2021 Advanced Regression Methods for Independent Data

BIOSTAT/STAT 570

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington
jonno@uw.edu

CHAPTER 6: GENERAL REGRESSION MODELS

Outline

Motivation

General Regression Models

In this chapter we consider the analysis of data that are not well-modeled by the linear models described in Chapter 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.

For a nonlinear model, 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 $\boldsymbol{\beta}$ and the errors ϵ independent with zero mean.

Motivating Example: PK of Theophylline

In Table 1 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 4.53mg/kg.

These data are plotted in Figure 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 $i = 1, \ldots, n = 10$.

Motivating Example: PK of Theophylline

Observation number i	Time (hours) x_i	Concentration (mg/liter) y_i
1	0.27	4.40
2	0.58	6.90
3	1.02	8.20
4	2.02	7.80
5	3.62	7.50
6	5.08	6.20
7	7.07	5.30
8	9.00	4.90
9	12.15	3.70
10	24.17	1.05

Table 1: Concentration (y) of the drug Theophylline as a function of time (x), obtained from a subject who was administered an oral dose of size 4.53mg/kg.

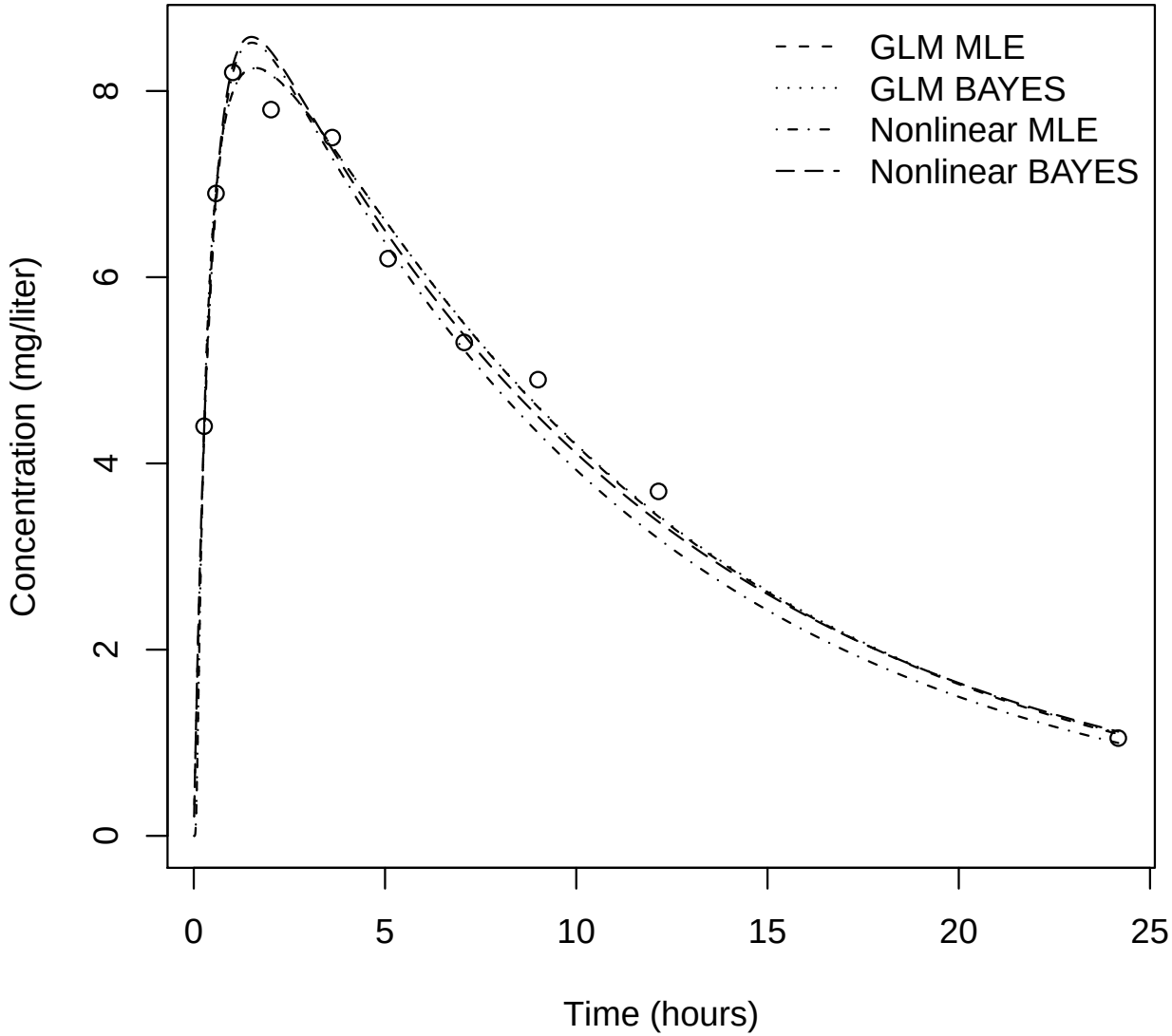


FIGURE 1: Theophylline data, along with fitted curves under various models and inferential approaches. Four curves are included, corresponding to MLE and Bayes analyses of GLM and nonlinear models. The two nonlinear curves are indistinguishable.

Motivating Example: PK of Theophylline

For the data considered here, a starting point for $\mu(x)$ is:

$$\mu(x) = \frac{Dk_a}{V(k_a - k_e)} \left[\exp(-k_e x) - \exp(-k_a x)\right] \tag{1}$$

where there are three unknown parameters:

- $k_a > 0$ 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).

Motivating Example: PK of Theophylline

A stochastic component may be added to (1) in a variety of ways, but one simple approach is via

$$y(x) = \mu(x) + \delta(x), \tag{2}$$

where $\mathrm{E}[\delta(x)] = 0$ and $\mathrm{var}[\delta(x)] = \sigma^2\mu(x)^2$ with $\delta(x)$ at different 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 (1) and (2) gives an example of a three parameter nonlinear model.

Motivating Example: PK of Theophylline

An approximately constant coefficient of variation can also be achieved by taking

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

with independent errors with $\mathsf{E}[\epsilon(x)] = 0$ and $\mathsf{var}[\epsilon(x)] = \sigma^2$.

In this case $\mu(x)$ represents the median concentration at time x.

Model (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 $[Vk_e/k_a, k_e, k_a]$.

Motivating Example: PK of Theophylline

Recall 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.

Motivating Example: PK of Theophylline

Model (1) may be written in the alternative form

$$\begin{aligned} \mu(x) &= \frac{Dk_a}{V(k_a - k_e)} \left[\exp(-k_e x) - \exp(-k_a x)\right] \\ &= \exp(\beta_0 + \beta_1 x) \left\{1 - \exp[-(k_a - k_e)x]\right\}, \end{aligned} \tag{3}$$

where

- $\beta_0 = \log[Dk_a/V(k_a - k_e)]$ and
- $\beta_1 = -k_e$.

Motivating Example: PK of Theophylline

As an alternative to the compartmental model, (1), we will also consider the fractional polynomial model (Nelder, 1966),

$$\mu(x) = \exp(\beta_0 + \beta_1 x + \beta_2/x). \tag{4}$$

Comparison with (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 (4) falls within the GLM class, as we see shortly.

Motivating Example: PK of Theophylline

In a pharmacokinetic study, 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 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.

Motivating Example: PK of Theophylline

With respect to the PK model (1) the derived parameters of interest, in terms of $[V, k_a, k_e]$ are:

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

where AUC is the area under the curve between 0 and ∞.

Motivating Example: PK of Theophylline

With respect to the GLM model (4) the derived parameters of interest, in terms of $[\beta_0, \beta_1, \beta_2]$ are:

$$\begin{aligned} x_{1/2} &= -\frac{\log 2}{\beta_1} \\ x_{\max} &= \left(\frac{\beta_2}{\beta_1}\right)^{1/2} \\ \mu(x_{\max}) &= D \exp\left[\beta_0 - 2(\beta_1\beta_2)^{1/2}\right] \\ Cl &= \frac{\sqrt{\beta_1/\beta_2}}{2\exp(\beta_0) K_1[2(\beta_1\beta_2)^{1/2}]}, \end{aligned} \tag{5}$$

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.

Generalized Linear Models

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

$$p(y_i \mid \theta_i, \alpha) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\alpha)} + c(y_i, \alpha)\right), \tag{6}$$

  for functions $a(\cdot)$, $b(\cdot)$, $c(\cdot,\cdot)$ and where θ_i and α are scalars.
- 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
  - $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^{\mathsf{T}}$ is a $(k+1) \times 1$ vector of regression parameters.

It is straightforward to show that

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

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

Further, it is assumed that $\mathsf{cov}(Y_i, Y_j \mid \theta_i, \theta_j, \alpha) = 0$, for $i \neq j$.

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) 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 $\theta_i = \boldsymbol{x}_i\beta$ and provide some advantages, including simplifications in computation.

GLMs are very useful pedagogically since they separate the deterministic and random 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".

Distribution	$\mathsf{N}(\mu, \sigma^2)$	$\mathsf{Poisson}(\mu)$	$\mathsf{Bern}(\mu)$	$\mathsf{Ga}(1/\alpha, 1/[\mu\alpha])$
Mean $\mathsf{E}[Y \mid \theta]$	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	$\frac{\log(y/\alpha)}{\alpha} - \log y + \log\Gamma(\alpha)$

Table 2: Characteristics of some common GLMs. The notation is as in equation (6). The canonical parameter is θ, the mean is $\mathsf{E}[Y] = \mu$ and the variance is $\mathrm{var}(Y) = \alpha V(\mu)$.

Example: Pharmacokinetics of Theophylline

Model (3) is an example of a GLM with a log link:

$$\log \mu(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{7}$$

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

Turning to the stochastic component, the error terms often display a constant coefficient of variation.

With this in mind we may combine (7) with a gamma distribution via:

$$Y(\boldsymbol{x}) \mid \boldsymbol{\beta}, \alpha \sim_{ind} \text{Ga}\left(\alpha^{-1}, [\mu(\boldsymbol{x})\alpha]^{-1}\right), \tag{8}$$

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.

Example: Pharmacokinetics of Theophylline

Note that for the gamma distribution the reciprocal transform is the canonical link but this option does not constrain the mean function to be positive, which is surprising.

In the pharmacokinetic context the reciprocal link also results in a concentration-time curve that is not integrable between 0 and ∞, so that the clearance parameter is undefined.

For these reasons, the log link is preferable.

Example: Lung Cancer and Radon

A starting model is

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

The exponential family representation[1] is easier to see if we examine the log probability distribution:

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

with

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

to give a GLM with a canonical log link.

[1] $\log p(y_i \mid \theta_i, \alpha) = \frac{y_i \theta_i - b(\theta_i)}{a(\alpha)} + c(y_i, \alpha)$

Example: Lung Cancer and Radon

The Poisson 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, for example:

- the negative binomial model, or
- the quasi-likelihood approach.

With unknown scale parameter (which we label as ϕ) the negative binomial is not a GLM.

We consider the case of known ϕ (which will rarely be of interest in a practical setting).

The negative binomial model

The probability distribution is

$$\Pr(Y = y) = \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\mu}{\mu+\phi}\right)^{y} \left(\frac{\phi}{\mu+\phi}\right)^{\phi},$$

for $\mu, \phi > 0$ and $y = 0, 1, 2, \ldots$.

One derivation is as $Y|\mu, \delta \sim \text{Poisson}(\mu\delta)$, $\delta|\phi \sim \text{Ga}(\phi, \phi)$.

We label the scale parameter of the negative binomial model as ϕ.

We now show that with ϕ known, the negative binomial is a member of the exponential family.

The negative binomial model

We reparameterize in terms of $p = \phi/(\mu + \phi)$:

$$\begin{aligned} \Pr(Y = y \mid p) &= \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}(1-p)^y p^\phi \\ &= \exp\left[y \log(1-p) + \phi \log p \right. \\ &+ \left. \log\Gamma(y+\phi) - \log\Gamma(y+1) - \log\Gamma(\phi)\right]. \end{aligned}$$

which is of the form (6) with

$$\begin{aligned} \theta &= \log(1-p) \\ b(\theta) &= \phi \log(1 - \mathrm{e}^\theta), \\ c(y) &= \log\Gamma(y+\phi) - \log\Gamma(y+1) - \log\Gamma(\phi) \\ \alpha &= 1 \\ a(\alpha) &= 1 \end{aligned}$$

Note that ϕ appears in $b(\theta)$ and $c(y)$ but it is a constant (the same occurs in the binomial distribution, where n appears).

The negative binomial model

The mean and variance are:

$$\begin{aligned} \mathsf{E}[Y \mid \mu] &= \mu = b'(\theta) = -\frac{\phi \mathrm{e}^{\theta}}{1 - \mathrm{e}^{\theta}}, \\ \mathrm{var}(Y \mid \mu) &= \phi \times b''(\theta) = -\frac{\phi \mathrm{e}^{\theta}}{1 - \mathrm{e}^{\theta}} - \frac{\phi \mathrm{e}^{2\theta}}{(1 - \mathrm{e}^{\theta})^2} = \mu + \mu^2/\phi. \end{aligned}$$

The canonical link is

$$\theta = \log\left(\frac{\mu}{\mu + \phi}\right) = \boldsymbol{x}\boldsymbol{\beta}.$$

A more natural link from a modeling/practical perspective is

$$\log \mu = \boldsymbol{x}\boldsymbol{\beta}.$$

Parameter Interpretation

Interpretation of the regression parameters in a GLM is link function specific.

The linear link was discussed and the log link was considered repeatedly, in the context of the lung cancer and radon data.

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.

Parameter Interpretation

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 – so more specific to the model form assumed.

For example, the nonlinear concentration-time curve (1) was defined in terms of the volume of distribution V and the absorption and elimination rates k_a and k_e.

Likelihood Inference for GLMs: Estimation

We now derive the score vector and information matrix.

For an independent sample from the exponential family, (6),

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

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

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{d\ell_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{Y_i - b'(\theta_i)}{\alpha} \frac{1}{V_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}, \end{aligned} \tag{9}$$

because ,

$$\frac{d^2 b}{d\theta_i^2} = \frac{d\mu_i}{d\theta_i} = V_i, \qquad i = 1, \ldots, n.$$

Since, $\mathrm{E}[Y_i \mid \boldsymbol{\beta}] = \mu_i$ and $\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V_i$,

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^{\mathsf{T}} \frac{Y_i - \mathrm{E}[Y_i \mid \mu_i]}{\mathrm{var}(Y_i \mid \mu_i)} \\ &= \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}) \right] / \alpha, \end{aligned} \tag{10}$$

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 i-th diagonal element V_i.

Consequently, an estimator $\widehat{\boldsymbol{\beta}}_n$ defined through $\boldsymbol{S}(\widehat{\boldsymbol{\beta}}_n) = \mathbf{0}$ will be consistent with respect to the assumed mean function, since the estimating function is unbiased.

Likelihood Inference for GLMs: Estimation

For canonical links, for which $\theta_i = \boldsymbol{x}_i\beta$,

$$\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{d\ell_i}{d\theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{1}{\alpha} \sum_{i=1} \boldsymbol{x}_i^{\mathsf{T}} \left[Y_i - \mu_i(\boldsymbol{\beta}) \right]$$

so that at the MLE, $\widehat{\boldsymbol{\beta}}$, the sufficient statistics,

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

are recovered.

From Result 2.1, the MLE has asymptotic distribution

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

where the expected information is

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

In practice we use,

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

where $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{D}}$ are evaluated at $\hat{\beta}_n$.

The variance of the estimator is

$$\widehat{\text{var}}(\hat{\beta}) = \alpha \left(\hat{\boldsymbol{D}}^{\mathsf{T}} \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{D}} \right)^{-1} \tag{11}$$

and is appropriately estimated (i.e, consistent) if the second moment is correctly specified, i.e., if,

$$\text{var}(Y_i \mid \beta) = \alpha V_i.$$

The expected 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, (9), may be written, for parameter j, $j = 0, 1, \ldots, k$, as

$$\begin{aligned} S_j(\beta) &= \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\alpha V_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\alpha V_i} \frac{d\mu_i}{d\eta_i} x_{ij}. \end{aligned} \tag{12}$$

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

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

so that

$$\boldsymbol{I}_{jj'}(\boldsymbol{\beta}) = -\mathsf{E}\left[\frac{\partial^2 l}{\partial\beta_j \partial\beta_{j'}}\right] = \sum_{i=1}^{n} \frac{x_{ij}x_{ij'}}{\alpha V_i}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.$$

The (expected) information matrix, therefore, takes the form

$$\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{x}, \tag{13}$$

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

$$w_i = \frac{(d\mu_i/d\eta_i)^2}{\alpha V_i}, \quad i = 1, \ldots, n.$$

Likelihood Inference for GLMs: Estimation

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

$$\widehat{\alpha} = \frac{1}{n-k-1} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{14}$$

where $\widehat{\mu}_i = \widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$.

This is 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 (10), i.e.,

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}) \right] / \alpha,$$

is appealing since it depends on only the first two moments.

Implications:

- The linear form means that $\mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})] = \boldsymbol{0}$ since $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{\mu}(\boldsymbol{\beta})$, and therefore we have consistency of $\widehat{\boldsymbol{\beta}}_n$ in whatever model we specify (not uniased, unless a linear model).
- The variance of the estimator is correct so long as $\mathrm{var}(\boldsymbol{Y})$ is correctly specified, i.e., correct specification of the mean-variance relationship – accurate asymptotic CIs in this case.
- These properties do not depend on the distribution of the data.

If the score is of the form (6), i.e., 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 also share consistency of estimation (so long as regularity conditions are satisfied).

Likelihood Inference for GLMs: 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 algorithm (a generic method for root-finding), known as Fisher scoring may be used to find the MLEs.

Consider the Newton-Raphson method: Let $\boldsymbol{S}(\beta)$ represent a $p \times 1$ vector of functions that are themselves functions of a $p \times 1$ vector β.

We wish to find β such that $\boldsymbol{S}(\beta) = \boldsymbol{0}$.

A first-order Taylor series expansion about initial giess $\beta^{(0)}$ gives:

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

Setting the left-hand side to zero yields

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

The Newton-Raphson method iterates the step:

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

for $t = 0, 1, 2, \ldots$

The Fisher scoring algorithm is the Newton-Raphson method applied to the score equation, but with the observed information, $\boldsymbol{S}'(\beta)$, replaced by the expected information $\mathrm{E}[\boldsymbol{S}'(\beta)] = -\boldsymbol{I}(\beta)$, to give

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

so that a new estimate is calculated based on the score and information evaluated at the previous estimate.

Recall that for a GLM, $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{x}^\top \boldsymbol{W}(\boldsymbol{\beta})\boldsymbol{x}$.

Using this form, and (12), i.e.,

$$S_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\alpha V_i} \frac{d\mu_i}{d\eta_i} x_{ij},$$

we write

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{x}^\top \boldsymbol{W}^{(t)} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W}^{(t)} \left[\boldsymbol{x}\boldsymbol{\beta}^{(t)} + (\boldsymbol{W}^{(t)})^{-1} \boldsymbol{u}^{(t)} \right] \\ &= (\boldsymbol{x}^\top \boldsymbol{W}^{(t)} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)} \end{aligned} \tag{15}$$

where $\boldsymbol{u}^{(t)}$ and $\boldsymbol{z}^{(t)}$ are $n \times 1$ vectors with i-th 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

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

The Fisher scoring updates (15) therefore have the form of a WLS:

$$(\boldsymbol{z}^{(t)} - \boldsymbol{x}\beta)^{\top} \boldsymbol{W}^{(t)} (\boldsymbol{z}^{(t)} - \boldsymbol{x}\beta) \tag{16}$$

with “working”, or “adjusted” response $\boldsymbol{z}$.

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.

Example: Pharmacokinetics of Theophylline

Fitting the gamma model (8) with mean function (7), gives MLEs for $[\beta_0, \beta_1, \beta_2]$ of

$$[2.42, -0.0959, -0.246].$$

The fitted curve is shown in Figure 11.

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 (10).

Example: Pharmacokinetics of Theophylline

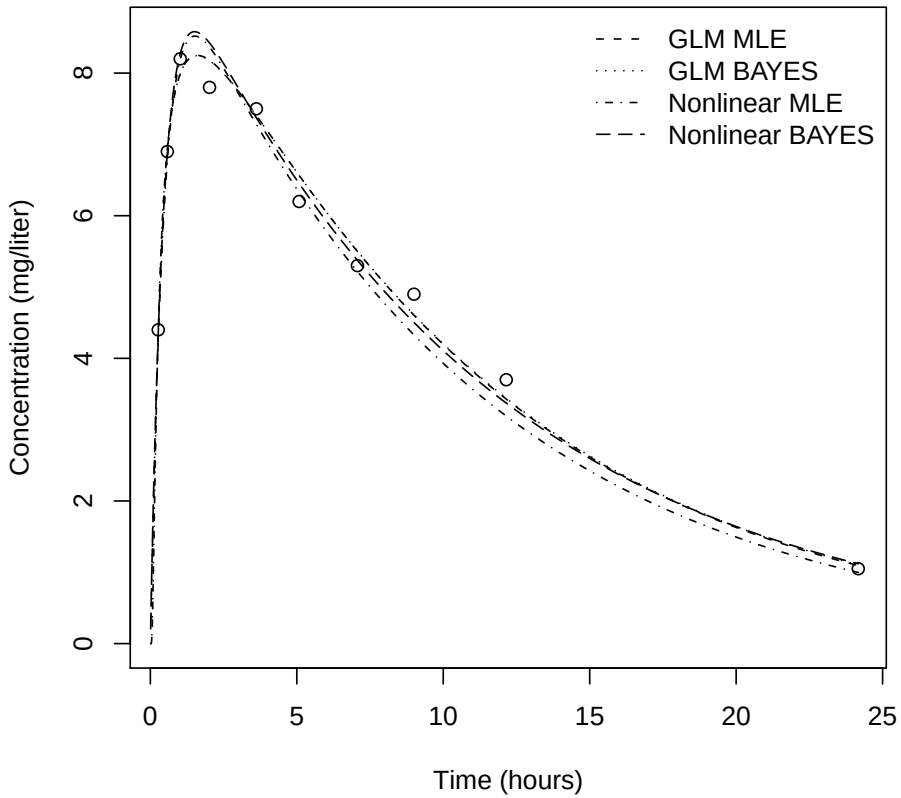


Figure 2: Theophylline data, along with fitted curves under various models and inferential approaches. Four curves are included, corresponding to MLE and Bayes analyses of GLM and nonlinear models. The two nonlinear curves are indistinguishable.

Example: Pharmacokinetics of Theophylline

The top row of Table 3 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

$$Cl = \frac{\sqrt{\beta_1/\beta_2}}{2\exp(\beta_0)K_1[2(\beta_1\beta_2)^{1/2}]},$$

using the delta method is more complex.

Working with $\theta = \log Cl$ and with $\boldsymbol{V} = \text{var}(\widehat{\boldsymbol{\beta}})$,

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

Example: Pharmacokinetics of Theophylline

The required derivatives are not friendly, but computable,

$$
\begin{aligned}
D_0 &= \frac{\partial \theta}{\partial \beta_0} = 1 \\
D_1 &= \frac{\partial \theta}{\partial \beta_1} = \frac{1}{\beta_1} + \sqrt{\frac{\beta_2}{\beta_1}} \frac{K_0\left(2\sqrt{\beta_1 \beta_2}\right)}{K_1\left(2\sqrt{\beta_1 \beta_2}\right)} \\
D_2 &= \frac{\partial \theta}{\partial \beta_2} = \sqrt{\frac{\beta_1}{\beta_2}} \frac{K_0\left(2\sqrt{\beta_1 \beta_2}\right)}{K_1\left(2\sqrt{\beta_1 \beta_2}\right)}.
\end{aligned}
$$

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: Pharmacokinetics of Theophylline

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 Sand	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]

Table 3: Point and 90% interval estimates for the Theophylline data of Table 1, 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.

Summary/Questions?

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},
- Y_{i1} represents the number of counts of a rare disease in an exposed group of size N_{i1}, all in area i,

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

In practice: we only observe the sum of the disease counts, $Y_i = Y_{i0} + Y_{i1}$, along with N_{i0} and N_{i1}.

Example: Poisson Data with a Linear Link

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 $\boldsymbol{\beta}^\star = [\beta_0^\star, \beta_1^\star]$.

Writing $x_i = N_{1i}/N_i$ as the proportion of exposed individuals, the distribution of the total disease counts is

$$Y_i \mid \boldsymbol{\beta}^\star \sim_{ind} \text{Poisson}\left\{N_i[(1 - x_i)\beta_0^\star + x_i\beta_1^\star]\right\}, \tag{17}$$

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, $\beta_1^\star/\beta_0^\star$.

Example: Poisson Data with a Linear Link

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.

Data:

- 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 (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.

Example: Poisson Data with a Linear Link

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

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

for area i, exposure group j and stratum k.

Example: Poisson Data with a Linear Link

This model contains far too many parameters, p_{ijk}, to estimate, and so we simplify by making the proportionality assumption:

$$p_{ijk} = \beta_j \times p_k, \tag{18}$$

across all areas i.

Consequently:

- p_k is the (marginal) 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 affect is the same across areas and across age stratum.

The age-specific probabilities p_k, are assumed known (being based on rates from a larger geographic region, for example).

Example: Poisson Data with a Linear Link

The numbers of exposed individuals in each age stratum are unknown and we therefore make the important additional assumption that the proportion of exposed and unexposed individuals is constant across age stratum, i.e., $N_{i0k} = N_{ik}(1 - x_i)$ and $N_{i1k} = N_{ik}x_i$.

This assumption is made since N_{i0k} and N_{i1k} are unavailable and is distinct from assumption (18) which concerns the underlying disease model.

Summing across stratum and exposure groups, gives

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

Example: Poisson Data with a Linear Link

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

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\left\{E_i[(1 - x_i)\beta_0 + x_i\beta_1]\right\}. \tag{19}$$

Example: Poisson Data with a Linear Link

Under this model,

$$\mathsf{E}\left[\frac{Y_i}{E_i}\right] = \beta_0 + (\beta_1 - \beta_0)x_i, \tag{20}$$

illustrating that the mean model for the standardized morbidity ratio (SMR), Y_i/E_i, is linear in x.

Figure 3 plots the standardized morbidity ratios (SMRs) Y_i/E_i versus x_i, with a linear fit added, and we see evidence of increasing SMR with increasing x.

Example: Poisson Data with a Linear Link

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 (20) is superimposed on Figure 3.

The estimate of the relative risk β_1/β_0 is 22.7 with asymptotic standard error 3.39.

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

Example: Poisson Data with a Linear Link

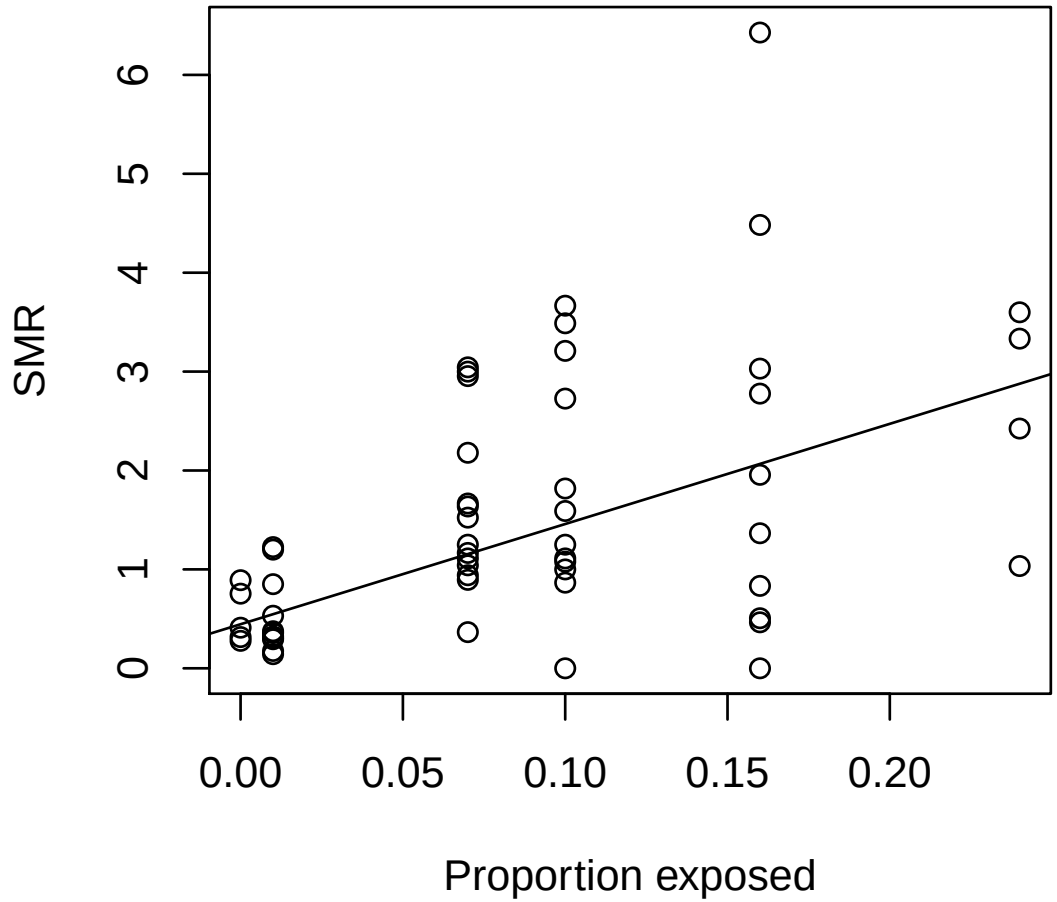


Figure 3: Plot of standardized morbidity ratio versus proportion exposed, for lip cancer incidence in 56 counties of Scotland. The linear model fit is indicated.

Hypothesis Testing

Suppose that $\dim(\boldsymbol{\beta}) = k + 1$ and let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ be a partition with

$$\boldsymbol{\beta}_1 = [\beta_0, \ldots, \beta_q] \qquad \text{and} \qquad \boldsymbol{\beta}_2 = [\beta_{q+1}, \ldots, \beta_k],$$

with $0 \leqslant q < k$.

Interest focuses on testing whether the subset of $k - q$ parameters are equal to zero via a test of the null

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_1 \text{ unrestricted}, \ \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20} \\ H_1 &: \boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2] \neq [\boldsymbol{\beta}_1, \boldsymbol{\beta}_{20}]. \end{aligned} \tag{21}$$

As outlined previously, 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, which formally accounts for estimation of the scale parameter.

Hypothesis Testing

The log likelihood is

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

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,

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

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

$$H_0 : \text{ Reduced model}$$

versus

$$H_1 : \text{ Saturated model},$$

is

$$2\left[\ell(\widetilde{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}})\right] = \frac{2}{\alpha}\sum_{i=1}^{n}\left[Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\right] = \frac{D}{\alpha}, \qquad (22)$$

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 modeled using binomial or Poisson distributions 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,

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}, \tag{23}$$

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, (21), i.e.,

$$\begin{aligned} H_0 &: \beta_1 \text{ unrestricted, } \beta_2 = \beta_{20} \\ H_1 &: \beta = [\beta_1, \beta_2] \neq [\beta_1, \beta_{20}]. \end{aligned}$$

We describe an attractive additivity property of the log 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:

$$\begin{aligned} 2\left[\ell(\widehat{\boldsymbol{\beta}}^{(1)}) - \ell(\widehat{\boldsymbol{\beta}}^{(0)})\right] &= 2\left\{\ell(\widehat{\boldsymbol{\beta}}^{(s)}) - \ell(\widehat{\boldsymbol{\beta}}^{(0)}) - [\ell(\widehat{\boldsymbol{\beta}}^{(s)}) - \ell(\widehat{\boldsymbol{\beta}}^{(1)})]\right\} \\ &= \frac{1}{\alpha}\left(D_0 - D_1\right) \rightarrow_d \chi^2_{q_1 - q_0}, \end{aligned}$$

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 the additivity property.

For a GLM, in contrast to the linear model, 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

Consider the model $\boldsymbol{Y} \mid \boldsymbol{\beta} \sim \mathrm{N}_n(\boldsymbol{x}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n)$ with log-likelihood

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

with α in the GLM formulation being replaced by σ^2.

Again, let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ where

$$\boldsymbol{\beta}_1 = [\beta_0, \ldots, \beta_q] \quad \text{and} \quad \boldsymbol{\beta}_2 = [\beta_{q+1}, \ldots, \beta_k],$$

and consider the null

$$H_0 : \boldsymbol{\beta}_1 \text{ unrestricted}, \ \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}.$$

Under this null, from (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 i-th case, based on the MLEs under the reduced model, H_0.

Example: Normal Linear Model

In this case, the asymptotic distribution is exact since

$$\frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}^{(0)})^2}{\sigma^2} \sim \chi^2_{n-q+1}. \tag{24}$$

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

Example: Normal Linear Model

In terms of comparing the nested hypotheses

$$H_0 : \beta_1 \text{ unrestricted}, \ \beta_2 = \beta_{20}$$

and

$$H_1 : \beta = [\beta_1, \beta_2] \neq [\beta_1, \beta_{20}]$$

the likelihood ratio statistic is

$$\begin{aligned} \frac{1}{\sigma^2}(D_0 - D_1) &= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\widehat{\beta}^{(0)})^2 - \sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\widehat{\beta}^{(1)})^2\right] \\ &= \frac{\text{RSS}_0 - \text{RSS}_1}{\sigma^2} = \frac{\text{FSS}_{01}}{\sigma^2} \end{aligned} \tag{25}$$

where $\boldsymbol{x}\widehat{\beta}^{(j)}$ 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 under H_1.

Example: Normal Linear Model

In practice if n is large, we may use (25) with σ^2 replaced by a consistent estimator $\widehat{\sigma}^2$.

Alternatively, the ratios of scaled versions of (25) and (24) may be taken to produce an F-statistic by which statistical significance may be assessed.

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

$$H_0 : \beta_0 \text{ unrestricted, } \beta_1 = 0 \text{ versus } H_1 : [\beta_0, \beta_1] \neq [\beta_0, 0],$$

in the model with mean $\mu_i = E_i \exp(\beta_0 + \beta_1 x_i)$.

Example: Lung Cancer and Radon

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 alternative and null 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 are -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%.

Quasi-Likelihood Inference for GLMs

GLMs that do not contain a scale parameter are particularly vulnerable to mean-variance model misspecification, specifically the presence of over-dispersion 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

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

with $\mathsf{cov}(Y_i, Y_j \mid \boldsymbol{\beta}) = 0$.

Quasi-Likelihood Inference for GLMs

From these specifications the quasi-score is defined as

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

and coincides with the score function of a GLM that we saw earlier, (10).

Hence, the maximum quasi-likelihood estimator $\widehat{\boldsymbol{\beta}}$ is identical to the MLE, due to the multiplicative form of the variance model.

Quasi-Likelihood Inference for GLMs

Estimation of α may be carried out using the method of moments form (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

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

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, as before define the quasi log likelihood as,

$$\ell(\boldsymbol{\beta}, \alpha) = \int_y^{\mu} \frac{y-t}{\alpha V(t)} dt.$$

Quasi-Likelihood Inference for GLMs

If $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}, \alpha = 1)$ represents the likelihood upon which the quasi-likelihood is based (for example, a Poisson or binomial likelihood),

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}, \alpha) \times \alpha \tag{26}$$

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

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

or equivalently

$$2\left[\ell(\widehat{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}}^{(0)})\right] \rightarrow_d \widehat{\alpha} \times \chi^2_{k-q}. \tag{27}$$

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

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}] &= E_i \exp(\beta_0 + \beta_1 x_i), && (28) \\ \mathrm{var}(Y_i \mid \boldsymbol{\beta}) &= \alpha \times \mathrm{E}[Y_i \mid \boldsymbol{\beta}] && (29) \end{aligned}$$

yields identical point estimates for $\boldsymbol{\beta}$ to the Poisson model, with scale parameter estimate $\widehat{\alpha} = 2.81$, obtained via (14).

Therefore, with respect to $H_0 : \beta_0$ unrestricted, $\beta_1 = 0$, the quasi log likelihood ratio statistic of 46.2/2.81 = 16.5, so that the significance level is vastly reduced, though still strongly suggestive of a non-zero slope.

Summary/Questions

Sandwich Estimation for GLMs

Sandwich Estimation for GLMs

The asymptotic variance-covariance for $\widehat{\beta}$, which is given by (11), is appropriate only if the first two moments are correctly specified.

In general, the variance of the estimator follows the sandwich form

$$\text{var}(\widehat{\beta}) = \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{\mathsf{T}})^{-1}$$

where

$$\begin{aligned} \boldsymbol{A} &= \mathsf{E}\left[\frac{\partial \boldsymbol{G}}{\partial \beta}\right] \\ &= \boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{D}, \end{aligned} \tag{30}$$

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

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

where $\boldsymbol{G}(\beta) = \boldsymbol{S}(\beta)/n$.

Sandwich Estimation for GLMs

Under the assumption of uncorrelated errors,

$$\widehat{\boldsymbol{B}} = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^{\mathsf{T}} \frac{\text{var}(Y_i)}{V_{ii}^2} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \tag{31}$$

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

$$\widehat{\sigma}_i^2 = (Y_i - \widehat{\mu}_i)^2, \tag{32}$$

which has finite-sample bias.

Combination of (31) and (32) provides a consistent estimator of the variance and therefore asymptotically correct confidence interval coverage (so long as independence of responses holds).

There can be great instability of the estimator if n is not large, however.

Bootstrap Methods for GLMs

Bootstrap methods 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 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, Section 7.2) discuss both resampling residuals and resampling cases in the context of GLMs.

Bayesian Inference for GLMs

We now consider Bayesian inference for the GLM.

The posterior is

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

Usually, prior independence between the regression coefficients $\boldsymbol{\beta}$ and the scale parameter α, is assumed, i.e.,

$$\pi(\boldsymbol{\beta}, \alpha) = \pi(\boldsymbol{\beta})\pi(\alpha).$$

Bayesian Inference for GLMs: 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 $\mathrm{N}(\boldsymbol{m}, \boldsymbol{V})$ for $\boldsymbol{\beta}$ is the obvious choice.

Furthermore, independent priors are frequently defined for each component.

As a limiting case ($\boldsymbol{V}^{-1} \to \mathbf{0}$), 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, but for non-canonical links, innocuous datasets may lead to impropriety, as the Poisson model with a linear link example considered below illustrates.

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 $\beta_0 > 0$, i.e.

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

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

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

where $\mu_i^{\star} = \mu_i^{\star}(\gamma)$.

We integrate out β_0 and consider the propriety of the marginal posterior $p(\gamma \mid \boldsymbol{y})$.

The marginal posterior for γ is

$$
\begin{aligned}
p(\gamma \mid \boldsymbol{y}) &= \int p(\beta_0, \gamma \mid \boldsymbol{y})\, d\beta_0 \\
&\propto \int \underbrace{L(\beta_0, \gamma)}_{\text{Poisson}}\, d\beta_0 \times \pi(\gamma) \\
&\propto \int \exp\left(-\beta_0 \sum_{i=1}^{n} \mu_i^{\star y_i}\right) \beta_0^{\sum_{i=1}^{n} y_i}\, d\beta_0 \prod_{i=1}^{n} \mu_i^{\star y_i} \times \pi(\gamma) \\
&\propto \prod_{i=1}^{n} \left(\frac{E_i[(1-x_i) + x_i e^{\gamma}]}{\sum_{i=1}^{n} E_i[(1-x_i) + x_i e^{\gamma}]}\right)^{y_i} \times \pi(\gamma) \qquad (33) \\
&= L(\gamma) \times \pi(\gamma), \qquad (34)
\end{aligned}
$$

where the last line follows from the previous on recognizing that the integrand is the kernel of a Ga $\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} \mu_i^{\star}\right)$ distribution.

Poisson Data with a Linear Link

The "likelihood", $L(\gamma)$ in (34), is of multinomial form with the total number of cases y_+ distributed amongst 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

$$\prod_{i=1}^{n} \left(\frac{E_i(1 - x_i)}{\sum_{i=1}^{n} E_i(1 - x_i)} \right)^{y_i} \tag{35}$$

as $\gamma \to -\infty$, showing that, in general, a proper prior is required (since the tail will be non-integrable).

Poisson Data with a Linear Link

The constant (35) is non-zero 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 (35) providing the limiting constant.

Figure 4 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 (35) is indicated with a dashed horizontal line.

The MLE $\widehat{\gamma} = \log(22.7)$ is indicated as a vertical dotted line.

Poisson Data with a Linear Link

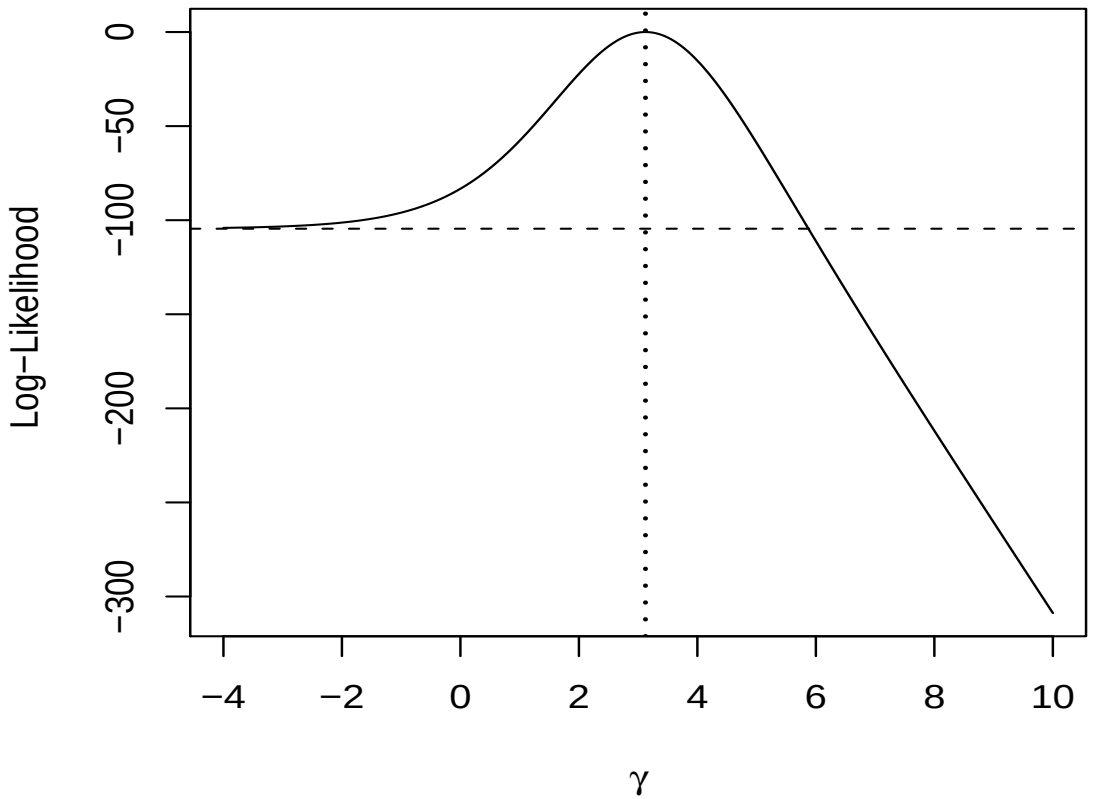


Figure 4: Log likelihood for the log relative risk parameter γ, for the Scottish lip cancer data. The dashed horizontal line is the constant to which the log likelihood tends to as $\gamma \to -\infty$.

Bayesian Computation for a GLM

Outside of the normal linear model with convenient priors, the required integrals for Bayesian analysis are not available for GLMs.

However, `INLA` is ideally suited to out the required computations. MCMC is obviously a candidate but INLA is much computationally efficient, and accurate in almost all cases – logistic regression models can be inaccurate when events are rare, see Fong *et al.* (2010).

As already described, there is asymptotic equivalence between the sampling distribution of the MLE and the posterior distribution (so long as the prior doesn't exclude the relevant part of the parameter space).

Hence, Bayes estimators for $\boldsymbol{\beta}$ are consistent due to the form of the likelihood, so long as the priors are non-zero in a neighborhood of the true values of $\boldsymbol{\beta}$.

Hypothesis Testing

A simple method for examining hypotheses involving a single parameter,

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0,$$

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 provide a more general tool for comparing hypotheses:

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

Great care is required in the specification of priors when model comparison is carried out using Bayes factors.

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, e.g., the negative binomial model which is derived via the introduction of gamma random effects into a Poisson model.

In general, normal random effects may be used, which allows flexibility to model different dependencies in the data – not conjugate, but computations for the negative binomial are not available in closed form, anyway.

Example: Lung Cancer and Radon

The Bayesian Poisson model was fitted previously using a Metropolis-Hastings implementation.

Here the use of the INLA method, 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 interpretable parameters, and convert back to obtain the implied priors on $\beta_0, \beta_1, \beta_2$.

In particular, we choose 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.

Example: Pharmacokinetics of Theophylline

For a generic parameter θ denote the prior by $\theta \sim \text{LogNormal}(\mu, \sigma)$.

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

$$\mu = \log(\theta_m), \quad \sigma = \frac{\log(\theta_u) - \mu}{1.645}. \tag{36}$$

Based on a literature search we assume prior 50% (95%) points of 8 (12), 1.5 (3), 9 (12), for $x_{1/2}$, $x_{\max}$, $\mu(x_{\max})$, respectively.

For the coefficient of variation the corresponding values are 0.05 (0.10).

Example: Pharmacokinetics of Theophylline

The third line of Table 3 summarizes these priors.

To examine the posterior we use a rejection algorithm.

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 β is:

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

Table 3 summarizes inference for the parameters of interest, via medians and 90% interval estimates.

Point and interval estimates show close correspondence with the frequentist summaries.

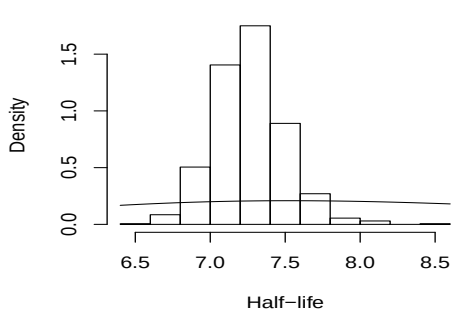
Example: Pharmacokinetics of Theophylline

Figure 5 gives the marginal posterior distributions for the half-life, the time to maximum concentration, the maximum concentration and the coefficient of variation (expressed as a percentage).

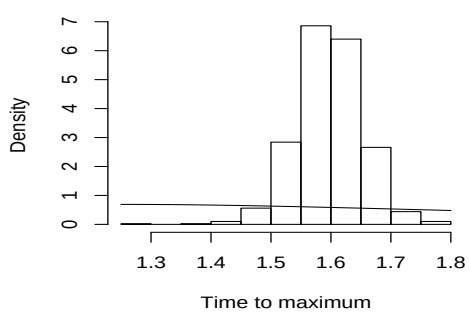
The prior distributions are also indicated as solid curves.

We see some skewness in each of the posteriors, which is common for nonlinear parameters unless the data are abundant.

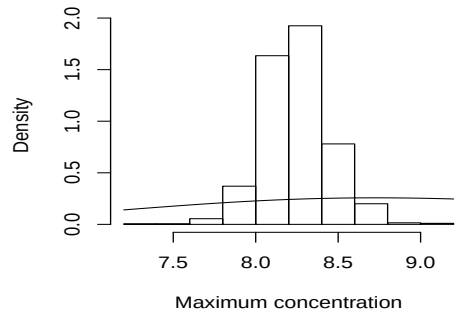
Example: Pharmacokinetics of Theophylline



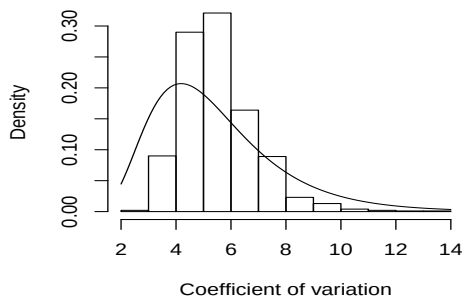
(a)



(b)



(c)



(d)

Figure 5: Histogram representations of posterior distributions from the GLM for the Theophylline data, for: (a) half-life (b) time to maximum, (c) maximum concentration, (d) coefficient of variation, with priors superimposed as solid lines.

Example: Pharmacokinetics of Theophylline

Inference for the clearance parameter is relatively straightforward, since one simply substitutes samples for $\boldsymbol{\beta}$ into (5), i.e.,

$$Cl = \frac{\sqrt{\beta_1/\beta_2}}{2\exp(\beta_0)K_1[2(\beta_1\beta_2)^{1/2}]}.$$

Figure 6 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 parameter is often found to be well-behaved (when there are data over the whole concentration-time curve), since it is a function of the area under the curve, which is reliably estimated.

Example: Pharmacokinetics of Theophylline

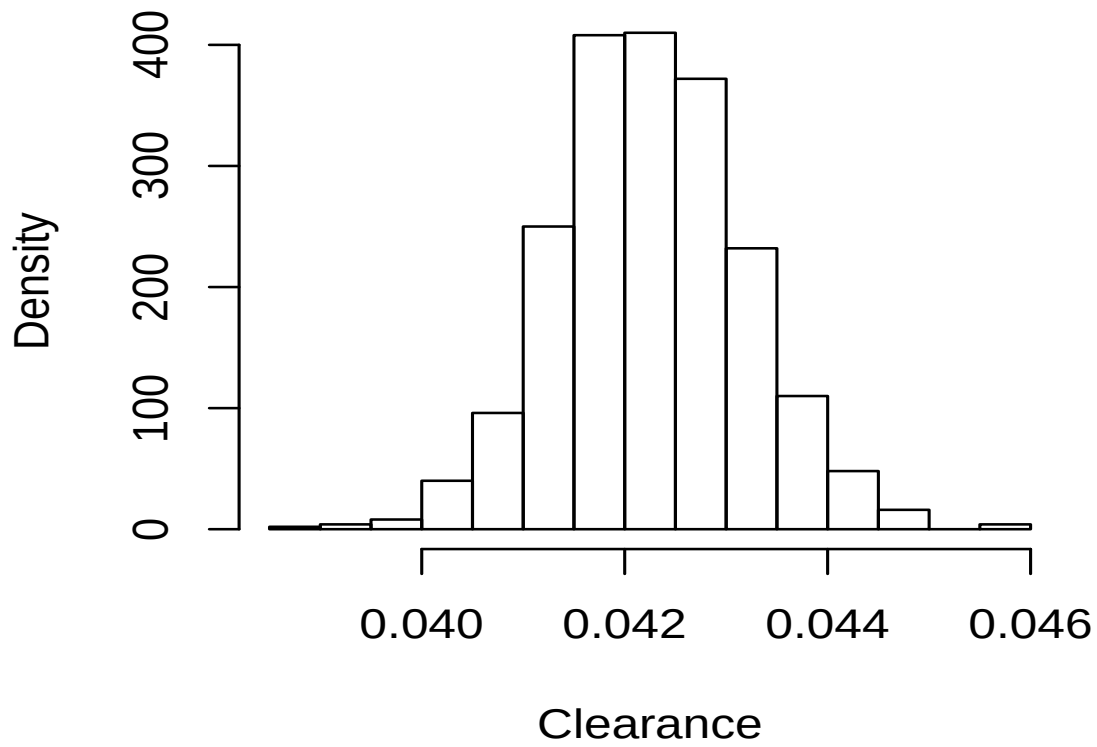


Figure 6: Posterior distribution of the clearance parameter from the GLM fitted to the Theophylline data.

Summmary/Comments

Assessment of Assumptions

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 (because $\text{var}(Y_i)$ is a function of μ_i), 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 reproduces Pearson's χ^2 statistic:

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

as previously introduced, (23).

For Pearson residuals, under the model, $\mathrm{E}[\widehat{\sigma}_i e_i^\star] = 0$ 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 $D = \sum_{i=1}^n e_i^{\star 2}$.

Assessment of Assumptions for GLMs

As an example, for a Poisson likelihood, the deviance residuals are

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

Notice that there are two measures of distance between y_i and $\widehat{\mu}_i$, here.

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., the same 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.

Assessment of Assumptions for GLMs

As with linear model residuals, 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.

Assessment of Assumptions for GLMs

Consideration of the updates (15) in the IRLS fitting algorithm, reveals that for a GLM we may define a hat matrix as

$$\boldsymbol{h} = \boldsymbol{w}^{1/2}\boldsymbol{x}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{w}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{w}^{1/2},$$

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 $i = 1, \ldots, n$.

Example: Pharmacokinetics of Theophylline

We fit the gamma GLM,

$$Y_i \mid \boldsymbol{\beta}, \alpha \sim_{ind} \text{Ga}\left[\alpha^{-1}, (\alpha\mu_i)^{-1}\right],$$

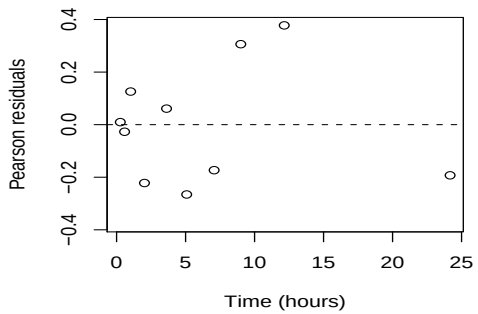
using MLE, and calculate Pearson residuals,

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

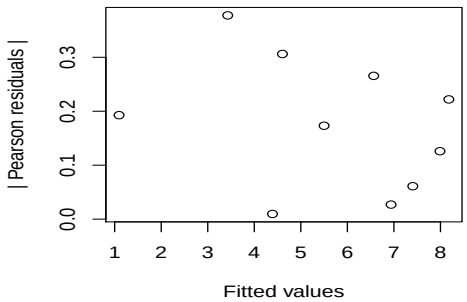
In Figure 7(a) these residuals are plotted versus time x_i and show no obvious systematic pattern, though interpretation is difficult, given the small number of data points and the spacing of these points over time.

Figure 7(b) plots $|e_i^\star|$ against fitted values to attempt to discover any unmodeled mean-variance relationship and again no strong signal is apparent.

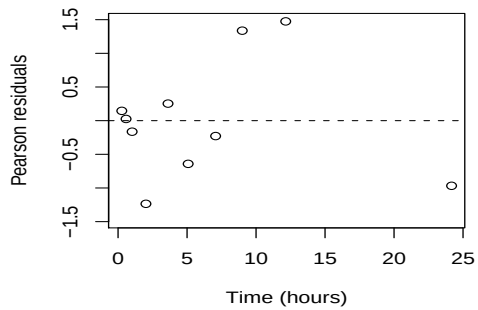
Example: Pharmacokinetics of Theophylline



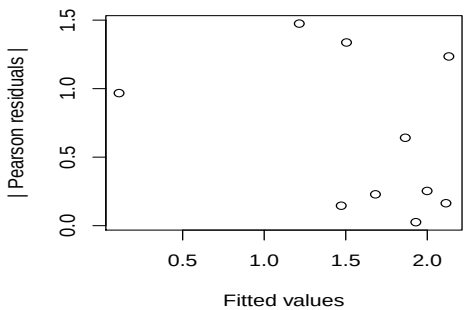
(a)



(b)



(c)



(d)

Figure 7: Pearson residual plots for Theophylline: (a) residuals vs time for the GLM, (b) absolute values of residuals vs fitted values for the GLM, (c) residuals vs time for the nonlinear compartmental model, (d) absolute values of residuals vs fitted values for the nonlinear compartmental model.

Example: Lung Cancer and Radon

As we have seen, fitting the quasi-likelihood model given by the mean and variance specifications (28) and (29) yields $\widehat{\alpha} = 2.76$, illustrating a large amount of over-dispersion.

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 mean-variance relationship,

$$\text{var}(Y_i) = \mu_i(1 + \mu_i/\phi). \tag{37}$$

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 $\widehat{\phi} = 61.3$ with standard error 17.3.

Example: Lung Cancer and Radon

Figure 8 shows the fitted quadratic relationship (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 ϕ (which corresponds to a large departure from the linear mean-variance model).

Example: Lung Cancer and Radon

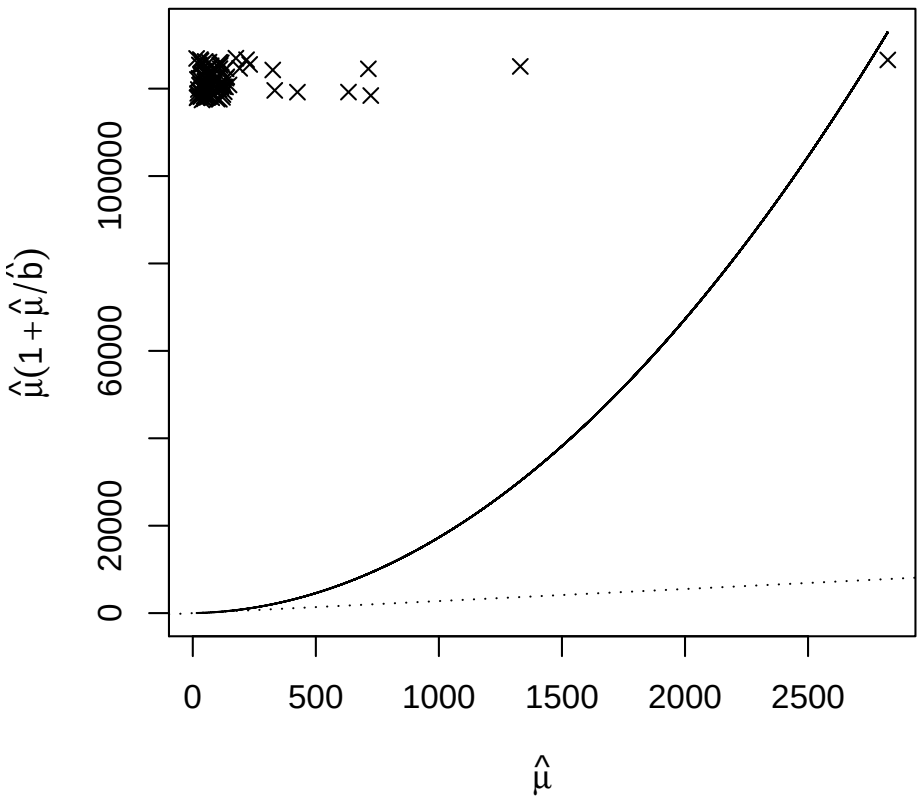


Figure 8: The solid line shows the fitted negative binomial variance function, $\widehat{\text{var}}(Y) = \hat{\mu}(1 + \hat{\mu}/\hat{\phi})$ plotted versus $\hat{\mu}$ for the lung cancer and radon data. The dotted line corresponds to the fitted quasi-likelihood model, $\widehat{\text{var}}(Y) = \hat{\kappa} \times \hat{\mu}$.

Example: Lung Cancer and Radon

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 $[\widehat{\beta}, \widehat{\phi}]$.

We form Pearson residuals,

$$e_i^\star = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i(1 + \widehat{\mu}_i/\widehat{\phi})}}.$$

Example: Lung Cancer and Radon

Procedure:

- Simulate multiple datasets from a negative binomial model.
- Fit a Poisson model (which provides identical fitted values as from a quasi-likelihood model).
- Form residuals

$$e^\star = \frac{y - \widehat{\mu}}{\sqrt{\widehat{\mu}}},$$

  i.e., residuals from a Poisson model.
- Plot $e^\star$ versus $\sqrt{\mu}$, to see if we can detect departures from linearity.

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.

Example: Lung Cancer and Radon

Figure 9 shows four representative plots and Figure 10 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.

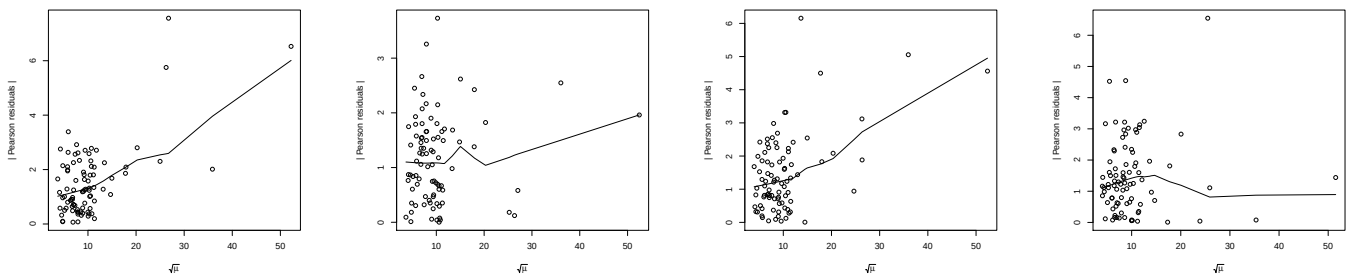


Figure 9: Absolute values of Pearson residuals versus $\sqrt{\mu}$ when the true mean-variance relationship is quadratic, but we analyze as if linear, for four simulated datasets with the same expected numbers and covariate values as in the lung cancer and radon data.

Example: Lung Cancer and Radon

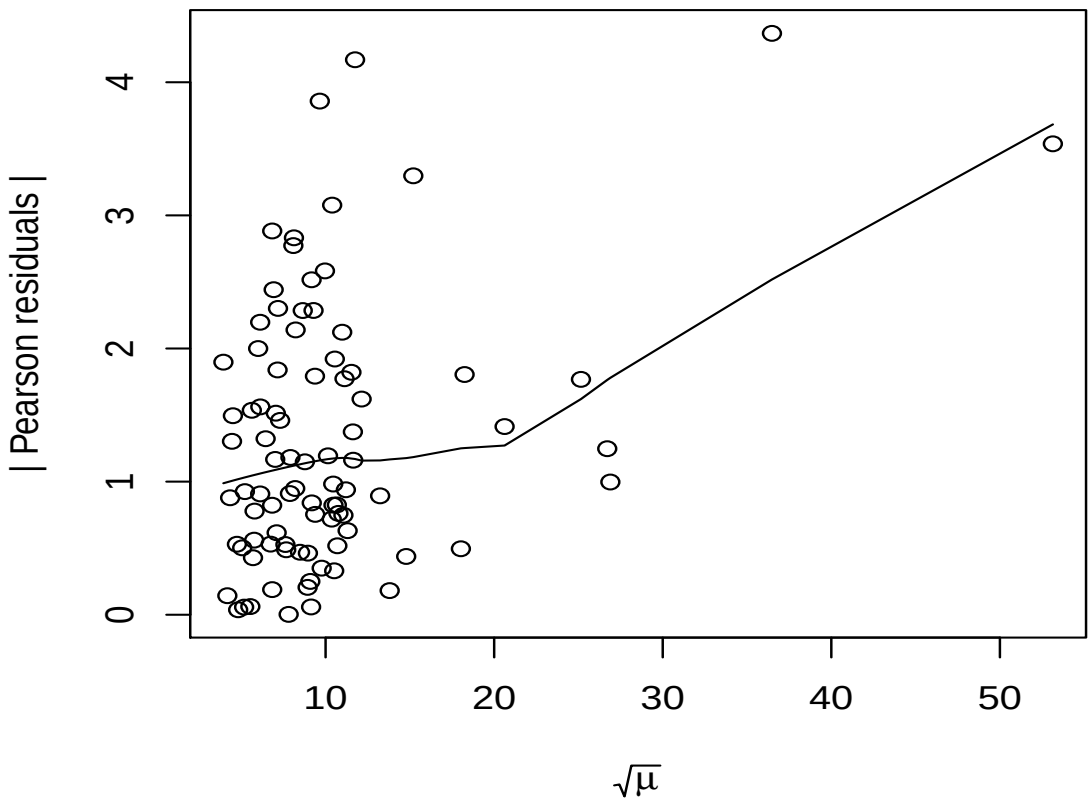


Figure 10: Absolute values of Pearson residuals versus $\sqrt{\mu}$ for the lung cancer and radon data.

Nonlinear Regression Models

We now consider models of the form

$$Y_i = \mu_i(\boldsymbol{\beta}) + \epsilon_i, \tag{38}$$

for $i = 1, \ldots, n$, where $\mu_i(\boldsymbol{\beta}) = \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$ is nonlinear in $\boldsymbol{\beta}$, which is assumed to be of dimension $k + 1$:

$$\mathsf{E}[\epsilon_i \mid \mu_i] = 0, \quad \mathsf{var}(\epsilon_i \mid \mu_i) = \sigma^2 f(\mu_i), \quad \mathsf{cov}(\epsilon_i, \epsilon_j \mid \mu_i) = 0.$$

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.

More generally we might assume that

$$\text{var}(\epsilon_i \mid \boldsymbol{\beta}, \boldsymbol{x}_i) = \sigma^2 g_1(\boldsymbol{\beta}, \boldsymbol{x}_i),$$

with

$$\text{cov}(\epsilon_i, \epsilon_j \mid \boldsymbol{\beta}, \boldsymbol{x}_i, \boldsymbol{x}_j) = g_2(\boldsymbol{\beta}, \boldsymbol{x}_i, \boldsymbol{x}_j).$$

When data are measured over time, serial correlation can be a particular problem.

We concentrate on the simpler uncorrelated second moment errr structure here.

Example: Michaelis-Menten Model

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

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

a nonlinear model, with $\alpha_0, \alpha_1 > 0$ and $\mu(0) = 0$.

We have

$$\frac{d\mu}{dx} = \frac{\alpha_0 \alpha_1}{(\alpha_1 + x)^{-2}} > 0,$$

so that the curve increases from 0.

Parameter interpretation:

- As $x \to \infty$, $\mu(x) \to \alpha_0$, so that α_0 is the asymptote.
- At α_1, $\mu(\alpha_1) = \alpha_0/2$, so that α_1 is the value of x at which the curve reaches half of its maximum value.

Example: Michaelis-Menten Model

An alternative to

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

is to write

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

where $z = 1/x$ and

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

which is a GLM with reciprocal link which is the canonical link with a gamma family.

The point of this is that with some ingenuity we can massage a preferred model into GLM form.

Identifiability

For many nonlinear models identifiability is an issue, by which we mean that the same curve may be obtained with different parameter values.

We have already seen one example of this for the nonlinear "flip=flop" model fitted to the Theophylline data – identifiability could be imposed through a substantive assumption such as $k_a > k_e > 0$..

As a second example, consider the popular sum-of-exponentials model

$$\mu(x) = \beta_0 \exp(-x\beta_1) + \beta_2 \exp(-x\beta_3), \tag{39}$$

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

$$[\beta_0, \beta_1, \beta_2, \beta_3] \qquad \text{and} \qquad [\beta_2, \beta_3, \beta_0, \beta_1]$$

and so we have non-identifiability.

For model (39) we may enforce (say) $\beta_3 > \beta_1 > 0$ and work with the set

$$\gamma = [\gamma_0, \gamma_1, \gamma_2, \gamma_3] = [\log \beta_0, \log(\beta_3 - \beta_1), \log \beta_2, \log \beta_1]$$

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.

This option is less appealing since it can lead to a discontinuity in the prior (not very elegant!).

Likelihood Inference for Nonlinear Models: Estimation

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

A popular choice is

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

for $i = 1, \ldots, n$, and with $r = 0, 1$ or 2 being common choices.

Note:

$$\frac{\mathsf{var}(Y_i)}{\mu_i^r} = \frac{\mathsf{E}[(Y_i - \mu_i)^2]}{\mu_i^r} = \sigma^2.$$

The corresponding log-likelihood function is

$$\ell(\boldsymbol{\beta}, \sigma) = -n \log \sigma - \frac{r}{2} \sum_{i=1}^{n} \log \mu_i(\boldsymbol{\beta}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})}. \quad (40)$$

Likelihood Inference for Nonlinear Models: Estimation

Differentiation with respect to β yields, with a little rearrangement, the score equation:

$$\begin{aligned} \boldsymbol{S}_1(\boldsymbol{\beta}, \sigma) &= \frac{\partial \ell}{\partial \boldsymbol{\beta}} \\ &= \frac{r}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{1}{\mu_i(\boldsymbol{\beta})} \left\{ \frac{[Y_i - \mu_i(\boldsymbol{\beta})]^2}{\mu_i^r(\boldsymbol{\beta})} - \sigma^2 \right\} + \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{[Y_i - \mu_i(\boldsymbol{\beta})]}{\mu_i(\boldsymbol{\beta})^r} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \end{aligned} \tag{41}$$

If $r = 0$ the first term of (41) disappears and we require the first moment only for consistency.

It is important to emphasize that if $r > 0$ we require the second moment to be correctly specified in order to ensure $\mathrm{E}[\boldsymbol{S}_1] = \mathbf{0}$, which is needed to produce a consistent estimator of β.

Likelihood Inference for Nonlinear Models: Estimation

Differentiation with respect σ yields the score equation:

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

$\mathrm{E}[S_2] = 0$ if the second moment is correctly specified, in which case consistency of σ results.

Likelihood Inference for Nonlinear Models: Estimation

In general, the MLEs $\widehat{\beta}$ are not available in closed form but numerical solutions are usually straightforward (via Gauss-Newton methods, or variants, for example) and are available in most statistical software.

The MLE for σ is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{[Y_i - \mu_i(\widehat{\beta})]^2}{\mu_i^r(\widehat{\beta})}, \tag{42}$$

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

$$\widetilde{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{[Y_i - \mu_i(\widehat{\beta})]^2}{\mu_i^r(\widehat{\beta})}. \tag{43}$$

For a nonlinear model, $\widetilde{\sigma}^2$ has finite sample bias, but is often preferred to (42) because of better small sample performance.

Likelihood Inference for Nonlinear Models: Estimation

Under the usual regularity conditions

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

where $\boldsymbol{\theta} = [\boldsymbol{\beta}, \sigma]$ and $\boldsymbol{I}(\boldsymbol{\theta})$ is Fisher's expected information.

In the case of $r = 0$ we obtain log-likelihood and score equations:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma) &= -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - \mu_i(\boldsymbol{\beta})]^2 \\ \boldsymbol{S}_1(\boldsymbol{\beta}, \sigma) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} [Y_i - \mu_i(\boldsymbol{\beta})] \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\ S_2(\boldsymbol{\beta}, \sigma) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} [Y_i - \mu_i(\boldsymbol{\beta})]^2 \end{aligned} \tag{44}$$

Likelihood Inference for Nonlinear Models: Estimation

And expected information:

$$
\begin{aligned}
\boldsymbol{I}_{11} &= -\mathrm{E}\left[\frac{\partial \boldsymbol{S}_1}{\partial \boldsymbol{\beta}}\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^{\mathsf{T}}\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right) = \frac{\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}}{\sigma^2} \\
\boldsymbol{I}_{12} &= -\mathrm{E}\left[\frac{\partial \boldsymbol{S}_1}{\partial \sigma}\right] = \boldsymbol{0} \\
\boldsymbol{I}_{21} &= -\mathrm{E}\left[\frac{\partial S_2}{\partial \boldsymbol{\beta}}\right] = \boldsymbol{0}^{\mathsf{T}} \\
\boldsymbol{I}_{22} &= -\mathrm{E}\left[\frac{\partial S_2}{\partial \sigma}\right] = \frac{2n}{\sigma^2}.
\end{aligned}
$$

Likelihood Inference for Nonlinear Models: Estimation

Asymptotically,

$$\frac{\sum_{i=1}^{n}[Y_i - \mu(\widehat{\beta})]^2}{\sigma^2} \rightarrow_d \chi^2_{n-k-1} \tag{45}$$

which may be used to construct approximate F tests – dependence on normality is a worry here.

If r is unknown then it may also be estimated, by deriving the score from the likelihood (40), though an abundance of data will be required.

Example: Pharmacokinetics of Theophylline

We let y_i represent the log concentration and assume the model

$$y_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \mathsf{N}[\mu_i(\boldsymbol{\beta}), \sigma^2],$$

$i = 1, \ldots, n$, where

$$\mu_i(\boldsymbol{\beta}) = \log\left\{ \frac{D k_a}{V(k_a - k_e)} \left[\exp(-k_e x) - \exp(-k_a x)\right] \right\} \tag{46}$$

with $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$ where $\beta_0 = V$, $\beta_1 = k_a$, $\beta_2 = k_e$.

Example: Pharmacokinetics of Theophylline

We fit this model using maximum likelihood estimation for β and the moment estimator (43) for σ^2.

The results are displayed in Table 3, with the fitted curve displayed on Figure 11.

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.

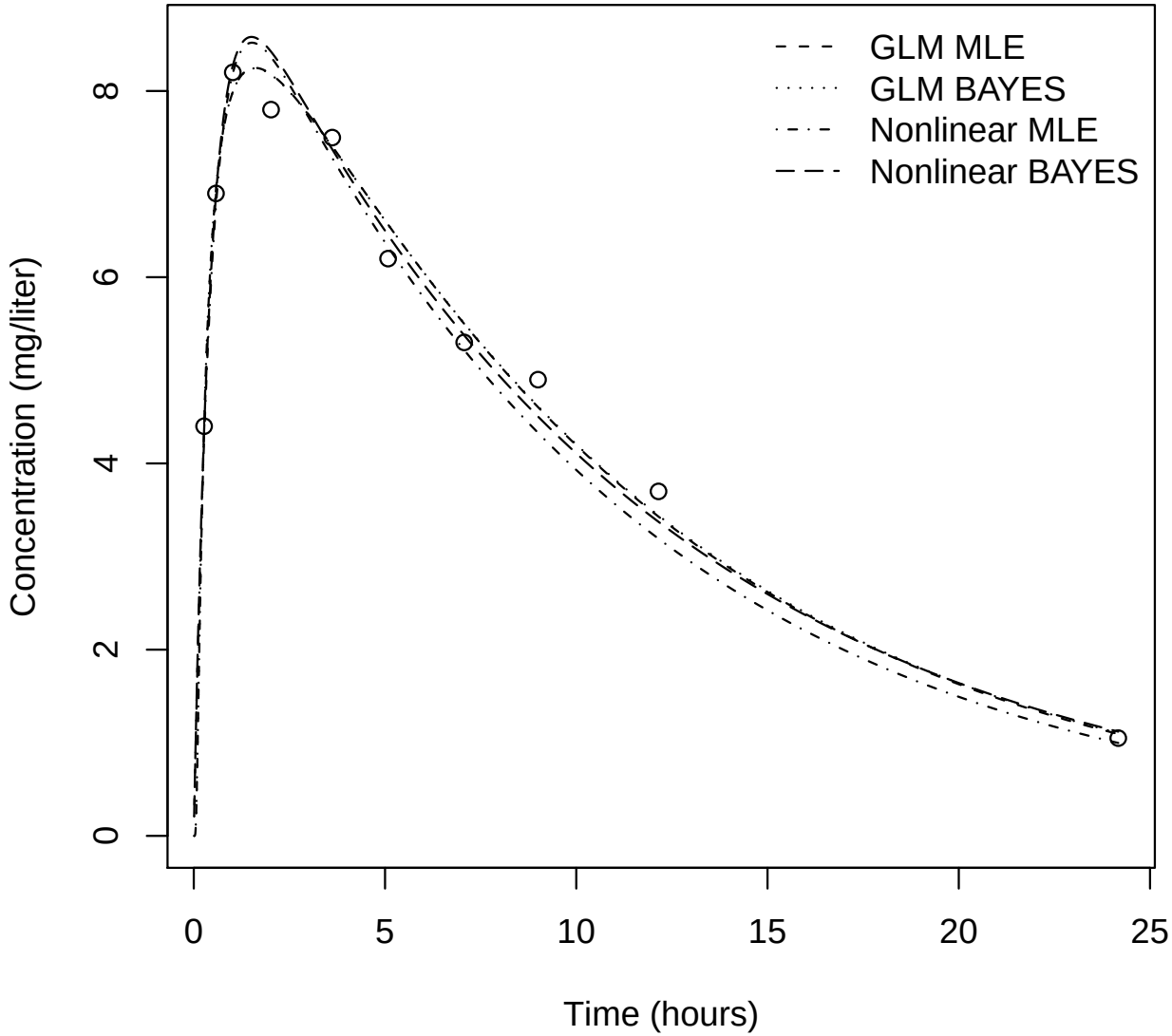


FIGURE 11: Theophylline data, along with fitted curves under various models and inferential approaches. Four curves are included, corresponding to MLE and Bayes analyses of GLM and nonlinear models. The two nonlinear curves are indistinguishable.

Example: Pharmacokinetics of Theophylline

In the left column of Figure 12 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.

In this plot 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 figures with a range of ± 30 standard errors, 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.

Example: Pharmacokinetics of Theophylline

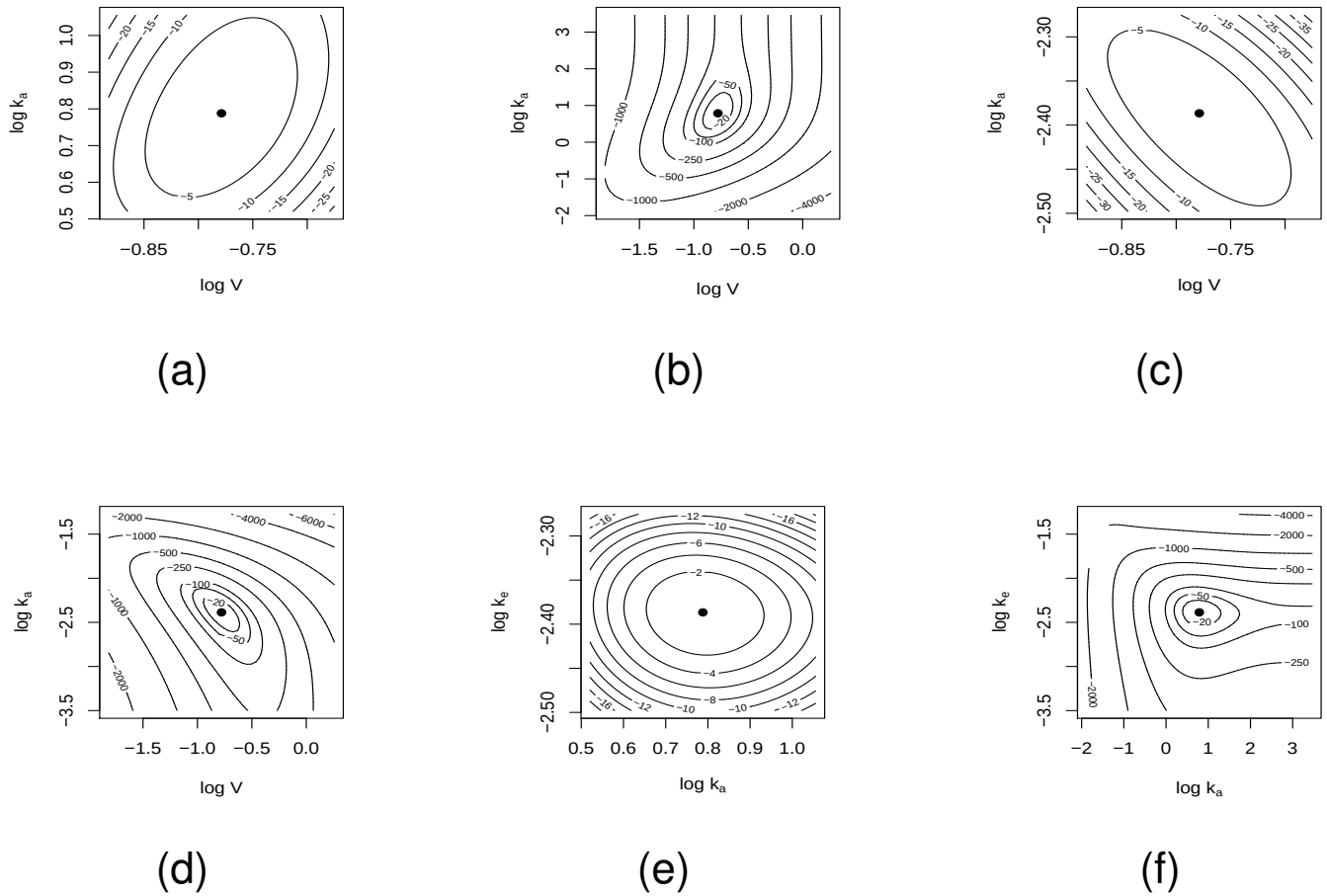


Figure 12: Likelihood contours for the Theophylline data with the range of each parameter being the MLE $\pm$ 3 standard errors in panels (a), (c), (e) and $\pm$ 30 standard errors in panels (b), (d), (f); y axis labels in (c), (d) should be log k_e.

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 $0 \leqslant q < k$.

Interest focuses on testing whether a subset of $k - q$ parameters are equal to zero via a test of the null

$$H_0 : \boldsymbol{\beta}_1 \text{ unrestricted, } \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$$

versus

$$H_1 : \boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2] \neq [\boldsymbol{\beta}_1, \boldsymbol{\beta}_{20}].$$

Asymptotically, and with known σ,

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

where $\widehat{\boldsymbol{\beta}}^{(0)}$ and $\widehat{\boldsymbol{\beta}}^{(1)}$ are the MLEs under null and alternative, respectively, and $\ell(\boldsymbol{\beta}, \sigma^2)$ is given by (40).

Unlike the normal linear model, this result is only asymptotically valid.

For the usual case of unknown σ^2 one may substitute an estimate.

Least Squares Inference

We first consider model (38) with

$$\mathrm{E}[\epsilon_i \mid \mu_i] = 0, \quad \mathrm{var}(\epsilon_i \mid \mu_i) = \sigma^2, \quad \mathrm{cov}(\epsilon_i, \epsilon_j \mid \mu_i, \mu_j) = 0.$$

In this case we may obtain ordinary least squares estimates, $\widehat{\boldsymbol{\beta}}$, that minimize the sum-of-squares

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

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

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

which is identical to (44), and is optimal within the class of linear estimating functions, under correct specification of the first two moments.

Least Squares Inference

If we now assume uncorrelated errors with

$$\text{var}(\epsilon_i \mid \mu_i) = \sigma^2 \mu_i^r(\boldsymbol{\beta})$$

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.

We therefore minimize

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

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

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

which is identical to that under quasi-likelihood (10).

Inference may be based on the asymptotic result

$$(\boldsymbol{D}^{\mathsf{T}} \boldsymbol{V}^{-1} \boldsymbol{D}/\sigma^2)^{1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightarrow_d \mathsf{N}_{k+1}(\mathbf{0}, \mathsf{I}_{k+1}). \tag{47}$$

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 second moment model misspecification.

Therefore, the approach that is followed should depend on how much faith we have in the assumed model.

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

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

with $\boldsymbol{D}$ an $n \times (k+1)$ matrix with elements $\partial \mu_i / \partial \beta_j$, $i = 1, \ldots, n$, $j = 0, \ldots, k+1$ and $\boldsymbol{V}$ the diagonal matrix with elements $V_{ii} = \mu_i(\beta)^r$ with $r \geqslant 0$ known.

This estimating equation arises from likelihood considerations if $r = 0$ or, more generally, from GLS.

With this form for $\boldsymbol{G}(\cdot)$, equations (30), (31) and (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 $i = 1, \ldots, n$, and assuming uncorrelated errors.

One possibility for fitting is weighted 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.

Example: Pharmacokinetics of Theophylline

Table 4 gives estimates of the above model under WLS and MLE, along with likelihood estimation for the model,

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

There are some differences in the table but overall the estimates and standard errors are in reasonable agreement.

Model	log V	log k_a	log k_e
MLE Log Scale	-0.78 (0.035)	0.79 (0.089)	-2.39 (0.037)
WLS 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)

Table 4: 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.

Example: Pharmacokinetics of Theophylline

Table 5 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 again a little shorter.

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 Sand	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]

Table 5: Point and 90% interval estimates for the Theophylline data of Table 1, 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.

Geometry of Least Squares

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

$$(\boldsymbol{y} - \boldsymbol{\mu})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{48}$$

where $\boldsymbol{y}$ and $\boldsymbol{\mu}$ are $n \times 1$ vectors.

We first examine the linear model, $\boldsymbol{\mu} = \boldsymbol{x}\beta$, where $\boldsymbol{x}$ is $n \times (k+1)$ and β is $(k+1) \times 1$.

The Geometry of Least Squares

For fixed $\boldsymbol{x}$, the so called solution locus maps out the fitted values $\boldsymbol{x}\widetilde{\beta}$ for all values of $\widetilde{\beta}$ and is a $(k+1)$-dimensional hyperplane of infinite extent.

Differentiation of (48) gives

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

where $\widehat{\beta} = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y}$ and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals.

The Geometry of Least Squares

So the sum of squares is minimized when the vector $(\boldsymbol{y} - \boldsymbol{x}\beta)$ is orthogonal to the hyperplane that constitutes the solution locus.

The fitted values are

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

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.

The Geometry of Least Squares

These observations have several implications.

In terms of inference, recall that for a linear model a $100(1-\alpha)\%$ confidence interval for $\boldsymbol{\beta}$ is the ellipsoid

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

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 disc.

For nonlinear models asymptotic inference for $\boldsymbol{\beta}$ results from

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

where $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2\widehat{\boldsymbol{V}}$, with $\widehat{\sigma}^2 = s^2$.

The Geometry of Least Squares

The approximation here 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 Geometry of Least Squares

The curvature and lack of equally-spaced points manifests itself in contours of equal likelihood being banana-shaped, and perhaps "open" (so that they do not join).

The ± 30 standard errors panels of Figure 13 give 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, while the aspect that is parameterization dependent is the parameter-effects curvature.

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.

Example: Pharmacokinetics of Theophylline

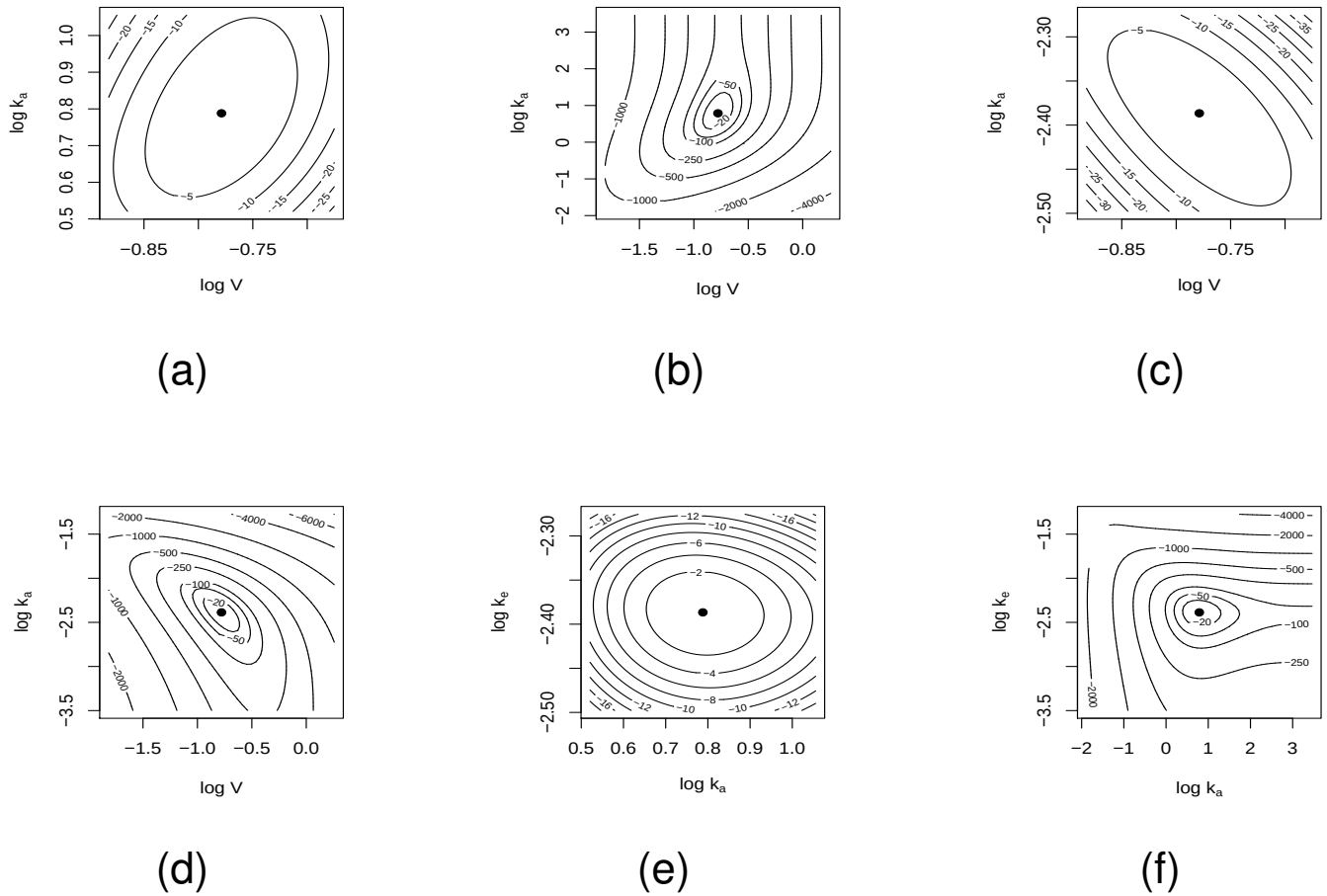


Figure 13: Likelihood contours for the Theophylline data with the range of each parameter being the MLE $\pm$ 3 standard errors in panels (a), (c), (e) and $\pm$ 30 standard errors in panels (b), (d), (f); y axis labels in (c), (d) should be log k_e.

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,

$$\mu = \boldsymbol{x}\beta, \qquad -\infty < \beta < \infty,$$

with the (simplified) nonlinear Michaelis-Menten model

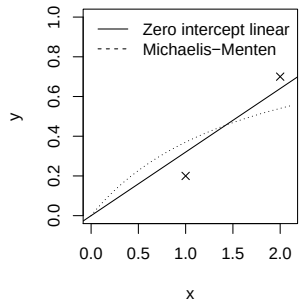
$$\mu = \frac{\boldsymbol{x}}{\boldsymbol{x} + \theta}, \qquad \theta > 0.$$

Figure 14(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 $k = 0$).

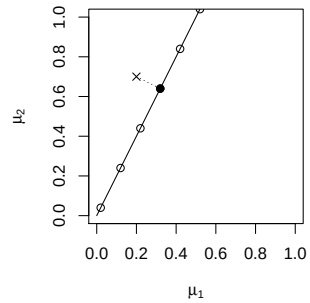
The point $[x_1\widehat{\beta}, x_2\widehat{\beta}]$ with least squares estimate

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

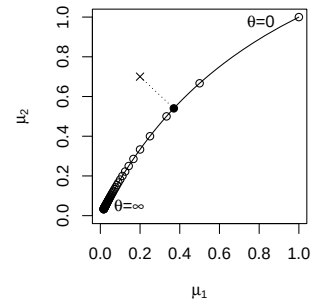
is the fitted point, and is indicated as a solid circle.



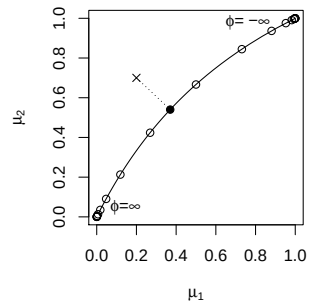
(a)



(b)



(c)



(d)

FIGURE 14: (a) Fictitious data with $\boldsymbol{x} = [1, 2]$ and $\boldsymbol{y} = [0.2, 0.7]$, (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, (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

The Geometry of Least Squares

The dashed line is the vector joining $[y_1, y_2]$ to the fitted point and is perpendicular to the 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 Figure 14 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 $\left[x_1/(x_1 + \widehat{\theta}), x_2/(x_2 + \widehat{\theta})\right]$ is perpendicular to the solution locus, but we see that points on the 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

$$\left[\frac{x_1}{x_1 + \exp(\widehat{\phi})}, \frac{x_2}{x_2 + \exp(\widehat{\phi})} \right],$$

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.

Bayesian Inference for Nonlinear Models

Bayesian Inference for Nonlinear Models: Prior Specification

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, estimation and hypothesis testing.

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 $a = b = 0$, to give 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 $\pi(\sigma^2 \mid r)$; the scale of σ^2 depends on the choice for r.

So far as a prior for β is concerned, great care must be taken to ensure that the resultant posterior is proper.

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

$$\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)^2\right] = c > 0$$

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

Bayesian 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 is not so appealing since the conditional distribution, $\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma$, will not have a familiar form.

MCMC via a Metropolis-Hastings algorithm is straightforward, however, and this model is easy to fit in `Stan` (for example).

If an informative prior is present direct sampling via a rejection algorithm, with the prior as a proposal, may present a viable option.

Example: Pharmacokinetics of Theophylline

As with GLMs posterior tail areas and Bayes factors are available to test hypotheses/compare models.

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
- back-solve for the parameters that parametrize the likelihood.

Example: Pharmacokinetics of Theophylline

For the compartmental model we transform back from $[x_{1/2}, x_{\max}, \mu(x_{\max})]$, to the original parameters $[V, k_a, k_e]$ via

$$\begin{aligned} k_e &= (\log 2)/x_{1/2} \\ 0 &= x_{\max}(k_a - k_e) - \log\left(\frac{k_a}{k_e}\right) \\ V &= \frac{D}{\mu(x_{\max})}\left(\frac{k_a}{k_e}\right)^{k_a/(k_a-k_e)} \end{aligned} \tag{49}$$

so that k_a is not directly available, but must be obtained as the root of (49).

Example: Pharmacokinetics of Theophylline

Table 3 summarizes inference for the parameters of interest with the interval estimates and medians being obtained as the sample quantiles.

Figure 15 shows the posteriors for functions of interest under the nonlinear model – the posteriors are skewed for all functions of interest.

These figures and Table 3 show that Bayesian inference under the GLM and nonlinear model is very similar.

Frequentist and Bayesian methods are also in close agreement for these data, which is reassuring.

Example: Pharmacokinetics of Theophylline

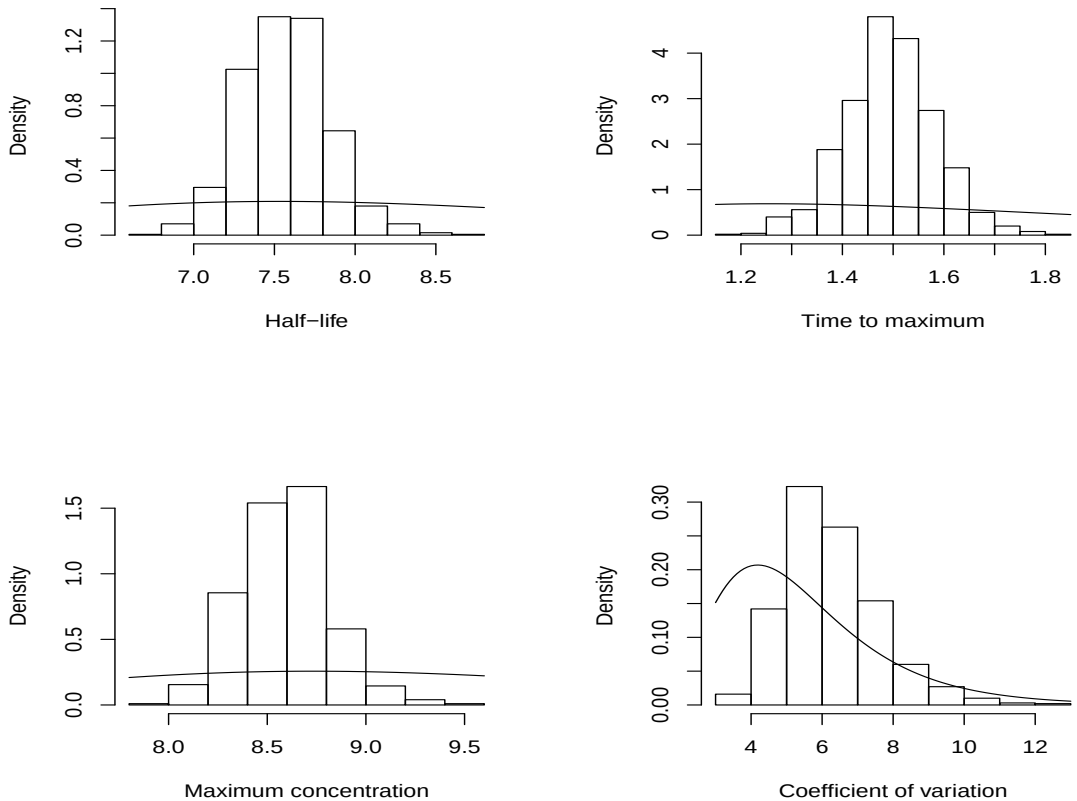


Figure 15: Histogram representations of posterior distributions from the nonlinear compartmental model for the Theophylline data for the: (a) half-life (b) time to maximum, (c) maximum concentration, (d) coefficient of variation, with priors superimposed as solid lines.

Example: Pharmacokinetics of Theophylline

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 (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 16 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.

Example: Pharmacokinetics of Theophylline

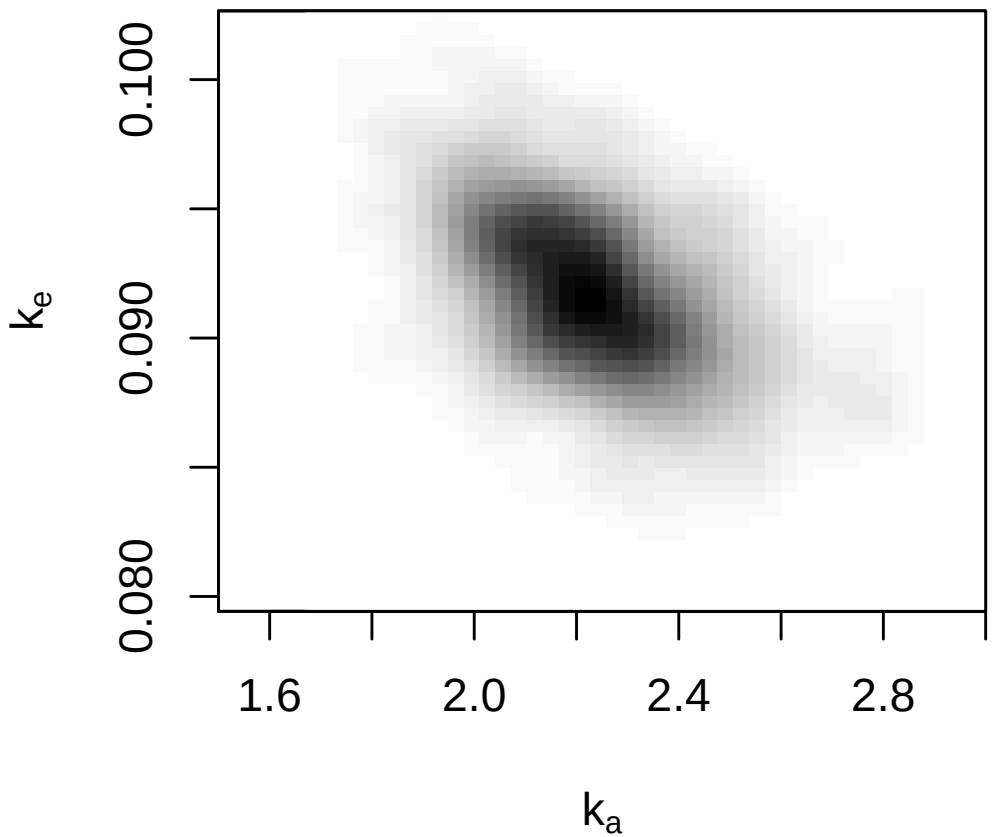


Figure 16: Image plot of samples from the joint posterior distribution of the absorption and elimination rate constants, k_a and k_e, for the Theophylline data.

Example: Pharmacokinetics of Theophylline

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.

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 proposal.

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.

Example: Pharmacokinetics of Theophylline

For these data, based on the above analyses, we 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.

Example: Pharmacokinetics of Theophylline

We fit the nonlinear model with the same priors as used previously and with computation carried out with the rejection algorithm.

Figure 17 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.

Example: Pharmacokinetics of Theophylline

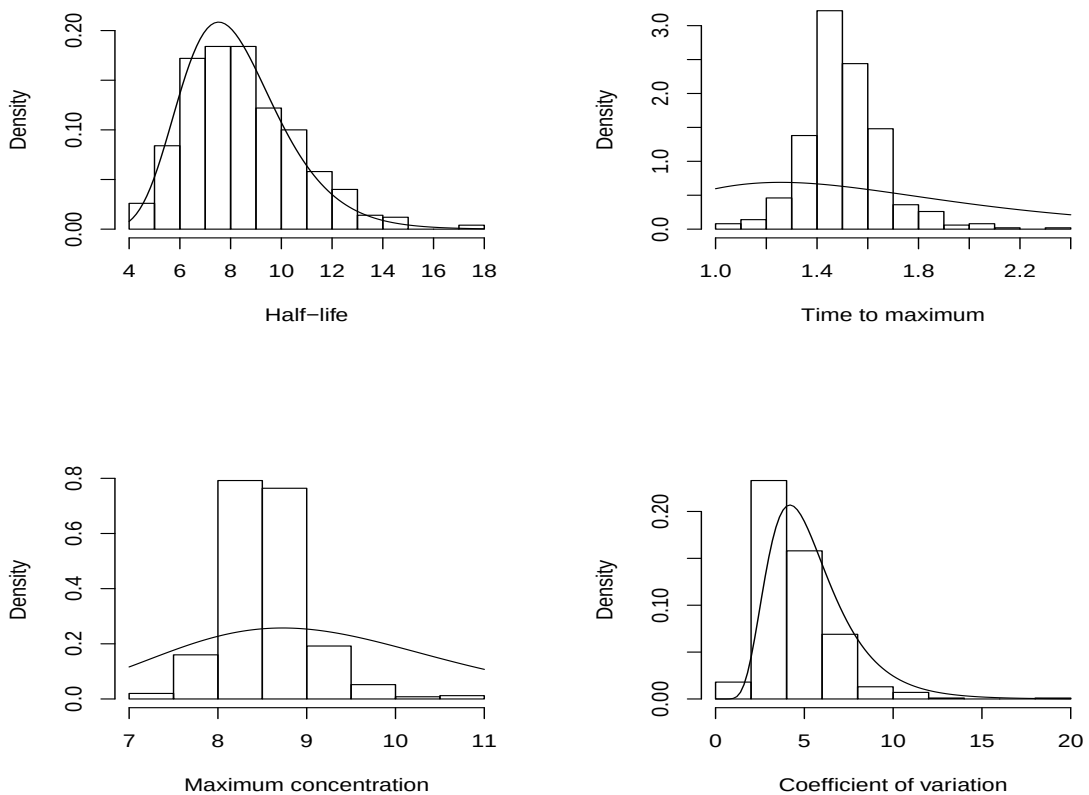


Figure 17: Histogram representations of posterior distributions from the nonlinear compartmental models for the reduced Theophylline dataset of $n = 3$ points, for the: (a) half-life (b) time to maximum, (c) maximum concentration, (d) coefficient of variation, with priors superimposed as solid lines.

Assessment of Assumptions for Nonlinear Models

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

$$e_i^\star = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\text{var}}(Y_i)}}, \tag{50}$$

which we refer to as Pearson residuals.

These residuals may be used in the usual ways.

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.

Example: Pharmacokinetics of Theophylline

Letting y_i represent the log concentration at time x_i, we examine the Pearson residuals, as given by (50), obtained following likelihood estimation with the model $y_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \mathrm{N}(\mu_i, \sigma^2)$, with μ_i given by (46), for $i = 1, \ldots, n$.

Figure 7(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 18 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.

Example: Pharmacokinetics of Theophylline

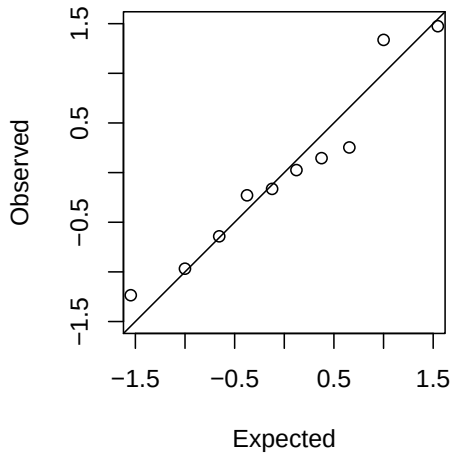


Figure 18: Normal QQ plot for the Theophylline data and model (46).

Discussion

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.

The construction of GLMs, in particular the score being linear in the response, is such that asymptotic inference is accurate for relatively small n.

Concluding Remarks

Care is required in the fitting of, and inference for, nonlinear models.

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.

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, for both asymptotics and computation, and prior specification for a Bayesian analysis.

References

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, **11**, 397–412.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

Nelder, J. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128–141.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.