Chapter 6
General Regression Models

6.1 Introduction

In this chapter we consider the analysis of data that are not well-modeled by the linear models described in Chap. 5. We continue to assume that the responses are (conditionally) independent. We describe two model classes, *generalized linear models* (GLMs) and what we refer to as *nonlinear models*. In the latter, a response Y is assumed to be of the form $Y = \mu(\boldsymbol{x}, \boldsymbol{\beta}) + \epsilon$ with $\mu(\boldsymbol{x}, \boldsymbol{\beta})$ nonlinear in $\boldsymbol{x}$ and the errors ϵ independent with zero mean.

In Sect. 6.2 we introduce a motivating pharmacokinetic dataset that we will subsequently analyze using both GLMs and nonlinear models. Section 6.3 considers GLMs, which were introduced as an extension to linear models and have received considerable attention due to their computational and mathematical convenience. While computational advances have unshackled the statistician from the need to restrict attention to GLMs, they still provide an extremely useful class. Parameter interpretation for GLMs is discussed in Sect. 6.4. Sections 6.5, 6.6, 6.7, and 6.8 describe, respectively, likelihood inference, quasi-likelihood inference, sandwich estimation, and Bayesian inference for the GLM. Section 6.9 considers the assessment of the assumptions required for reliable inference in GLMs. In Sect. 6.10, we introduce nonlinear regression models, with identifiability discussed in Sect. 6.11. We then describe likelihood and least squares approaches to inference in Sects. 6.12 and 6.13 and sandwich estimation in Sect. 6.14. A geometrical comparison of linear and nonlinear least squares is provided in Sect. 6.15. Bayesian inference is described in Sect. 6.16 and Sect. 6.17 concentrates on the examination of assumptions. Concluding comments appear in Sect. 6.18 with bibliographic notes in Sect. 6.19.

In Chap. 7 we discuss models for binary data; models for such data could have been included in this chapter but are considered separately since there are a number of wrinkles that deserve specific attention.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_6,

6.2 Motivating Example: Pharmacokinetics of Theophylline

In Table 1.2 we displayed pharmacokinetic data on the sampling times and measured concentrations of the drug theophylline, collected from a subject who received an oral dose of 4.53 mg/kg. These data are plotted in Fig. 6.1, along with fitted curves from various approaches to modeling that we describe subsequently. We will fit both a nonlinear (so-called, compartmental) model to these data and a GLM. Let x_i and y_i represent the sampling time and concentration in sample i, respectively, for $i = 1, \ldots, n = 10$.

In Sect. 1.3.4, we detailed the aims of a pharmacokinetic study and described in some detail compartmental models that have been successfully used for modeling concentration–time data. Let $\mu(x)$ represent the deterministic model relating the response to time, x; $\mu(x)$ will usually be the mean response, though may correspond to the median response, depending on the assumed error structure. Notationally we have suppressed the dependence of $\mu(x)$ on unknown parameters. For the data considered here, a starting point for $\mu(x)$ is

$$\mu(x) = \frac{Dk_a}{V(k_a - k_e)} [\exp(-k_e x) - \exp(-k_a x)] \tag{6.1}$$

where $k_a > 0$ is the absorption rate constant, $k_e > 0$ is the elimination rate constant, and $V > 0$ is the (apparent) volume of distribution (that converts total amount of drug into concentration). This model was motivated in Sect. 1.3.4. A stochastic component may be added to (6.1) in a variety of ways, but one simple approach is via

$$y(x) = \mu(x) + \delta(x), \tag{6.2}$$

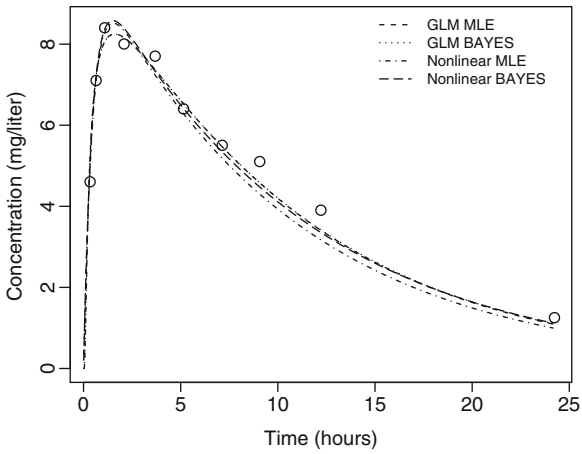

Fig. 6.1 Theophylline data, along with fitted curves under various models and inferential approaches. Four curves are included, corresponding to MLE and Bayes analyses of GLM and nonlinear models. The two nonlinear curves are indistinguishable

where $\mathrm{E}[\delta(x)] = 0$ and $\mathrm{var}[\delta(x)] = \sigma^2\mu(x)^2$ with $\delta(x)$ at different times x being independent. The variance model produces a constant coefficient of variation (defined as the ratio of the standard deviation to the mean), which is often observed in practice for pharmacokinetic data. Combining (6.1) and (6.2) gives an example of a three parameter nonlinear model. An approximately constant coefficient of variation can also be achieved by taking

$$\log y(x) = \log \mu(x) + \epsilon(x),$$

with $\mathrm{E}[\epsilon(x)] = 0$ and $\mathrm{var}[\epsilon(x)] = \sigma^2$. In this case, $\mu(x)$ represents the median concentration at time x (Sect. 5.5.3).

Model (6.1) is sometimes known as the *flip-flop* model, because there is an identifiability problem in that the same curve is achieved with each of the parameter sets $[V, k_a, k_e]$ and $[Vk_e/k_a, k_e, k_a]$. Recall from Sect. 2.4.1 that identifiability is required for consistency and asymptotic normality of the MLE. Often, identifiability is achieved by enforcing $k_a > k_e > 0$, since the absorption rate is greater than the elimination rate for most drugs. Such identifiability issues are not a rare phenomenon for nonlinear models, and will receive further attention in Sect. 6.11.

Model (6.1) may be written in the alternative form

$$\begin{aligned}\mu(x) &= \frac{Dk_a}{V(k_a - k_e)}\left[\exp(-k_e x) - \exp(-k_a x)\right] \\ &= \exp(\beta_0 + \beta_1 x)\left\{1 - \exp[-(k_a - k_e)x]\right\}, \end{aligned} \tag{6.3}$$

where $\beta_0 = \log[Dk_a/V(k_a - k_e)]$ and $\beta_1 = -k_e$. As an alternative to the compartmental model, (6.1), we will also consider the fractional polynomial model (as introduced by Nelder 1966) given by

$$\mu(x) = \exp\left(\beta_0 + \beta_1 x + \beta_2/x\right). \tag{6.4}$$

Comparison with (6.3) shows that β_2 is the parameter that is determining the absorption phase. This model only makes sense if it produces both an increasing absorption phase and a decreasing elimination phase, which correspond, retrospectively, to $\beta_2 < 0$ and $\beta_1 < 0$. When combined with an appropriate choice for the stochastic component, model (6.4) falls within the GLM class, as we see shortly.

In a pharmacokinetic study, as discussed in Sect. 1.3.4, interest often focuses on certain derived parameters. Of specific interest are $x_{1/2}$, the elimination half-life, which is the time it takes for the drug concentration to drop by 50% (for times sufficiently large for elimination to be the dominant process); $x_{\max}$, the time to maximum concentration; $\mu(x_{\max})$, the maximum concentration; and Cl, the clearance, which is the amount of blood cleared of drug in unit time.

With respect to model (6.1), the derived parameters of interest, in terms of $[V, k_a, k_e]$, are

$$
\begin{aligned}
x_{1/2} &= \frac{\log 2}{k_e} \\
x_{\max} &= \frac{1}{k_a - k_e} \log\left(\frac{k_a}{k_e}\right) \\
\mu(x_{\max}) &= \frac{Dk_a}{V(k_a - k_e)} \left[\exp(-k_e x_{\max}) - \exp(-k_a x_{\max})\right] \\
&= \frac{D}{V}\left(\frac{k_a}{k_e}\right)^{k_a/(k_a - k_e)} \\
Cl &= \frac{D}{\text{AUC}} \\
&= V \times k_e
\end{aligned}
$$

where AUC is the area under the concentration–time curve between 0 and ∞. With respect to model (6.4), as functions of $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$,

$$
\begin{aligned}
x_{1/2} &= -\frac{\log 2}{\beta_1} \\
x_{\max} &= \left(\frac{\beta_2}{\beta_1}\right)^{1/2} \\
\mu(x_{\max}) &= D \exp\left[\beta_0 - 2(\beta_1 \beta_2)^{1/2}\right] \\
Cl &= \frac{\sqrt{\beta_1/\beta_2}}{2\exp(\beta_0) K_1[2(\beta_1\beta_2)^{1/2}]},
\end{aligned} \tag{6.5}
$$

where $K_s(x)$ denotes a modified Bessel function of the second kind of order s. Consequently, for both models, the quantities of interest are nonlinear functions of the original parameters, which has implications for inference.

6.3 Generalized Linear Models

Generalized linear models (GLMs) were introduced by Nelder and Wedderburn (1972) and provide a class with relatively broad applicability and desirable statistical properties. For a GLM:

- The responses y_i follow an exponential family, so that the distribution is of the form

$$
p(y_i \mid \theta_i, \alpha) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha)\right), \tag{6.6}
$$

Table 6.1 Characteristics of some common GLMs. The notation is as in (6.6). The canonical parameter is θ, the mean is $\mathrm{E}[Y] = \mu$, and the variance is $\mathrm{var}(Y) = \alpha V(\mu)$

Distribution	$\mathrm{N}(\mu, \sigma^2)$	$\mathrm{Poisson}(\mu)$	$\mathrm{Bernoulli}(\mu)$	$\mathrm{Ga}(1/\alpha, 1/[\mu\alpha])$
Mean $\mathrm{E}[Y \mid \theta]$	θ	$\exp(\theta)$	$\frac{\exp(\theta)}{1+\exp(\theta)}$	$-\frac{1}{\theta}$
Variance $V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1+\mathrm{e}^{\theta})$	$-\log(-\theta)$
$c(y, \alpha)$	$-\frac{1}{2}\left[\frac{y^2}{2} + \log(2\pi\alpha)\right]$	$-\log y!$	1	$\frac{\log(y/\alpha)}{\alpha} - \log y + \log \Gamma(\alpha)$

for functions $b(\cdot)$, $c(\cdot, \cdot)$ and where θ_i and α are scalars. It is straightforward to show (using the results of Sect. 2.4) that

$$\begin{aligned} \mathrm{E}[Y_i \mid \theta_i, \alpha] &= \mu_i \\ &= b'(\theta_i) \end{aligned}$$

and

$$\begin{aligned} \mathrm{var}(Y_i \mid \theta_i, \alpha) &= \alpha b''(\theta_i) \\ &= \alpha V(\mu_i), \end{aligned}$$

for $i = 1, \ldots, n$. We assume $\mathrm{cov}(Y_i, Y_j \mid \theta_i, \theta_j, \alpha) = 0$, for $i \neq j$ (Chap. 9 provides the extension to dependent data).

- A *link function* $g(\cdot)$ provides the connection between the mean function $\mu_i = \mathrm{E}[Y_i \mid \theta_i, \alpha]$ and the *linear predictor* $\boldsymbol{x}_i\boldsymbol{\beta}$ via

$$g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta},$$

where $\boldsymbol{x}_i$ is a $(k+1) \times 1$ vector of explanatory variables (including a 1 for the intercept) and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^{\mathrm{T}}$ is a $(k+1) \times 1$ vector of regression parameters.

To summarize, a GLM assumes a linear relationship on a transformed mean scale (which, as we shall see, offers certain computational and statistical advantages) and an exponential family form for the distribution of the response.

If α is known, then (6.6) is a one-parameter exponential family model. If α is unknown, then the distribution may or may not be a two-parameter exponential family model. So-called *canonical links* have $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ and provide simplifications in terms of computation.

GLMs are very useful pedagogically since they separate the deterministic and stochastic components of the model, and this aspect was emphasized in the abstract of Nelder and Wedderburn (1972): "The implications of the approach in designing statistics courses are discussed."

Table 6.1, adapted from Table 2.1 of McCullagh and Nelder (1989), characterizes a number of common GLMs. Another example which is not listed in the table, is the inverse Gaussian distribution; Exercise 6.1 derives the detail for this case.

Example: Pharmacokinetics of Theophylline

Model (6.3) is an example of a GLM with a log link:

$$\log \mu(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{6.7}$$

where $\boldsymbol{x} = [1, x_1, x_2]$ and $x_2 = 1/x_1$.

Turning to the stochastic component, as noted in Sect. 6.2, the error terms often display a constant coefficient of variation. With this in mind, we may combine (6.7) with a gamma distribution via

$$Y(\boldsymbol{x}) \mid \boldsymbol{\beta}, \alpha \sim_{ind} \text{Ga}\{\alpha^{-1}, [\mu(\boldsymbol{x})\alpha]^{-1}\}, \tag{6.8}$$

to give $\text{E}[Y(\boldsymbol{x})] = \mu(\boldsymbol{x})$ and $\text{var}[Y(\boldsymbol{x})] = \alpha\mu(\boldsymbol{x})^2$ so that $\alpha^{1/2}$ is the coefficient of variation. Lindsey et al. (2000) examine various distributional choices for pharmacokinetic data and found the gamma assumption to be reasonable in their examples. It is interesting to note that for the gamma distribution, the reciprocal transform is the canonical link, but this option is not statistically appealing since it does not constrain the mean function to be positive. In the pharmacokinetic context the reciprocal link also results in a concentration–time curve that is not integrable between 0 and ∞ so that the fundamental clearance parameter is undefined. One disadvantage of the loglinear GLM defined above, compared to the nonlinear compartmental model we discuss later, is that if multiple doses are considered, the mean function does not correspond to a GLM.

Example: Lung Cancer and Radon

In Sect. 1.3.3 we described data on lung cancer incidence in counties in Minnesota, with Y_i the number of cases, x_i the average radon, and E_i the expected number of cases, in area i, $i = 1, \ldots, n$. These data were examined repeatedly in Chaps. 2 and 3.

A starting model is $Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\,[E_i \exp(\beta_0 + \beta_1 x_i)]$, which we write as

$$\log \Pr(Y = y_i \mid \boldsymbol{\beta}) = y_i \log \mu_i - \mu_i - \log y_i!$$

with $\log \mu_i = \log E_i + \beta_0 + \beta_1 x_i$, to give a GLM with a (canonical) log link. As discussed in Chaps. 2 and 3, this model is fundamentally inadequate because $\alpha = 1$, and so there is no parameter to allow for excess-Poisson variation. The latter can be modeled using the negative binomial model of Sect. 6.3 or the quasi-likelihood approach described in Sect. 6.6.

With unknown scale parameter, the negative binomial is not a GLM. We consider the case of known b (which will rarely be of interest in a practical setting). For

consistency with its use in Chap. 2, we label the scale parameter of the negative binomial model as b. In the following, care should therefore be taken to discriminate between $b(\cdot)$, as in (6.6), and the scale parameter, b. From (2.40),

$$\log \Pr(Y = y_i \mid \mu_i) = b^{-1}\left[y_i b \log\left(\frac{\mu_i}{\mu_i + b}\right) - b^2 \log(\mu_i + b)\right] + \log \Gamma(y_i + b) - \log \Gamma(b) - \log y_i! - b(b+1)\log b$$

which is of the form (6.6) with

$$\begin{aligned} \theta_i &= b \log\left(\frac{\mu_i}{\mu_i + b}\right), \\ b(\theta_i) &= b^2 \log(\mu_i + b), \\ c(y_i, b) &= \log \Gamma(y_i + b) - \log \Gamma(b) - \log y_i! - b(b+1)\log b, \end{aligned}$$

so that

$$\begin{aligned} \mathrm{E}[Y_i \mid \mu_i] &= \mu_i = b'(\theta_i) \\ &= \frac{b\mathrm{e}^{\theta_i/b}}{1 - \mathrm{e}^{\theta_i/b}}, \\ \mathrm{var}(Y_i \mid \mu_i) &= b \times b''(\theta_i) \\ &= \mu_i + \mu_i^2/b. \end{aligned}$$

The canonical link is

$$\theta_i = b \log\left(\frac{\mu_i}{\mu_i + b}\right) = \boldsymbol{x}\beta,$$

which depends on b. The negative binomial distribution is described in detail by Cameron and Trivedi (1998).

6.4 Parameter Interpretation

Interpretation of the regression parameters in a GLM is link function specific. The linear link was discussed in Chap. 5, and the log link was considered repeatedly (in the context of the lung cancer and radon data) in Chaps. 2 and 3. We provide an interpretation of binary data link functions, such as the logistic, in Chap. 7. Linearity on some scale offers advantages, as illustrated by the following example.

Consider the log linear model:

$$\log \mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The parameter $\exp(\beta_1)$ has a relatively straightforward interpretation, being the multiplicative change in the average response associated with a one-unit increase in x_1, with x_2 held constant.

In contrast, for general nonlinear models, the parameters often define particular functions of the response covariate curve or fundamental quantities that define the system under study. We saw an example of this in Sect. 6.2, in which the nonlinear concentration–time curve (6.1) was defined in terms of the volume of distribution V and the absorption and elimination rate constants k_a and k_e. Alternatively, we could define the model in terms of characteristics of the curve, for example, the half-life, $x_{1/2}$, the time to maximum concentration, $x_{\max}$, and the maximum concentration, $\mu(x_{\max})$. We now discuss inference for the GLM.

6.5 Likelihood Inference for GLMs

6.5.1 Estimation

We first derive the score vector and information matrix. For an independent sample from the exponential family (6.6)

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha),$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta}) = [\theta_1(\boldsymbol{\beta}), \ldots, \theta_n(\boldsymbol{\beta})]$ is the vector of canonical parameters. Using the chain rule, the score function is

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{dl_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{Y_i - b'(\theta_i)}{\alpha}\frac{1}{V_i}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}, \end{aligned} \tag{6.9}$$

where $\text{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V_i$ and

$$\frac{d^2 b}{d\theta_i^2} = \frac{d\mu_i}{d\theta_i} = V_i,$$

for $i = 1, \ldots, n$. Hence,

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} \frac{[Y_i - \mathrm{E}(Y_i \mid \mu_i)]}{\text{var}(Y_i \mid \mu_i)} \\ &= \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right]/\alpha, \end{aligned} \tag{6.10}$$

where $\boldsymbol{D}$ is the $n \times (k+1)$ matrix with elements $\partial \mu_i / \partial \beta_j$, $i = 1, \ldots, n$, $j = 0, \ldots, k$, and $\boldsymbol{V}$ is the $n \times n$ diagonal matrix with ith diagonal element V_i. Consequently, an estimator $\widehat{\boldsymbol{\beta}}_n$ defined through $\boldsymbol{S}(\widehat{\boldsymbol{\beta}}_n) = \mathbf{0}$ will be consistent so long as the mean function is correctly specified, since the estimating function is unbiased in this case. For canonical links, for which $\theta_i = \boldsymbol{x}_i \boldsymbol{\beta}$,

$$\sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{dl_i}{d\theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{1}{\alpha} \sum_{i=1} \boldsymbol{x}_i^{\mathrm{T}} [Y_i - \mu_i(\boldsymbol{\beta})]$$

so that the sufficient statistics

$$\sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} Y_i = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}} \mu_i(\widehat{\boldsymbol{\beta}})$$

are recovered at the MLE, $\widehat{\boldsymbol{\beta}}$.

From Result 2.1, the MLE has asymptotic distribution

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

where the expected information is

$$\boldsymbol{I}_n(\boldsymbol{\beta}) = \mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta}) \boldsymbol{S}(\boldsymbol{\beta})^{\mathrm{T}}] = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{D} / \alpha.$$

In practice we use

$$\boldsymbol{I}_n(\widehat{\boldsymbol{\beta}}_n) = \widehat{\boldsymbol{D}}^{\mathrm{T}} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}} / \alpha,$$

where $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{D}}$ are evaluated at $\widehat{\boldsymbol{\beta}}_n$. The variance of the estimator is

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \alpha \left(\widehat{\boldsymbol{D}}^{\mathrm{T}} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}} \right)^{-1} \tag{6.11}$$

and is consistently estimated if the second moment is correctly specified.

The information matrix may be written in a particularly simple and useful form, as we now show. We first let $\eta_i = g(\mu_i)$ denote the linear predictor. The score, (6.9), may be written, for parameter j, $j = 0, 1, \ldots, k$, as

$$\begin{aligned} S_j(\boldsymbol{\beta}) = \frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\alpha V_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\alpha V_i} \frac{d\mu_i}{d\eta_i} x_{ij}. \end{aligned} \tag{6.12}$$

Hence, element (j, j') of the expected information is

$$
\begin{aligned}
-\sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial^2 l_i}{\partial\beta_j \partial\beta_{j'}}\right] &= \sum_{i=1}^{n} \mathrm{E}\left[\left(\frac{\partial l_i}{\partial\beta_j}\right)\left(\frac{\partial l_i}{\partial\beta_{j'}}\right)\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[\frac{(Y_i-\mu_i)x_{ij}}{\alpha V_i}\frac{d\mu_i}{d\eta_i}\frac{(Y_i-\mu_i)x_{ij'}}{\alpha V_i}\frac{d\mu_i}{d\eta_i}\right] \\
&= \sum_{i=1}^{n} \frac{x_{ij}x_{ij'}}{\alpha V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2.
\end{aligned}
$$

The information matrix therefore takes the form

$$
\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{x} \tag{6.13}
$$

where $\boldsymbol{W}$ is the diagonal matrix with elements

$$
w_i = \frac{(d\mu_i/d\eta_i)^2}{\alpha V_i},
$$

$i = 1, \ldots, n$.

When α is unknown, it may be estimated using maximum likelihood or the method of moments estimator

$$
\widehat{\alpha} = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{6.14}
$$

where $\widehat{\mu}_i = \widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$. Section 2.5 contained the justification for this estimator, which has the advantage of being, in general, a consistent estimator in a broader range of circumstances than the MLE. The method of moments approach is routinely used for normal and gamma data. As usual, there will be an efficiency loss when compared to the use of the MLE if the distribution underlying the derivation of the latter is "true."

The use of (6.10) is appealing since it depends on only the first two moments so that consistency of $\widehat{\boldsymbol{\beta}}_n$ does not depend on the distribution of the data. Accurate asymptotic confidence interval coverage depends only on correct specification of the mean–variance relationship. Section 6.7 describes how the latter requirement may be relaxed.

If the score is of the form (6.6), that is, if the score arises from an exponential family, it is not necessary to have a mean function of GLM form (i.e., a linear predictor on some scale). So, for example, the nonlinear models considered later in the chapter, when embedded within an exponential family, also share consistency of estimation (so long as regularity conditions are satisfied).

6.5.2 Computation

Computation is relatively straightforward for GLMs, since the form of a GLM yields a log-likelihood surface that is well behaved, for all but pathological datasets. In particular, a variant of the Newton–Raphson method (a generic method for root-finding), known as *Fisher scoring*, may be used to find the MLEs. We briefly digress to describe the Newton–Raphson method. Let $\boldsymbol{S}(\boldsymbol{\beta})$ represent a $p \times 1$ vector of functions that are themselves functions of a $p \times 1$ vector $\boldsymbol{\beta}$. We wish to find $\boldsymbol{\beta}$ such that $\boldsymbol{S}(\boldsymbol{\beta}) = \mathbf{0}$. A first-order Taylor series expansion about $\boldsymbol{\beta}^{(0)}$ gives

$$\boldsymbol{S}(\boldsymbol{\beta}) \approx \boldsymbol{S}(\boldsymbol{\beta}^{(0)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^{\mathrm{T}} \boldsymbol{S}'(\boldsymbol{\beta}^{(0)}).$$

Setting the left-hand side to zero yields

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} - \boldsymbol{S}'(\boldsymbol{\beta}^{(0)})^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(0)}).$$

The Newton–Raphson method iterates the step:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{S}'(\boldsymbol{\beta}^{(t)})^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(t)}),$$

for $t = 0, 1, 2, \ldots$ The Fisher scoring method is the Newton–Raphson method applied to the score equation, but with the observed information, $\boldsymbol{S}'(\boldsymbol{\beta})$, replaced by the expected information $\mathrm{E}[\boldsymbol{S}'(\boldsymbol{\beta})] = -\boldsymbol{I}(\boldsymbol{\beta})$ to give

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \boldsymbol{I}(\boldsymbol{\beta}^{(t)})^{-1} \boldsymbol{S}(\boldsymbol{\beta}^{(t)}),$$

so that a new estimate is calculated based on the score and information evaluated at the previous estimate. Recall that for a GLM, $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}(\boldsymbol{\beta}) \boldsymbol{x}$. Using this form, and (6.12), we write

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}^{(t)} \boldsymbol{x})^{-1} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}^{(t)} \left[\boldsymbol{x} \boldsymbol{\beta}^{(t)} + (\boldsymbol{W}^{(t)})^{-1} \boldsymbol{u}^{(t)} \right] \\ &= (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}^{(t)} \boldsymbol{x})^{-1} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)} \end{aligned} \tag{6.15}$$

where $\boldsymbol{u}^{(t)}$ and $\boldsymbol{z}^{(t)}$ are $n \times 1$ vectors with ith elements

$$u_i^{(t)} = \frac{(Y_i - \mu_i^{(t)})}{\alpha V_i^{(t)}} \left. \frac{d\mu_i}{d\eta_i} \right|_{\boldsymbol{\beta}^{(t)}},$$

and

$$z_i^{(t)} = \boldsymbol{x}_i \boldsymbol{\beta}^{(t)} + (Y_i - \mu_i^{(t)}) \left. \frac{d\eta_i}{d\mu_i} \right|_{\boldsymbol{\beta}^{(t)}},$$

Table 6.2 Point and 90% interval estimates for the theophylline data of Table 1.2, under various models and estimation techniques. CV is the coefficient of variation and is expressed as a percentage. The Bayesian point estimates correspond to the posterior medians

Model	$x_{1/2}$	$x_{\max}$	$\mu(x_{\max})$	CV ($\times 100$)
GLM MLE	7.23 [6.89,7.59]	1.60 [1.52,1.69]	8.25 [7.95,8.56]	4.38 [3.04,6.33]
GLM sandwich	7.23 [6.97,7.50]	1.60 [1.57,1.64]	8.25 [8.02,8.48]	4.38 [3.04,6.33]
Nonlinear MLE	7.54 [7.09,8.01]	1.51 [1.36,1.66]	8.59 [7.99,9.24]	6.32 [4.38,9.13]
Nonlinear sandwich	7.54 [7.11,7.98]	1.51 [1.43,1.58]	8.59 [8.11,9.10]	6.32 [4.38,9.13]
Prior	8.00 [5.30,12.0]	1.50 [0.75,3.00]	9.00 [6.80,12.0]	5.00 [2.50,10.0]
GLM Bayes	7.26 [6.93,7.74]	1.60 [1.51,1.68]	8.24 [7.89,8.54]	5.21 [3.72,7.86]
Nonlinear Bayes	7.57 [7.15,8.04]	1.50 [1.36,1.66]	8.59 [8.22,8.94]	6.01 [4.34,8.93]

respectively. The Fisher scoring updates (6.15) therefore have the form of a weighted least squares solution to

$$(\boldsymbol{z}^{(t)} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W}^{(t)} (\boldsymbol{z}^{(t)} - \boldsymbol{x}\boldsymbol{\beta}) \tag{6.16}$$

with "working" or "adjusted" response $\boldsymbol{z}^{(t)}$. This method is therefore known as *iteratively reweighted least squares* (IRLS). For canonical links, the observed and expected information coincide so that the Fisher scoring and Newton–Raphson methods are identical.

The existence and uniqueness of estimates have been considered by a number of authors; early references are Wedderburn (1976) and Haberman (1977).

Example: Pharmacokinetics of Theophylline

Fitting the gamma model (6.8) with mean function (6.7) gives MLEs for $[\beta_0, \beta_1, \beta_2]$ of $[2.42, -0.0959, -0.246]$. The fitted curve is shown in Fig. 6.1. The method of moments estimate of the coefficient of variation, $100\sqrt{\alpha}$, is 5.3%, while the MLE is 4.4%. Asymptotic standard errors for $[\beta_0, \beta_1, \beta_2]$, based on the method of moments estimator for α, are $[0.033, 0.0028, 0.018]$. The point estimates of $\boldsymbol{\beta}$ are identical, regardless of the estimate used for α, because the root of the score is independent of α in a GLM, as is clear from (6.10).

The top row of Table 6.2 gives MLEs for the derived parameters, along with asymptotic 90% confidence intervals, derived using the delta method. All are based upon the method of moments estimator for α. The parameters of interest are all positive, and so the intervals were obtained on the log scale and then exponentiated. Deriving an interval estimate for the clearance parameter using the delta method is more complex. Working with $\theta = \log Cl$, we have

$$\mathrm{var}(\widehat{\theta}) = [D_0 \; D_1 \; D_2] \boldsymbol{V}^{\star} \begin{bmatrix} D_0 \\ D_1 \\ D_2 \end{bmatrix}$$

where, from (6.5),

$$D_0 = \frac{\partial \theta}{\partial \beta_0} = 1$$

$$D_1 = \frac{\partial \theta}{\partial \beta_1} = \frac{1}{\beta_1} + \sqrt{\frac{\beta_2}{\beta_1}} \frac{K_0\left(2\sqrt{\beta_1\beta_2}\right)}{K_1\left(2\sqrt{\beta_1\beta_2}\right)}$$

$$D_2 = \frac{\partial \theta}{\partial \beta_2} = \sqrt{\frac{\beta_1}{\beta_2}} \frac{K_0\left(2\sqrt{\beta_1\beta_2}\right)}{K_1\left(2\sqrt{\beta_1\beta_2}\right)},$$

and $\boldsymbol{V}^\star$ is the variance–covariance matrix of $\widehat{\boldsymbol{\beta}}$ as given by (6.11). For the theophylline data, the MLE is $\widehat{Cl} = 0.042$ with asymptotic 90% confidence interval [0.041,0.044]. Inference for the clearance parameter using the sampling-based Bayesian approach that we describe shortly is straightforward, once samples are generated from the posterior.

Example: Poisson Data with a Linear Link

We now describe a GLM that is a little more atypical and reveals some of the subtleties of modeling that can occur. In the context of a spatial study, suppose that, in a given time period, Y_{i0} represents the number of counts of a (statistically) rare disease in an unexposed group of size N_{i0}, while Y_{i1} represents the number of counts of a rare disease in an exposed group of size N_{i1}, all in area i, $i = 1, \ldots, n$. Suppose also that we only observe the sum of the disease counts, $Y_i = Y_{i0} + Y_{i1}$, along with N_{i0} and N_{i1}. If we had observed Y_{i0}, Y_{i1}, we would fit the model $Y_{ij} \mid \boldsymbol{\beta}^\star \sim_{ind} \text{Poisson}(N_{ij}\beta_j^\star)$ so that $0 < \beta_j^\star < 1$ is the probability of disease in exposure group j, with $j = 0/1$ representing unexposed/exposed and $\boldsymbol{\beta}^\star = [\beta_0^\star, \beta_1^\star]$. Then, writing $x_i = N_{1i}/N_i$ as the proportion of exposed individuals, the distribution of the total disease counts is

$$Y_i \mid \boldsymbol{\beta}^\star \sim_{ind} \text{Poisson}\left\{N_i[(1 - x_i)\beta_0^\star + x_i\beta_1^\star]\right\}, \tag{6.17}$$

so that we have a Poisson GLM with a linear link function. Since the parameters $\beta_0^\star$ and $\beta_1^\star$ are the probabilities (or risks) of disease for unexposed and exposed individuals, respectively, a parameter of interest is the relative risk, $\beta_1^\star/\beta_0^\star$.

We illustrate the fitting of this model using data on the incidence of lip cancer in men in $n = 56$ counties of Scotland over the years 1975–1980. These data were originally reported by Kemp et al. (1985) and have been subsequently reanalyzed by numerous others, see, for example, Clayton and Kaldor (1987). The covariate x_i is the proportion of individuals employed in agriculture, fishing, and farming in county i. We let Y_i represent the number of cases in county i. Model (6.17) requires some

adjustment, since the only available data here, in addition to x_i, are the expected numbers E_i that account for the age breakdown in county i (see Sect. 1.3.3). We briefly describe the model development in this case, since it requires care and reveals assumptions that may otherwise be unapparent.

Let Y_{ijk} be the number of cases, from a population of N_{ijk} in county i, exposure group j, and age stratum k, $i = 1, \ldots, n$, $j = 0, 1$, $k = 1, \ldots, K$. An obvious starting model for a rare disease is

$$Y_{ijk} \mid p_{ijk} \sim_{ind} \text{Poisson}(N_{ijk} p_{ijk}).$$

This model contains far too many parameters, p_{ijk}, to estimate, and so we simplify by assuming

$$p_{ijk} = \beta_j \times p_k, \tag{6.18}$$

across all areas i. Consequently, p_k is the probability of disease in age stratum k and $\beta_j > 0$ is the relative risk adjustment in exposure group j, and we are assuming that the exposure effect is the same across areas and across age stratum. The age-specific probabilities p_k are assumed known (e.g., being based on rates from a larger geographic region). The numbers of exposed individuals in each age stratum are unknown, and we therefore make the important assumption that the proportion of exposed and unexposed individuals is constant across age stratum, that is, $N_{i0k} = N_{ik}(1 - x_i)$ and $N_{i1k} = N_{ik} x_i$. This assumption is made since N_{i0k} and N_{i1k} are unavailable and is distinct from assumption (6.18) which concerns the underlying disease model. Summing across stratum and exposure groups gives

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\left(\beta_0(1 - x_i) \sum_{k=1}^{K} N_{ik} p_k + \beta_1 x_i \sum_{k=1}^{K} N_{ik} p_k\right).$$

Letting $E_i = \sum_{k=1}^{K} N_{ik} p_k$ represent the expected number of cases, and simplifying the resultant expression gives

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\left\{E_i[(1 - x_i)\beta_0 + x_i \beta_1]\right\}. \tag{6.19}$$

Under this model,

$$\text{E}\left[\frac{Y_i}{E_i}\right] = \beta_0 + (\beta_1 - \beta_0) x_i, \tag{6.20}$$

illustrating that the mean model for the standardized morbidity ratio (SMR), Y_i/E_i, is linear in x. Figure 6.2 plots the SMRs Y_i/E_i versus x_i, with a linear fit added, and we see evidence of increasing SMR with increasing x.

Fitting the Poisson linear link model gives estimates (asymptotic standard errors) for β_0 and β_1 of 0.45 (0.043) and 10.1 (0.77). The fitted line (6.20) is superimposed on Fig. 6.2. The estimate of the relative risk β_1/β_0 is 22.7 with asymptotic standard

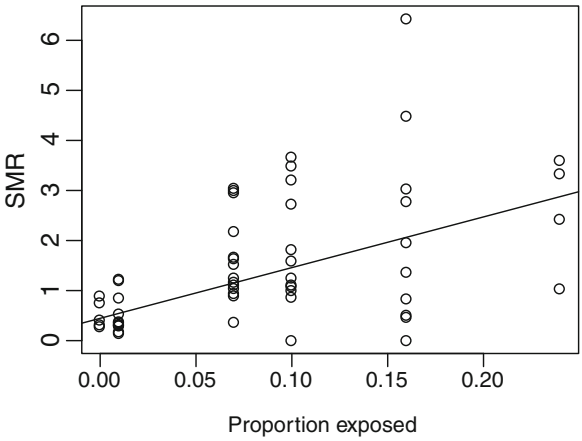

Fig. 6.2 Plot of standardized morbidity ratio versus proportion exposed for lip cancer incidence in 56 counties of Scotland. The linear model fit is indicated

error 3.39. The latter is a model-based estimate and in particular depends on there being no excess-Poisson variation, which is highly dubious for applications such as this, because of all of the missing auxiliary information, including data on smoking.

6.5.3 Hypothesis Testing

Suppose that $\dim(\boldsymbol{\beta}) = k + 1$ and let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ be a partition with $\boldsymbol{\beta}_1 = [\beta_0, \ldots, \beta_q]$ and $\boldsymbol{\beta}_2 = [\beta_{q+1}, \ldots, \beta_k]$, with $0 \leq q < k$. Interest focuses on testing whether the subset of $k - q$ parameters are equal to zero via a test of the null

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_1 \text{ unrestricted, } \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20} \\ H_1 &: \boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2] \neq [\boldsymbol{\beta}_1, \boldsymbol{\beta}_{20}]. \end{aligned} \tag{6.21}$$

As outlined in Sect. 2.9, there are three main frequentist approaches to hypothesis testing, based on Wald, score, and likelihood ratio tests. We concentrate on the latter. For the linear model, the equivalent approach is based on an F test (Sect. 5.6.1), which formally accounts for estimation of the scale parameter.

The log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha),$$

with α the scale parameter. We let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta}) = [\theta_1(\boldsymbol{\beta}), \ldots, \theta_n(\boldsymbol{\beta})]$ denote the vector of canonical parameters. Under the null, from Sect. 2.9.5,

$$2\left[l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}^{(0)})\right] \rightarrow_d \chi^2_{k-q},$$

where $\widehat{\boldsymbol{\beta}}$ is the unrestricted MLE and $\widehat{\boldsymbol{\beta}}^{(0)} = [\widehat{\boldsymbol{\beta}}_{10}, \boldsymbol{\beta}_{20}]$ is the MLE under the null.

In some circumstances, one may assess the *overall* fit of a particular model via comparison of the likelihood of this model with the maximum attainable log-likelihood which occurs under the *saturated model*. We write $\widetilde{\boldsymbol{\theta}} = [\widetilde{\theta}_1, \ldots, \widetilde{\theta}_n]$ to represent the MLEs under the saturated model. Similarly, let $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1, \ldots, \widehat{\theta}_n]$ denote the MLEs under a reduced model containing $q+1$ parameters. The log-likelihood ratio statistic of $H_0:$ reduced model, $H_1:$ saturated model is

$$2\left[l(\widetilde{\boldsymbol{\theta}}) - l(\widehat{\boldsymbol{\theta}})\right] = \frac{2}{\alpha}\sum_{i=1}^{n}\left[Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\right] = \frac{D}{\alpha}, \tag{6.22}$$

where D is known as the *deviance* (associated with the saturated model) and D/α is the *scaled deviance*. If the saturated model has a fixed number of parameters, p, then, under the reduced model,

$$\frac{D}{\alpha} \;\rightarrow_d\; \chi^2_{p-q-1}.$$

In general, this result is rarely used, though cross-classified discrete data provide one instance in which the overall fit of a model can be assessed in this way. An alternative measure of the overall fit is the Pearson statistic

$$X^2 = \sum_{i=1}^{n}\frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{6.23}$$

with $X^2 \rightarrow_d \chi^2_{p-q-1}$ under the null. Again, the saturated model should contain a fixed number of parameters (as $n \rightarrow \infty$).

Consider again the nested testing situation with hypotheses, (6.21). We describe an attractive additivity property of the likelihood ratio test statistic for nested models. Let $\widehat{\boldsymbol{\beta}}^{(0)}$, $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(s)}$ represent the MLEs of $\boldsymbol{\beta}$ under the null, alternative, and saturated models, respectively. Suppose that the dimensionality of $\widehat{\boldsymbol{\beta}}^{(j)}$ is q_j with $0 < q_0 < q_1 < p$. Under H_0,

$$2\left[l(\widehat{\boldsymbol{\beta}}^{(1)}) - l(\widehat{\boldsymbol{\beta}}^{(0)})\right] = 2\left\{l(\widehat{\boldsymbol{\beta}}^{(s)}) - l(\widehat{\boldsymbol{\beta}}^{(0)}) - [l(\widehat{\boldsymbol{\beta}}^{(s)}) - l(\widehat{\boldsymbol{\beta}}^{(1)})]\right\}$$

$$= \frac{1}{\alpha}\left(D_0 - D_1\right) \;\rightarrow_d\; \chi^2_{q_1-q_0},$$

where D_j is the deviance representing the fit under hypothesis j, relative to the saturated model, $j = 0, 1$. The Pearson statistic does not share this additivity property.

For a GLM, in contrast to the linear model (see Sect. 5.8), even if a covariate is orthogonal to all other covariates, its significance will still depend on which covariates are currently in the model.

Example: Normal Linear Model

We consider the model $\boldsymbol{Y} \mid \boldsymbol{\beta} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. The log-likelihood is

$$l(\boldsymbol{\beta}, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}),$$

with α in the GLM formulation being replaced by σ^2. Again, let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ where $\boldsymbol{\beta}_1 = [\beta_0, \ldots, \beta_q]$ and $\boldsymbol{\beta}_2 = [\beta_{q+1}, \ldots, \beta_k]$, and consider the null H_0 : $\boldsymbol{\beta}_1$ unrestricted, $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$. Under this null, from (6.22),

$$D = \sum_{i=1}^{n} \left(Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}^{(0)} \right)^2$$

where $\boldsymbol{x}_i \widehat{\boldsymbol{\beta}}^{(0)}$ are the fitted values for the ith case, based on the MLEs under the reduced model, H_0. In this case, the asymptotic distribution is exact since

$$\frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}^{(0)})^2}{\sigma^2} \sim \chi^2_{n-q+1}. \tag{6.24}$$

This result is almost never directly useful, however, since σ^2 is rarely known.

In terms of comparing the nested hypotheses $H_0 : \boldsymbol{\beta}_1$ unrestricted, $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$, and $H_1 : \boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2] \neq [\boldsymbol{\beta}_1, \boldsymbol{\beta}_{20}]$, the likelihood ratio statistic is

$$\begin{aligned} \frac{1}{\sigma^2}(D_0 - D_1) &= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}^{(0)})^2 - \sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}^{(1)})^2 \right] \\ &= \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\sigma^2} = \frac{\mathrm{FSS}_{01}}{\sigma^2} \end{aligned} \tag{6.25}$$

where $\boldsymbol{x}\widehat{\boldsymbol{\beta}}^{(j)}$ are the fitted values corresponding to the MLEs under model j, RSS_j is the residual sum of squares for model j, $j = 0, 1$, and FSS_{01} is the fitted sum of squares due to the additional parameters present in H_1.

In practice if n is large, we may use (6.25) with σ^2 replaced by a consistent estimator $\widehat{\sigma}^2$. Alternatively, the ratios of scaled versions of (6.25) and (6.24) may be taken to produce an F-statistic by which statistical significance may be assessed, as described in Sect. 5.6.1.

Example: Lung Cancer and Radon

Under a Poisson model, the deviance and scaled deviance are identical since $\alpha = 1$. For a Poisson model with MLE $\widehat{\boldsymbol{\beta}}$, the deviance is

$$2\sum_{i=1}^{n}\left[(\mu_i(\widehat{\boldsymbol{\beta}}) - y_i) + y_i \log\left(\frac{y_i}{\mu_i(\widehat{\boldsymbol{\beta}})}\right)\right]$$

and if the sum of the observed and fitted counts agree, then we obtain the intuitive distance measure

$$2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\mu_i(\widehat{\boldsymbol{\beta}})}\right).$$

For the Minnesota data, suppose we wish to test $H_0 : \beta_0$ unrestricted, $\beta_1 = 0$ versus $H_1 : [\beta_0, \beta_1] \neq [\beta_0, 0]$, in the model $\mu_i = E_i \exp(\beta_0 + \beta_1 x_i)$. The likelihood ratio statistic is

$$T = 2\sum_{i=1}^{n} y_i \log\left(\frac{\mu_i(\widehat{\boldsymbol{\beta}})}{\mu_i(\widehat{\boldsymbol{\beta}}^{(0)})}\right),$$

since $\sum_{i=1}^{n} \mu_i(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \mu_i(\widehat{\boldsymbol{\beta}}^{(0)})$, and where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^{(0)}$ are the MLEs under the null and alternative hypotheses. Under H_0, $T \to_d \chi^2_1$.

For the Minnesota data $T = 46.2$ to give an extremely small p-value. The estimate (standard error) of β_1 is -0.036 (0.0054) so that for a one-unit increase in average radon, there is an associated drop in relative risk of lung cancer of 3.6%.

6.6 Quasi-likelihood Inference for GLMs

Section 2.5 provided an extended discussion of quasi-likelihood, and here we recap the key points. GLMs that do not contain a scale parameter are particularly vulnerable to variance model misspecification, specifically the presence of overdispersion in the data. The Poisson and binomial models are especially susceptible in this respect.

Rather than specify a complete probability model for the data, quasi-likelihood proceeds by specifying the mean and variance as

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \mathrm{var}(Y_i \mid \boldsymbol{\beta}) &= \alpha V(\mu_i), \end{aligned}$$

with $\mathrm{cov}(Y_i, Y_j \mid \boldsymbol{\beta}) = 0$. From these specifications, the quasi-score is defined as in (2.30) and coincides with the score function (6.10). Hence, the maximum quasi-likelihood estimator $\widehat{\boldsymbol{\beta}}$ is identical to the MLE due to the multiplicative form of the variance model. Estimation of α may be carried out using the form (6.14) or via

$$\widehat{\alpha} = \frac{D}{n-k-1},$$

where D is the deviance and $\dim(\boldsymbol{\beta}) = k+1$. Asymptotic inference is based on

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

In practice, $\boldsymbol{D}$ and $\boldsymbol{V}$ are evaluated at $\widehat{\boldsymbol{\beta}}_n$, and $\widehat{\alpha}$ replaces α.

Hypothesis tests follow in an obvious fashion, with adjustment for $\widehat{\alpha}$. Specifically, if as before

$$l(\boldsymbol{\beta}, \alpha) = \int_y^{\mu} \frac{y-t}{\alpha V(t)} \mathrm{d}t,$$

then if $l(\boldsymbol{\beta}) = l(\boldsymbol{\beta}, \alpha = 1)$ represents the likelihood upon which the quasi-likelihood is based (e.g., a Poisson or binomial likelihood),

$$l(\boldsymbol{\beta}) = l(\boldsymbol{\beta}, \alpha) \times \alpha \tag{6.26}$$

and to test $H_0 : \boldsymbol{\beta}_1$ unrestricted, $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$, we may use the quasi-likelihood ratio test statistic

$$2\left[l(\widehat{\boldsymbol{\beta}}, \widehat{\alpha}) - l(\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\alpha})\right] \to_d \chi^2_{k-q-1},$$

or equivalently

$$2\left[l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}^{(0)})\right] \to_d \widehat{\alpha} \times \chi^2_{k-q}. \tag{6.27}$$

If, as is usually the case, $\widehat{\alpha} > 1$, then larger differences in the log-likelihood are required to attain the same level of significance, as compared to the $\alpha = 1$ case.

Example: Lung Cancer and Radon

Fitting the quasi-likelihood model

$$\mathrm{E}[Y_i \mid \boldsymbol{\beta}] = E_i \exp(\beta_0 + \beta_1 x_i) \tag{6.28}$$

$$\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha \mathrm{E}[Y_i \mid \boldsymbol{\beta}], \tag{6.29}$$

yields identical point estimates for $\boldsymbol{\beta}$ to the Poisson model, with scale parameter estimate $\widehat{\alpha} = 2.81$, obtained via (6.14). Therefore, with respect to $H_0 : \beta_0$ unrestricted, $\beta_1 = 0$, the quasi log-likelihood ratio statistic is $46.2/2.81 = 16.5$ so that the significance level is vastly reduced, though still strongly suggestive of a nonzero slope.

6.7 Sandwich Estimation for GLMs

The asymptotic variance–covariance for $\widehat{\boldsymbol{\beta}}$, which is given by (6.11), is appropriate only if the first two moments are correctly specified. In general, as detailed in Sect. 2.6, $\text{var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{\text{T}})^{-1}$ where

$$\boldsymbol{A} = \text{E}\left[\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\beta}}\right] = \boldsymbol{D}^{\text{T}}\boldsymbol{V}^{-1}\boldsymbol{D}, \tag{6.30}$$

regardless of the distribution of the data (so long as the mean is correctly specified), and

$$\boldsymbol{B} = \text{var}\left[\boldsymbol{G}(\boldsymbol{\beta})\right] = \boldsymbol{D}^{\text{T}}\boldsymbol{V}^{-1}\text{var}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D},$$

where $\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{S}(\boldsymbol{\beta})/n$. Under the assumption of uncorrelated errors,

$$\widehat{\boldsymbol{B}} = \sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\text{T}}\frac{\text{var}(Y_i)}{V_{ii}^2}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right) \tag{6.31}$$

where a naive estimator of $\text{var}(Y_i)$ is

$$\widehat{\sigma}_i^2 = (Y_i - \widehat{\mu}_i)^2, \tag{6.32}$$

which has finite sample bias. Combination of (6.31) and (6.32) provides a consistent estimator of the variance and therefore asymptotically corrects confidence interval coverage (so long as independence of responses holds).

Bootstrap methods (Sect. 2.7.2) may also be used to provide inference that is robust to certain aspects of model misspecification, provided n is sufficiently large. The resampling residuals method may be applied, but the meaning of residuals is ambiguous in GLMs (Sect. 6.9), and this method does not correct for mean–variance misspecification, which is a major drawback. The resampling cases approach corrects for this aspect. Davison and Hinkley (1997, Sect. 7.2) discuss both resampling residuals and resampling cases in the context of GLMs.

Example: Pharmacokinetics of Theophylline

Table 6.2 gives confidence intervals for $x_{1/2}$, $x_{\max}$ and $\mu(x_{\max})$, based on sandwich estimation. In each case, the interval estimates are a little shorter than the model-based estimates. This could be due to either instability in the sandwich estimates with a small sample size ($n = 10$) or to the gamma mean–variance assumption being inappropriate.

6.8 Bayesian Inference for GLMs

We now consider Bayesian inference for the GLM. The posterior is

$$p(\boldsymbol{\beta}, \alpha \mid \boldsymbol{y}) \propto l(\boldsymbol{\beta}, \alpha)\pi(\boldsymbol{\beta}, \alpha)$$

where it is usual to assume prior independence between the regression coefficients $\boldsymbol{\beta}$ and the scale parameter α, that is, $\pi(\boldsymbol{\beta}, \alpha) = \pi(\boldsymbol{\beta})\pi(\alpha)$.

6.8.1 Prior Specification

Recall that $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]$. Often, β_j, $j = 0, 1, \ldots, k$, is defined on $\mathbb{R}$, and so a multivariate normal prior for $\boldsymbol{\beta}$ is the obvious choice. Furthermore, independent priors are frequently defined for each component. As a limiting case, the improper prior $\pi(\boldsymbol{\beta}) \propto 1$ results. However, care should be taken with this choice since it may lead to an improper posterior. With canonical links, impropriety only occurs for pathological datasets (see the binomial model example of Sect. 3.4), but for noncanonical links, innocuous datasets may lead to impropriety, as the Poisson data with a linear link example described below illustrates. If the scale parameter $\alpha > 0$ is unknown, gamma or lognormal distributions provide obvious choices.

Poisson Data with a Linear Link

Recall the Poisson model with a linear link function

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\{E_i[(1 - x_i)\beta_0 + x_i\beta_1]\}$$

and suppose we assume an improper uniform prior for $\beta_0 > 0$, that is,

$$\pi(\beta_0) \propto 1.$$

We define $\mathrm{e}^{\gamma} = \beta_1/\beta_0 > 0$ as the parameter of interest and write

$$\mu_i = \beta_0 E_i[(1 - x_i) + x_i \exp(\gamma)] = \beta_0 \mu_i^{\star}.$$

The marginal posterior for γ is

$$\begin{aligned} p(\gamma \mid \boldsymbol{y}) &= \int p(\beta_0, \gamma \mid \boldsymbol{y})\, d\beta_0 \\ &\propto \int l(\beta_0, \gamma)\, d\beta_0 \times \pi(\gamma) \\ &\propto \int \exp\left(-\beta_0 \sum_{i=1}^{n} \mu_i^{\star y_i}\right) \beta_0^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} \mu_i^{\star y_i}\, d\beta_0 \times \pi(\gamma) \\ &\propto \prod_{i=1}^{n} \left(\frac{E_i[(1-x_i) + x_i \mathrm{e}^{\gamma}]}{\sum_{i=1}^{n} E_i[(1-x_i) + x_i \mathrm{e}^{\gamma}]}\right)^{y_i} \times \pi(\gamma) \qquad (6.33) \\ &= l(\gamma) \times \pi(\gamma), \qquad (6.34) \end{aligned}$$

where the last line follows from the previous on recognizing that the integrand is the kernel of a $\mathrm{Ga}\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} \mu_i^{\star}\right)$ distribution. The "likelihood," $l(\gamma)$ in (6.34), is of multinomial form with the total number of cases y_+ distributed among the n areas with probabilities proportional to $E_i[(1-x_i)+x_i \exp(\gamma)]$ so that, for example, larger E_i and larger x_i (if $\gamma > 0$) lead to a larger allocation of cases to area i. The likelihood contribution to the posterior tends to the constant

$$\prod_{i=1}^{n} \left(\frac{E_i(1-x_i)}{\sum_{i=1}^{n} E_i(1-x_i)}\right)^{y_i} \qquad (6.35)$$

as $\gamma \to -\infty$, showing that, in general, a proper prior is required (since the tail will be non-integrable). The constant (6.35) is nonzero unless $x_i = 1$ in any area with $y_i \neq 0$. The reason for the impropriety is that in the limit as $\gamma \to -\infty$, the relative risk $\exp(\gamma) \to 0$ so that exposed individuals cannot get the disease, which is not inconsistent with the observed data, unless all individuals in area i are exposed, $x_i = 1$, and $y_i \neq 0$ in that area since then clearly (under the assumed model) the cases are due to exposure. A similar argument holds as $\gamma \to \infty$, with replacement of $1 - x_i$ by x_i in (6.35) providing the limiting constant.

Figure 6.3 illustrates this behavior for the Scottish lip cancer example, for which $x_i = 0$ in five areas. The log- likelihood has been scaled to have maximum 0, and the constant (6.35) is indicated with a dashed horizontal line. The MLE $\widehat{\gamma} = \log(22.7)$ is indicated as a vertical dotted line.

6.8.2 Computation

Unfortunately, when continuous covariates are present in the model, conjugate analysis is unavailable. However, sampling-based approaches are relatively easy to implement. In particular, if informative priors are available, then the rejection algorithm of Sect. 3.7.6 is straightforward to implement with sampling from the prior.

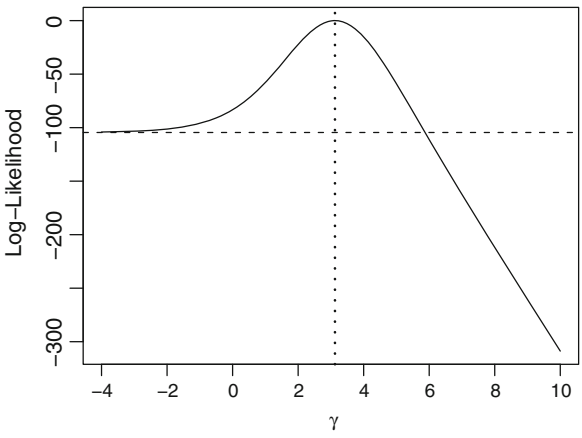

Fig. 6.3 Log-likelihood for the log relative risk parameter γ, for the Scottish lip cancer data. The *dashed horizontal line* is the constant to which the log-likelihood tends to as $\gamma \to -\infty$

MCMC (Sect. 3.8) is obviously a candidate for computation and was illustrated for Poisson and negative binomial models in Chap. 3. The INLA method described in Sect. 3.7.4 may also be used.

As described in Sect. 3.3, there is asymptotic equivalence between the sampling distribution of the MLE and the posterior distribution. Hence, Bayes estimators for $\boldsymbol{\beta}$ are consistent due to the form of the likelihood, so long as the priors are nonzero in a neighborhood of the true values of $\boldsymbol{\beta}$.

6.8.3 Hypothesis Testing

A simple method for examining hypotheses involving a single parameter, $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, with any remaining parameters unrestricted, is to evaluate the posterior tail probability $\Pr(\beta_j > 0 \mid \boldsymbol{y})$, with values close to 0 or 1 indicating that the null is unlikely to be true. Bayes factors (which were discussed in Sects. 3.10 and 4.3) provide a more general tool for comparing hypotheses (by analogy with the likelihood ratio statistic, though of course, as usual, interpretation is very different):

$$\text{BF} = \frac{p(\boldsymbol{y} \mid H_0)}{p(\boldsymbol{y} \mid H_1)}.$$

The use of Bayes factors will be illustrated in Sect. 6.16.3. As discussed in Sect. 4.3.2, great care is required in the specification of priors when model comparison is carried out using Bayes factors.

6.8.4 Overdispersed GLMs

Quasi-likelihood provides a simple procedure by which frequentist inference may accommodate overdispersion in GLMs. No such simple remedy exists within the Bayesian framework. An alternative method of increasing the flexibility of GLMs is through the introduction of random effects. We have already seen an example of this in Sect. 2.5 when the negative binomial model was derived via the introduction of gamma random effects into a Poisson model.

Example: Lung Cancer and Radon

The Bayesian Poisson model was fitted in Chap. 3 using a Metropolis–Hastings implementation. Here the use of the INLA method of Sect. 3.7.4, with improper flat priors on β_0, β_1, gives a 95% interval estimate for the relative risk $\exp(\beta_1)$ of [0.954,0.975] which is identical to that based on asymptotic likelihood inference (the posterior mean and MLE both equal -0.036, and the posterior standard deviation and standard error both equal 0.0054).

Example: Pharmacokinetics of Theophylline

With respect to the gamma GLM with $\mu(x) = \exp(\beta_0 + \beta_1 x + \beta_2/x)$, the interpretation of β_0 and β_2 in particular is not straightforward, which makes prior specification difficult. As an alternative, we specify prior distributions on the half-life $x_{1/2}$, time to maximum $x_{\max}$, maximum concentration $\mu(x_{\max})$, and coefficient of variation, $\sqrt{\alpha}$. We choose independent lognormal priors for these four parameters. For a generic parameter θ, denote the prior by $\theta \sim \text{LogNorm}(\mu, \sigma)$. To obtain the moments of these distributions, we specify the prior median θ_m and the 95% point of the prior θ_u. We then solve for the moments via

$$\mu = \log(\theta_m), \quad \sigma = \frac{\log(\theta_u) - \mu}{1.645}, \tag{6.36}$$

as described in Sect. 3.4.2. Based on a literature search, we assume prior 50% (95%) points of 8 (12), 1.5 (3), and 9 (12) for $x_{1/2}$, $x_{\max}$, and $\mu(x_{\max})$, respectively. For the coefficient of variation, the corresponding values are 0.05 (0.10). The third line of Table 6.2 summarizes these priors. To examine the posterior, we use a rejection algorithm, as described in Sect. 3.7.6. We sample from the prior on the parameters of interest and then back-solve for the parameters that describe the likelihood. For the loglinear model, the transformation to $\boldsymbol{\beta}$ is via

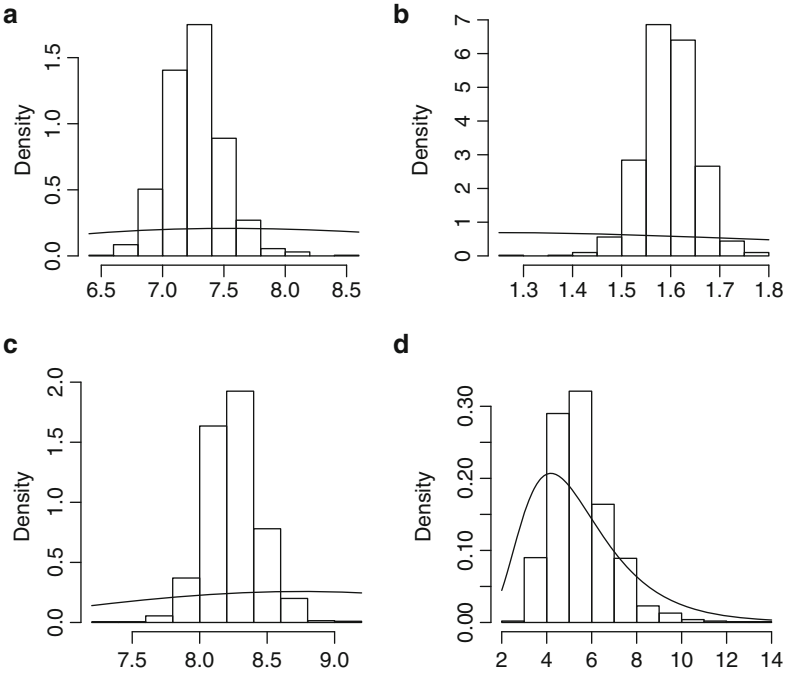

Fig. 6.4 Histogram representations of posterior distributions from the GLM for the theophylline data for (**a**) half-life, (**b**) time to maximum, (**c**) maximum concentration, and (**d**) coefficient of variation, with priors superimposed as *solid lines*

$$\beta_1 = -\frac{\log 2}{x_{1/2}}$$
$$\beta_2 = \beta_1 x_{\max}^2$$
$$\beta_0 = \log \mu(x_{\max}) + 2(\beta_1\beta_2)^{1/2}.$$

Table 6.2 summarizes inference for the parameters of interest, via medians and 90% interval estimates. Point and interval estimates show close correspondence with the frequentist summaries. Figure 6.4 gives the posterior distributions for the half-life, the time to maximum concentration, the maximum concentration, and the coefficient of variation (expressed as a percentage). The prior distributions are also indicated as solid curves. We see some skewness in each of the posteriors, which is common for nonlinear parameters unless the data are abundant.

Inference for the clearance parameter is relatively straightforward, since one simply substitutes samples for $\boldsymbol{\beta}$ into (6.5). Figure 6.5 gives a histogram representation of the posterior distribution. The posterior median of the clearance is 0.042 with 90% interval [0.041,0.044]; these summaries are identical to the likelihood-based counterparts. We see that the posterior shows little skewness; the clearance

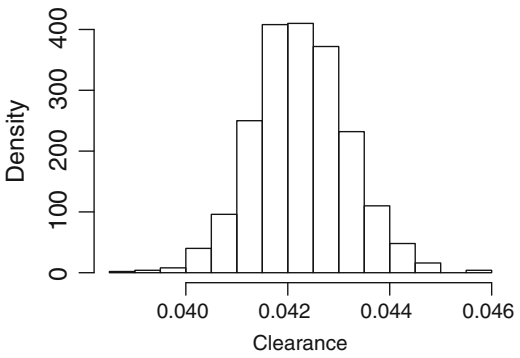

Fig. 6.5 Posterior distribution of the clearance parameter from the GLM fitted to the theophylline data

parameter is often found to be well behaved, since it is a function of the area under the curve, which is reliably estimated so long as the tail of the curve is captured.

6.9 Assessment of Assumptions for GLMs

The assessment of assumptions for GLMs is more difficult than with linear models. The definition of a residual is more ambiguous, and for discrete data in particular, the interpretation of residuals is far more difficult, even when the model is correct. Various attempts have been made to provide a general definition of residuals that possess zero mean, constant variance, and a symmetric distribution. In general, the latter two desiderata are in conflict.

When first examining the data, one may plot the response, transformed to the linear predictor scale, against covariates. For example, with Poisson data and canonical log link, one may plot $\log y$ versus covariates x.

The obvious definition of a residual is

$$e_i = Y_i - \widehat{\mu}_i$$

but clearly in a GLM, such residuals will generally have unequal variances so that some form of standardization is required. Pearson residuals, upon which we concentrate, are defined as

$$e_i^\star = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{\text{var}}(Y_i)}} = \frac{Y_i - \widehat{\mu}_i}{\widehat{\sigma}_i},$$

where $\widehat{\text{var}}(Y_i) = \widehat{\alpha} V(\widehat{\mu}_i)$ and $\widehat{\mu}_i$ are the fitted values from the model. Squaring and summing these residuals reproduce Pearson's χ^2 statistic:

$$X^2 = \sum_{i=1}^{n} e_i^{\star 2},$$

as previously introduced, (6.23). For Pearson residuals, $\mathrm{E}[\widehat{\sigma}_i e_i^\star] = 0$ and $\mathrm{E}[e_i^{\star 2}] = 1$, but the third moment is not equal to zero in general so that the residuals are skewed. As an example, for Poisson data, $\mathrm{E}[e^{\star 3}] = \mu^{-1/2}$. Clearly for normal data, Pearson residuals have zero skewness.

Deviance residuals are given by

$$e_i^\star = \mathrm{sign}(Y_i - \widehat{\mu}_i)\sqrt{D_i}$$

so that $D = \sum_{i=1}^n e_i^{\star 2}$, as defined in Sect. 6.5.3. As an example, for a Poisson likelihood, the deviance residuals are

$$e_i^\star = \mathrm{sign}(y_i - \widehat{\mu}_i)\{2[y_i \log(y_i/\widehat{\mu}_i) - y_i + \widehat{\mu}_i]\}^{1/2}.$$

For discrete data with small means, residuals are extremely difficult to interpret since the response can only take on a small number of discrete values. One strategy to aid in interpretation is to simulate data with the same design (i.e., x values) and under the parameter estimates from the fitted model. One may then examine residual plots to see their form when the model is known.

As with linear model residuals (Sect. 5.11), Pearson or deviance residuals can be plotted against covariates to suggest possible model forms. They may also be plotted against fitted values or some function of the fitted values to access mean–variance relationships. If the spread is not constant, then this suggests that the assumed mean–variance relationship is not correct. McCullagh and Nelder (1989, p. 398–399) recommend plotting against the fitted values transformed to the "constant-information" scale. For example, for Poisson data, the suggestion is to plot the residuals against $2\sqrt{\widehat{\mu}}$. Residuals can also be examined for outliers/points of high influence.

For the linear model, the diagonal elements of the hat matrix, $\boldsymbol{h} = \boldsymbol{x}(\boldsymbol{x}^\mathrm{T}\boldsymbol{x})^{-1}\boldsymbol{x}^\mathrm{T}$, correspond to the *leverage* of response i, with $h_{ii} = 1$ if $\widehat{y}_i = \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}$ (Sect. 5.11.2). Consideration of (6.15) reveals that for a GLM we may define a hat matrix as $\boldsymbol{h} = \boldsymbol{w}^{1/2}\boldsymbol{x}(\boldsymbol{x}^\mathrm{T}\boldsymbol{w}\boldsymbol{x})^{-1}\boldsymbol{x}^\mathrm{T}\boldsymbol{w}^{1/2}$, from which the diagonal elements may be extracted and, once again, large values of h_{ii} indicate that the fit is sensitive to y_i in some way. As with the linear model, responses with h_{ii} close to 1 have high influence. Unlike the linear case, $\boldsymbol{h}$ depends on the response through $\boldsymbol{w}$. Another useful standardized version of residuals is

$$e_i^\star = \frac{Y_i - \widehat{\mu}_i}{\sqrt{(1 - h_{ii})\widehat{\mathrm{var}}(Y_i)}},$$

for $i = 1, \ldots, n$.

It is approximately true that

$$\boldsymbol{V}^{-1/2}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \approx \boldsymbol{h}\boldsymbol{V}^{-1/2}(\boldsymbol{Y} - \boldsymbol{\mu})$$

(McCullagh and Nelder 1989, p. 397), and so

$$\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\boldsymbol{\mu}) \approx (\mathbf{I}-\boldsymbol{h})\boldsymbol{V}^{-1/2}(\boldsymbol{Y}-\widehat{\boldsymbol{\mu}})$$

which shows the effect of estimation of $\boldsymbol{\mu}$ on properties of the residuals.

Example: Pharmacokinetics of Theophylline

We fit the gamma GLM $Y_i \mid \boldsymbol{\beta}, \alpha \sim_{ind} \text{Ga}\left[\alpha^{-1}, (\alpha\mu_i)^{-1}\right]$ using MLE and calculate Pearson residuals

$$e_i^\star = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{\alpha}}\,\widehat{\mu}_i}.$$

In Fig. 6.6(a), these residuals are plotted versus time x_i and show no obvious systematic pattern, though interpretation is difficult, given the small number of data points and the spacing of these points over time. Figure 6.6(b) plots $|e_i^\star|$ against fitted values to attempt to discover any unmodeled mean–variance relationship, and again no strong signal is apparent.

Example: Lung Cancer and Radon

As we have seen, fitting the quasi-likelihood model given by the mean and variance specifications (6.28) and (6.29) yields $\widehat{\alpha} = 2.76$, illustrating a large amount of overdispersion. The quasi-MLE for β_1 is -0.035, with standard error 0.0088. We compare with a negative binomial model having the same loglinear mean model and

$$\text{var}(Y_i) = \mu_i(1 + \mu_i/b). \tag{6.37}$$

Previously, a negative binomial model was fitted to these data using a frequentist approach in Sect. 2.5 and a Bayesian approach in Sect. 3.8 The negative binomial MLE is -0.029, with standard error 0.0082, illustrating that there is some sensitivity to the model fitted.

For these data, the MLE is $\widehat{b} = 61.3$ with standard error 17.3. Figure 6.7 shows the fitted quadratic relationship (6.37) for these data. We also plot the quasi-likelihood fitted variance function. At first sight, it is surprising that the latter is not steeper, but the jittered fitted values included at the top of the plot are mostly concentrated on smaller values. The few larger values are very influential in producing a small estimated value of b (which corresponds to a large departure from the linear mean–variance model).

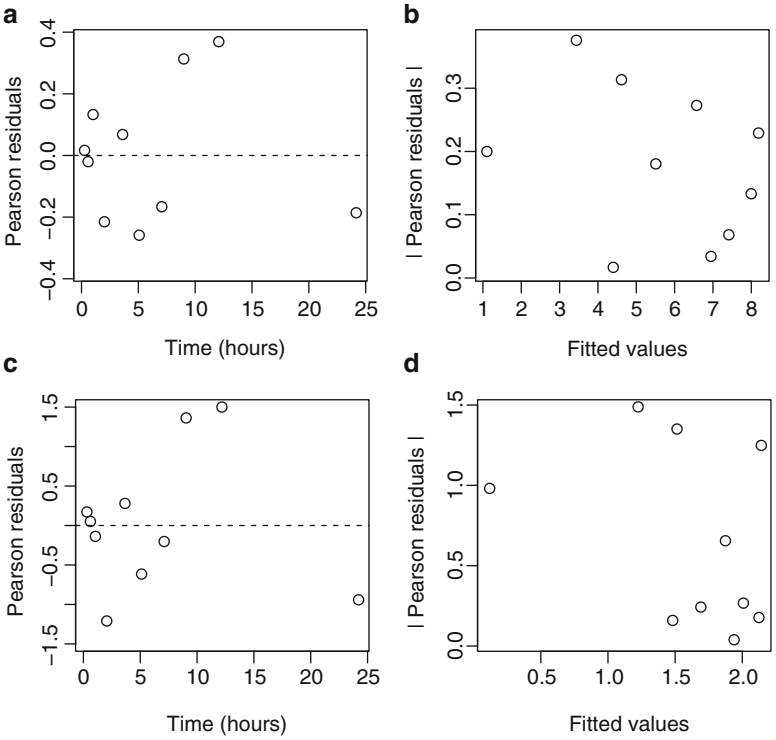

Fig. 6.6 Pearson residual plots for the theophylline data: (**a**) residuals versus time for the GLM, (**b**) absolute values of residuals versus fitted values for the GLM, (**c**) residuals versus time for the nonlinear compartmental model, and (**d**) absolute values of residuals versus fitted values for the nonlinear compartmental model

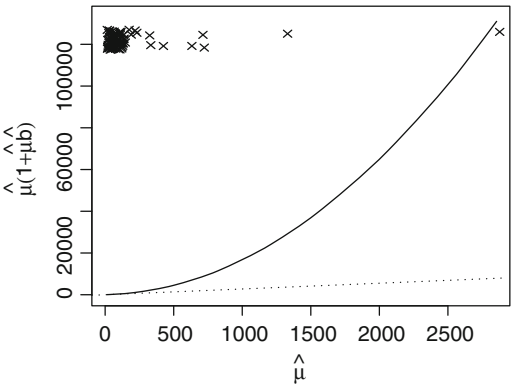

Fig. 6.7 The *solid line* shows the fitted negative binomial variance function, $\widehat{\text{var}}(Y) = \widehat{\mu}(1 + \widehat{\mu}/\widehat{b})$ plotted versus $\widehat{\mu}$ for the lung cancer and radon data. The *dotted line* corresponds to the fitted quasi-likelihood model, $\widehat{\text{var}}(Y) = \widehat{b} \times \widehat{\mu}$

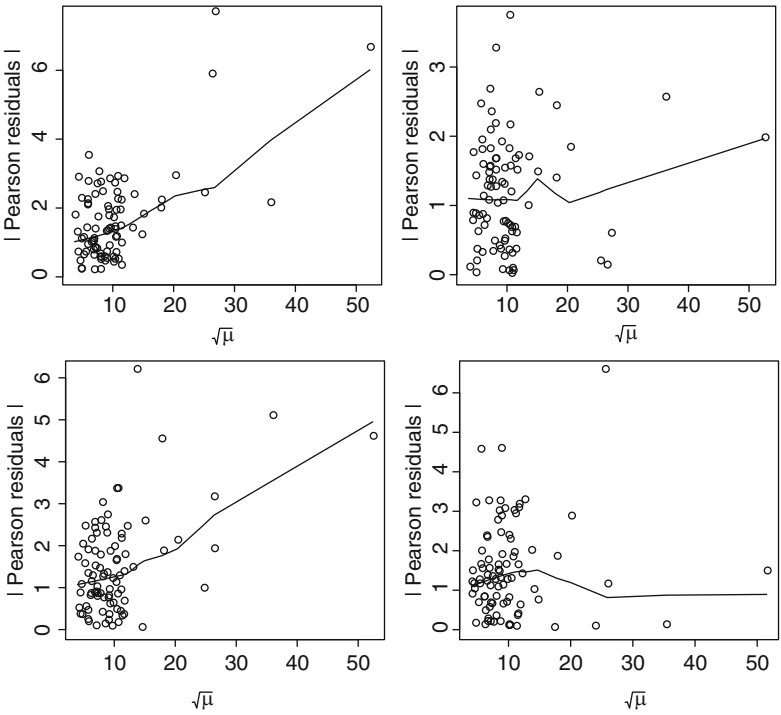

Fig. 6.8 Absolute values of Poisson Pearson residuals versus $\sqrt{\mu}$ when the true mean–variance relationship is quadratic, but we analyze as if linear, for four simulated datasets with the same expected numbers and covariate values as in the lung cancer and radon data

To attempt to determine which variance function is more appropriate, we simulate data under the negative binomial model using $\{E_i, x_i, i = 1, \ldots, n\}$ and $[\widehat{\boldsymbol{\beta}}, \widehat{b}]$.

We then fit a Poisson model (which provides identical fitted values as from a quasi-likelihood model), form residuals $(y - \widehat{\mu})/\sqrt{\widehat{\mu}}$, that is, residuals from a Poisson model, and then plot the absolute value versus $\sqrt{\mu}$ to see if we can detect a trend. In the majority of simulations, the inadequacy of assuming the variance is proportional to the mean is apparent; this endeavor is greatly helped by having just a few points with very large fitted values. Specifically, the upward trend indicates that the Poisson linear mean–variance assumption is not strong enough. Figure 6.8 shows four representative plots. Figure 6.9 gives the equivalent plot from the real data. This plot shows a similar behavior to the simulated data, and so we tentatively conclude that the quadratic mean–variance relationship is more appropriate for these data. Cox (1983) provides further discussion of the effects on estimation of different forms of overdispersion, including an extended discussion of excess-Poisson variation.

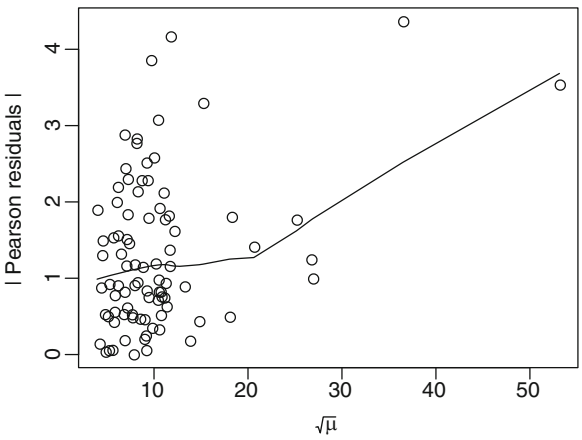

Fig. 6.9 Absolute values of Poisson Pearson residuals versus $\sqrt{\mu}$ for the lung cancer and radon data

6.10 Nonlinear Regression Models

We now consider models of the form

$$Y_i = \mu_i(\boldsymbol{\beta}) + \epsilon_i, \tag{6.38}$$

for $i = 1, \ldots, n$, where $\mu_i(\boldsymbol{\beta}) = \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$ is nonlinear in $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ is assumed to be of dimension $k + 1$, $\mathrm{E}[\epsilon_i \mid \mu_i] = 0$, $\mathrm{var}(\epsilon_i \mid \mu_i) = \sigma^2 f(\mu_i)$, and $\mathrm{cov}(\epsilon_i, \epsilon_j \mid \mu_i, \mu_j) = 0$. Such models are often used for positive responses, and if such data are modeled on the original scale, it is common to find that the variance is of the form $f(\mu) = \mu$ or $f(\mu) = \mu^2$. An alternative approach that is appropriate for the latter case is to assume constant errors on the log-transformed response scale (see Sect. 5.5.3). More generally, we might assume that $\mathrm{var}(\epsilon_i \mid \boldsymbol{\beta}, \boldsymbol{x}_i) = \sigma^2 g_1(\boldsymbol{\beta}, \boldsymbol{x}_i)$, with $\mathrm{cov}(\epsilon_i, \epsilon_j \mid \boldsymbol{\beta}, \boldsymbol{x}_i, \boldsymbol{x}_j) = g_2(\boldsymbol{\beta}, \boldsymbol{x}_i, \boldsymbol{x}_j)$. When data are measured over time, serial correlation can be a particular problem. We concentrate on the simpler second moment structure here.

Example: Michaelis–Menten Model

A nonlinear form that is used to model the kinetics of many enzymes has mean

$$\mu(z) = \frac{\alpha_0 z}{\alpha_1 + z},$$

a nonlinear model. Parameter interpretation is obtained by recognizing that as $z \to \infty$, $\mu(z) \to \alpha_0$ and at α_1, $\mu(\alpha_1) = \alpha_0/2$. A possible model for such data is

$$Y(z) = \mu(z) + \epsilon(z),$$

with $\mathrm{E}[\epsilon(z)] = 0$, $\mathrm{var}[\epsilon(z)] = \sigma^2\mu(z)^r$, with $r = 0, 1$, or 2. An alternative approach is to write

$$\frac{1}{\mu(x)} = \beta_0 + \beta_1 x$$

where

$$\begin{aligned} x &= 1/z \\ \beta_0 &= 1/\alpha_0 \\ \beta_1 &= \alpha_1/\alpha_0, \end{aligned}$$

which is a GLM with reciprocal link.

6.11 Identifiability

For many nonlinear models, *identifiability* is an issue, by which we mean that the same curve may be obtained with different sets of parameter values. We have already seen one example of this for the nonlinear model fitted to the theophylline data (Sect. 6.2). As a second example, consider the sum-of-exponentials model

$$\mu(x, \boldsymbol{\beta}) = \beta_0 \exp(-x\beta_1) + \beta_2 \exp(-x\beta_3), \tag{6.39}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3]$ and $\beta_j > 0$, $j = 0, 1, 2, 3$. The same curve results under the parameter sets $[\beta_0, \beta_1, \beta_2, \beta_3]$ and $[\beta_2, \beta_3, \beta_0, \beta_1]$, and so we have non-identifiability. In the previous "flip-flop" model (Sect. 6.2), identifiability could be imposed through a substantive assumption such as $k_a > k_e > 0$, and for model (6.39), we may enforce (say) $\beta_3 > \beta_1 > 0$ and work with the set

$$\boldsymbol{\gamma} = [\log \beta_0, \log(\beta_3 - \beta_1), \log \beta_2, \log \beta_1]$$

which constrains $\beta_0 > 0$, $\beta_2 > 0$, and $\beta_1 > \beta_3 > 0$. If a Bayesian approach is followed, a second possibility is to retain the original parameter set, but assign one set of curves zero mass in the prior. The latter option is less appealing since it can lead to a discontinuity in the prior.

6.12 Likelihood Inference for Nonlinear Models

6.12.1 Estimation

To obtain the likelihood function, a probability model for the data must be fully specified. A common choice is

$$Y_i \mid \boldsymbol{\beta}, \sigma \sim_{ind} \mathrm{N}[\mu_i(\boldsymbol{\beta}), \sigma^2 \mu_i(\boldsymbol{\beta})^r],$$

for $i = 1, \ldots, n$, and with $r = 0, 1$, or 2 being common choices. The corresponding likelihood function is

$$l(\boldsymbol{\beta}, \sigma) = -n \log \sigma - \frac{r}{2} \sum_{i=1}^{n} \log \mu_i(\boldsymbol{\beta}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})}. \tag{6.40}$$

Differentiation with respect to $\boldsymbol{\beta}$ and σ yields, with a little rearrangement, the score equations

$$\begin{aligned} \boldsymbol{S}_1(\boldsymbol{\beta}, \sigma) &= \frac{\partial l}{\partial \boldsymbol{\beta}} \\ &= \frac{r}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{1}{\mu_i(\boldsymbol{\beta})} \left\{ \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})} - \sigma^2 \right\} + \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]}{\mu_i(\boldsymbol{\beta})^r} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \end{aligned} \tag{6.41}$$

$$\begin{aligned} S_2(\boldsymbol{\beta}, \sigma) &= \frac{\partial l}{\partial \sigma} \\ &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})}. \end{aligned}$$

Notice that this pair of quadratic estimating functions (Sect. 2.8) are such that $\mathrm{E}[\boldsymbol{S}_1] = \mathbf{0}$ and $\mathrm{E}[S_2] = 0$ if the first two moments are correctly specified, in which case consistency of $\boldsymbol{\beta}$ results. It is important to emphasize that if $r > 0$, we require the second moment to be correctly specified in order to produce a consistent estimator of $\boldsymbol{\beta}$. If $r = 0$, the first term of (6.41) disappears, and we require the first moment only for consistency. In general, the MLEs $\widehat{\boldsymbol{\beta}}$ are not available in closed form, but numerical solutions are usually straightforward (e.g.,via Gauss–Newton methods or variants thereof) and are available in most statistical software. The MLE for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\widehat{\boldsymbol{\beta}})]^2}{\mu_i^r(\widehat{\boldsymbol{\beta}})}, \tag{6.42}$$

but, by analogy with the linear model case, it is more usual to use the degrees of freedom adjusted estimator

$$\widetilde{\sigma}^{\,2} = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{[Y_i - \mu_i(\widehat{\boldsymbol{\beta}})]^2}{\mu_i^r(\widehat{\boldsymbol{\beta}})}. \tag{6.43}$$

For a nonlinear model, $\widetilde{\sigma}^{\,2}$ has finite sample bias but is often preferred to (6.42) because of better small sample performance.

Under the usual regularity conditions,

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

where $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma]$ and $\boldsymbol{I}(\boldsymbol{\theta})$ is Fisher's expected information. In the case of $r = 0$, we obtain

$$\begin{aligned}
l(\boldsymbol{\beta}, \sigma) &= -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[Y_i - \mu_i(\boldsymbol{\beta})]^2 \\
\boldsymbol{S}_1(\boldsymbol{\beta}, \sigma) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}[Y_i - \mu_i(\boldsymbol{\beta})]\frac{\partial\mu_i}{\partial\boldsymbol{\beta}} \\
S_2(\boldsymbol{\beta}, \sigma) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}[Y_i - \mu_i(\boldsymbol{\beta})]^2 \\
\boldsymbol{I}_{11} &= -\mathrm{E}\left[\frac{\partial\boldsymbol{S}_1}{\partial\boldsymbol{\beta}}\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right) \\
\boldsymbol{I}_{12} &= -\mathrm{E}\left[\frac{\partial\boldsymbol{S}_1}{\partial\sigma}\right] = \mathbf{0} \\
\boldsymbol{I}_{21} &= -\mathrm{E}\left[\frac{\partial S_2}{\partial\boldsymbol{\beta}}\right] = \mathbf{0}^{\mathrm{T}} \\
\boldsymbol{I}_{22} &= -\mathrm{E}\left[\frac{\partial S_2}{\partial\sigma}\right] = \frac{2n}{\sigma^2}.
\end{aligned} \tag{6.44}$$

Asymptotically,

$$\frac{\sum_{i=1}^{n}[Y_i - \mu(\widehat{\boldsymbol{\beta}})]^2}{\sigma^2} \to_d \chi^2_{n-k-1} \tag{6.45}$$

which may be used to construct approximate F tests, as described in Sect. 6.12.2. If r is unknown, then it may also be estimated by deriving the score from the likelihood (6.40), though an abundance of data will be required. Estimation of the power in a related variance model is carried out in the example at the end of Sect. 9.20.

Example: Pharmacokinetics of Theophylline

We let y_i represent the log concentration and assume the model $y_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \mathrm{N}[\mu_i(\boldsymbol{\beta}), \sigma^2]$, $i = 1, \ldots, n$, where

$$\mu_i(\boldsymbol{\beta}) = \log\left\{\frac{Dk_a}{V(k_a - k_e)}\left[\exp(-k_e x) - \exp(-k_a x)\right]\right\} \tag{6.46}$$

with $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$ and $\beta_0 = V$, $\beta_1 = k_a$, $\beta_2 = k_e$. We fit this model using maximum likelihood estimation for $\boldsymbol{\beta}$ and the moment estimator (6.43) for σ^2. The results are displayed in Table 6.2, with the fitted curve displayed on Fig. 6.1. Confidence intervals, based on the asymptotic distribution of the MLE, were calculated for the parameters of interest using the delta method. These parameters are all positive, and so the intervals were obtained on the log-transformed scale and then exponentiated.

In Fig. 6.10, slices through the three-dimensional likelihood surface are displayed. The two-dimensional surfaces are evaluated at the MLE of the third variable. A computationally expensive alternative would be to profile with respect to the third parameter, as described in Sect. 2.4.2. In the left column the range of each variable is taken as three times the asymptotic standard errors, and the surfaces are very well behaved. By contrast, in the right column of the figure, the range is ± 30 standard errors, and here we see very irregular shapes, with some of the contours remaining open. Such shapes are typical when nonlinear models are fitted and are not in general only apparent at points far from the maximum of the likelihood.

6.12.2 Hypothesis Testing

As usual, hypothesis tests may be carried out using Wald, score, or likelihood ratio statistics, and again we concentrate on the latter. Suppose that $\dim(\boldsymbol{\beta}) = k + 1$ and let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ be a partition with $\boldsymbol{\beta}_1 = [\beta_0, \ldots, \beta_q]$ and $\boldsymbol{\beta}_2 = [\beta_{q+1}, \ldots, \beta_k]$, with $0 \leq q < k$. Interest focuses on testing whether a subset of $k - q$ parameters are equal to zero via a test of the null

$$H_0 : \boldsymbol{\beta}_1 \text{ unrestricted, } \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20} \quad \text{versus} \quad H_1 : \boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2] \neq [\boldsymbol{\beta}_1, \boldsymbol{\beta}_{20}].$$

Asymptotically, and with known σ,

$$2\left[l(\widehat{\boldsymbol{\beta}}^{(1)}, \sigma^2) - l(\widehat{\boldsymbol{\beta}}^{(0)}, \sigma^2)\right] \rightarrow_d \chi^2_{k-q-1}$$

where $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\boldsymbol{\beta}}^{(1)}$ are the MLEs under null and alternative, respectively, and $l(\boldsymbol{\beta}, \sigma^2)$ is given by (6.40). Unlike the normal linear model, this result is only

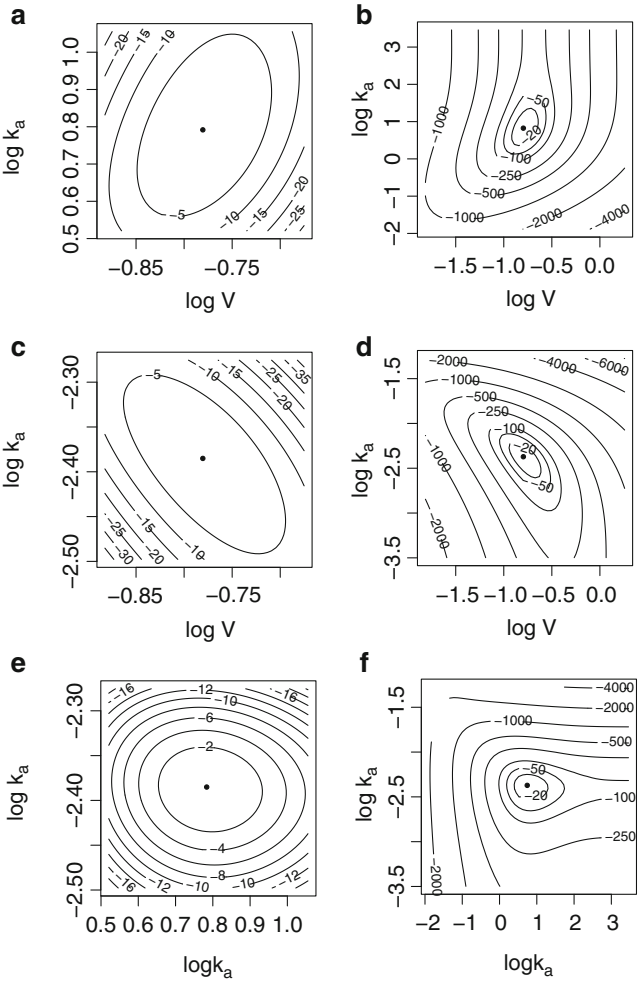

Fig. 6.10 Likelihood contours for the theophylline data with the range of each parameter being the MLE $\pm$ 3 standard errors in the *left column* and $\pm$ 30 standard errors in the *right column*; (**a**) and (**b**) $\log k_a$ versus $\log V$, (**c**) and (**d**) $\log k_e$ versus $\log V$, and (**e**) and (**f**) $\log k_e$ versus $\log k_a$. On each plot, the *filled circle* represents the MLE. In each panel, the third variable is held at its maximum value

asymptotically valid for a normal nonlinear model. For the usual case of unknown σ^2, one may substitute an estimate or use an F test with degrees of freedom $k-q-1$ and $n-k-1$, though the numerator and denominator sums of squares are only asymptotically independent. The denominator sum of squares is given in (6.45). More cautiously, one may assess the significance using Monte Carlo simulation under the null.

6.13 Least Squares Inference

We first consider model (6.38) with $\mathrm{E}[\epsilon_i \mid \mu_i] = 0$, $\mathrm{var}(\epsilon_i \mid \mu_i) = \sigma^2$, and $\mathrm{cov}(\epsilon_i, \epsilon_j \mid \mu_i, \mu_j) = 0$. In this case we may obtain ordinary least squares estimates, $\widehat{\boldsymbol{\beta}}$, that minimize the sum of squares

$$\sum_{i=1}^{n} [Y_i - \mu_i(\boldsymbol{\beta})]^2 = [\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})]^{\mathrm{T}} [\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})].$$

Differentiation with respect to $\boldsymbol{\beta}$, and letting $\boldsymbol{D}$ be the $n \times (k+1)$ dimensional matrix with element (i, j), $\partial \mu_i / \partial \beta_j$, yields the estimating function

$$\sum_{i=1}^{n} [Y_i - \mu_i(\boldsymbol{\beta})] \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \boldsymbol{D}^{\mathrm{T}} (\boldsymbol{Y} - \boldsymbol{\mu})$$

which is identical to (6.44) and is optimal within the class of linear estimating functions, under correct specification of the first two moments.

If we now assume uncorrelated errors with $\mathrm{var}(\epsilon_i \mid \mu_i) = \sigma^2 \mu_i^r(\boldsymbol{\beta})$, then the method of *generalized least squares* estimates $\widehat{\boldsymbol{\beta}}$ by temporarily forgetting that the variance depends on $\boldsymbol{\beta}$. This is entirely analogous to the motivation for quasi-likelihood; see the discussion centered around (2.28) in Sect. 2.5.1. We therefore minimize

$$\sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})} = [\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})]^{\mathrm{T}} \boldsymbol{V}(\boldsymbol{\beta})^{-1} [\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})],$$

where $\boldsymbol{V}$ is the $n \times n$ diagonal matrix with diagonal elements $\mu_i^r(\boldsymbol{\beta})$, $i = 1, \ldots, n$. The estimating function is

$$\sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}),$$

which is identical to that under quasi-likelihood (6.10). Inference may be based on the asymptotic result

$$(\boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{D} / \sigma^2)^{1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathrm{N}_{k+1}(\boldsymbol{0}, \mathbf{I}_{k+1}). \tag{6.47}$$

If the normal model is true, then the GLS estimator is not as efficient as that obtained from a likelihood approach but is more reliable under model misspecification. Therefore, the approach that is followed should depend on how much faith we have in the assumed model.

In Sect. 9.10, we will discuss further the trade-offs encountered when one wishes to exploit the additional information concerning $\boldsymbol{\beta}$ contained within the variance function.

6.14 Sandwich Estimation for Nonlinear Models

The sandwich estimator of the variance is again available and takes exactly the same form as with the GLM. In particular, consider the estimating function

$$\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}),$$

with $\boldsymbol{D}$ an $n \times (k+1)$ matrix with elements $\partial\mu_i/\partial\boldsymbol{\beta}_j$, $i = 1, \ldots, n$, $j = 0, \ldots, k+1$ and $\boldsymbol{V}$ the diagonal matrix with elements $V_{ii} = \mu_i(\boldsymbol{\beta})^r$ with $r \geq 0$ known. This estimating equation arises from likelihood considerations if $r = 0$ or, more generally, from GLS. With this form for $\boldsymbol{G}(\cdot)$, (6.30), (6.31), and (6.32) all hold.

Example: Pharmacokinetics of Theophylline

We now let y_i be the concentration and consider the model with first two moments

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}, \sigma^2] &= \mu_i(\boldsymbol{\beta}) = \frac{Dk_a}{V(k_a - k_e)} \left[\exp(-k_e x) - \exp(-k_a x)\right], \\ \mathrm{var}(Y_i \mid \boldsymbol{\beta}, \sigma^2) &= \sigma^2 \mu_i(\boldsymbol{\beta})^2, \end{aligned}$$

for $i = 1, \ldots, n$. One possibility for fitting is generalized least squares. As an alternative, we may assume $Y_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \mathrm{N}[\mu_i(\boldsymbol{\beta}), \sigma^2\mu_i(\boldsymbol{\beta})^2]$, $i = 1, \ldots, n$ and proceed with maximum likelihood estimation. Table 6.3 gives estimates of the above model under GLS and MLE, along with likelihood estimation for the model,

$$\log y_i \mid \boldsymbol{\beta}, \tau^2 \sim_{ind} \mathrm{N}\left\{\log[\mu_i(\boldsymbol{\beta})], \tau^2\right\}.$$

There are some differences in the table, but overall the estimates and standard errors are in reasonable agreement. Table 6.2 gives confidence intervals for $x_{1/2}$, $x_{\max}$, and $\mu(x_{\max})$ based on sandwich estimation. As with the GLM analysis, the interval estimates are a little shorter.

Table 6.3 Point estimates and asymptotic standard errors for the theophylline data, under various models and estimation techniques. In all cases the coefficient of variation is approximately constant

Model	$\log V$	$\log k_a$	$\log k_e$
MLE log scale	−0.78 (0.035)	0.79 (0.089)	−2.39 (0.037)
GLS original scale	−0.77 (0.030)	0.81 (0.055)	−2.39 (0.032)
MLE original scale	−0.74 (0.025)	0.85 (0.069)	−2.45 (0.044)

6.15 The Geometry of Least Squares

In this section we briefly discuss the geometry of least squares to gain insight into the fundamental differences between linear and nonlinear fitting.

We consider minimization of

$$(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{6.48}$$

where $\boldsymbol{y}$ and $\boldsymbol{\mu}$ are $n \times 1$ vectors. We first examine the linear model, $\boldsymbol{\mu} = \boldsymbol{x}\boldsymbol{\beta}$, where $\boldsymbol{x}$ is $n \times (k+1)$ and $\boldsymbol{\beta}$ is $(k+1) \times 1$. For fixed $\boldsymbol{x}$, the so-called *solution locus* maps out the fitted values $\boldsymbol{x}\widetilde{\boldsymbol{\beta}}$ for all values of $\widetilde{\boldsymbol{\beta}}$ and is a $(k+1)$-dimensional hyperplane of infinite extent. Differentiation of (6.48) gives

$$\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{e} = \boldsymbol{0}$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals. So the sum of squares is minimized when the vector $(\boldsymbol{y}-\boldsymbol{x}\boldsymbol{\beta})$ is orthogonal to the hyperplane that constitutes the solution locus. The fitted values are

$$\widehat{\boldsymbol{y}} = \boldsymbol{x}\widehat{\boldsymbol{\beta}} = \boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{h}\boldsymbol{y},$$

and are the orthogonal projection of $\boldsymbol{y}$ onto the plane spanned by the columns of $\boldsymbol{x}$, with $\boldsymbol{h}$ the matrix that represents this projection.

For a nonlinear model, the solution locus is a curved $(k+1)$-dimensional surface, possibly with finite extent. In contrast to the linear model, equally spaced points on lines in the parameter space do not map to equally spaced points on the solution locus but rather to unequally spaced points on curves.

These observations have several implications. In terms of inference, recall from Sect. 5.6.1, in particular equation (5.27) with $q = -1$, that for a linear model, a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\beta}$ is the ellipsoid

$$(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \le (k+1)s^2 F_{k+1,n-k-1}(1-\alpha).$$

Geometrically, the region has this form because the solution locus is a plane and the residual vector is orthogonal to the plane so that values of $\boldsymbol{\beta}$ map onto a disk. For nonlinear models, asymptotic inference for $\boldsymbol{\beta}$ results from

$$(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\mathrm{T}} \widehat{\boldsymbol{V}}^{-1} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq (k+1) s^2 F_{k+1,n-k-1}(1-\alpha),$$

where $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2 \widehat{\boldsymbol{V}}$, with $\widehat{\sigma}^2 = s^2$. The approximation occurs because the solution locus is curved, and equi-spaced points in the parameter space map to unequally spaced points on curved lines on the solution locus. Intuitively, inference will be more accurate if the relevant part of the solution locus is flat and if parallel equi-spaced lines in the parameter space map to parallel equi-spaced lines on the solution locus. The curvature and lack of equally spaced points manifest itself in contours of equal likelihood being banana-shaped and perhaps "open" (so that they do not join). The right column of Fig. 6.10 gives examples of this behavior. Another important aspect is that reparameterization of the model can alter the behavior of points mapped onto the solution locus, but cannot affect the curvature of the locus. Hence, the curvature of the solution locus has been referred to as the *intrinsic curvature* (Beale 1960; Bates and Watts 1980), while the aspect that is parameterization dependent is the *parameter-effects curvature* (Bates and Watts 1980). We note that the solution locus does not depend on the observed data but only on the model and design. As $n \to \infty$, the surface becomes increasingly locally linear and inference correspondingly more accurate.

We illustrate with a simple fictitious example with $n = 2$, $\boldsymbol{x} = [1, 2]$, and $\boldsymbol{y} = [0.2, 0.7]$. We compare two models, each with a single parameter, the linear zero intercept model

$$\mu = x\beta, \qquad -\infty < \beta < \infty,$$

and the (simplified) nonlinear Michaelis–Menten model

$$\mu = x/(x+\theta), \qquad \theta > 0.$$

Figure 6.11(a) plots the data versus the two fitted curves (obtained via least squares), while panel (b) plots the solution locus for the linear model, which in this case is a line (since $k = 0$). The point $[x_1\widehat{\beta}, x_2\widehat{\beta}]$ with least squares estimate

$$\widehat{\beta} = \sum_{i=1}^{2} x_i y_i / \sum_{i=1}^{2} x_i^2 = 0.32,$$

is the fitted point and is indicated as a solid circle. The dashed line is the vector joining $[y_1, y_2]$ to the fitted point and is perpendicular to the curved solution locus. The circles indicated on the solution locus correspond to changes in β of 0.1 and are equi-spaced on the locus. The final aspect to note is that the locus is of infinite extent.

Panel (c) of Fig. 6.11 plots the solution locus for the Michaelis–Menten model, for which $\widehat{\theta} = 1.70$. The vector joining $[y_1, y_2]$ to the fitted values $[x_1/(x_1+\widehat{\theta}), x_2/(x_2+\widehat{\theta})]$ is perpendicular to the curved solution locus, but we see that points on the

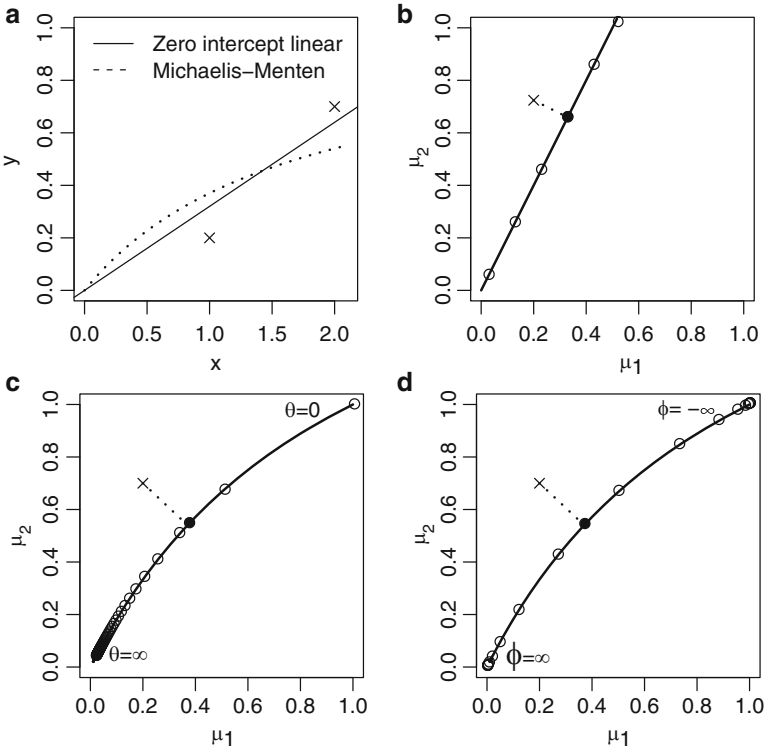

Fig. 6.11 (**a**) Fictitious data with $\boldsymbol{x} = [1, 2]$ and $\boldsymbol{y} = [0.2, 0.7]$, and fitted lines (**b**) solution locus for the zero intercept linear model with the observed data indicated as a cross and the fitted value as a *filled circle*, (**c**) solution locus for the Michaelis–Menten model with the observed data indicated as a *cross* and the fitted value as a *filled circle*, and (**d**) solution locus for the Michaelis–Menten model under a second parametrization with the observed data indicated as a *cross* and the fitted value as a *filled circle*

latter are not equally spaced. Also, the solution locus is of finite extent moving from the point $[0, 0]$ for $\theta = \infty$ to the point (1,1) for $\theta = 0$ (these are the asymptotes of the model). Finally, panel (d) reproduces panel (c) with the Michaelis–Menten model reparameterized as $\left[x_1/[x_1 + \exp(\widehat{\phi})], x_2/[x_2 + \exp(\widehat{\phi})]\right]$, with $\phi = \log\theta$. The spacing of points on the solution locus is quite different under the new parameterization. The points are more equally spaced close to the fitted value, indicating that asymptotic standard errors are more likely to be accurate under this parametrization.

6.16 Bayesian Inference for Nonlinear Models

Bayesian inference for nonlinear models is based on the posterior distribution

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto l(\boldsymbol{\beta})\pi(\boldsymbol{\beta}, \sigma^2).$$

We discuss in turn prior specification, computation, and hypothesis testing.

6.16.1 Prior Specification

We begin by assuming independent priors on $\boldsymbol{\beta}$ and σ^2:

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta})\pi(\sigma^2).$$

The prior on σ^2 is a less critical choice, and $\sigma^{-2} \sim \text{Ga}(a, b)$ is an obvious candidate. The choice $a = b = 0$, which gives the improper prior $\pi(\sigma^2) \propto 1/\sigma^2$, will often be a reasonable option. If the variance model is of the form $\text{var}(Y_i) = \sigma^2 \mu_i(\boldsymbol{\beta})^r$, then clearly substantive prior beliefs will depend on r so that we must specify the conditional form $\pi(\sigma^2 \mid r)$, since the scale of σ^2 depends on the choice for r.

So far as a prior for $\boldsymbol{\beta}$ is concerned, great care must be taken to ensure that the resultant posterior is proper; Sect. 3.4 provided an example of the problems that can arise with a nonlinear model. In general, models must be considered on a case-by-case basis. However, a parameter, θ (say), corresponding to an asymptote (so that $\mu \to a$ as $\theta \to \infty$), will generally require proper priors because the likelihood tends to the constant

$$\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)^2\right]$$

as $\theta \to \infty$ and not zero as is necessary to ensure propriety.

6.16.2 Computation

Unfortunately, closed-form posterior distributions do not exist with a nonlinear mean function, but sampling-based methods are again relatively straightforward to implement. A pure Gibbs sampling strategy (Sect. 3.8.4) is not so appealing since the conditional distribution, $\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma$, will not have a familiar form. However, Metropolis–Hastings algorithms (Sect. 3.8.2) will be easy to construct. If an informative prior is present, direct sampling via a rejection algorithm, with the prior as a proposal, may present a viable option.

6.16.3 Hypothesis Testing

As with GLMs (Sect. 6.8.3), posterior tail areas and Bayes factors are available to test hypotheses/compare models.

Example: Pharmacokinetics of Theophylline

We report a Bayesian analysis of the theophylline data and specify lognormal priors for $x_{1/2}$, $x_{\max}$, and $\mu(x_{\max})$ using the same specification as with the GLM analysis. Samples from the posterior for $[V, k_a, k_e]$ are obtained from the rejection algorithm. Specifically, we sample from the prior on the parameters of interest and then backsolve for the parameters that describe the likelihood. For the compartmental model, we transform back to the original parameters via

$$\begin{aligned} k_e &= (\log 2)/x_{1/2} \\ 0 &= x_{\max}(k_a - k_e) - \log\left(\frac{k_a}{k_e}\right) \\ V &= \frac{D}{\mu(x_{\max})}\left(\frac{k_a}{k_e}\right)^{k_a/(k_a-k_e)} \end{aligned} \tag{6.49}$$

so that k_a is not directly available but must be obtained as the root of (6.49).

Table 6.2 summarizes inference for the parameters of interest with the interval estimates and medians being obtained as the sample quantiles. Figure 6.12 shows the posteriors for functions of interest under the nonlinear model. The posteriors are skewed for all functions of interest. These figures and Table 6.2 show that Bayesian inference for the GLM and nonlinear model are very similar. Frequentist and Bayesian methods are also in close agreement for these data, which is reassuring.

Recall that the parameter sets $[V, k_a, k_e]$ and $[Vk_e/k_a, k_e, k_a]$ produce identical curves for the compartmental model (6.1). One solution to this identifiability problem is to enforce $k_a > k_e > 0$, for example, by parameterizing in terms of $\log k_e$ and $\log(k_a - k_e)$. Pragmatically, not resorting to this parameterization is reasonable, so long as k_a and k_e are not close. Figure 6.13 shows the bivariate posterior distribution $p(k_a, ke \mid \boldsymbol{y})$, and we see that $k_a \gg k_e$ for these data, and so there is no need to address the identifiability issue.

Another benefit of specifying the prior in terms of model-free parameters is that models may be compared using Bayes factors on an "even playing field," in the sense that the prior input for each model is identical. For more discussion of this issue, see Pérez and Berger (2002). To illustrate, we compare the GLM and nonlinear compartmental models. The normalizing constants for these models are 0.00077 and 0.00032, respectively, as estimated via importance sampling with the prior as

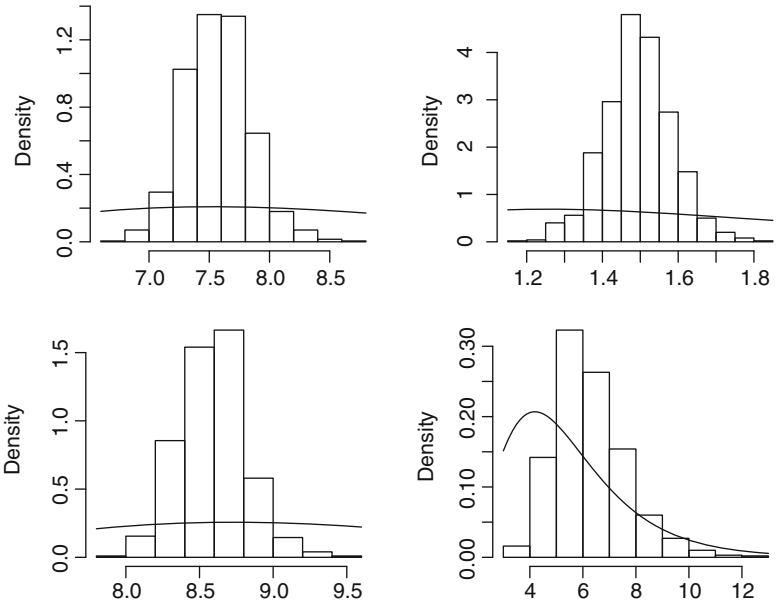

Fig. 6.12 Histogram representations of posterior distributions from the nonlinear compartmental model for the theophylline data for the (**a**) half-life, (**b**) time to maximum, (**c**) maximum concentration, and (**d**) coefficient of variation, with priors superimposed as *solid lines*

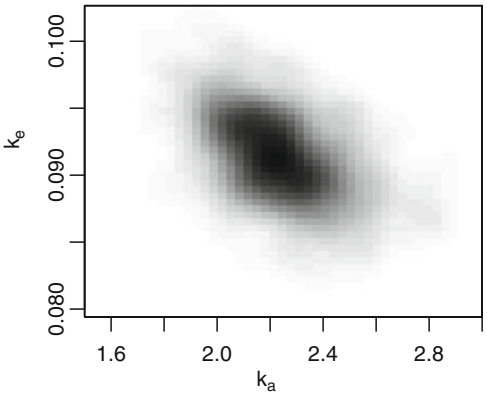

Fig. 6.13 Image plot of samples from the joint posterior distribution of the absorption and elimination rate constants, k_a and k_e, for the theophylline data

proposal and using (3.28). Consequently, the Bayes factor comparing the GLM to the nonlinear model is 2.4 so that the data are just over twice as likely under the GLM, but this is not strong evidence. For these data, based on the above analyses, we

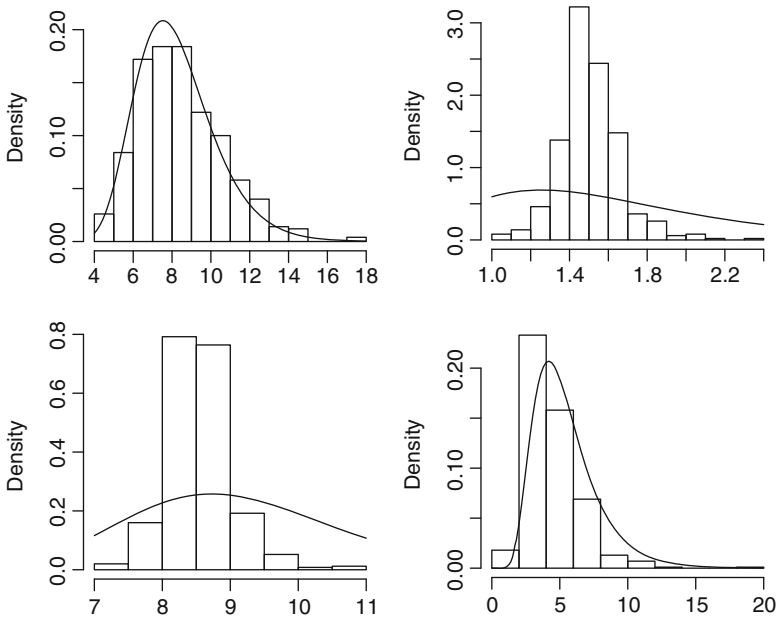

Fig. 6.14 Histogram representations of posterior distributions from the nonlinear compartmental models for the reduced theophylline dataset of $n = 3$ points for the (**a**) half-life, (**b**) time to maximum, (**c**) maximum concentration, and (**d**) coefficient of variation, with priors superimposed as *solid lines*

conclude that both the GLM and the nonlinear models provide adequate fits to the data, and there is little difference between the frequentist and Bayesian approaches to inference.

We now demonstrate the benefits of a Bayesian approach with substantive prior information, when the data are sparse. To this end, we consider a reduced dataset consisting of the first $n = 3$ concentrations only. Clearly, a likelihood or least squares approach is not possible in this case, since the number of parameters (three regression parameters plus a variance) is greater than the number of data points. We fit the nonlinear model with the same priors as used previously and with computation carried out with the rejection algorithm. Figure 6.14 shows the posterior distributions, with the priors also indicated. As we might expect, there is no/little information in the data concerning the terminal half-life $\log k_e/2$ or the standard deviation σ. In contrast, the data are somewhat informative with respect to the time to maximum concentration, and the maximum concentration.

6.17 Assessment of Assumptions for Nonlinear Models

In contrast to GLMs, residuals are unambiguously defined for nonlinear models as

$$e_i^\star = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\text{var}}(Y_i)}}, \tag{6.50}$$

which we refer to as Pearson residuals. These residuals may be used in the usual ways; see Sects. 5.11.3 and 6.9. In particular, the residuals may be plotted versus covariates to assess the mean model, and the absolute values of the residuals may be plotted versus the fitted values $\widehat{\mu}_i$ to assess the appropriateness of the mean–variance model. For a small sample size, normality of the errors will aid in accurate asymptotic inference and may be assessed via a normal QQ plot, as described in Sect. 5.11.3.

Example: Pharmacokinetics of Theophylline

Letting y_i represent the log concentration at time x_i, we examine the Pearson residuals, as given by (6.50), obtained following likelihood estimation with the model $y_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \text{N}(\mu_i, \sigma^2)$, with μ_i given by (6.46), for $i = 1, \ldots, n$. Figure 6.6(c) plots $e_i^\star$ versus x_i and shows no gross inadequacy of the mean model. Panel (d), which plots $|e_i^\star|$ versus x_i, similarly shows no great problem with the mean–variance relationship. Figure 6.15 gives a normal QQ plot of the residuals and indicates no strong violation of normality. In all cases, interpretation is hampered by the small sample size.

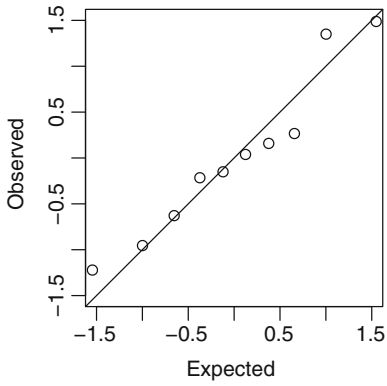

Fig. 6.15 Normal QQ plot for the theophylline data and model (6.46)

6.18 Concluding Remarks

Within the broad class of general regression models, the use of GLMs offers certain advantages in terms of computation and interpretation, though one should not restrict attention to this class. Many results and approaches used for linear models hold approximately for GLMs. For example, the influence of points was defined through the weight matrix used in the "working response" approach implicit in the IRLS algorithm (Sect. 6.5.2). The form of GLMs, in particular the linearity of the score with respect to the responses, is such that asymptotic inference is accurate for relatively small n.

Care is required in the fitting of, and inference for, nonlinear models. For example, models must be examined to see if the parameters are uniquely identified. For both GLMs and nonlinear models, the examination of residual plots is essential to determine whether the assumed model is appropriate, but such plots are difficult to interpret because the behavior of residuals is not always obvious, even if the fitted model is correct. The use of a distribution from the exponential family is advantageous in that results on consistency of estimators follow easily, as discussed in Sect. 6.5.1. The identifiability of nonlinear models should always be examined, and one should be wary of the accuracy of asymptotic inference for small sample sizes. The parameterization adopted is also important, as discussed in Sect. 6.15.

6.19 Bibliographic Notes

The most comprehensive and interesting description of GLMs remains McCullagh and Nelder (1989). An excellent review is also given by Firth (1993). Sandwich estimation for GLMs is discussed by Kauermann and Carroll (2001).

Nonlinear models are discussed by Bates and Watts (1988) and Chap. 2 of Davidian and Giltinan (1995), with an emphasis on generalized least squares. Book-length treatments on nonlinear models are provided by Gallant (1987); Seber and Wild (1989); see also Carroll and Ruppert (1988).

Gibaldi and Perrier (1982) provide a comprehensive account of pharmacokinetic models and principles and Godfrey (1983) an account of compartmental modeling in general. Wakefield et al. (1999) provide a review of pharmacokinetic and pharmacodynamic modeling including details on both the biological and statistical aspects of such modeling. The model given by (6.7) and (6.8) was suggested by Wakefield (2004) and was developed more extensively in Salway and Wakefield (2008).

6.20 Exercises

6.1 A random variable Y is inverse Gaussian if its density is of the form

$$p(y \mid \lambda, \delta) = \left(\frac{\delta}{2\pi y^3}\right)^{1/2} \exp\left[\frac{-\delta(y-\lambda)^2}{2\lambda^2 y}\right],$$

for $y > 0$.

(a) Show that the inverse Gaussian distribution is a member of the exponential family and identify θ, α, $b(\theta)$, $a(\alpha)$, and $c(y, \alpha)$.
(b) Give forms for $\mathrm{E}[Y \mid \theta, \alpha]$ and $\mathrm{var}(Y \mid \theta, \alpha)$ and determine the canonical link function.

6.2 Table 6.4 reproduces data, from Altham (1991), of counts of T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and from 20 additional patients in remission from disseminated malignancies. A question of interest here is whether there is a difference in the distribution of cell counts between the two diseases. A quantitative assessment of any difference is also desirable.

(a) Carry out an exploratory examination of these data and provide an informative graphical summary of the two distributions of responses.
(b) These data may be examined: (1) on their original scale, (2) $\log_e$ transformed, and (3) square root transformed. Carefully define a difference in location parameter in each of the designated scales. What are the considerations when choosing a scale? Obtain 90% confidence interval for each of the difference parameters.
(c) Fit Poisson, gamma, and inverse Gaussian models to the cell count data, assuming canonical links in each case.
(d) Using the asymptotic distribution of the MLE, give 90% confidence intervals for the difference parameters in each of the three models. Under each of the models, would you conclude that the means of the two groups are equal?

6.3 The data in Table 6.5, taken from Wakefield et al. (1994), were collected following the administration of a single 30 mg dose of the drug cadralazine

Table 6.4 Counts of T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 other patients in remission from disseminated malignancies

Hodgkin's disease	396	568	1,212	171	554	1,104	257	435	295	397
Non-Hodgkin's disease	375	375	752	208	151	116	736	192	315	1,252
Hodgkin's disease	288	1,004	431	795	1,621	1,378	902	958	1,283	2,415
Non-Hodgkin's disease	675	700	440	771	688	426	410	979	377	503

Table 6.5 Concentrations y_i of the drug cadralazine as a function of time x_i, obtained from a subject who was administered a dose of 30 mg. These data are from Wakefield et al. (1994)

Observation number i	Time (hours) x_i	Concentration (mg/liter) y_i
1	2	1.63
2	4	1.01
3	6	0.73
4	8	0.55
5	10	0.41
6	24	0.01
7	28	0.06
8	32	0.02

to a cardiac failure patient. The response y_i represents the drug concentration at time x_i, $i = 1, \ldots, 8$. The most straightforward model for these data is to assume

$$\log y_i = \mu(\boldsymbol{\beta}) + \epsilon_i = \log\left[\frac{D}{V}\exp(-k_e x_i)\right] + \epsilon_i,$$

where $\epsilon_i \mid \sigma^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$, $\boldsymbol{\beta} = [V, k_e]$ and the dose is $D = 30$. The parameters are the volume of distribution $V > 0$ and the elimination rate k_e.

(a) For this model, obtain expressions for:

  (i) The log-likelihood function $L(\boldsymbol{\beta}, \sigma^2)$
  (ii) The score function $\boldsymbol{S}(\boldsymbol{\beta}, \sigma^2)$
  (iii) The expected information matrix $\boldsymbol{I}(\boldsymbol{\beta}, \sigma^2)$

(b) Obtain the MLE and provide an asymptotic 95% confidence interval for each element of $\boldsymbol{\beta}$.
(c) Plot the data, along with the fitted curve.
(d) Using residuals, examine the appropriateness of the assumptions of the above model. Does the model seem reasonable for these data?
(e) The clearance $Cl = V \times k_e$ and elimination half-life $x_{1/2} = \log 2/k_e$ are parameters of interest in this experiment. Find the MLEs of these parameters along with asymptotic 95% confidence intervals.

A Bayesian analysis will now be carried out, assuming independent lognormal priors for V, k_e and an independent inverse gamma prior for σ^2. For the latter, assume the improper prior $\pi(\sigma^2) \propto \sigma^{-2}$.

(f) Assume that the 50% and 90% points for V are 20 and 40 and that for k_e, these points are 0.12 and 0.25. Solve for the lognormal parameters using the method of moments equations (6.36).
(g) Implement an MCMC Metropolis–Hastings algorithm (Sect. 3.8.2). Report the median and 90% interval estimates for each of V, k_e, Cl, and $x_{1/2}$. Pro-

vide graphical summaries of each of the univariate and bivariate posterior distributions.

6.4 Let Y_i represent a count and $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ik}]$ a covariate vector for individual i, $i = 1, \ldots, n$. Assume that $Y_i \mid \mu_i \sim_{iid} \text{Poisson}(\mu_i)$, with

$$\mu_i = \text{E}[Y_i \mid \gamma_{0i}, \gamma_1, \ldots, \gamma_k] = \exp(\gamma_{0i} + \gamma_1 x_{i1} + ... + \gamma_k x_{ik}), \tag{6.51}$$

where the intercept is a *random effect* (see Chap. 9) that varies according to

$$\gamma_{0i} \mid \gamma_0, \tau^2 \sim_{iid} \text{N}(\gamma_0, \tau^2).$$

(a) Give an interpretation of each of the parameters γ_0 and γ_1.
(b) Suppose we fit an alternative Poisson model with mean

$$\mu_i^\star = \text{E}[Y_i \mid \beta_0, \beta_1, \ldots, \beta_k] = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}). \tag{6.52}$$

Evaluate

$$\text{E}[Y_i \mid \tau^2, \gamma_0, \gamma_1, \ldots, \gamma_k],$$

and hence, by comparison with $\text{E}[Y_i \mid \beta_0, \beta_1, \ldots, \beta_k]$, equate γ_j to β_j, $j = 0, 1, \ldots, k$.
(c) Evaluate $\text{var}(Y_i \mid \tau^2, \gamma_0, \gamma_1, \ldots, \gamma_k)$ and compare this expression with $\text{var}(Y_i \mid \beta_0, \beta_1, \ldots, \beta_k)$.
(d) Suppose one is interested in the parameters $\gamma_1, \ldots, \gamma_k$. Use your answers to the previous two parts to discuss the implications of fitting model (6.52) when the true model is (6.51).
(e) Now consider an alternative random effects structure in which

$$\delta_i \mid a, b \sim_{iid} \text{Ga}(a, b),$$

where $\delta_i = \exp(\gamma_{0i})$. Evaluate the marginal mean $\text{E}[Y_i \mid a, b, \gamma_1, \ldots, \gamma_k]$ and marginal variance $\text{var}(Y_i \mid a, b, \gamma_1, \ldots, \gamma_k)$.
(f) Compare the expressions for the mean and variance under the normal and gamma formulations.
(g) For the Poisson-Gamma model, calculate the form of the likelihood

$$L(\gamma_1, \ldots, \gamma_k, a, b) = \prod_{i=1}^{n} \int \Pr(y_i \mid \gamma_{0i}, \gamma_1, \ldots, \gamma_k) \pi(\gamma_{0i} \mid a, b)\, d\gamma_{0i}.$$

Derive expressions for the score and information matrix and hence describe how inference may be performed from a likelihood standpoint.

Table 6.6 Concentrations y_i of the drug theophylline as a function of time x_i obtained from a subject who was administered an oral dose of size 4.40 mg/kg

Observation number i	Time (hours) x_i	Concentration (mg/liter) y_i
1	0.27	1.72
2	0.52	7.91
3	1.00	8.31
4	1.92	8.33
5	3.50	6.85
6	5.02	6.08
7	7.03	5.40
8	9.00	4.55
9	12.00	3.01
10	24.30	0.90

6.5 Table 6.6 gives concentration–time data for an individual who was given a dose of 4.40 mg/kg of the drug theophylline. In this chapter we have analyzed the data from another of the individuals in the same trial.

(a) For the data in Table 6.6,[1] fit the gamma GLM given by (6.7) and (6.8) using maximum likelihood and report the MLEs and standard errors.
(b) Obtain MLEs and standard errors for the parameters of interest $x_{1/2}$, $x_{\max}$, $\mu(x_{\max})$, and Cl.
(c) Let z_i represent the log concentration and consider the model $z_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \mathrm{N}[\mu_i(\boldsymbol{\beta}), \sigma^2]$, $i = 1, \ldots, n$, where $\mu_i(\boldsymbol{\beta})$ is given by the compartmental model (6.46). Fit this model using maximum likelihood and report the MLEs and standard errors.
(d) Obtain the MLEs and standard errors for the parameters of interest $x_{1/2}$, $x_{\max}$, $\mu(x_{\max})$, and Cl.
(e) Compare these summaries with those obtained under the GLM.
(f) Examine the fit of the two models and discuss which provides the better fit.

[1] These data correspond to individual 2 in the `Theoph` data, which are available in `R`.