Math 35200
Autumn term 2003

Generalised linear models

by Peter Green (University of Bristol,
P.J.Green@bristol.ac.uk).

- motivation: basic ideas
- connection to linear models
- principles of statistical model-building
- maximum likelihood as a general approach to model-fitting
- applications, including to survival analysis
- demonstrations and practical work using **R**

1. Generalised linear models

Linear model in brief

Responses are known linear functions of unknown parameters, plus an error term. Matrix formulation:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y}$ is $n \times 1$, and X is $n \times p$, where $p < n$. We usually assume X is full-rank (i.e. rank of X is p, columns of X are l.i., $X^T X$ is nonsingular – these are equivalent).

(a) No assumptions

Least-squares estimator $\widehat{\boldsymbol{\beta}}$ minimises $S(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = ||\boldsymbol{Y} - X\boldsymbol{\beta}||^2$. In full-rank case, $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$.

(b) Mean and variance assumptions

If $E(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 I$ then least-squares estimator is unbiased and BLUE (minimum variance unbiased estimator) – this is the Gauss-Markov theorem. $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$.

(c) Normal theory assumptions

If $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$, $\widehat{\boldsymbol{\beta}}$ is also the MLE, we have:
$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$,
$S(\widehat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2_{n-p}$, and
$\widehat{\boldsymbol{\beta}}$ and $S(\widehat{\boldsymbol{\beta}})$ are independent.

This leads to confidence intervals for linear functions of $\boldsymbol{\beta}$, tests of hypotheses, analysis of variance, etc.

1.1 Motivation

We are going to extend the ideas of normal linear modelling:

- model specification
- estimation
- testing via ANOVA, etc.

to non-normal and nonlinear cases.

Some examples that will become familiar:

- Quantal response models (such as logit and probit regression)
- Contingency tables
- Survival analysis models

Some generic problems/challenges:

- non-normality
- unequal variances
- nonlinearity
- extrapolation
- bounds on responses
- least-squares not relevant?

Let us first revisit the normal linear model. Instead of writing

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

we will separate the definition into three pieces:

1. Random component: $Y_i \sim N(\mu_i, \sigma^2)$, independently, $i = 1, 2, \ldots, n$
2. Systematic component: $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \equiv \boldsymbol{x}_i^T \boldsymbol{\beta}$
3. Link: $\mu_i = \eta_i$.

We will consider changing (1) and (3). Keeping (2) means that we can keep all the model-building ideas from Linear Models (model-building in the sense of selecting explanatory variables and factors going into the design/model matrix X).

Note that the unknown parameters are $\boldsymbol{\beta}$ and σ^2, and these appear in *different* parts of the model now.

1.2 Exponential families

This will be what we need to generalise part (1), the random component, of the model. First consider a single random variable Y. We say that Y has a distribution in the exponential family if its density or probability function can be written in the canonical form

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Here θ and ϕ are parameters, a, b and c are specified functions, and we assume the range of Y does not depend on θ and ϕ.

Example P The Poisson(λ) distribution has probability function

$$\frac{e^{-\lambda}\lambda^y}{y!} = \exp\{y \log \lambda - \lambda - \log y!\}$$

for $y = 0, 1, \ldots$. This is in the right form with $\theta = \log \lambda$, $b(\theta) = \lambda = e^\theta$, ϕ omitted, $c(y, \phi) = -\log y!$, and $a(\phi) \equiv 1$.

Example N The normal distribution $N(\mu, \sigma^2)$.

Example B The binomial distribution $Bin(m, \pi)$.

1.3 Moments

It is hard to ‘see’ the interpretation of the functions a, b and c. It helps to find the moments of Y in terms of them.

Recall two results about derivatives of log-likelihoods.

Let $L(\theta) = L(\theta; Y)$ be the likelihood, and $\ell(\theta) = \ell(\theta; Y) = \log L(\theta)$ be the log-likelihood.

Then

$$E_\theta\left(\frac{d\ell}{d\theta}(\theta; Y)\right) = 0$$

and

$$\mathrm{var}_\theta\left(\frac{d\ell}{d\theta}(\theta; Y)\right) = E_\theta\left(-\frac{d^2\ell}{d\theta^2}(\theta; Y)\right)$$

in regular cases where the range of the distribution does not depend on θ.

Proof. We have

$$\frac{d\ell}{d\theta} = \frac{1}{L}\frac{dL}{d\theta}$$

so

$$E_\theta\left(\frac{d\ell}{d\theta}(\theta;Y)\right) = \int \frac{1}{L}\frac{dL}{d\theta}L dy = \int \frac{dL}{d\theta}dy$$

$$\stackrel{?}{=} \frac{d}{d\theta}\int L dy = \frac{d}{d\theta}1 = 0$$

Rather similarly,

$$\frac{d^2\ell}{d\theta^2} = \frac{1}{L}\frac{d^2L}{d\theta^2} - \frac{1}{L^2}\left(\frac{dL}{d\theta}\right)^2 = \frac{1}{L}\frac{d^2L}{d\theta^2} - \left(\frac{d\ell}{d\theta}\right)^2$$

so

$$E_\theta\left(\frac{d^2\ell}{d\theta^2}\right) = \int \frac{1}{L}\frac{d^2L}{d\theta^2}L dy - E_\theta\left(\frac{d\ell}{d\theta}\right)^2$$

$$\stackrel{?}{=} \frac{d^2}{d\theta^2}\int L dy - E_\theta\left(\frac{d\ell}{d\theta}\right)^2 = 0 - E_\theta\left(\frac{d\ell}{d\theta}\right)^2$$

Moments of an exponential family distribution

$$\ell(\theta) = \frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)$$

We treat ϕ as constant.

$$\frac{d\ell}{d\theta} = \frac{Y - b'(\theta)}{a(\phi)}$$

$$\frac{d^2\ell}{d\theta^2} = \frac{-b''(\theta)}{a(\phi)}$$

Thus from the standard results above

$$E_\theta(Y) = b'(\theta) \quad \text{and} \quad \text{var}_\theta\left(\frac{Y - b'(\theta)}{a(\phi)}\right) = \frac{b''(\theta)}{a(\phi)}$$

so that $\text{var}_\theta(Y) = a(\phi)b''(\theta)$.

We usually write $\mu = b'(\theta)$; usually b' is 1-1. Then $b''(\theta)$ is implicitly a function of μ - we write $b''(\theta) = V(\mu)$; V is called the *variance function*.

Example P

$$b(\theta) = e^{\theta} \quad \text{and} \quad a(\phi) \equiv 1$$

so that $E_\theta(Y) = b'(\theta) = e^\theta = \lambda$ and $\text{var}_\theta(Y) = b''(\theta) = e^\theta = \lambda$. So $V(\mu) = \mu$.

1.4 The link function

Let's complete the specification of a generalised linear model. If we have responses Y_i and explanatory variables $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$:

1. Random component: Y_i has density or probability function

   $$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

   independently for $i = 1, 2, \ldots, n$.

2. Systematic component: $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$

3. $g(\mu_i) = \eta_i$, $i = 1, 2, \ldots, n$, for some specified *link function* g.

We suppose that g and b' are sufficiently well-behaved functions that their inverses exists, and that both functions and their inverses are differentiable. Then there are 3 alternative choices of parameter for each Y_i in a GLM:

- the *natural parameter* θ_i,
- the *mean parameter* μ_i and
- the *linear predictor* η_i.

We have $\eta = g(\mu)$ and $\mu = b'(\theta)$, both functions operating component-by-component.

Examples of GLM's

Example P-log Suppose

$$Y_i \sim \text{Poisson}(e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) \quad i = 1, 2, \ldots, n,$$

independently (the exponential transformation here is a natural choice, since it ensures the Poisson means are positive whatever the values of the $\boldsymbol{x}$ and $\boldsymbol{\beta}$). We have
$\mu_i = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} = e^{\eta_i}$ so $\eta_i = \log \mu_i$. g is the log function; we say this model has the log link function. In this case $\theta_i \equiv \eta_i (= \log \mu_i)$

Example N-id

$$Y_i \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2) \quad i = 1, 2, \ldots, n,$$

independently. Now all parameters are identical: $\theta_i = \eta_i = \mu_i$.

Fitting GLM's in R

This is very easy, a natural extension of the way that linear models are fitted: we use the function `glm()`:

`glm(`*formula*`,`*family*`,`*data*`)`

where *formula* and *data* are as in `lm()`:

- *formula*specifies the response variable, and describes the explanatory variables that are to be included, and
- *data*(optional) is a dataframe where the variables are to be found.

The new ingredient is

- *family* which specifies the error distribution and link function in a mnemonic way: examples are `normal(link=identity)` or `poisson(link=log)`.

1.5 Link functions for quantal response data

For $i = 1, 2, \ldots, n$, m_i animals are given a log-dose $x_i$of a poison: suppose Y_i are 'affected' by it. All animals are independent, and it is natural to suppose they are 'exchangeable'. This implies

$$Y_i \sim \text{Bin}(m_i, \pi_i) \quad i = 1, 2, \ldots, n,$$

where π_i is some function of x_i, to be estimated.

On approach is to postulate the existence of a *tolerance* random variable: each animal has a tolerance T. This is defined to be the maximum log-dose the animal can tolerate without being affected. Thus

$$P\{\text{animal affected}\} = P\{T \leq x\}$$

The distribution of tolerance T in the population of animals is specified by its distribution function

$$P\{T \leq t\} = \Psi(\beta_0 + \beta_1 t)$$

where Ψ is a specified distribution function (e.g. Normal), and β_0 and β_1 are parameters to be estimated.

Then

$$P\{\text{animal affected}\} = P\{T \leq x\}$$

$$= \Psi(\beta_0 + \beta_1 x) = \Psi(\boldsymbol{x}^T \boldsymbol{\beta})$$

We see that $\pi_i = \Psi(\boldsymbol{x}_i^T \boldsymbol{\beta})$ and
$Y_i \sim \text{Bin}(m_i, \Psi(\boldsymbol{x}_i^T \boldsymbol{\beta})) \quad i = 1, 2, \ldots, n.$

This is a generalised linear model, with Binomial error distribution.

What is the link function?

Since $\mu_i = E(Y_i) = m_i \pi_i = m_i \Psi(\boldsymbol{x}_i^T \boldsymbol{\beta})$, we have $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \Psi^{-1}(\mu_i / m_i)$. The link function is $g = \Psi^{-1}(\cdot / \mu_i)$.

The link function is $g = \Psi^{-1}(\cdot/m_i)$.....

Note that (unless the m_i are all equal, $i = 1, 2, \ldots, n$), the link functions are really different for different i - this is a minor issue that does not cause any difficulty.

Examples

(a) Normal distribution

$$\Psi(\eta) = \Phi(\eta) \equiv \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

$\Psi^{-1}(\pi)$ has no closed form: use tables or numerical algorithms. We call this the *probit* link function: *family* is `binomial(link=probit)`.

(b) Logistic distribution

$$\Psi(\eta) = \frac{e^\eta}{1 + e^\eta}$$

$$\Psi^{-1}(\pi) = \log \frac{\pi}{1 - \pi}$$

$$\Psi^{-1}(\mu/m) = \log \frac{\mu}{m - \mu}$$

This is called the *logit* link function: specified as `binomial(link=logit)`.

(c) 'Extreme value' distribution

$$\Psi(\eta) = 1 - \exp(-e^\eta)$$

$$\Psi^{-1}(\pi) = \log\{-\log(1 - \pi)\}$$

– the 'complementary log-log' link function, specified as `binomial(link=cloglog)`.

The proportional odds model

The Binomial GLM with the logit link function is sometimes referred to as the ‘proportional odds model’. This is because

$$\eta = \Psi^{-1}(\pi) = \log \frac{\pi}{1-\pi}$$

means that equal additive increments to each explanatory variable x_j, giving equal additive increments to $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$ results in equal multiplicative changes to the *odds*, $\pi/(1-\pi)$.

Taking this a little further, suppose animals come from several groups indexed by s(for example, sex). Then we would include a factor `Sex` say as well as the variable `Logdose` (x). The *formula* in the `glm()` call might then be `Kill~ Sex +Logdose`.

`Kill~Sex+Logdose`

$$\eta = \eta(s,x) = \gamma_s + \beta_0 + \beta_1 x$$

Let $\pi(s,x)$ be the response probability. Then with the logit link function,

$$\frac{\pi(s,x)}{1-\pi(s,x)} = \exp(\eta(s,x)) = \exp(\gamma_s + \beta_0 + \beta_1 x)$$

If we compare two groups s, s', then the *odds ratio*, i.e. the ratio of the odds on response for group $s$compared to group s' is

$$\frac{\pi(s,x)}{1-\pi(s,x)} \div \frac{\pi(s',x)}{1-\pi(s',x)} = \frac{\exp(\eta(s,x))}{\exp(\eta(s',x))} = \exp(\gamma_s - \gamma_{s'})$$

for all x– that is, we have 'proportional odds', the odds ratio does not depend on the log-dose x.

1.6 Canonical link functions

For each exponential family distribution, one particular choice of link function has special *mathematical* status – which does not imply anything about its statistical relevance to a particular data set.

Recall the mappings between the three key parameters, θ, μ and η:

Suppose we set g to be the inverse of the function b', then

$$\eta = g(b'(\theta)) \equiv \theta$$

i.e. the linear predictor is the same as the natural parameter. This choice of g is called the *canonical link function* for the distribution.

To understand the mathematical consequence of this, suppose for simplicity that $a_i(\phi) = \phi$ for all i. Then it is easily checked that the log-likelihood can be written

$$\phi^{-1} \sum_{j=1}^{p} \beta_j \left(\sum_{i=1}^{n} Y_i x_{ij} \right) - B(\boldsymbol{\beta}, \phi) + C(\boldsymbol{Y}, \phi)$$

for certain functions B and C. This implies that $X^T \boldsymbol{Y}$ is sufficient for $\boldsymbol{\beta}$, for each ϕ. (This means that if you know the value of ϕ and $X^T \boldsymbol{Y}$, then you can make the same inference about $\boldsymbol{\beta}$ as you can if you know ϕ and $\boldsymbol{Y}$ itself. This is because the conditional distibution of $\boldsymbol{Y}$ given ϕ and $X^T \boldsymbol{Y}$ does not depend on $\boldsymbol{\beta}$.)

We will not make any use of this, but simply regard the canonical link function as the default - in fact, that is exactly what **R** does.

Example P For the Poisson distribution, $b(\theta) = e^{\theta}$, so the inverse of b' is the log function: thus $\eta = g(\mu) = \log \mu$ is the canonical link function for the Poisson distribution.

Using `poisson`for *family* in the `glm()` command in **R** is the same as `poisson(link=log)`.

2. Fitting a generalised linear model

What statistical principle is appropriate for fitting a GLM (that is, estimating $\boldsymbol{\beta}$ and ϕ), and how do we go about actually doing it?

Since we have retained much of the formulation of the linear model, it is tempting to think we could just continue to use least squares.

Idea 1. Use least squares directly: choose $\boldsymbol{\beta}$ to minimise

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 = ||\boldsymbol{Y} - X\boldsymbol{\beta}||^2$$

Problem! $\boldsymbol{Y}$ and $X\boldsymbol{\beta}$ are on different scales – we have ignored the link function, which maps the Y and μ scale onto the η scale.

Idea 2. Use least squares after transformation: choose $\boldsymbol{\beta}$ to minimise either

$$\sum_{i=1}^{n}(Y_i - g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))^2$$

or

$$\sum_{i=1}^{n}(g(Y_i) - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

Problem! Neither $\mathrm{var}(Y_i)$ nor $\mathrm{var}(g(Y_i))$ are constant, they depend on $\boldsymbol{x}_i^T\boldsymbol{\beta}$ in general.

Idea 3. Use weighted least squares, after transformation: e.g. choose $\boldsymbol{\beta}$ to minimise

$$\sum_{i=1}^{n} \frac{(Y_i - g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}))^2}{V(g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}))}$$

Problem! No longer a quadratic in $\boldsymbol{\beta}$, so we would need numerical methods to solve. If we have to go to that trouble, can we find a better principle in the first place?

2.1 Maximum likelihood estimation in GLM's

We know that MLE is a principle that very often gives good estimates in statistical models, and we know that in the normal linear model, MLE is the same as least squares (see slide 3). So let's try it here, for GLM's.

From now on, we will suppose that $a_i(\phi) \equiv \phi$ for all i; much of the theory remains true without this restriction, but this makes the derivations a bit cleaner, and is enough for all the models we will encounter.

With this substitution, recall from slide 6 that the log-likelihood $\ell(\boldsymbol{\theta}, \phi) = \log L(\boldsymbol{\theta}, \phi)$ is given by:

$$\ell(\boldsymbol{\theta}, \phi) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi) \right\} = \sum_{i=1}^{n} \ell_i,$$

say, and that $\boldsymbol{\theta}$is a function of $\boldsymbol{\beta}$ through the relationship $g(b'(\theta_i)) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ (slides 11 and 12).

For the MLE of $\boldsymbol{\beta}$, we need to solve the *likelihood equations*

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0 \quad \text{for all } j = 1, 2, \ldots, p.$$

(For the moment, we ignore the parameter $\phi$and $\partial \ell / \partial \phi$.)

By the chain rule for differentiation,

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \frac{d\ell_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \frac{d\ell_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \frac{d\ell_i}{d\theta_i}\left(\frac{d\mu_i}{d\theta_i}\right)^{-1}\left(\frac{d\eta_i}{d\mu_i}\right)^{-1}\frac{\partial \eta_i}{\partial \beta_j} \\
&= \frac{Y_i - b'(\theta_i)}{\phi}(b''(\theta_i))^{-1}(g'(\mu_i))^{-1}x_{ij} \\
&= \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)}
\end{aligned}$$

Thus the likelihood equations we have to solve for the MLE $\widehat{\boldsymbol{\beta}}$ are

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = 0,$$

$j = 1, 2, \ldots, p$, where $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$.

Likelihood equations, simplified

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{(Y_i-\mu_i)x_{ij}}{V(\mu_i)g'(\mu_i)}=0,$$

– a very similar structure to the Normal equations arising in ordinary least squares.

Example N-id Here, we have $g(\mu)=\mu$ so $g'(\mu)\equiv 1$; also $V(\mu)\equiv 1$, $\phi=\sigma^2$ and $\mu_i=\boldsymbol{x}_i^T\boldsymbol{\beta}$, so the equations become

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})x_{ij}=0$$

Ignoring the factor σ^2, the LHS is the j^{th} element of the vector $X^T(\boldsymbol{Y}-X\boldsymbol{\beta})$, so the equations reduce to the Normal equations of least squares:

$X^T(\boldsymbol{Y}-X\boldsymbol{\beta})=0$ or equivalently $X^TX\boldsymbol{\beta}=X^TY$.

Example P-log This time, $g(\mu) = \log \mu$, so $g'(\mu) = 1/\mu$, while $V(\mu) = \mu$ and ϕ is omitted. Substituting $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$, the equations become

$$\sum_{i=1}^{n} (Y_i - e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}) x_{ij} = 0,$$

which are not linear in $\boldsymbol{\beta}$ and no explicit solution is possible in general.

2.2 Iterative methods

Iterative numerical methods to solve an equation for a vector variable $\boldsymbol{\beta}$, say, start with an initial estimate $\boldsymbol{\beta}^{(0)}$ and successively improve it: $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots$, continuing until the difference

$$||\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}||$$

is sufficiently small that we can assume "convergence has occurred".

Ideally, we should like to be able to prove that $\boldsymbol{\beta}^{(t)} \to \widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}}$ is the desired solution (in our case, the MLE), but we will often be unable to prove this, and be content with a less stringent 'fixed point justification', that is, proving that

if $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$, then $\boldsymbol{\beta}^{(t)} = \widehat{\boldsymbol{\beta}}$.

The Newton-Raphson procedure

This is a simple numerical method for solving equations, like our likelihood equations $\partial\ell/\partial\beta_j = 0$ for all $j = 1, 2, \dots, p$. It consists of the following iteration:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (H^{(t)})^{-1}\boldsymbol{u}^{(t)}$$

where the $p \times 1$ gradient vector $\boldsymbol{u}^{(t)}$ is

$$\boldsymbol{u}^{(t)} = \left(\frac{\partial\ell}{\partial\beta_1}, \dots, \frac{\partial\ell}{\partial\beta_p}\right)^T$$

and the $p \times p$ Hessian matrix $H^{(t)}$ is given by

$$H_{jk}^{(t)} = \frac{\partial^2\ell}{\partial\beta_j\partial\beta_k},$$

for $j, k = 1, 2, \dots, p$, both $\boldsymbol{u}^{(t)}$ and $H^{(t)}$ being evaluated at the current estimate $\boldsymbol{\beta}^{(t)}$.

Newton-Raphson for GLM's

Omitting the superscript (t)s:

$$u_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{d\ell_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{d\ell_i}{d\eta_i} x_{ij}$$

and

$$H_{jk} = \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} \left(\frac{\partial \ell_i}{\partial \beta_j} \right) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} \left(\frac{d\ell_i}{d\eta_i} x_{ij} \right)$$

$$= \sum_{i=1}^{n} \frac{d}{d\eta_i} \left(\frac{d\ell_i}{d\eta_i} x_{ij} \right) x_{ik} = \sum_{i=1}^{n} \frac{d^2 \ell_i}{d\eta_i^2} x_{ij} x_{ik}$$

Thus if we write $\boldsymbol{s}$ for the $n \times 1$ vector with

$$s_i = \frac{d\ell_i}{d\eta_i}$$

and $\widetilde{W}$ for the *diagonal* $n \times n$ matrix with

$$\widetilde{W}_{ii} = -\frac{d^2\ell_i}{d\eta_i^2}$$

we have $\boldsymbol{u} = X^T\boldsymbol{s}$ and $H = -X^T\widetilde{W}X$ and the Newton-Raphson iteration is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (H^{(t)})^{-1}\boldsymbol{u}^{(t)}$$

$$= \boldsymbol{\beta}^{(t)} + (X^T\widetilde{W}^{(t)}X)^{-1}X^T\boldsymbol{s}^{(t)}$$

Note that we *do* have a fixed-point justification for this algorithm: if $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$ then we must have

$$X^T\boldsymbol{s}^{(t)} = \boldsymbol{u}^{(t)} = 0,$$

that is, the likelihood equations $\partial\ell/\partial\beta_j = 0$ are satisfied for all $j = 1, 2, \ldots, p$, when $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$.

Fisher scoring for GLM's

Typically, partly for reasons of tradition, we use a modification of this in fitting statistical models. The matrix $\widetilde{W}$ is replaced by W, another diagonal $n \times n$ matrix with

$$W_{ii} = E(\widetilde{W}_{ii}) = E\left(-\frac{d^2\ell_i}{d\eta_i^2}\right)$$

Why? The point is that this matrix is guaranteed to have non-negative (typically positive) diagonal entries, since

$$W_{ii} = E\left(-\frac{d^2\ell_i}{d\eta_i^2}\right) = \mathrm{var}\left(\frac{d\ell_i}{d\eta_i}\right)$$

(see slides 8 and 9), so that $W = \mathrm{var}(\boldsymbol{s})$, so the matrix is non-negative-definite.

The iteration is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (X^T W^{(t)} X)^{-1} X^T \boldsymbol{s}^{(t)}$$

and we still have the fixed-point justification.

Iteratively reweighted least squares

The iteration

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (X^T W^{(t)} X)^{-1} X^T \boldsymbol{s}^{(t)} \qquad (1)$$

is formally similar to the familiar solution for least squares estimates in linear models:

$$\boldsymbol{\beta} = (X^T X)^{-1} X^T \boldsymbol{Y}$$

or more particularly the related *weighted least squares* estimates:

$$\boldsymbol{\beta} = (X^T W X)^{-1} X^T W \boldsymbol{Y}$$

In fact, (1) can be re-arranged to have exactly this form, which means that the iteration can be computed using a weighted least squares calculation: the algorithm is sometimes known as iteratively reweighted least squares (IRLS): note that the weight matrix and even the 'response vector' $X\boldsymbol{\beta}^{(t)} + (W^{(t)})^{-1}\boldsymbol{s}^{(t)}$ taking the place of $\boldsymbol{Y}$ vary from one iteration to the next.

2.3 Estimating ϕ

Recall that in the linear model case, the variance σ^2 did not affect the estimation of $\boldsymbol{\beta}$– the MLE/LSE of $\boldsymbol{\beta}$ was the same for all σ^2.

In the general GLM case, continuing to assume that $a_i(\phi) = \phi$ for all $i = 1, 2, \ldots, n$, we have

$$s_i = \frac{d\ell_1}{d\eta_i} = \frac{d\ell}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}$$

$$= \frac{Y_i - \mu_i}{\phi V(\mu_i) g'(\mu_i)}$$

as in slide 28. Also,

$$W_{ii} = \operatorname{var}(s_i) = \frac{\operatorname{var}(Y_i)}{\{\phi V(\mu_i) g'(\mu_i)\}^2}$$

$$= \frac{\phi V(\mu_i)}{\{\phi V(\mu_i) g'(\mu_i)\}^2} = \frac{1}{\phi V(\mu_i)(g'(\mu_i))^2}$$

so that $1/\phi$ appears as a scale factor in W and $\boldsymbol{s}$ but otherwise does not appear in the estimating equations or iteration for $\widehat{\boldsymbol{\beta}}$.

As in the Normal/identity case, (a) we are less interested in ϕ, and (b) we see that ϕ can be separately estimated from $\boldsymbol{\beta}$.

Recall that $\text{var}(Y_i) = \phi V(\mu_i)$ (as used above). In other words,

$$\frac{E((Y_i - \mu_i)^2)}{V(\mu_i)} = \phi$$

We can use this to suggest a simple estimator for ϕ:

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu_i})^2}{V(\widehat{\mu_i})}$$

where

$$\widehat{\mu_i} = g^{-1}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}),$$

the 'fitted values'.

2.4 Using R to fit a GLM

Let us look at the poisons data in the **R** dataframe `rotenone`. Our first model is a logistic regression – that is, a Binomial GLM with the logit link function. Algebraically:

$$Y_i \sim \text{Bin}(m_i, \pi_i)$$

where

$$\log \frac{\pi_i}{1-\pi_i} = \alpha_{s_i} + \delta x_i$$

for $i = 1, 2, \ldots, n = 17$. Here $s_i = 1, 2, 3$ is the identifier of the poison used (a factor), and x_i is the log-dose, for animal group i. The parameters to be estimated are $\alpha_1, \alpha_2, \alpha_3, \delta$ (cf. revision sheet question 1(b)).

The **R** command to fit this model is

```
fit<-glm(Kill~Poison+Logdose,
                  binomial,rotenone)
```

which stores the results in the variable `fit`.

A note on the parameters used by R

Although this can be changed by using a more complicated command, the default option in **R** model fitting of factors uses a rather strange choice of parameters. For a three-level factor, instead of $\alpha_1, \alpha_2, \alpha_3$, it estimates $\omega_1, \omega_2, \omega_3$, where

$$\omega_1 = \alpha_1$$

$$\omega_2 = \alpha_2 - \alpha_1$$

$$\omega_3 = \alpha_3 - \alpha_1$$

whence

$$\begin{aligned} \alpha_1 &= \omega_1 \\ \alpha_2 &= \omega_1 + \omega_2 \\ \alpha_3 &= \omega_1 + \omega_3 \end{aligned}$$

3. Inference for Generalised linear models

What about assessing goodness of fit, providing standard errors, and constructing confidence intervals?

3.1 The residual deviance

Instead of *minimising* a sum of squares, $S(\boldsymbol{\beta}) = ||\boldsymbol{Y} - X\boldsymbol{\beta}||^2$, we have been *maximising* a log-likelihood $\ell(\boldsymbol{\theta}, \phi)$(where $g(b'(\theta_i)) = \boldsymbol{x}_i^T\boldsymbol{\beta}$).

$S(\widehat{\boldsymbol{\beta}})$ was a simple numerical summary of how well the linear model fitted; $S(\widehat{\boldsymbol{\beta}}) = 0$means a perfect fit.

What is the *best* we can do in fitting a GLM? Recall

$$\ell_i = \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi)$$

so

$$\frac{d\ell_i}{d\theta_i} = 0 \iff Y_i - b'(\theta_i) = 0$$

$$\frac{d\ell_i}{d\theta_i} = 0 \iff Y_i - b'(\theta_i) = 0$$

A model that achieves this equality for all $i = 1, 2, \ldots, n$ is called *saturated*: we will need one free parameter for each observation, otherwise $g(b'(\theta_i)) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ would be a restriction. Denote the corresponding θ_i by $\widetilde{\theta}_i$, i.e. the solution of $b'(\widetilde{\theta}_i) = Y_i$.

Now look at

$$2\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\} = \frac{2}{\phi}\{Y_i(\widetilde{\theta}_i - \theta_i) - b(\widetilde{\theta}_i) + b(\theta_i)\}.$$

Maximising the likelihood is the same as *minimising* this quantity, which is always non-negative, and is 0 only if there is a perfect fit to the i^{th} observation.

Thus $\ell(\widetilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \ell_i(\widetilde{\theta}_i)$ provides a baseline value for the log-likelihood.

Example N-id $b(\theta) = \frac{1}{2}\theta^2$, $b'(\theta) = \theta = \mu$, $\widetilde{\theta}_i = Y_i$ and $\phi = \sigma^2$ so

$$2\{\ell_i(\widetilde{\theta}_i) - \ell_i(\theta_i)\}$$

$$= \frac{2}{\sigma^2}\{Y_i(Y_i - \mu_i) - \frac{1}{2}Y_i^2 + \frac{1}{2}\mu_i^2\} = (Y_i - \mu_i)^2/\sigma^2.$$

In general, let

$$D_i = 2\{Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\}$$

We call $D = \sum_{i=1}^n D_i$ the *deviance* of the model, and if ϕ is present,

$$\frac{D}{\phi} = 2\{\ell(\widetilde{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}})\}$$

is the *scaled deviance*.

In **R**, D is called the *residual* deviance for emphasis, and it is printed out by the function `summary()`.

The *null model* is the model fitting only a constant ('intercept') terms, so $\eta_i =$ constant for all i, and the *null deviance* is the value of the residual deviance corresponding to this model.

In the Normal case, the null deviance is just $\sum_{i=1}^n (Y_i - \overline{Y})^2$.

Thus the residual deviance plays the same role for GLM's as does the residual sum of squares for linear models.

3.2 Residuals

The analogy, and the significance of the normal example can be taken further. The square roots of the individual terms in the residual sum of squares are of course the residuals, $Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}$. We use the square roots of the individual terms in the residual deviances in the same way. What sign to use?
$+$ if Y_i is more than the fitted value $\widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})$,
$-$ otherwise. Thus we call the quantities
$\sqrt{D_i}\text{sign}(Y_i - \widehat{\mu}_i)$ the *deviance residuals*.

Interpreting D_i

Use Taylor's theorem to expand b and b' about $\widehat{\theta_i}$:

$$b(\widetilde{\theta_i}) \doteq b(\widehat{\theta_i}) + (\widetilde{\theta_i} - \widehat{\theta_i})b'(\widehat{\theta_i}) + \frac{1}{2}(\widetilde{\theta_i} - \widehat{\theta_i})^2 b''(\widehat{\theta_i})$$

$$b'(\widetilde{\theta_i}) \doteq b'(\widehat{\theta_i}) + (\widetilde{\theta_i} - \widehat{\theta_i})b''(\widehat{\theta_i})$$

But $b'(\widetilde{\theta_i}) = Y_i$, $b'(\widehat{\theta_i}) = \widehat{\mu}_i$ and $b''(\widehat{\theta_i}) = V(\widehat{\mu}_i)$, so (a):

$$b(\widetilde{\theta_i}) \doteq b(\widehat{\theta_i}) + (\widetilde{\theta_i} - \widehat{\theta_i})\widehat{\mu}_i + \frac{1}{2}(\widetilde{\theta_i} - \widehat{\theta_i})^2 V(\widehat{\mu}_i)$$

and (b):

$$Y_i \doteq \widehat{\mu}_i + (\widetilde{\theta_i} - \widehat{\theta_i})V(\widehat{\mu}_i)$$

Now

$$D_i = 2\{Y_i(\widetilde{\theta_i} - \widehat{\theta_i}) - b(\widetilde{\theta_i}) + b(\widehat{\theta_i})\}$$

$$\doteq 2\{Y_i(\widetilde{\theta_i} - \widehat{\theta_i}) - (\widetilde{\theta_i} - \widehat{\theta_i})\widehat{\mu}_i - \frac{1}{2}(\widetilde{\theta_i} - \widehat{\theta_i})^2 V(\widehat{\mu}_i)\}$$

$$\doteq (\widetilde{\theta_i} - \widehat{\theta_i})^2 V(\widehat{\mu}_i) \doteq \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

using (a), (b) and (b) again.

So D_i is like a standardised squared residual. The signed square root of this approximation is called the *Pearson residual*. Following fitting of a GLM in **R**, and assignment of the result to a variable `fit`, say, you can extract the deviance and Pearson residuals by `resid(fit)` and `resid(fit,type="pearson")` respectively.

The distribution theory for this is very approximate, but a rule of thumb is that *if the model fits*, the scaled deviance D/ϕ (or in practice $D/\widehat{\phi}$) $\approx \chi^2_{n-p}$, and that if $\phi = 1$, the deviance residuals $\approx N(0, 1)$.

Diagnostic plots

These are useful diagnostics: the command `plot(fit)` produces 4 diagnostic plots, which have exactly the same form as with linear models, except that deviance residuals are used instead of ordinary residuals, and linear predictor values instead of fitted values.

3.3 Standard errors of estimates

Recall from Statistics 2 that in large samples, maximum likelihood estimators are approximately normally distributed:

$$\widehat{\theta} \approx N\left(\theta, \frac{1}{nI(\theta)}\right)$$

where $I(\theta)$ is (Fisher's) information about θ *in a single observation*.

The general form of this result applied to GLM's says that

$$\widehat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, (I(\boldsymbol{\beta}))^{-1})$$

where now $I(\boldsymbol{\beta})$ is the information *in the whole sample*, a $p \times p$ matrix. We have

$$(I(\boldsymbol{\beta}))_{jk} = E\left(-\frac{\partial^2 \ell}{\partial\beta_j \partial\beta_k}\right) = X^T W X.$$

(see slides 33/34). Thus

$$\widehat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, (X^T W X)^{-1})$$

where W is evaluated at the MLE $\widehat{\boldsymbol{\beta}}$.

This is the basis for the estimated standard errors in the **R** output.

3.4 The analysis of deviance

How can we test hypotheses about models, and in particular decide which explanatory variables to include? We need an analogue of the analysis of variance (ANOVA).

Let us concentrate on the simplest case, of testing a full vs. a reduced model. Partition the model matrix X and the parameter vector $\boldsymbol{\beta}$ conformably as

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},$$

where X_1 is $n \times q$ and $\boldsymbol{\beta}_1$ is $q \times 1$. The full model is $\boldsymbol{\eta} = X\boldsymbol{\beta} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2$, the reduced model is $\boldsymbol{\eta} = X_1\boldsymbol{\beta}_1$. We wish to test $H_0 : \boldsymbol{\beta}_2 = 0$, i.e. that the reduced model is adequate for the data.

Let the residual deviances for the full and reduced models be D_F and D_R, respectively. Note that

$D_F \leq D_R$; how much less indicates how inadequate the reduced model is. This suggests that $D_R - D_F$ could be the basis for a test statistic for testing H_0.

Now

$$\frac{D_R}{\phi} = 2\{\ell(\widetilde{\boldsymbol{\theta}}) - \sup_{\boldsymbol{\beta}_2=0} \ell(\boldsymbol{\theta})\}$$

while

$$\frac{D_F}{\phi} = 2\{\ell(\widetilde{\boldsymbol{\theta}}) - \sup \ell(\boldsymbol{\theta})\}$$

so

$$\frac{D_R - D_F}{\phi} = 2\log\left\{\frac{\sup L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\beta}_2=0} L(\boldsymbol{\theta})}\right\} = \; -2\log\Lambda$$

where Λ is the likelihood ratio statistic.

It can be shown (Statistics 2) that under the reduced model,

$$-2\log\Lambda \approx \chi^2_{p-q}$$

So we can conduct a test of the adequacy of the reduced model by referring

$$-2 \log \Lambda = \frac{D_R - D_F}{\phi}$$

to χ^2_{p-q}, rejecting H_0 if the statistic is too large (P-value too small).

If ϕ is not present in the model – that's it. If ϕ is present, we either substitute $\widehat{\phi} = D_F/(n-p)$ (or use the estimator on slide 38) and continue to use χ^2_{p-q} or (better) test

$$\frac{D_R - D_F}{p-q} \div \frac{D_F}{n-p} \quad \text{against} \quad F_{(p-q),(n-p)}$$

just as in the normal case.

In **R** these two tests can be performed respectively by `anova(fit,test="Chisq")` and `anova(fit,test="F")`, where `fit` is the variable output by the `glm` command.

Analysis of Deviance in R

As with linear models, whenever you fit a GLM in **R**, the `anova()` function also computes the residual deviance and residual degrees of freedom for each of the simpler models obtained by adding the terms in the linear predictor one-by-one.

Thus, the successive differences give the "$D_R - D_F$" differences relevant to testing the significance of adding that single term. Note that the order matters in interpreting the results. The table is interpreted sequentially, and tests are only valid between nested pairs of models. Each test assumes that all terms earlier in the sequence (higher up in the table) are already included.

Some possible models for `rotenone`

	R		no.	residual	
Words	syntax	η	pars	d.f.	dev.
Null		α	1	16	350.07
single line	`Logdose`	$\alpha + \delta x_i$	2	15	71.78
3 constants	`Poison`	α_{s_i}	3	14	332.07
parallel lines	`Poison+Logdose`	$\alpha_{s_i} + \delta x_i$	4	13	31.97
separate lines	`Poison*Logdose`	$\alpha_{s_i} + \delta_{s_i} x_i$	6	11	22.72
Saturated		no restriction	17	0	0

In this case, every test is significant: in particular, testing the 4 vs. 6 parameter models: $D_R - D_F = 31.97 - 22.72 = 9.25$ which is very significant (P=0.01) when referred to the χ^2_2 distribution. Interestingly, the F test gives a different answer in this case: P=0.13.

No model less complicated than the separate lines model is adequate for the data.

Goodness of fit. Checking $D_F/\phi = D_F = 22.72$ against χ^2_{n-p} (slide 46), we get evidence that the 6-parameter model still does not fit: P=0.025, which is a possible explanation for the contradiction between the χ^2 and F tests above.

4. Tables of counts, and loglinear models

This is a very common type of data in the social sciences, especially: a contingency table, individuals cross-classified according to 2 or more categorical variables, which may each be thought of as explanatory or response variables.

4.1 Sampling designs and distributions

How can such data arise? Take the 2-dimensional case, an $I \times J$ table, with cells indexed by (i, j).

(a) Subjects arise at random, the study continues until a fixed time elapses. Each subject is categorised according to two variables. Suppose the number in cell (i, j) is Y_{ij}, then it is reasonable to assume $Y_{ij} \sim \text{Poisson}(\mu_{ij})$ for some $\{\mu_{ij}\}$, which will be the focus of study.

(b) The total number of subjects is fixed at n, say.

The numbers in cells follow a multinomial distribution: $(Y_{ij}) \sim M(n; (\pi_{ij}))$:

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}$$

if $\sum_i \sum_j y_{ij} = n$, 0 otherwise.

(c) One margin is fixed: say $\{y_{+j}\}$ for each $j = 1, 2, \ldots, J$. In each column, we have an independent multinomial sample

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \prod_{j=1}^{J} \left(\frac{y_{+j}!}{\prod_{i=1}^{I} y_{ij}!} \prod_{i=1}^{I} \rho_{ij}^{y_{ij}} \right)$$

where $\rho_{ij}$is the probability that a column-j individual is in row i (so $\rho_{+j} = 1$).

There are more, but similar, options for a higher-order table (3 or more dimensions).

You cannot tell which of these models is appropriate by just looking at the numbers – you need to know how they were collected! But fortunately the 3 possibilities lead to essentially the same analysis.

Theorem. Let $Y_1, Y_2, \ldots, Y_k$ be random variables with $Y_+ = \sum_{i=1}^k Y_i$.

Then (i) $Y_1, Y_2, \ldots, Y_k$ are independent Poisson(λ_i)

$\iff$ (ii) $Y_+ \sim$ Poisson(λ_+) and

$$(Y_1, Y_2, \ldots, Y_k | Y_+ = n) \sim M(n; (\pi_1, \pi_2, \ldots, \pi_k))$$

where $\lambda_+ = \sum_{i=1}^k \pi_i$ and $\pi_i = \lambda_i / \lambda_+$.

Proof. Suppose (i). Recall $\sum_i Y_i \sim$ Poisson$(\sum_i \lambda_i)$. Then

$$P(Y_i = y_i \forall i | Y_+ = n) = \frac{P(Y_i = y_i \forall i, Y_+ = n)}{P(Y_+ = n)}$$

$$= \left(\prod_i \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} I[y_+ = n] \right) \div \left(\frac{e^{-\lambda_+} \lambda_+^n}{n!} \right)$$

$$= \frac{n!}{\prod_i y_i!} \prod_i \pi_i^{y_i} I[y_+ = n]$$

as required.

Now suppose (ii). Then

$$P(Y_i = y_i \forall i) = \sum_n \frac{e^{-\lambda_+}\lambda_+^n}{n!} I[y_+ = n] \frac{n!}{\prod_i y_i!} \prod_i \pi_i^{y_i}$$

$$= \frac{e^{-\lambda_+}\lambda_+^{y_+}}{y_+!} \frac{y_+!}{\prod_i y_i!} \prod_i \pi_i^{y_i} = \prod_{i=1}^{k} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

as required.

4.2 Loglinear models for means

In the three sampling scenarios above:

(a) $E(Y_{ij}) = \mu_{ij}$
If row and column categories were independent, we would have

$$\mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{\mu_{++}}$$

(b) $E(Y_{ij}) = n\pi_{ij}$
If factors were independent, $\pi_{ij} = \pi_{i+}\pi_{+j}$

(c) $E(Y_{ij}) = y_{+j}\rho_{ij}$

If the distribution of the response (i) were the same for all j, we would have $\rho_{ij} = \rho_i$, say, so $E(Y_{ij}) = y_{+j}\rho_i$.

In *each* case, a *multiplicative* structure for the means μ_{ij} is natural – a simplifying hypothesis. That is the same as an additive model for $\log \mu_{ij}$. So we are interested in whether the *loglinear model*

$$\log \mu_{ij} = \theta + \alpha_i + \beta_j$$

(where $\sum_{i=1}^{I} \alpha_i = 0 = \sum_{j=1}^{J} \beta_j$) is a good fit to the data. We call θ the overall mean, and $\alpha_i$and $\beta_j$the main effects.

What about interaction? We can always fit any table of numbers $\log \mu_{ij}$ perfectly if we allow interaction as well:

$$\log \mu_{ij} = \theta + \alpha_i + \beta_j + \gamma_{ij},$$

by setting

$$\gamma_{ij} = \log \mu_{ij} - I^{-1} \sum_i \log \mu_{ij} - J^{-1} \sum_j \log \mu_{ij}$$

$$+ (IJ)^{-1} \sum_{ij} \log \mu_{ij}$$

(and θ, α_i and β_j accordingly), so this is the *saturated model*.

We look at higher order tables (3 or more dimensions) a little later.

4.3 GLM formulation

In scenario (a) we explicitly have a Poisson distribution for the data, and a loglinear model gives us a log link to the linear predictor, so this is a GLM (`family=poisson(log)`). A generic example would be
`fit_glm(Counts~Row+Column,poisson,data)`
where `Counts` are the actual counts, and `Row` and `Column` are the factors classifying rows and columns, all in the dataframe `data`.

What about scenarios (b) and (c)?

Because of the Poisson/multinomial connection (Theorem on slide 56), we can do *exactly the same*!

The rule is that we must include in the linear predictor factors (and interactions) corresponding to all marginal totals that are fixed by design. This ensures that all such totals fit perfectly, and that estimates for all other effects are correct, even for the multinomial model. E.g. if we had a 3-way $I \times J \times K$ table `Age×Sex×Smoking` and we had sampled separately for each `Age×Sex` combination, so that Y_{ij+} was fixed for each (i, j), we would have to include `Age*Sex` in the model.

4.4 Deviance for a Poisson loglinear model

$$D_i = 2\{Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\}$$
$$= 2\{Y_i(\log Y_i - \log \widehat{\mu}_i) - Y_i + \widehat{\mu}_i\}$$

since $b(\theta) = \exp(\theta) = \mu$, $\theta_i = \log\mu_i$, and $\widetilde{\theta}_i = \log Y_i$. So

$$D_i = 2\{Y_i \log\frac{Y_i}{\widehat{\mu}_i} - Y_i + \widehat{\mu}_i\}$$

If you expand the logarithm about 1 in powers of $(Y_i - \widehat{\mu}_i)/\widehat{\mu}_i$, you find

$$D_i \doteq \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

a special case of the approximation on slide 45.

$$\sum_i \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i} = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

is known as Pearson's χ^2 statistic, a familiar tool from elementary statistics for testing for 'independence' in a two-way table. Thus we have a new justification for this procedure – the test is a quadratic approximation to the likelihood ratio test for goodness-of-fit in a Poisson loglinear model.

4.5 Models for higher order tables

In a 2-way table, if you fit an interaction term, you immediately get a saturated model – so no very interesting structure can be uncovered.

But in 3 or more dimensions, some low-order interactions can be fitted, without over-fitting the data. For example, in 3-way tables, consider the loglinear model:

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik} + \phi_{jk} + \psi_{ijk}$$

subject to all the obvious constraints, for example $\epsilon_{i+} = 0 \forall i$ and $\psi_{+jk} = 0 \forall j, k$. This is saturated. But if we fit only

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik}$$

then we are making a statement about the structure of the independence in the table.

For definiteness, suppose we are in sampling scenario (b) – that is $n = Y_{+++}$ is fixed.

Consider all three factors, Y_1, Y_2, Y_3 say, to be random responses. Then

$$P(Y_1 = i, Y_2 = j, Y_3 = k) = \pi_{ijk} = \frac{\mu_{ijk}}{n}$$

$$= \frac{1}{n}\exp(\theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik})$$

Now look at

$$P(Y_2 = j, Y_3 = k | Y_1 = i) = \frac{P(\boldsymbol{Y} = (i,j,k))}{\sum_j \sum_k P(\boldsymbol{Y} = (i,j,k))}$$

$$= \frac{e^{\beta_j + \delta_{ij} + \gamma_k + \epsilon_{ik}}}{\sum_j \sum_k e^{\beta_j + \delta_{ij} + \gamma_k + \epsilon_{ik}}} = \frac{e^{\beta_j + \delta_{ij}}}{\sum_j e^{\beta_j + \delta_{ij}}} \times \frac{e^{\gamma_k + \epsilon_{ik}}}{\sum_k e^{\gamma_k + \epsilon_{ik}}}$$

(factoring out $n^{-1}\exp(\theta + \alpha_i)$)

$$= P(Y_2 = j | Y_1 = i) \times P(Y_3 = k | Y_1 = i)$$

that is, Y_2 and Y_3 are *conditionally independent* given Y_1. This would not have been true if either ϕ_{jk} or ψ_{ijk} had been included in the linear predictor.

5. Survival analysis

For each of a set of individuals, we record a *failure time* or *lifetime*, e.g.

- lifetime of machine component
- time until patient's cure, remission, death
- time for subject to perform a test
- duration of an economic cycle

– may be measured in non-time units:

- length of run of particle track
- amount of usage of machine (work/fuel)

For different individuals, time origins may differ, and because of missing observations, histories may be incomplete, either:

- censored (exact time not recorded), or
- truncated (the individual is not recorded at all)

5.1 Notation

Index observations by $i = 1, 2, \ldots, n$. For each i, let T_i be the *true* lifetime, and c_i the *potential* censoring time. What we observe are

$$Y_i = \min(T_i, c_i),$$

together with the indicator

$$\delta_i = \begin{cases} 1 & T_i \leq c_i \text{ (uncensored)} \\ 0 & T_i > c_i \text{ (censored)} \end{cases}$$

The general approach is to model the distribution of T_i, including its dependence on explanatory variables, and use observations on (Y_i, δ_i) to make inference about unknown parameters in this model.

5.2 Distribution theory

Because of the kinds of question we have to answer, it is useful to develop some non-standard notation to describe the distributions of lifetimes, non-negative random variables T.

Survivor function

$$\mathcal{F}(t) = P\{T \geq t\}$$

(note $\geq$ not $>$, so unless the distribution is continuous, this is not quite $1 - F$)

Density function

In the continuous case,

$$f(t) = -\frac{d}{dt}\mathcal{F}(t)$$

Hazard function

(also called the age-specific failure rate)

$$\begin{aligned} h(t) &= \lim_{\Delta \to 0} \frac{P\{t \le T < t + \Delta | T \ge t\}}{\Delta} \\ &= \lim_{\Delta \to 0} \frac{P\{t \le T < t + \Delta\}}{\Delta P\{T \ge t\}} \\ &= \frac{f(t)}{\mathcal{F}(t)} = -\frac{d}{dt} \log \mathcal{F}(t) \end{aligned}$$

Cumulative hazard function

Now, $\log \mathcal{F}(t) = \log \mathcal{F}(t) - \log \mathcal{F}(0) = -\int_0^t h(u)du$, in other words: $\mathcal{F}(t) = \exp(-H(t))$, where $H(t)$ is the *cumulative* or *integrated hazard function* $\int_0^t h(u)du$. Thus $f(t) = h(t)\exp(-H(t))$.

Each of these 4 functions determine the others.

Examples

Exponential (ρ) **distribution** $\quad \mathcal{F}(t) = e^{-\rho t}$, $f(t) = \rho e^{-\rho t}$, $h(t) = \rho$, $H(t) = \rho t$, all for $t > 0$.

Note the constant hazard rate ($\equiv$ lack of memory). The coefficient of variation, =standard deviation $\div$ mean, is $\rho^{-1}/\rho^{-1} = 1$.

Very widely used basic model, especially in industrial applications, but it is very inflexible, only one parameter giving overall scale, no variation in shape or relative spread, cannot model 'aging'.

Gamma (κ, ρ) **distribution**

$$f(t) = \rho^{\kappa} t^{\kappa-1} e^{-\rho t}/\Gamma(\kappa), \quad t > 0.$$

Mean is κ/ρ, variance κ/ρ^2, so coefficient of variation is $\kappa^{-1/2}$.

More flexible. However, there is no explicit form for $\mathcal{F}(t)$, $h(t)$ or $H(t)$, except for small integer κ, so rarely used in survival analysis.

Weibull (κ, ρ) distribution

$$\mathcal{F}(t) = \exp\{-(\rho t)^{\kappa}\}$$

$$f(t) = \kappa\rho(\rho t)^{\kappa-1}\exp\{-(\rho t)^{\kappa}\}$$

$$h(t) = \kappa\rho(\rho t)^{\kappa-1} \quad \text{and} \quad H(t) = (\rho t)^{\kappa}$$

– the hazard function showing that this is a natural generalisation of the Exponential distribution.

5.3 Parametric statistical inference

Suppose, initially, that we observe i.i.d. pairs $(Y_i, \delta_i), i = 1, 2, \ldots, n$, from the same distribution given by f, $\mathcal{F}$, h or H, depending on a parameter θ. How can we estimate θ? MLE?

What is the likelihood?

If $\delta_i = 1$, then $Y_i = T_i$, and the contribution to the likelihood from this observation is $f(T_i; \theta) = f(Y_i; \theta)$.

If $\delta_i = 0$, then $Y_i = c_i$, and the contribution to the likelihood comes from the probability of the event that $\{T_i > c_i\}$, i.e. $P\{T_i > c_i\} = \mathcal{F}(c_i; \theta) = \mathcal{F}(Y_i; \theta)$.

So the overall likelihood L is

$$\prod_{i:\delta_i=1} f(Y_i;\theta) \prod_{i:\delta_i=0} \mathcal{F}(Y_i;\theta)$$

$$= \prod_{i:\delta_i=1} h(Y_i;\theta) \prod_{i=1}^{n} \mathcal{F}(Y_i;\theta)$$

and the log-likelihood is

$$\ell(\theta) = \sum_{i:\delta_i=1} \log h(Y_i;\theta) - \sum_{i=1}^{n} H(Y_i;\theta)$$

We will write $\sum_{i:\delta_i=1}$ as $\sum_U$.

Exponential case

$$\ell(\rho) = \sum_U \log \rho - \sum_{i=1}^n \rho Y_i$$

so $\ell'(\rho) = m/\rho - \sum Y_i$, giving $\widehat{\rho} = m/\sum Y_i$, where m is the number of *un*censored observations. When $m = n$ this is the usual answer.

Asymptotic distributions and standard errors
As usual (e.g. slide 47), we have

$$\widehat{\theta} \approx N\left(\theta, \frac{1}{I(\theta)}\right)$$

where $I(\theta)$ is Fisher's information *in the whole sample*. In this example, $\ell''(\rho) = -m/\rho^2$, so $I(\rho) = m/\rho^2$, and

$$\widehat{\rho} \approx N(\rho, \frac{\rho^2}{m})$$

whence by the usual (S2) argument, we have an approximate 95% confidence interval for ρ given by

$$\left\{\rho : \sqrt{m}\left|\frac{\widehat{\rho}-\rho}{\rho}\right| < 1.96\right\}$$

$$= \left[\frac{\widehat{\rho}}{1+1.96/\sqrt{m}}, \frac{\widehat{\rho}}{1-1.96/\sqrt{m}}\right]$$

For example, for the leukaemia data (data set A), we have:
Control group: $n = 21$, $\sum Y_i = 182$, no censoring,
$\widehat{\rho} = 21/182 = 0.115$, s.e. = $0.115/\sqrt{m} = 0.025$.
CI = [0.081,0.201].
Treatment group: $m = 9$, $\sum Y_i = 359$,
$\widehat{\rho} = 9/359 = 0.025$, s.e. = $0.025/\sqrt{m} = 0.008$.
CI = [0.015,0.072].

Although this has not been done as a formal test, there is obviously significant evidence that the drug is effective.

5.4 Explanatory variables

Let us move from 'single sample' to 'regression' case: T_i now has a distribution specified by, say, a hazard function $h(t; \ldots)$ depending on both parameters and explanatory variables $\boldsymbol{x}$, and we wish to analyse the dependency on $\boldsymbol{x}$.

As usual, explanatory variables may be discrete (factors) or continuous, and may represent:

- treatment variables (e.g. drug)
- individual characteristics (e.g. age, sex)
- environmental conditions (e.g. hospital)

We will not discuss this here, but survival analysis also has to deal with *time-dependent* explanatory variables – e.g. treatment variables that change with time.

There are many models for regression analysis for survival data; in general we want

$$T_i \sim h(t; \boldsymbol{x}_i, \boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ is to be estimated, to quantify the dependence. Two main approaches:

- Proportional hazards (PH) – the effect of $\boldsymbol{x}$ is to scale up or down the hazard function
- Accelerated life (AL) – the effect of $\boldsymbol{x}$ is to speed up or slow down time

5.5 Proportional hazards model

Suppose the effect of $\boldsymbol{x}$ is summarised by a one-dimensional non-negative *hazard ratio* function $\psi(\boldsymbol{x})$: we suppose

$$h(t; \boldsymbol{x}) = \psi(\boldsymbol{x}) h_0(t)$$

where $h_0(t)$ is a fully-specified *baseline* hazard. We choose the scale of measurement for $\boldsymbol{x}$ so that $\psi(0) = 1$, i.e. $h_0(t) = h(t; 0)$.

It follows that

$$
\begin{aligned}
H(t;\boldsymbol{x}) &= \psi(\boldsymbol{x})H_0(t)\\
\mathcal{F}(t;\boldsymbol{x}) &= (\mathcal{F}_0(t))^{\psi(\boldsymbol{x})}\\
f(t;\boldsymbol{x}) &= \psi(\boldsymbol{x})(\mathcal{F}_0(t))^{\psi(\boldsymbol{x})-1}f_0(t)
\end{aligned}
$$

MLE for the PH model with exponential hazard ratio

A common choice is $\psi(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ to be estimated.

As in single-sample case, the log-likelihood is:

$$\ell(\boldsymbol{\beta}) = \sum_U \log h - \sum H$$

Substituting $h(t;\boldsymbol{x}) = e^{\boldsymbol{x}^T\boldsymbol{\beta}}h_0(t)$ and $H(t;\boldsymbol{x}) = e^{\boldsymbol{x}^T\boldsymbol{\beta}}H_0(t)$, we get

$$\ell(\boldsymbol{\beta}) = \sum_U \boldsymbol{x}_i^T\boldsymbol{\beta} + \sum_U \log h_0(Y_i) - \sum_{i=1}^n e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}H_0(Y_i)$$

whence the likelihood equations are

$$\frac{\partial \ell}{\partial \beta_j} = \sum_U x_{ij} - \sum_{i=1}^{n} x_{ij} e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} H_0(Y_i) = 0$$

There is no explicit solution in general, but there is in special cases. For example, suppose the observations fall into k disjoint groups with $x_{ij} = 1$ if i is in group j, 0 otherwise. Let m_j be the number of uncensored observations in group j, that is $m_j = \sum_U x_{ij}$. Then the likelihood equations become

$$\frac{\partial \ell}{\partial \beta_j} = m_j - \sum_{i:x_{ij}=1} e^{\beta_j} H_0(Y_i) = 0$$

so

$$\widehat{\beta}_j = \log \left[m_j / \sum_{i:x_{ij}=1} H_0(Y_i) \right].$$

cf. slide 71.

Another case that can be solved explicitly is where there is a single explanatory variable x that takes only non-negative integer values. Then $d\ell/d\beta$ is just a polynomial in e^{β} and may be solvable.

But in general, we need to use numerical methods.

5.6 An artificial GLM

Recall $h(t;\boldsymbol{x}) = e^{\boldsymbol{x}^T\boldsymbol{\beta}} h_0(t)$, and that

$$\ell(\boldsymbol{\beta}) = \sum_U \boldsymbol{x}_i^T\boldsymbol{\beta} + \sum_U \log h_0(Y_i) - \sum_{i=1}^n e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} H_0(Y_i)$$

Imagine now an artificial experiment, with responses $\delta_i, i = 1, 2, \ldots, n$ that happen to be all 0 or 1, but are modelled as independent Poisson with means $\mu_i = e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} H_0(Y_i)$. Note we are using the same symbols and notation for a *different* setup.

In the artificial experiment, the log-likelihood is

$$\begin{aligned}
\ell^{\star}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (\delta_i \log \mu_i - \mu_i - \log \delta_i!) \\
&= \sum_{U} \boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_{U} \log H_0(Y_i) - \sum_{i=1}^{n} e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} H_0(Y_i) \\
&= \ell(\boldsymbol{\beta}) - \sum_{U} \log \frac{h_0(Y_i)}{H_0(Y_i)}
\end{aligned}$$

So ℓ and $\ell^{\star}$ differ only by something that does not depend on $\boldsymbol{\beta}$. Thus

$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell^{\star}}{\partial \beta_j}$$

and the likelihood equations and hence the MLE's are the same in the two models!

How can we fit the artificial model?

The responses are assumed Poisson distributed with

$$\log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \log H_0(Y_i)$$

– it is a Poisson GLM with log link, but an extra term $\log H_0(Y_i)$, which is completely known once the data are collected, in the linear predictor. Such an extra term is called an *offset*, and the **R** `glm()` function allows an offset to be included, through using the `offset` function in the model formula.

For example, if we have a dataframe *data* with variables `Age`, `Sex`, and the observed (Y_i, δ_i) are in `Time`and `Delta`, we could write

`glm(Delta~Age+Sex+offset(log(`H_0`(Time))),`
`poisson,`*data*`)`

where of course H_0 is replaced by a suitable **R** expression, to fit the artificial model.

From the derivation above, we have also fitted the PH model (with exponential hazard ratio).

5.7 Accelerated life model

An individual with explanatory variables $\boldsymbol{x}$ is assumed to experience time speeded up by a non-negative factor $\xi(\boldsymbol{x})$, where we suppose $\xi(0) = 1$, i.e. $\boldsymbol{x} = 0$ represents the baseline again. Thus:

$$\begin{aligned} \mathcal{F}(t;\boldsymbol{x}) &= \mathcal{F}_0(\xi(\boldsymbol{x})t) \\ f(t;\boldsymbol{x}) &= \xi(\boldsymbol{x})f_0(\xi(\boldsymbol{x})t) \\ h(t;\boldsymbol{x}) &= \xi(\boldsymbol{x})h_0(\xi(\boldsymbol{x})t) \end{aligned}$$

If there were only a small number of possible values for $\xi(\boldsymbol{x})$, either through $\boldsymbol{x}$ being very discrete, or because of the assumed form for ξ, we could just take the unique values of $\xi(\boldsymbol{x})$ as parameters, and estimate these. (True also in PH case).

Otherwise, we would assume a parametric form for ξ and estimate the parameters. As with the PH model, a natural choice is $\xi(\boldsymbol{x}) = e^{\boldsymbol{x}^T\boldsymbol{\beta}}$.

Fitting the AL model with exponential speed-up

$$\mathcal{F}(t;\boldsymbol{x}) = \mathcal{F}_0(\xi(\boldsymbol{x})t)$$

If T is the true lifetime for an individual with explanatory variables $\boldsymbol{x}$, define a new random variable W by $W = \log T + \boldsymbol{x}^T\boldsymbol{\beta}$. Its distribution function is:

$$\begin{aligned} P\{W < w\} &= P\{\log T < w - \boldsymbol{x}^T\boldsymbol{\beta}\} \\ &= 1 - P\{T \geq \exp(w - \boldsymbol{x}^T\boldsymbol{\beta})\} \\ &= 1 - \mathcal{F}(e^{w-\boldsymbol{x}^T\boldsymbol{\beta}}) = 1 - \mathcal{F}_0(e^w) \end{aligned}$$

Thus, W has a distribution that is independent of $\boldsymbol{x}$, and indeed is completely known if we assume the baseline is fully specified.

Suppose $E(W) = \mu_0$, then we can write this as

$$\log T_i = \mu_0 - \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$$

where the ε_i are iid random variables with mean 0.

This is just a linear model, so in the case where there is no censoring, we could just use least squares to estimate $\boldsymbol{\beta}$ – Gauss-Markov says this is optimal (MVLUE) whatever the distribution of W.

If there *is* censoring, there are more complicated algorithms for censored linear models, or use Newton-Raphson for solving the likelihood equations – this time there is no connection to GLM.

5.8 The Proportional Odds model

Suppose that the baseline distribution is the so-called *log-logistic*:

$$\mathcal{F}_0(t) = \frac{1}{1 + (\rho t)^\kappa}$$

and that we have an AL model with $\xi(\boldsymbol{x}) = e^{\boldsymbol{x}^T \boldsymbol{\beta}}$. For some fixed threshold time t, suppose we call $\{T < t\}$ a 'success', and let us write π for the probability of success. Then the log-odds satisfy:

$$\begin{aligned} \log \frac{\pi}{1-\pi} &= \log \frac{P\{T < t\}}{P\{T \geq t\}} \\ = \log \frac{1 - \mathcal{F}(t; \boldsymbol{x})}{\mathcal{F}(t; \boldsymbol{x})} &= \log \frac{1 - \mathcal{F}_0(\xi(\boldsymbol{x})t)}{\mathcal{F}_0(\xi(\boldsymbol{x})t)} \\ = \log(\xi(\boldsymbol{x})\rho t)^\kappa &= \kappa \boldsymbol{x}^T \boldsymbol{\beta} + \kappa \log \rho t \end{aligned}$$

Thus if we only recorded whether survival times were less than t or not, we would have a Binomial/logit GLM! The only parameter of the GLM that depends on the threshold t is the intercept $\kappa \log \rho t$.

5.9 When are PH and AL the same?

The survivor functions under the two models are:
PH with hazard ratio function ψ:

$$\mathcal{F}(t;\boldsymbol{x}) = (\mathcal{F}_0(t))^{\psi(\boldsymbol{x})}$$

AL with speed-up function ξ:

$$\mathcal{F}(t;\boldsymbol{x}) = \mathcal{F}_0(\xi(\boldsymbol{x})t)$$

In the case where the baseline distribution $\mathcal{F}_0$ is Weibull (ρ, κ), we have (slide 69):

$$\mathcal{F}_0(t) = \exp\{-(\rho t)^\kappa\}$$

So we have

$$\mathcal{F}_{\mathrm{PH}}(t;\boldsymbol{x}) = \mathcal{F}_{\mathrm{AL}}(t;\boldsymbol{x}) = \exp\{-(\rho t)^\kappa \psi(\boldsymbol{x})\}$$

if $\xi(\boldsymbol{x}) \equiv (\psi(\boldsymbol{x}))^{1/\kappa}$.

In fact, it is quite easy to show that this is the *only* case where the two models coincide.