsymbol{\theta}) = -\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta}), \tag{2.44}$$

so that

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}_n) = \frac{\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{\mathrm{T}})^{-1}}{n} = \frac{\boldsymbol{I}_1(\boldsymbol{\theta})^{-1}}{n}.$$

The sandwich estimator (2.43) is based on

$$\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}} l_i(\boldsymbol{\theta})\bigg|_{\widehat{\boldsymbol{\theta}}}$$

and

$$\boldsymbol{B} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta})\right)^{\mathrm{T}}\bigg|_{\widehat{\boldsymbol{\theta}}}.$$

The sandwich method can be applied to general estimating functions, not just those arising from a score equation (in Sect. 2.4.3, we considered the latter in the context of model misspecification).

Suppose we assume $\mathrm{E}[Y_i] = \mu_i$ and $\mathrm{var}(Y_i) = \alpha V(\mu_i)$, and $\mathrm{cov}(Y_i, Y_j) = 0$, $i, j = 1, \ldots, n$, $i \neq j$, as a *working* covariance model. Under this specification, it is natural to take the quasi-score function (2.30) as an estimating function, and in this case, the variance of the resultant estimator is

$$\mathrm{var}_{\mathrm{S}}(\widehat{\boldsymbol{\beta}}_n) = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathrm{var}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}.$$

The appropriate variance is obtained by substituting in the correct form for $\mathrm{var}(\boldsymbol{Y})$. The latter is, of course, unknown but a simple “sandwich” estimator of the variance is given by

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_n) = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathrm{diag}(\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}})\boldsymbol{V}^{-1}\boldsymbol{D}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1},$$

where $\boldsymbol{R} = [R_1, \ldots, R_n]^{\mathrm{T}}$ is the $n \times 1$ vector of (unstandardized) residuals

$$R_i = Y_i - \mu_i(\widehat{\boldsymbol{\beta}}),$$

so that $\mathrm{diag}(\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}})$ is the $n \times n$ diagonal matrix with diagonal elements $\left[Y_i - \mu_i(\widehat{\boldsymbol{\beta}})\right]^2$ for $i = 1, \ldots, n$. This estimator is consistent for the variance of $\widehat{\boldsymbol{\beta}}$, under correct

Table 2.2 Components of estimation under the assumption of independent outcomes and for one-dimensional β

	Likelihood	Quasi-likelihood
$G(\beta) = \sum_i G_i(\beta)$	$\sum_i \frac{\partial}{\partial \beta} \log L_i$	$\frac{1}{\alpha} \sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right) \frac{Y_i - \mu_i}{V_i}$
$A = \sum_i \mathrm{E}\left[\frac{\partial G_i}{\partial \beta} \right]$	$\sum_i \mathrm{E}\left[\frac{\partial^2}{\partial \beta^2} \log L_i \right]$	$-\frac{1}{\alpha} \sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{1}{V_i}$
$\widehat{A} = \sum_i \frac{\partial G_i}{\partial \beta} \Big|_{\widehat{\beta}}$	$\sum_i \frac{\partial^2}{\partial \beta^2} \log L_i$	$-\frac{1}{\alpha} \sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{1}{V_i}$
$B = \sum_i \mathrm{E}[G_i(\beta)^2]$	$\sum_i \mathrm{E}\left[\left(\frac{\partial}{\partial \beta} \log L_i \right)^2 \right]$	$\frac{1}{\alpha} \sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{1}{V_i}$
$\widehat{B} = \sum_i G_i(\widehat{\beta})^2$	$\sum_i \left(\frac{\partial}{\partial \beta} \log L_i \right)^2$	$\frac{1}{\alpha^2} \sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{(Y_i - \widehat{\mu}_i)^2}{V_i^2}$
Model-based variance	$\left\{ \sum_i \mathrm{E}\left[\frac{\partial^2}{\partial \beta^2} \log L_i \right] \right\}^{-1}$	$\alpha \left\{ \sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right) \frac{1}{V_i} \right\}^{-1}$
Sandwich variance	$\dfrac{\sum_i \left(\frac{\partial}{\partial \beta} \log L_i \right)^2}{\left[\sum_i \frac{\partial^2}{\partial \beta^2} \log L_i \right]^2}$	$\dfrac{\sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{(Y_i - \widehat{\mu}_i)^2}{V_i^2}}{\left[\sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{1}{V_i} \right]^2}$

The likelihood model is $p(\boldsymbol{y} \mid \beta) = \prod_i L_i(\beta)$, and the quasi-likelihood model has $\mathrm{E}[Y_i \mid \beta] = \mu_i(\beta)$, $\mathrm{var}(Y_i \mid \beta) = \alpha V_i(\beta)$, $i = 1, \ldots, n$, and $\mathrm{cov}(Y_i, Y_j \mid \beta) = 0$, $i \neq j$. The expected information is $-\sum_i \mathrm{E}\left[\frac{\partial^2}{\partial \beta^2} \log L_i \right]$, and the observed information is $-\sum_i \frac{\partial^2}{\partial \beta^2} \log L_i$. The sandwich estimator is $\widehat{A}^{-1} \widehat{B} \widehat{A}^{-1}$ which simplifies to $-\widehat{A}^{-1}$ under the model

specification of the mean, and with uncorrelated data. There is finite sample bias in R_i as an estimate of $Y_i - \mu_i(\boldsymbol{\beta})$ and versions that adjust for the estimation of the parameters $\boldsymbol{\beta}$ are available; see Kauermann and Carroll (2001).

The great advantage of sandwich estimation is that it provides a consistent estimator of the variance in very broad situations and the use of the empirical residuals is very appealing. There are two things to bear in mind when one considers the use of the sandwich technique, however. The first is that, unless the sample size is sufficiently large, the sandwich estimator may be highly unstable; in terms of mean squared error, model-based estimators may be preferable for small- to medium-sized n (for small samples one would want to avoid the reliance on the asymptotic distribution anyway). Consequently, *empirical* is a better description of the estimator than *robust*. The second consideration is that if the assumed mean–variance model is correct, then a model-based estimator is more efficient.

In many cases, quasi-likelihood with a model-based variance estimate may be viewed as an intermediary between the full model specification and sandwich estimation, in that the form of the variance function separates estimation of $\boldsymbol{\beta}$ and α, to give consistency of $\boldsymbol{\beta}$ in broad circumstances, though the standard error will not be consistently estimated unless the variance function is correct. Table 2.2 provides a summary and comparison of the various elements of the likelihood and quasi-likelihood methods, with sandwich estimators for each.

Example: Poisson Mean

We report the results of a small simulation study to illustrate the efficiency-robustness trade-off of variance estimation. Data were simulated from the model $Y_i \mid \delta \sim \text{Poisson}(\delta)$, $i = 1, \dots, n$, where $\delta \sim_{iid} \text{Gamma}(\theta b, b)$. This setup gives marginal moments

$$\begin{aligned} \text{E}[Y_i] &= \theta \\ \text{var}(Y_i) &= \text{E}[Y_i] \times \left(1 + \frac{1}{b}\right) = \text{E}[Y_i] \times \alpha. \end{aligned}$$

We take $\theta = 10$ and $\alpha = 1, 2, 3$ corresponding to no excess-Poisson variability, and variability that is two and three times the mean. We estimate θ and then form an asymptotic confidence interval based on a Poisson likelihood, quasi-likelihood, and sandwich estimation.

For a univariate estimator $\widehat{\theta}$ arising from a generic estimating function $G(\theta, Y)$:

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d \text{N}\left(0, \frac{B}{A^2}\right).$$

where

$$A = \text{E}\left[\frac{d^2}{d\theta^2} G(\theta)\right], \quad B = \text{E}\left[\left(\frac{d}{d\theta} G(\theta)\right)^2\right].$$

Under the Poisson model

$$l_i(\theta) = -\theta + Y_i \log \theta$$

and

$$\begin{aligned} G(\theta, Y_i) = S_i(\theta) &= \frac{dl_i}{d\theta} = \frac{Y_i - \theta}{\theta} \\ \frac{d^2 l_i}{d\theta^2} &= -\frac{Y_i}{\theta^2}, \end{aligned}$$

to give the familiar MLE, $\widehat{\theta} = \overline{Y}$. As we already know

$$I_1(\theta) = -A = -\text{E}\left[\frac{d^2 l}{d\theta^2}\right] = B = \text{var}\left(\frac{(Y - \theta)^2}{\theta^2}\right) = \frac{\text{var}(Y)}{\theta^2} = \frac{1}{\theta},$$

under the assumption that $\text{var}(Y) = \theta$. The Poisson model-based variance estimator is therefore

$$\widehat{\text{var}}(\widehat{\theta}) = \frac{1}{n I_1(\widehat{\theta})} = \frac{\overline{Y}}{n}.$$

Under the Poisson model, the variance equals the mean, and given the efficiency of the latter, it makes sense to estimate the variance by the sample mean.

The quasi-likelihood estimator is derived from the quasi-score

$$G(\theta, Y_i) = U_i(\theta) = \frac{Y_i - \theta}{\alpha\theta},$$

and

$$\text{var}(\widehat{\theta}) = (\widehat{\boldsymbol{D}}^{\text{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}})^{-1}\widehat{\alpha}$$

where the scale parameter is estimated using the method of moments

$$\widehat{\alpha} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\theta})^2}{\widehat{\theta}}.$$

The quasi-likelihood estimator of the variance is

$$\widehat{\text{var}}(\widehat{\theta}) = \frac{s^2}{n},$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{\theta})^2.$$

For sandwich estimation based on the score

$$\widehat{A} = -\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\widehat{\theta}^2} = -\frac{1}{\overline{\overline{Y}}},$$

and

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\theta})^2}{\widehat{\theta}^2} = \frac{(n-1)s^2}{n\widehat{\theta}^2}.$$

Hence,

$$\widehat{\text{var}}(\widehat{\theta}) = \frac{s^2(n-1)/n}{n}. \tag{2.45}$$

Estimation of $\text{var}(Y_i)$ by $(Y_i - \overline{Y})^2$ produces the variance estimator (2.45). Estimating $\text{var}(Y_i)$ by $n(Y_i - \overline{Y})^2/(n-1)$ would reproduce the degrees of freedom adjusted quasi-likelihood estimator.

Table 2.3 gives the 95% confidence interval coverage for the model-based, quasi-likelihood, and sandwich estimator variance estimates as a function of the sample size n and overdispersion/scalar parameter α. We see that when the Poisson model is correct ($\alpha = 1$), the model-based standard errors produce accurate coverage for all values of n. For small n, the quasi-likelihood and sandwich estimators have low coverage, due to the instability in variance estimation, with sandwich estimation being slightly poorer in performance. As the level of overdispersion increases, the performance of the model-based approach starts to deteriorate as the standard error is underestimated, resulting in low coverage. For $\alpha = 2, 3$, the quasi-likelihood and

Table 2.3 Percent confidence interval coverage for the Poisson mean example, based on 100,000 simulations

	Overdispersion $\alpha = 1$			$\alpha = 2$			$\alpha = 3$		
n	Model	Quasi	Sand	Model	Quasi	Sand	Model	Quasi	Sand
5	95	87	84	83	87	84	74	86	83
10	94	92	90	83	91	90	73	91	89
15	95	93	92	84	92	92	75	92	91
20	95	93	93	83	93	93	73	93	92
25	95	94	93	83	94	93	74	93	93
50	95	94	94	83	94	94	74	94	94
100	95	95	94	83	95	94	74	95	94

The nominal coverage is 95%. The overdispersion is given by $\alpha = \text{var}(Y)/\text{E}[Y]$

sandwich estimators again give low coverage for small values of n, due to instability, but for larger values, the coverage quickly improves. The adjusted degrees of freedom used by quasi-likelihood give slightly improved estimation over the naive sandwich estimator.

This example shows the efficiency-robustness trade-off. If the model is correct (which corresponds here to $\alpha = 1$), then the model-based approach performs well. The sandwich and quasi-likelihood approaches are more robust to variance misspecification, but can be unstable when the sample size is small. The choice of which variance model to use depends crucially on our faith in the model. The use of a Poisson model is a risky enterprise, however, since it does not contain an additional variance parameter.

Example: Lung Cancer and Radon

Returning to the lung cancer and radon example, we calculate sandwich standard errors, assuming that counts in different areas are uncorrelated. We take as "working model" a Poisson likelihood, with maximum likelihood estimation of $\boldsymbol{\beta}$. The estimating function is

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{D}^{\text{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{x}^{\text{T}}\left(\boldsymbol{Y} - \boldsymbol{\mu}\right),$$

as derived previously, (2.20). Under this model

$$(\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{\text{T}})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \;\to_d\; \text{N}_2(\boldsymbol{0}, \text{I}_2),$$

with sandwich ingredients

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{D}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{D} \\ \boldsymbol{B} &= \boldsymbol{D}^{\text{T}}\boldsymbol{V}^{-1}\text{var}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D}, \end{aligned}$$

estimators

$$\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}}$$

$$\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\begin{bmatrix} \widehat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & \widehat{\sigma}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & \widehat{\sigma}_n^2 \end{bmatrix}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}}$$

and with $\widehat{\sigma}_i^2 = (Y_i - \widehat{\mu}_i)^2$, for $i = 1, \dots, n$. Substitution of the required data quantities yields the variance–covariance matrix

$$\begin{bmatrix} 0.043^2 & -0.87 \times 0.043 \times 0.0080 \\ -0.87 \times 0.043 \times 0.0080 & 0.0080^2 \end{bmatrix}.$$

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are 0.043 and 0.0080, respectively, and are 60% and 49% larger than their likelihood counterparts, though slightly smaller than the quasi-likelihood versions. An asymptotic 95% confidence interval for the relative risk associated with a one-unit increase in radon is $[0.949, 0.980]$.

We have a linear exponential family likelihood and so a consistent estimator of the loglinear association between lung cancer incidence and radon, as is clear from (2.20). If the outcomes are independent, then a consistent sandwich variance estimator is obtained and the large sample size indicates asymptotic inference is appropriate. However, in the context of these data, independence is a little dubious as we may have residual spatial dependence, particularly since we have not controlled for confounders such as smoking which may have spatial structure (and hence will induce spatial dependence). Sandwich standard errors do not account for such dependence (unless we can lean on replication across time). In Sect. 9.7, we describe a model that allows for residual spatial dependence in the counts. Although the loglinear association is consistently estimated, this of course says nothing about causality or about the appropriateness of the mean model.

2.7 Bootstrap Methods

With respect to estimation and hypothesis testing, the fundamental frequentist inferential summary is the distribution of an estimator under hypothetical repeated sampling from the distribution of the data. So far we have concentrated on the use of the asymptotic distribution of the estimator under an assumed model, though sandwich estimation (and to a lesser extent quasi-likelihood) provided one method by which we could relax the reliance on the assumed model. The bootstrap is a computational technique for alleviating some forms of model misspecification. The bootstrap may also be used, to some extent, to account for a "non-asymptotic" sample size. We first describe its use in single parameter settings before moving to a regression context.

2.7.1 The Bootstrap for a Univariate Parameter

Suppose $Y_1, \dots, Y_n$, are an independent and identically distributed sample from a distribution function F that depends on a univariate parameter θ. Let $\widehat{\theta}(\boldsymbol{Y})$ represent an estimator of θ. We may be interested in estimation of

(i) $\text{var}_F[\widehat{\theta}(\boldsymbol{Y})]$

(ii) $\Pr_F[a < \widehat{\theta}(\boldsymbol{Y}) < b]$

where we have emphasized that these summaries are evaluated under the sampling distribution of the data F. Estimation of (i) is of particular interest if the sampling distribution of $\widehat{\theta}$ is approximately normal, in which case a $100(1-\alpha)\%$ confidence interval is

$$\widehat{\theta}(\boldsymbol{Y}) + \text{bias}_F\left[\widehat{\theta}(\boldsymbol{Y})\right] \pm z_{1-\alpha/2}\sqrt{\text{var}_F(\widehat{\theta})} \tag{2.46}$$

where $\text{bias}_F\left[\widehat{\theta}(\boldsymbol{Y})\right]$ is the bias of the estimator, and $z_{1-\alpha/2}$ the $(1-\alpha/2)$ quantile of an $\text{N}(0,1)$ random variable. More generally, interest may focus on a function of interest $T(F)$.

The bootstrap is an idea that is so simple it seems, at first sight, like cheating but it turns out to be statistically valid in many circumstances, so long as care is taken in its implementation. The idea is to first draw B *bootstrap samples* of size n, $\boldsymbol{Y}_b^\star = [Y_{b1}^\star, \dots, Y_{bn}^\star]$, $b = 1, \dots, B$, from an estimate of F, $\widehat{F}$. In the *nonparametric* bootstrap, the estimate of F is F_n, the empirical estimate of the distribution function that places a mass of $1/n$ at each of the observed Y_i, $i = 1, \dots, n$. Bootstrap samples are obtained by sampling a new dataset $Y_{bi}^\star, i = 1, \dots, n$, from $\widehat{F}_n$, *with replacement*. If one has some faith in the assumed model, then $\widehat{F}$ may be based upon this model, which we call $F_{\hat{\theta}}$ where $\widehat{\theta} = \widehat{\theta}(\boldsymbol{y})$, to give a second implementation. In this case, bootstrap samples are obtained by sampling $Y_{bi}^\star, i = 1, \dots, n$, as independent and identically distributed samples from $F_{\hat{\theta}}$, to give a *parametric* bootstrap estimator.

Intuitively, we are replacing the distribution of

$$\widehat{\theta}_n - \theta$$

with

$$\widehat{\theta}_n^\star - \widehat{\theta}_n.$$

Much theory is available to support the use of the bootstrap; early references are Bickel and Freedman (1981) and Singh (1981); see also van der Vaart (1998). Further references to the bootstrap are given in Sect. 2.11. As a simple example of the sort of results that are available, we quote the following, a proof of which may be found in Bickel and Freedman (1981).

Result. Consider a bootstrap estimator of the sample mean, μ, of the distribution F and assume $\text{E}[Y^2] < \infty$ and let the variance of F be σ^2. Then we know that $\sqrt{n}(\overline{Y}_n - \mu) \to_d \text{N}(0, \sigma^2)$, and for almost every sequence $Y_1, Y_2, \dots,$

$$\sqrt{n}(\overline{Y}^{\star}_n - \overline{Y}_n) \to_d \text{N}(0, \sigma^2).$$

The distribution of other functions of interest can be obtained via the delta method; see van der Vaart (1998). There are two approximations that are being used in the bootstrap. First, we are estimating F by $\widehat{F}$, and second, we are estimating the quantity of interest, for example, (i) or (ii), using B samples from $\widehat{F}$. For example, if (i) is of interest, an obvious estimator of $\text{var}_F(\widehat{\theta})$ is

$$\widehat{\text{var}}_F(\widehat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left[\widehat{\theta}(\boldsymbol{Y}_b^{\star}) - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}(\boldsymbol{Y}_b^{\star})\right]^2. \tag{2.47}$$

In this case, the two approximations are

$$\text{var}_F\left(\widehat{\theta}\right) \approx \text{var}_{\widehat{F}}\left(\widehat{\theta}^{\star}\right) \approx \widehat{\text{var}}_{\widehat{F}}\left(\widehat{\theta}^{\star}\right)$$

and the first approximation may be poor if the estimate $\widehat{F}$ is not close to $\widehat{F}$, but we can control the second approximation by choosing large B. For the nonparametric bootstrap, we could, in principle, enumerate all possible samples, but there are n^n of these, of which $\begin{pmatrix} 2n-1 \\ n \end{pmatrix}$ are distinct, which is far too large a number to evaluate in practice.

There are many possibilities for computation of confidence limits, as required in (ii). If normality of $\widehat{\theta}$ is reasonable, then (2.46) is straightforward to use with the variance estimated by (2.47) and the bias by

$$\widehat{\text{bias}}_F\left[\widehat{\theta}(\boldsymbol{Y})\right] = \widehat{\theta}(\boldsymbol{y}) - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}(\boldsymbol{Y}_b^{\star}).$$

As a simple alternative, the *bootstrap percentile interval* for a confidence interval of coverage $1 - \alpha$ is

$$\left[\widehat{\theta}^{\star}_{\alpha/2}, \widehat{\theta}^{\star}_{1-\alpha/2}\right]$$

where $\widehat{\theta}^{\star}_{\alpha/2}$ and $\widehat{\theta}^{\star}_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap estimates $\widehat{\theta}(\boldsymbol{Y}_b^{\star})$, $b = 1, \dots, B$. More refined bootstrap confidence interval procedures are described in Davison and Hinkley (1997). For example, Exercise 2.9 outlines the derivation of a confidence interval based on a pivot. In Sect. 2.7.3, we illustrate the close links between bootstrap variance estimation and sandwich estimation.

The bootstrap method does not work for all functions of interest. In particular, it fails in situations when the tail behavior is not well behaved, for example, a bootstrap for the maximum $Y_{(n)}$ will be disastrous.

2.7.2 The Bootstrap for Regression

The parametric and nonparametric methods provide two distinct versions of the bootstrap, and in a regression context, another important distinction is between *resampling residuals* and *resampling cases*. We illustrate the difference by considering the model

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i, \tag{2.48}$$

where the residuals ϵ_i are such that $\mathrm{E}[\epsilon_i] = 0$, $i = 1, \ldots, n$ and are assumed uncorrelated. The two methods are characterized according to whether we take F to be the distribution of Y only or of $\{Y, \boldsymbol{X}\}$. In the resampling residuals approach, the covariates $\boldsymbol{x}_i$ are considered as fixed, and bootstrap datasets are formed as

$$Y_i^{(b)} = f(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) + \epsilon_{bi},$$

where a number of options are available for sampling ϵ_{bi}, $b = 1, \ldots, B$, $i = 1, \ldots, n$. The simplest, nonparametric, version is to sample ϵ_{bi} with replacement from

$$e_i = y_i - f(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}) - \frac{1}{n}\sum_{i=1}^{n}\left[y_i - f(\boldsymbol{x}_i, \widehat{\boldsymbol{\beta}})\right].$$

Various refinements of this simple approach are possible. If we are willing to assume (say) that $\epsilon_i \mid \sigma^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$, then a parametric resampling residuals method samples $\epsilon_{bi} \sim \mathrm{N}(0, \widehat{\sigma}^2)$ based on an estimate $\widehat{\sigma}^2$. In a model such as (2.48), the meaning of residuals is clear, but in generalized linear models (Chap. 6), for example, this is not the case and many alternative definitions exist.

The resampling residuals method has the advantage of respecting the "design," that is, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. A major disadvantage, however, is that we are leaning heavily on the assumed mean–variance relationship, and we would often prefer to protect ourselves against an assumed model. The resampling case method forms bootstrap datasets by sampling with replacement from $\{Y_i, \boldsymbol{X}_i,\ i = 1, \ldots, n\}$ and does not assume a mean–variance model. Again parametric and nonparametric versions are available, but the latter is preferred since the former requires a model for the joint distribution of the response and covariates which is likely to be difficult to specify. When cases are resampled, the design in each bootstrap sample will not in general correspond to that in the original dataset which, though not ideal (since it leads to wider confidence intervals than necessary), will have little impact on inference, except when there are outliers in the data; if the outliers are sampled multiple times, then instability may result.

2.7.3 Sandwich Estimation and the Bootstrap

In this section we heuristically show why we would often expect sandwich and bootstrap variance estimates to be in close correspondence. For simplicity, we consider a univariate parameter θ, and let $\widehat{\theta}_n$ denote the MLE arising from a sample of size n. In a change of notation, we denote the score by $\boldsymbol{S}(\theta) = [S_1(\theta), \ldots, S_n(\theta)]^{\mathrm{T}}$, where $S_i(\theta) = dl_i/d\theta$ is the contribution to the score from observation Y_i, $i = 1, \ldots, n$. Hence,

$$S(\theta) = \sum_{i=1}^{n} S_i(\theta) = \boldsymbol{S}(\theta)^{\mathrm{T}}\mathbf{1}$$

where $\mathbf{1}$ is an $n \times 1$ vector of 1's. The sandwich form of the asymptotic variance of $\widehat{\theta}_n$ is

$$\mathrm{var}(\widehat{\theta}_n) = \frac{1}{n}\frac{B}{A^2}$$

where

$$A(\theta) = \mathrm{E}\left[\frac{dS}{d\theta}\right], \quad B(\theta) = \mathrm{E}\left[S(\theta)^2\right].$$

These quantities may be empirically estimated via

$$\widehat{A}_n = \frac{1}{n}\frac{dS}{d\theta}\bigg|_{\widehat{\theta}_n} = \frac{1}{n}\sum_{i=1}^{n}\frac{dS_i}{d\theta}\bigg|_{\widehat{\theta}_n}$$

$$\widehat{B}_n = \frac{1}{n}\boldsymbol{S}(\theta)^{\mathrm{T}}\boldsymbol{S}(\theta)\bigg|_{\widehat{\theta}_n} = \frac{1}{n}\sum_{i=1}^{n}S_i(\theta)^2\bigg|_{\widehat{\theta}_n}.$$

A convenient representation of a bootstrap sample is $\boldsymbol{Y}^\star = \boldsymbol{Y} \times \boldsymbol{D}$ where $\boldsymbol{D} = \mathrm{diag}(D_1, \ldots, D_n)$ is a diagonal matrix consisting of a multinomial random variable

$$\begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix} \sim \text{Multinomial}\left[n, \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)\right]$$

with

$$\mathrm{E}\left([D_1, \ldots, D_n]^{\mathrm{T}}\right) = \mathbf{1}$$

$$\mathrm{var}\left([D_1, \ldots, D_n]^{\mathrm{T}}\right) = \mathrm{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}} \to \mathrm{I}_n$$

as $n \to \infty$. The MLE of θ in the bootstrap sample is denoted $\widehat{\theta}_n^\star$ and satisfies $S^\star(\widehat{\theta}_n^\star) = 0$, where $S^\star(\theta)$ is the score corresponding to $\boldsymbol{Y}^\star$. Note that

$$S^\star(\theta) = \sum_{i=1}^{n} S_i^\star(\theta) = \sum_{i=1}^{n} S_i(\theta)D_i.$$

We consider a one-step Newton–Raphson approximation (see Sect. 6.5.2 for a more detailed description of this method) to $\widehat{\theta}_n^\star$ and show that this leads to a bootstrap variance estimate that is approximately equal to the sandwich variance estimate. The following informal derivation is carried out without stating regularity conditions. It is important to emphasize that throughout we are conditioning on $\boldsymbol{Y}$ and therefore on $\widehat{\theta}_n$. A first-order Taylor series approximation

$$0 = S^\star(\widehat{\theta}_n^\star) \approx S^\star(\widehat{\theta}_n) + (\widehat{\theta}_n^\star - \widehat{\theta}_n) \left. \frac{dS^\star}{d\theta} \right|_{\widehat{\theta}_n}$$

leads to the one-step approximation

$$\widehat{\theta}_n^\star \approx \widehat{\theta}_n - \frac{S^\star(\widehat{\theta}_n)}{\frac{d}{d\theta} S^\star(\theta)|_{\widehat{\theta}_n}}.$$

The bootstrap score evaluated at $\widehat{\theta}_n$ is

$$\sum_{i=1}^n S_i^\star(\widehat{\theta}_n) = \sum_{i=1}^n S_i(\widehat{\theta}_n) D_i \neq 0,$$

unless the bootstrap sample coincides with the original sample, that is, unless $\boldsymbol{D} = \mathrm{I}_n$. We replace $\left[\frac{d}{d\theta} S^\star(\theta)|_{\widehat{\theta}_n}\right]$ by its limit

$$\mathrm{E}\left[\left. \frac{d}{d\theta} S^\star(\theta) \right|_{\widehat{\theta}_n} \right] = \mathrm{E}\left[\sum_{i=1}^n \left. \frac{d}{d\theta} S_i(\theta) D_i \right|_{\widehat{\theta}_n} \right] = \sum_{i=1}^n \left. \frac{d}{d\theta} S_i(\theta) \right|_{\widehat{\theta}_n} \mathrm{E}[D_i] = n \times \widehat{A}_n$$

where $\widehat{A}_n = \frac{1}{n} \frac{d}{d\theta} S(\theta)\big|_{\widehat{\theta}_n}$. Therefore, the one-step bootstrap estimator is approximated by

$$\widehat{\theta}_n^\star \approx \widehat{\theta}_n - \frac{\boldsymbol{S}(\widehat{\theta}_n)^{\mathrm{T}} \boldsymbol{D}}{n \widehat{A}_n}$$

and is approximately unbiased as an estimator since

$$\mathrm{E}[\widehat{\theta}_n^\star - \widehat{\theta}_n] \approx -\frac{\boldsymbol{S}(\widehat{\theta}_n)^{\mathrm{T}} \mathrm{E}[\boldsymbol{D}]}{n \widehat{A}_n} = -\frac{\boldsymbol{S}(\widehat{\theta}_n)^{\mathrm{T}} \mathbf{1}}{n \widehat{A}_n} = 0$$

and, recall, $\widehat{\theta}_n$ is being held constant. The variance is

$$\begin{aligned} \mathrm{var}(\widehat{\theta}_n^\star - \widehat{\theta}_n) &\approx \frac{\boldsymbol{S}(\widehat{\theta}_n)^{\mathrm{T}} \mathrm{var}([D_1, \ldots, D_n]^{\mathrm{T}}) \boldsymbol{S}(\widehat{\theta}_n)}{(n \widehat{A}_n)^2} = \frac{\boldsymbol{S}(\widehat{\theta}_n)^{\mathrm{T}} \left(\mathrm{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^{\mathrm{T}}\right) \boldsymbol{S}(\widehat{\theta}_n)}{(n \widehat{A}_n)^2} \\ &\approx \frac{\boldsymbol{S}(\widehat{\theta}_n)^{\mathrm{T}} \mathrm{I} \boldsymbol{S}(\widehat{\theta}_n)}{(n \widehat{A}_n)^2} = \frac{n \widehat{B}_n}{(n \widehat{A}_n)^2} = \frac{\widehat{B}_n}{n \widehat{A}_n^2}, \end{aligned}$$

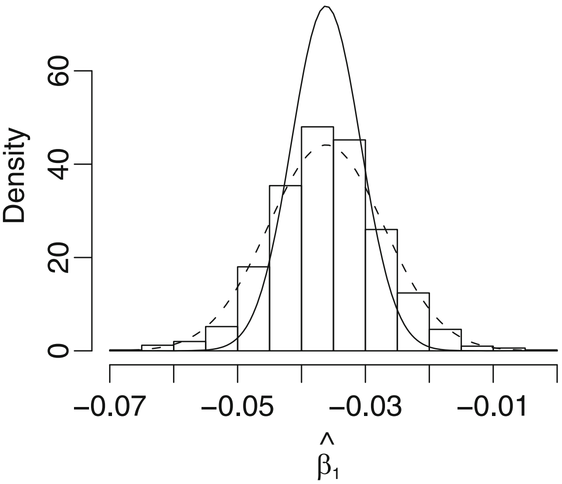

Fig. 2.3 Sampling distribution of $\widehat{\beta}_1$ arising from the nonparametric bootstrap samples. The *solid curve* is the asymptotic distribution of the MLE under the Poisson model, and the *dashed line* is the asymptotic distribution under the quasi-Poisson model

which is the sandwich estimator. Hence, $\text{var}(\widehat{\theta}_n^\star - \widehat{\theta}_n)$ approximates $\text{var}(\widehat{\theta}_n - \theta)$, which is a fundamental link in the bootstrap. For a more theoretical treatment, see Arcones and Giné (1992) and Sect. 10.3 of Kosorok (2008).

Example: Lung Cancer and Radon

For the lung cancer and radon example, we implement the nonparametric bootstrap resampling $B = 1{,}000$ sets of n case triples $[Y_{bi}^\star, E_{bi}^\star, x_{bi}^\star]$, $b = 1, \ldots, B$, $i = 1, \ldots, n$. Figure 2.3 displays the histogram of estimates arising from the bootstrap samples, along with the asymptotic normal approximations to the sampling distribution of the estimator under the Poisson and quasi-Poisson models. We see that the distribution under the quasi-likelihood model is much wider than that under the Poisson model. This is not surprising since we have already seen that the lung cancer data are overdispersed relative to a Poisson distribution. The bootstrap histogram and quasi-Poisson sampling distribution are very similar, however.

Table 2.4 summarizes inference for β_1 under a number of different methods and again confirms the similarity of asymptotic inference under the quasi-Poisson model and nonparametric bootstrap. In this example the similarity in the intervals from quasi-likelihood, sandwich estimation, and the nonparametric bootstrap is reassuring. The point estimates from the Poisson, quasi-likelihood, and sandwich approaches are identical. The point estimate from the quadratic variance model (that arises from a negative binomial model) is slightly closer to zero for these data, due to the difference in the variance models over the large range of counts in these data.

Table 2.4 Comparison of inferential summaries over various approaches, for the lung cancer and radon example

Inferential method	$\widehat{\beta}_1$	s.e.$(\widehat{\beta}_1)$	95% CI for $\exp(\beta_1)$
Poisson	−0.036	0.0054	$0.954,\ 0.975$
Quasi-likelihood	−0.036	0.0090	$0.947,\ 0.982$
Quadratic variance	−0.030	0.0085	$0.955,\ 0.987$
Sandwich estimation	−0.036	0.0080	$0.949,\ 0.980$
Bootstrap normal	−0.036	0.0087	$0.948,\ 0.981$
Bootstrap percentile	−0.036	0.0087	$0.949,\ 0.981$

The last two lines refer to nonparametric bootstrap approaches, with intervals based on normality of the sampling distribution of the estimator ("Normal") and on taking the 2.5% and 97.5% points of this distribution ("Percentile")

2.8 Choice of Estimating Function

The choice of estimating function is driven by the conflicting aims of *efficiency* and *robustness to model misspecification*. If the likelihood corresponds to the true model, then MLEs are asymptotically efficient so that asymptotic confidence intervals have minimum length. However, if the assumed model is incorrect, then there are no guarantees of even consistency of estimation.

Basing estimating functions on simple model-free functions of the data often provides robustness. As we discuss in Sect. 5.6.3, the classic Gauss–Markov theorem states, informally, that among estimators that are linear in the data, the least squares estimator has smallest variance, and this result is true for fixed sample sizes. There is also a Gauss–Markov theorem for estimating functions. Suppose $\text{E}[Y_i \mid \boldsymbol{\beta}] = \mu_i(\boldsymbol{\beta})$, $\text{var}(Y_i) = \sigma_i^2$ and $\text{cov}(Y_i, Y_j) = 0$, $i \neq j$, and consider the class of *linear unbiased estimating functions* (of zero) that are of the form

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{n} a_i(\boldsymbol{\beta}) \left[Y_i - \mu_i(\boldsymbol{\beta}) \right], \tag{2.49}$$

where $a_i(\boldsymbol{\beta})$ are specified nonrandom functions, subject to $\sum_{i=1}^{n} a_i(\boldsymbol{\beta}) = c$, a constant (this is to avoid obtaining an arbitrarily small variance by multiplying the estimating function by a constant). The estimating function (2.49) provides a consistent estimator $\widehat{\boldsymbol{\beta}}$ so long as the mean $\mu_i(\boldsymbol{\beta})$ is correctly specified. It can be shown, for example, Godambe and Heyde (1987), that

$$\text{E}[\boldsymbol{U}\boldsymbol{U}^{\text{T}}] \leq \text{E}[\boldsymbol{G}\boldsymbol{G}^{\text{T}}], \tag{2.50}$$

where

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{D}^{\text{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})/\alpha,$$

so that this estimating function has the smallest variance. Quasi-likelihood estimators are therefore asymptotically optimal in the class of linear estimating functions and will be asymptotically efficient if the quasi-score functions correspond to the score of the likelihood of the true data-generating model. Of course a superior estimator (in terms of efficiency) may result from an estimating function that is not linear in the data, if the data arise from a model for which the score function is not linear. The consideration of *quadratic* estimating functions illustrates the efficiency-robustness trade-off.

Result (2.50) is true for an *estimating function* based on a finite sample size n, though there is no such result for the derived *estimator*. However, the estimator derived from the estimating function is asymptotically efficient (e.g., McCullagh 1983). The optimal estimating equation is that which has minimum expected distance from the score equation corresponding to the true model. We reemphasize that a consistent estimator of the parameters in the assumed regression model is obtained from the quasi-score (2.50), and the variance of the estimator will be appropriate so long as the second moment of the data has been specified correctly.

To motivate the class of quadratic estimating functions suppose

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} N\left[\mu_i(\boldsymbol{\beta}), \sigma_i^2(\boldsymbol{\beta})\right],$$

$i = 1, \dots, n$. The log-likelihood is

$$l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \log \sigma_i(\boldsymbol{\beta}) - \frac{1}{2}\sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\sigma_i(\boldsymbol{\beta})^2},$$

which gives the quadratic score equations

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \frac{\partial l}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{\{Y_i - \mu_i(\boldsymbol{\beta})\}}{\sigma_i(\boldsymbol{\beta})^2}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \frac{\left\{[Y_i - \mu_i(\boldsymbol{\beta})]^2 - \sigma_i(\boldsymbol{\beta})^2\right\}}{\sigma_i(\boldsymbol{\beta})^3}\frac{\partial \sigma_i}{\partial \boldsymbol{\beta}}. \end{aligned} \qquad (2.51)$$

If the first two moments are correctly specified, then $\mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})] = \boldsymbol{0}$, so that a consistent estimator is obtained.

In general, we may consider

$$\sum_{i=1}^{n} a_i(\boldsymbol{\beta})\left[Y_i - \mu_i(\boldsymbol{\beta})\right] + b_i(\boldsymbol{\beta})\left\{[Y_i - \mu_i(\boldsymbol{\beta})]^2 - \sigma_i(\boldsymbol{\beta})^2\right\},$$

where $a_i(\boldsymbol{\beta}), b_i(\boldsymbol{\beta})$ are specified nonrandom functions. With this estimating function, the information in the variance concerning the parameters $\boldsymbol{\beta}$ is being used to improve efficiency. Among quadratic estimating functions, it can be shown that (2.51) is optimal in the sense of producing estimators that are asymptotic efficient (Crowder 1987). In general, to choose the optimal estimating function, the first four moments of the data must be known, which may seem unlikely, but this approach may be contrasted with the use of the score as estimating function which effectively requires all of the moments to be known. There are two problems with using quadratic estimating functions. First, consistency requires the first two moments to be correctly specified. Second, to estimate the covariance matrix of the estimator, the skewness and kurtosis must be estimated, and these may be highly unstable. We return to this topic in Sect. 9.10.

2.9 Hypothesis Testing

Throughout the book, we emphasize estimation over hypothesis testing, for reasons discussed in Chap. 4, but in this section describe the rationale and machinery of frequentist hypothesis testing.

2.9.1 *Motivation*

A common aim of statistical analysis is to judge the evidence from the data in support of a particular hypothesis, defined through specific parameter values. Hypothesis tests have historically been used for various purposes, including:

- Determining whether a set of data is *consistent* with a particular hypothesis
- Making a *decision* as to which of two hypotheses is best supported by the data

We assume there exists a test statistic $T = T(\boldsymbol{Y})$ with large values of T suggesting departures from H_0. In Sects. 2.9.3–2.9.5, three specific recipes are described, namely, score, Wald, and likelihood ratio test statistics. We define the p-value, or *significance level*, as

$$p = p(\boldsymbol{Y}) = \Pr\left[\, T(\boldsymbol{Y}) > T(\boldsymbol{y}) \mid H_0 \,\right],$$

so that, intuitively, if this probability is "small," the data are inconsistent with H_0. If $T(\boldsymbol{Y})$ is continuous, then under H_0, the p-value $p(\boldsymbol{Y})$ follows the distribution $\mathrm{U}(0,1)$. Consequently, the significance level is the observed $p(\boldsymbol{y})$. The distribution of $T(\boldsymbol{Y})$ under H_0 may be known analytically or may be simulated to produce a Monte Carlo or bootstrap test.

The nomenclature associated with the broad topic of hypothesis testing is confusing, but we distinguish three procedures:

1. A *pure significance test* calculates p but does not reject H_0 and is often viewed as an exploratory tool.
2. A *test of significance* sets a cutoff value α (e.g., $\alpha = 0.05$) and rejects H_0 if $p < \alpha$ corresponding to $T > T_\alpha$. The latter is known as the *critical region*.
3. A *hypothesis test* goes one step further and specifies an *alternative hypothesis*, H_1. One then reports whether H_0 is rejected or not. The null hypothesis has

special position as the "status quo," and conventionally the phrase "accept H_0" is not used because not rejecting may be due to low power (perhaps because of a small sample size) as opposed to H_0 being true.

Rejecting H_0 when it is true is known as a type I error, and a type II error occurs when H_0 is not rejected when it is in fact false. To evaluate the probability of a type II error, specific alternative values of the parameters need to be considered. The *power* is defined as the probability of rejecting H_0 when it is false. We emphasize that a test of significance may reject H_0 for general departures, while a hypothesis test rejects in the specific direction of H_1.

A key point is that the consistency of the data with H_0 is being evaluated, and there is no reference to the probability of the null hypothesis being true. As usual in frequentist inference, H_0 is a fixed unknown and probability statements cannot be assigned to it.[4]

2.9.2 Preliminaries

We consider a p-dimensional vector of parameters $\boldsymbol{\theta}$ and consider two testing situations. In the first, we consider the *simple* null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus the alternative $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. In the second, we consider a partition of the parameter vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$, where the dimensions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are $p - r$ and r, respectively, and a *composite* null. Specifically, in the composite case, we compare the hypotheses:

$$\begin{aligned} H_0 &: \boldsymbol{\theta}_1 \text{ unrestricted, } \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}, \\ H_1 &: \boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2] \neq [\boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}]. \end{aligned}$$

As a simple example, in a regression context, let $\boldsymbol{\theta} = [\theta_1, \theta_2]$ with θ_1 the intercept and θ_2 the slope. We may then be interested in $H_0 : \theta_2 = 0$ with θ_1 unspecified. In both the simple and composite situations, the unrestricted MLE under the alternative is denoted $\widehat{\boldsymbol{\theta}}_n = [\widehat{\boldsymbol{\theta}}_{n1}, \widehat{\boldsymbol{\theta}}_{n2}]$.

For simplicity of exposition, unless stated otherwise, we suppose that the responses Y_i, $i = 1, \ldots, n$, are independent and identically distributed. Consequently we have $p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta})$. The extension to the nonidentically distributed situation, as required for regression, is straightforward. The $p \times 1$ score vector is

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

[4] As described in Chap. 3, in the Bayesian approach to hypothesis testing, a prior distribution is placed on the alternatives (and on the null), allowing the calculation of the probability of H_0 given the data, relative to other hypotheses under consideration.

where $l_i(\boldsymbol{\theta})$ is the log-likelihood contribution from observation i, $i = 1, \dots, n$. Let $\boldsymbol{S}_n(\boldsymbol{\theta}) = [\boldsymbol{S}_{n1}(\boldsymbol{\theta}), \boldsymbol{S}_{n2}(\boldsymbol{\theta})]^{\mathrm{T}}$ be a partition of the score vector with $\boldsymbol{S}_{n1}(\boldsymbol{\theta})$ of dimension $(p-r) \times 1$ and $\boldsymbol{S}_{n2}(\boldsymbol{\theta})$ of dimension $r \times 1$. Under the composite null, let $\widehat{\boldsymbol{\theta}}_n^0 = [\widehat{\boldsymbol{\theta}}_{n10}, \boldsymbol{\theta}_{20}]$ denote the MLE, where $\widehat{\boldsymbol{\theta}}_{n10}$ is found from the estimating equation

$$\boldsymbol{S}_{n1}(\widehat{\boldsymbol{\theta}}_{n10}, \boldsymbol{\theta}_{20}) = \mathbf{0}.$$

In general, $\widehat{\boldsymbol{\theta}}_{n10} \neq \widehat{\boldsymbol{\theta}}_{n1}$.

In the independent and identically distributed case, $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}_1(\boldsymbol{\theta})$ is the information in a sample of size n. Suppressing the dependence on $\boldsymbol{\theta}$, let

$$\boldsymbol{I}_1 = \begin{bmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{bmatrix}$$

denote a partition of the expected information matrix, where $\boldsymbol{I}_{11}$, $\boldsymbol{I}_{12}$, $\boldsymbol{I}_{21}$, and $\boldsymbol{I}_{22}$ are of dimensions $(p-r) \times (p-r)$, $(p-r) \times r$, $r \times (p-r)$, and $r \times r$, respectively. The inverse of $\boldsymbol{I}_1$ is

$$\boldsymbol{I}_1^{-1} = \begin{bmatrix} \boldsymbol{I}_{11\cdot 2}^{-1} & -\boldsymbol{I}_{11\cdot 2}^{-1}\boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1} \\ -\boldsymbol{I}_{22\cdot 1}^{-1}\boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1} & \boldsymbol{I}_{22\cdot 1}^{-1} \end{bmatrix}$$

where

$$\boldsymbol{I}_{11\cdot 2} = \boldsymbol{I}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21}$$
$$\boldsymbol{I}_{22\cdot 1} = \boldsymbol{I}_{22} - \boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1}\boldsymbol{I}_{12}$$

using results from Appendix B.

2.9.3 Score Tests

We begin with the simple null $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Recall the asymptotic distribution of the score, given in (2.17):

$$n^{-1/2}\boldsymbol{S}_n(\boldsymbol{\theta}) \ \to_d \ \mathrm{N}_p\left[\mathbf{0}, \boldsymbol{I}_1(\boldsymbol{\theta})\right].$$

Therefore, under the null hypothesis

$$\boldsymbol{S}_n(\boldsymbol{\theta}_0)^{\mathrm{T}}\boldsymbol{I}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{S}_n(\boldsymbol{\theta}_0)/n \ \to_d \ \chi_p^2. \tag{2.52}$$

Intuitively, if the elements of $\boldsymbol{S}_n(\boldsymbol{\theta}_0)$ are large, this means that the components of the gradient at $\boldsymbol{\theta}_0$ are large. The latter occurs when $\boldsymbol{\theta}_0$ is "far" from the estimator $\widehat{\boldsymbol{\theta}}_n$ for which $\boldsymbol{S}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. In (2.52), the matrix $\boldsymbol{I}_1^{-1}(\boldsymbol{\theta}_0)$ is scaling the gradient distance. The information may be evaluated at the MLE, $\widehat{\boldsymbol{\theta}}_n$, rather than at $\boldsymbol{\theta}_0$, since $\boldsymbol{I}_1(\widehat{\boldsymbol{\theta}}_n) \to_p \boldsymbol{I}_1(\boldsymbol{\theta}_0)$, by the weak law of large numbers.

Under the composite null hypothesis, $H_0 : \boldsymbol{\theta}_1$ unrestricted, $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$:

$$\boldsymbol{S}_n(\widehat{\boldsymbol{\theta}}_n^0)^{\mathrm{T}} \boldsymbol{I}_1^{-1}(\widehat{\boldsymbol{\theta}}_n^0) \boldsymbol{S}_n(\widehat{\boldsymbol{\theta}}_n^0)/n \to_d \chi_r^2.$$

As a simplification, we can express this statistic in terms of partitioned information matrices. Since r elements of the score vector are zero, that is, $\boldsymbol{S}_{n2}(\widehat{\boldsymbol{\theta}}_n^0) = \boldsymbol{0}$, we have

$$\boldsymbol{S}_{n1}(\widehat{\boldsymbol{\theta}}_n^0)^{\mathrm{T}} \boldsymbol{I}_{11\cdot 2}^{-1}(\widehat{\boldsymbol{\theta}}_n^0) \boldsymbol{S}_{n1}(\widehat{\boldsymbol{\theta}}_n^0)/n \to_d \chi_r^2.$$

Hence, the model only needs to be fitted under the null. Each of the score statistics remains asymptotically valid on replacement of the expected information by the observed information.

2.9.4 Wald Tests

Under the simple null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, the Wald statistic is based upon the asymptotic distribution

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \ \to_d \ \mathrm{N}_p \left[\boldsymbol{0}, \boldsymbol{I}_1(\boldsymbol{\theta}_0)^{-1}\right], \tag{2.53}$$

and the Wald statistic is the quadratic form based on (2.53):

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}} \boldsymbol{I}_1(\boldsymbol{\theta}_0) \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \to_d \quad \chi_p^2. \tag{2.54}$$

An alternative form that is often used in practice is

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}} \boldsymbol{I}_1(\widehat{\boldsymbol{\theta}}_n) \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \to_d \quad \chi_p^2,$$

which again follows because $\boldsymbol{I}_1(\widehat{\boldsymbol{\theta}}_n) \to_p \boldsymbol{I}_1(\boldsymbol{\theta}_0)$, by the weak law of large numbers.

Under a composite null hypothesis, the Wald statistic is based on the marginal distribution of $\widehat{\boldsymbol{\theta}}_{n2}$:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{n2} - \boldsymbol{\theta}_{20})^{\mathrm{T}} \boldsymbol{I}_{11\cdot 2}(\widehat{\boldsymbol{\theta}}_n^0) \sqrt{n}(\widehat{\boldsymbol{\theta}}_{n2} - \boldsymbol{\theta}_{20}) \ \to_d \ \chi_r^2.$$

The observed information may replace the expected information in either form of the Wald statistic.

2.9.5 Likelihood Ratio Tests

Finally, we consider the likelihood ratio statistic which, under a simple null, is

$$2\left[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)\right].$$

Unlike the score and Wald statistics, the asymptotic distribution is not an obvious quadratic form, and so we provide a sketch proof of the asymptotic distribution under H_0. A second-order Taylor expansion of $l_n(\boldsymbol{\theta}_0)$ about $\widehat{\boldsymbol{\theta}}_n$ gives

$$l_n(\boldsymbol{\theta}_0) = l_n(\widehat{\boldsymbol{\theta}}_n) + (\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n)^{\mathrm{T}} \left.\frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\widehat{\boldsymbol{\theta}}_n} + \frac{1}{2}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n)^{\mathrm{T}} \left.\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\widetilde{\boldsymbol{\theta}}} (\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_n),$$

where $\widetilde{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$. The middle term on the right-hand side is zero, and

$$\frac{1}{n} \left.\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\widetilde{\boldsymbol{\theta}}} \to_p -\boldsymbol{I}_1(\boldsymbol{\theta}_0).$$

Hence,

$$\begin{aligned} -2\left[l_n(\boldsymbol{\theta}_0) - l_n(\widehat{\boldsymbol{\theta}}_n)\right] &= 2\left[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)\right] \\ &\approx n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathrm{T}} \boldsymbol{I}_1(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \end{aligned}$$

and so

$$2\left[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}_0)\right] \to_d \chi_p^2. \tag{2.55}$$

Similarly, under a composite null hypothesis:

$$2\left[l_n(\widehat{\boldsymbol{\theta}}_n) - l_n(\widehat{\boldsymbol{\theta}}_n^0)\right] \to_d \chi_r^2.$$

2.9.6 *Quasi-likelihood*

We briefly consider the quasi-likelihood model described in Sect. 2.5. The score test can be based on the quasi-score statistic $\boldsymbol{U}_n(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})/\alpha$, with the information in a sample of size n being $\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha$. The latter is also used in the calculation of a Wald statistic since it supplies the required standard errors. Similarly, a quasi-likelihood ratio test can be performed using $l_n(\widehat{\boldsymbol{\theta}}_n, \alpha)$, the form of which is given in (2.32). Unknown α can be accommodated by substitution of a consistent estimator $\widehat{\alpha}$. For example, we might estimate α via the Pearson statistic estimator (2.31).

If one wished to account for estimation of α, then one possibility is to assume that $(n - p) \times \widehat{\alpha}$ follows a χ^2_{n-p} distribution and then evaluate significance based on the ratio of scaled χ^2-squared random variables, to give an F distribution under the null (see Appendix B). Outside of the normal linear model, this seems a dubious exercise, however, since the numerator and denominator will not be independent, and either of the χ^2 approximations could be poor. The use of an F statistic is conservative, however (so that significance will be reduced over the use of the plug-in χ^2 approximation).

2.9.7 Comparison of Test Statistics

The score test statistic is invariant under reparameterization, provided that the expected, rather than the observed, information is used. The score statistic may also be evaluated without second derivatives if $\boldsymbol{S}_n(\boldsymbol{\theta}_0)\boldsymbol{S}_n(\boldsymbol{\theta}_0)^{\mathrm{T}}$ is used, which may be useful if these derivatives are complex, or unavailable. The score statistic requires the value of the score at the null, but the MLE under the alternative is not required.

Confidence intervals can be derived directly from the Wald statistic so that there is a direct link between estimation and testing. Interpretation is also straightforward; in particular, statistical versus practical significance can be immediately considered. A major drawback of the Wald statistic is that it is not invariant to the parameterization chosen, which ties in with our earlier observation (Sect. 2.3) that asymptotic confidence intervals are more accurate on some scales than on others. The Wald statistic uses the MLE but not the value of the maximized likelihood.

The likelihood ratio statistic is invariant under reparameterization. Confidence intervals derived from likelihood ratio tests always preserve the support of the parameter, unlike score- and Wald-based intervals (unless a suitable parameterization is adopted). Similar to the attainment of the Cramér–Rao lower bound (Appendix F), there is an elegant theory under which the likelihood ratio test statistic emerges as the *uniformly most powerful* (UMP) test, via the famous Neyman–Pearson lemma; see, for example, Schervish (1995). The likelihood ratio test requires the fitting of two models.

The score, Wald, and likelihood ratio test statistics are asymptotically equivalent but are not equally well behaved in finite samples. In general, and by analogy with the asymptotic optimality of the MLE, the likelihood ratio statistic is often recommended for use in regular models. If $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$ are close, then the three statistics will tend to agree.

Chapter 4 provides an extended discussion and critique of hypothesis testing.

Example: Poisson Mean

We illustrate the use of the three statistics in a simple context. Suppose we have data $Y_i \mid \lambda \sim_{iid} \text{Poisson}(\lambda)$, $i = 1, \ldots, n$, and we are interested in $H_0 : \lambda = \lambda_0$. The log-likelihood, score, and information are

$$\begin{aligned} l_n(\lambda) &= -n\lambda + n\overline{Y}\log\lambda, \\ S_n(\lambda) &= -n + \frac{n\overline{Y}}{\lambda} = \frac{n(\overline{Y} - \lambda)}{\lambda}, \\ I_n(\lambda) &= \frac{n}{\lambda}. \end{aligned}$$

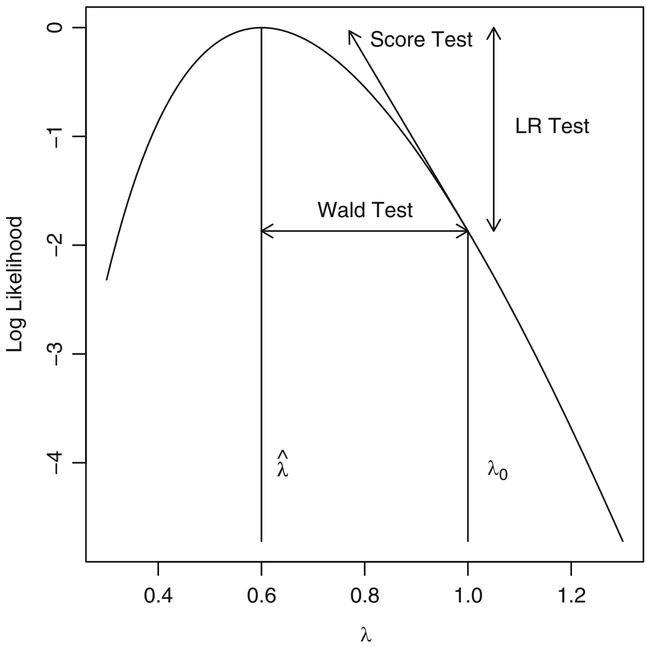

Fig. 2.4 Geometric interpretation of score, Wald, and likelihood ratio (LR) statistics, for Poisson data and a test of $H_0 : \lambda_0 = 1$, with data resulting in $\widehat{\lambda} = \overline{y} = 0.6$

The score and Wald statistics follow from (2.52) and (2.54) and both lead to

$$\frac{n(\overline{Y} - \lambda_0)^2}{\lambda_0} \rightarrow_d \chi_1^2$$

under the null. From (2.55), the likelihood ratio statistic is

$$2n\left[\overline{Y}(\log \overline{Y} - \log \lambda_0) - (\overline{Y} - \lambda_0)\right] \rightarrow_d \chi_1^2.$$

Suppose we observe $\sum_{i=1}^{20} y_i = 12$ events in $n = 20$ trials so that $\widehat{\lambda} = \overline{y} = 0.6$. Assume we are interested in testing the null hypothesis $H_0 : \lambda_0 = 1.0$. The score and Wald statistics are 3.20 and the likelihood ratio statistic is 3.74, with associated observed significance levels of 7.3% and 5.4%, respectively. Figure 2.4 plots the log-likelihood against λ for these data. The (unscaled) statistics are indicated on the figure. The score test is based on the gradient at λ_0, the Wald statistic is the squared horizontal distance between $\widehat{\lambda}$ and λ_0, and the likelihood ratio test statistic is two times the vertical distance between $l(\widehat{\lambda})$ and $l(\lambda_0)$.

We now reparameterize to $\theta = \log \lambda$, so that the null becomes $H_0 : \theta = \theta_0 = 0$. The likelihood ratio statistic is invariant to parameterization, and the score statistic turns out to be the same as previously in this example, since the observed and expected information are equal. The forms of the Wald, score, and likelihood ratio statistics, for general θ_0, are

$$n(\log \overline{Y} - \theta_0)^2 \exp(\theta_0)$$

$$n\left[\overline{Y} - \exp(\theta_0)\right]^2 \exp(-\theta_0)$$

$$2n\left\{\overline{Y}(\widehat{\theta} - \theta_0) - [\exp(\widehat{\theta}) - \exp(\theta_0)]\right\}$$

with numeric values of 5.22, 3.20 and 3.74, respectively, in the example.

2.10 Concluding Remarks

In Sect. 1.2, we emphasized that model formulation should begin with the model that is felt most appropriate for the context, before proceeding to determine the behavior of inferential procedures under this model. In this chapter we have seen that likelihood-based inference is asymptotically efficient *if* the model is correct. Hence, if one has strong belief in the assumed model, then a likelihood approach is appealing, particularly if the score equations are of linear exponential family form, since in this case consistent estimators of the parameters in the assumed regression model are obtained. If the likelihood is not of linear exponential form, then there are no guarantees of consistency under model misspecification. So far as estimation of the standard error is concerned, in situations in which n is sufficiently large for asymptotic inference to be accurate, sandwich estimation or the bootstrap may be used to provide consistent model-free standard errors, so long as the observations are uncorrelated. The relevance of asymptotic calculations for particular sample sizes may be investigated via simulation. In general, sandwich estimation is a very simple, broadly applicable and appealing technique.

In many instances the context and/or questions of interest may determine the mean function and perhaps give clues to the mean–variance relationship. The form of the data may suggest viable candidates for the full probability model. A caveat to this is that models such as the Poisson or exponential for which there is no dispersion parameter should be used with extreme caution since there is no mechanism to "soak up" excess variability. In practice, if the data exhibit overdispersion, as is often the case, then this will lead to confidence intervals that are too short. Information on the mean and variance may be used within a quasi-likelihood approach to define an estimator, and if n is sufficiently large, sandwich estimation can provide reliable standard errors. Experience of particular models may help to determine whether the assumption of a particular likelihood with the desired mean and variance functions is likely to be much less reliable than a quasi-likelihood approach. The choice of how parametric one wishes to be will often come down to personal taste.

We finally note that the efficiency-robustness trade-off will be weighted in different directions depending on the nature of the analysis. In an exploratory setting, one may be happy to proceed with a likelihood analysis, while in a confirmatory setting, one may want to be more conservative.

2.11 Bibliographic Notes

Numerous accounts of the theory behind frequentist inference are available, Cox and Hinkley (1974) remains a classic text. Casella and Berger (1990) also provides an in-depth discussion of frequentist estimation and hypothesis testing. A mathematically rigorous treatment of the estimating functions approach is provided by van der Vaart (1998). A gentler and very readable presentation of a reduced amount of material is Ferguson (1996). Further discussion of estimating functions, particularly for quasi-likelihood, may be found in Heyde (1997) and Crowder (1986).

Likelihood was introduced by Fisher (1922, 1925b), and quasi-likelihood by Wedderburn (1974). Asymptotic details for quasi-likelihood are described in McCullagh (1983), while Gauss–Markov theorems detailing optimality are described in Godambe and Heyde (1987) and Heyde (1997). Firth (1993) provides an excellent review of quasi-likelihood.

Crowder (1987) gives counterexamples that reveal situations in which quasi-likelihood is unreliable. Linear and quadratic estimating functions are described by Firth (1987) and Crowder (1987). Firth (1987) also investigates the efficiency of quasi-likelihood estimators and concludes that such estimators are robust to "moderate departures" from the likelihood corresponding to the score.

The form of the sandwich estimator was given in Huber (1967). White (1980) implemented the technique for the linear model, and Royall (1986) provides a clear and simple account with many examples. Carroll et al. (1995, Appendix A.3) gives a very readable review of sandwich estimation.

Efron (1979) introduced the bootstrap, and subsequently there has been a huge literature on its theoretical properties and practical use. Bickel and Freedman (1981) and Singh (1981) provide early theoretical discussions; see also van der Vaart (1998). Book-length treatments include Efron and Tibshirani (1993) and Davison and Hinkley (1997).

The score test was introduced in Rao (1948) as an alternative to the likelihood ratio and Wald tests introduced in Neyman and Pearson (1928) and Wald (1943), respectively. Consequently, the score test is sometimes known as the Rao score test. Cox and Hinkley (1974) provide a general discussion of hypothesis testing. Peers (1971) compares the power of score, Wald, and likelihood ratio tests. An excellent expository article on the three statistics, emphasizing a geometric perspective, may be found in Buse (1982).

2.12 Exercises

2.1 Suppose $Y_1, Y_2 \mid \theta \sim_{iid} \text{U}\,(\theta - 0.5, \theta + 0.5)$. Show that $\Pr(\min\{Y_1, Y_2\} < \theta < \max\{Y_1, Y_2\} \mid \theta) = 0.5$, so that $[\,\min\{Y_1, Y_2\}, \max\{Y_1, Y_2\}\,]$ is a 50% confidence interval for θ. Suppose we observe a particular interval with

$\max\{Y_1, Y_2\} - \min\{Y_1, Y_2\} \geq 0.5$. Show that in this case we know with probability 1 that this interval contains θ.[5]

2.2 Consider a single observation from a Poisson distribution: $Y \mid \theta \sim \text{Poisson}(\theta)$.

(a) Suppose we wish to estimate $\exp(-3\theta)$. Show that the UMVUE is $(-2)^y$ for $y = 0, 1, 2, \ldots$ Is this a reasonable estimator?

(b) Suppose we wish to estimate θ^2. Show that $T(T-1)/n^2$ is the UMVUE for θ^2. By examining the case $T = 1$ comment on whether this is a sensible estimator.

2.3 Let $Y_i \mid \sigma^2 \sim_{iid} \text{N}(\mu, \sigma^2)$ with μ known.

(a) Show that the distribution $p(y \mid \sigma^2)$ is a one-parameter exponential family member.

(b) Show that $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \mu)^2$ is an unbiased estimator of σ^2 and evaluate its variance.

(c) Consider estimators of the form $\widetilde{\sigma}_a^2 = a\sum_{i=1}^n (Y_i - \mu)^2$. Determine the value of a that minimizes the mean squared error.

(d) The use of mean squared error to judge an estimator is appropriate for a quadratic loss function, in this case $L(\widetilde{\sigma}_a^2, \sigma^2) = (\widetilde{\sigma}_a^2 - \sigma^2)^2$. Since $\sigma^2 > 0$, there is an asymmetry in this loss function. Hence, explain why downward bias in an estimator of σ^2 can be advantageous.

(e) Show that $\widehat{\sigma}^2$ is optimal amongst estimators $\widetilde{\sigma}_a^2$ with respect to the *Stein* loss function

$$L_s(\widetilde{\sigma}_a^2, \sigma^2) = \left(\frac{\widetilde{\sigma}_a^2}{\sigma^2}\right) - \log\left(\frac{\widetilde{\sigma}_a^2}{\sigma^2}\right) - 1.$$

2.4 Suppose $Y_i \mid \theta_i \sim_{ind} \text{Poisson}(\theta_i)$ with $\theta_i \sim_{ind} \text{Ga}(\mu_i b, b)$ for $i = 1, \ldots, n$.

(a) Show that $\text{E}[Y_i] = \mu_i$ and $\text{var}(Y_i) = \mu_i(1 + b^{-1})$.

(b) Show that the marginal distribution of $Y_i \mid \mu_i, b$ is negative binomial.

(c) Suppose $\log \mu_i = \beta_0 + \beta_1 x_i$. Write down the likelihood function $L(\boldsymbol{\beta}, b)$, log- likelihood function $l(\boldsymbol{\beta}, b)$, score function $\boldsymbol{S}(\boldsymbol{\beta}, b)$, and expected information matrix $\boldsymbol{I}(\boldsymbol{\beta}, b)$.

2.5 Consider the exponential regression problem with independent responses

$$p(y_i \mid \lambda_i) = \lambda_i e^{-\lambda_i y_i}, \quad y_i > 0$$

and $\log \lambda_i = -\beta_0 - \beta_1 x_i$ for given covariates x_i, $i = 1, \ldots, n$. We wish to estimate the 2×1 regression parameter $\boldsymbol{\beta} = [\beta_0, \beta_1]^{\text{T}}$ using MLE.

[5] This exercise shows that although the confidence interval has the correct frequentist coverage when averaging over all possible realizations of data, for some data we know with probability 1 that the *specific* interval created contains the parameter. The probability distribution of the data in this example is not regular (since the support of the data depends on the unknown parameter), and so we might anticipate difficulties. Conditioning on an ancillary statistic resolves the problems; see Davison (2003, Example 12.3).

Table 2.5 Survival times y_i and concentrations of a contaminant x_i for $i = 1, \ldots, 15$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	6.1	4.2	0.5	8.8	1.5	9.2	8.5	8.7	6.7	6.5	6.3	6.7	0.2	8.7	7.5
y_i	0.8	3.5	12.4	1.1	8.9	2.4	0.1	0.4	3.5	8.3	2.6	1.5	16.6	0.1	1.3

(a) Find expressions for the likelihood function $L(\boldsymbol{\beta})$, log-likelihood function $l(\boldsymbol{\beta})$, score function $\boldsymbol{S}(\boldsymbol{\beta})$, and Fisher's information matrix $\boldsymbol{I}(\boldsymbol{\beta})$.

(b) Find expressions for the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$. If no closed-form solution exists, then instead provide a functional form that could be simply implemented.

(c) For the data in Table 2.5, numerically maximize the likelihood function to obtain estimates of $\boldsymbol{\beta}$. These data consist of the survival times (y) of rats as a function of concentrations of a contaminant (x). Find the asymptotic covariance matrix for your estimate using the information $\boldsymbol{I}(\boldsymbol{\beta})$. Provide a 95% confidence interval for each of β_0 and β_1.

(d) Plot the log-likelihood function $l(\beta_0, \beta_1)$ and compare with the log of the asymptotic normal approximation to the sampling distribution of the MLE.

(e) Find the maximum likelihood estimate $\widehat{\beta_0}$ under the null hypothesis $H_0 : \beta_1 = 0$.

(f) Perform score, likelihood ratio, and Wald tests of the null hypothesis $H_0 : \beta_1 = 0$ with $\alpha = 0.05$. In each case, explicitly state the formula you use to compute the test statistic.

(g) Summarize the results of the estimation and hypothesis testing carried out above. In particular, address the question of whether increasing concentrations of the contaminant are associated with a rat's life expectancy.

2.6 Consider the so-called Neyman–Scott problem (Neyman and Scott 1948) in which $Y_{ij} \mid \mu_i, \sigma^2 \sim_{ind} \mathrm{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, 2$. Obtain the MLE of σ^2 and show that it is inconsistent. Why does the inconsistency arise in this example?

2.7 Consider the example discussed at the end of Sect. 2.4.3 in which the true distribution is gamma, but the assumed likelihood is exponential.

(a) Evaluate the form of the sandwich estimator of the variance, and compare with the form of the model-based estimator.

(b) Simulate data from Ga(4,2) and Ga(10,2) distributions, for $n = 10$ and $n = 30$, and obtain the MLEs and sandwich and model-based variance estimates. Compare these variances with the empirical variances observed in the simulations.

(c) Provide figures showing the log of the gamma densities of the previous part, plotted against y, along with the "closest" exponential densities.

2.8 Consider the Poisson-gamma random effects model given by (2.33) and (2.34), which leads to a negative binomial marginal model with the variance a quadratic function of the mean. Design a simulation study, along the lines of that which produced Table 2.3, to investigate the efficiency and robustness under the Poisson model, quasi-likelihood (with variance proportional to the mean), the negative binomial model, and sandwich estimation. Use a loglinear model

$$\log \mu_i = \beta_0 + \beta_1 x_i,$$

with $x_i \sim_{iid} \mathrm{N}(0,1)$, for $i = 1, \ldots, n$, and $\beta_0 = 2$, $\beta_1 = \log 2$. You should repeat the simulation for different values of both n and the negative binomial overdispersion parameter b. Report the 95% confidence interval coverages for β_0 and β_1, for each model.

2.9 A *pivotal* bootstrap interval is evaluated as follows. Let $R_n = \widehat{\theta}_n - \theta$ be a *pivot*, and $H(r) = \Pr_F(R_n \leq r)$ be the distribution function of the pivot. Now define an interval $C_n = [\, a_n, b_n \,]$ where

$$\begin{aligned} a_n &= \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \\ b_n &= \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right). \end{aligned}$$

(a) Show that

$$\Pr(a_n \leq \theta_n \leq b_n) = 1 - \alpha$$

so that C_n is an exact $100(1-\alpha)\%$ confidence interval for θ.

(b) Hence, show that the confidence interval is $C_n = [\, \widehat{a}_n, \widehat{b}_n \,]$ where

$$\begin{aligned} \widehat{a}_n &= \widehat{\theta}_n - \widehat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \widehat{\theta}_n - r^\star_{1-\alpha/2} \\ &= 2\widehat{\theta}_n - \theta^\star_{1-\alpha/2} \\ \widehat{b}_n &= \widehat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right) = \widehat{\theta}_n - r^\star_{\alpha/2} \\ &= 2\widehat{\theta}_n - \theta^\star_{\alpha/2} \end{aligned}$$

where $r^\star_\gamma$ denotes the γ sample quantile of the B bootstrap samples $[R^\star_{n1}, \ldots, R^\star_{nB}]$ and $\theta^\star_\gamma$ the γ sample quantile of $[\widehat{\theta}^\star_{n1}, \ldots, \widehat{\theta}^\star_{nB}]$.

[Hint: To evaluate a_n and b_n, we need to know H, which is unknown, but may be estimated based on the bootstrap estimates

$$\widehat{H}(r) = \frac{1}{B} \sum_{b=1}^{B} I(R^\star_{nb} \leq r)$$

where $R^\star_{nb} = \widehat{\theta}^\star_{nb} - \widehat{\theta}_n$, $b = 1, \ldots, B$.]