university of groningen

faculty of economics and business

Generalised Linear Models

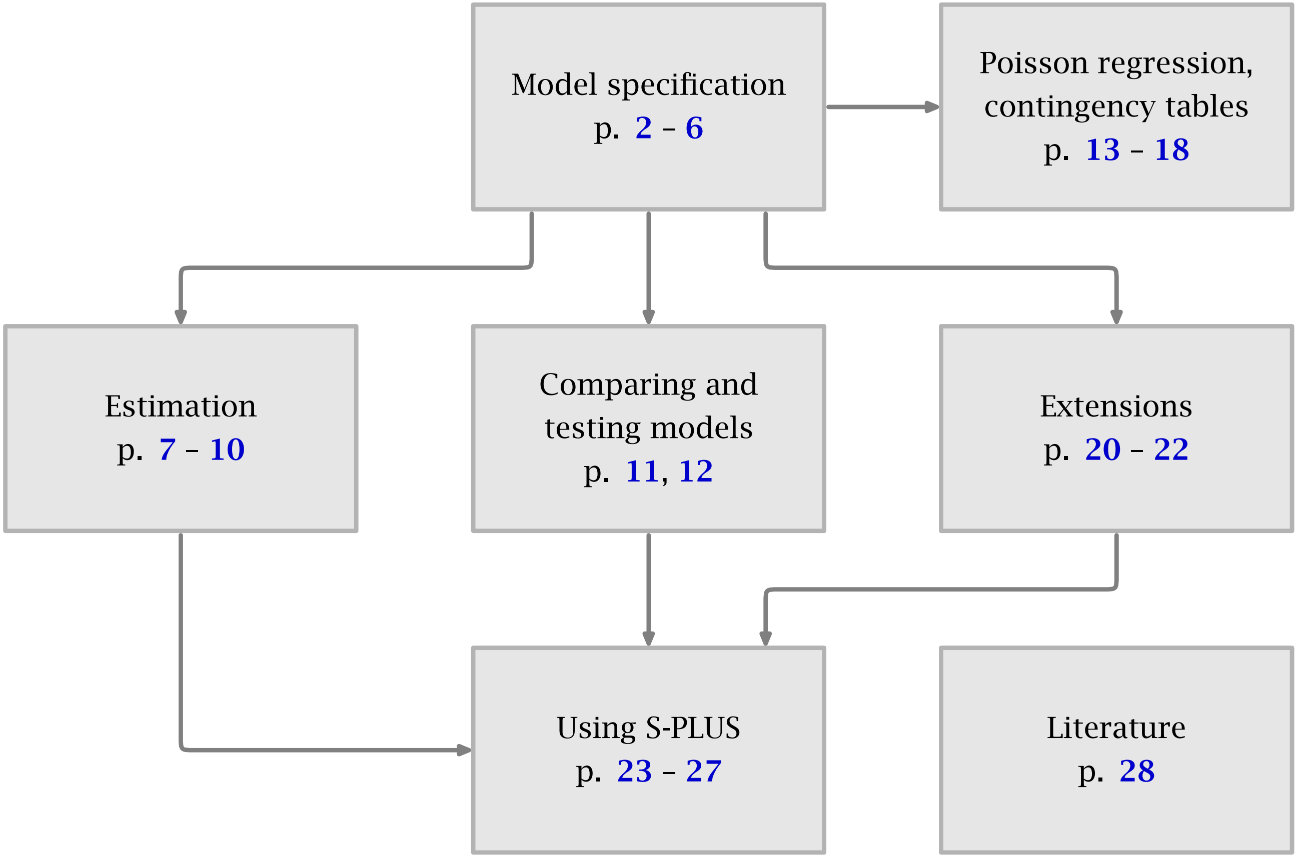
Sytse Knypstra

2008

Version: 090309

⏪ ⏩ ⏮ ◀ ▶ ⏭ ■

Contents



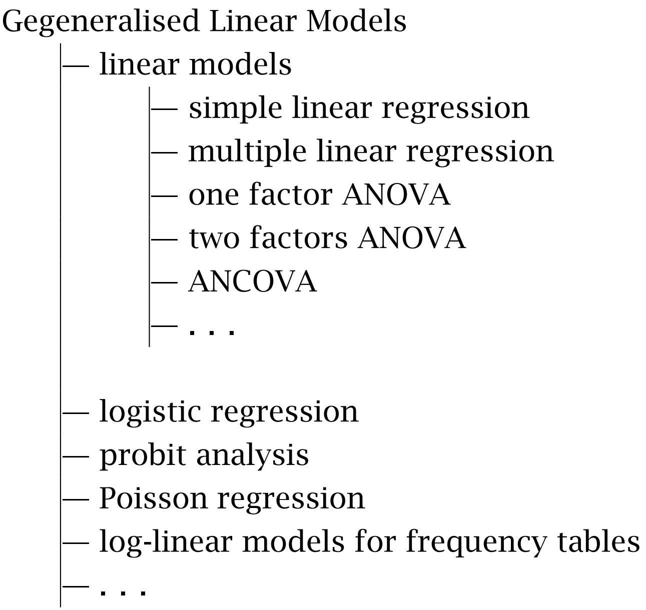
Hierarchy of models

Generalised linear models (GLM) not only encompass **linear models**, but also logistic regression, log-linear and other models.

Inference based on linear models is exact, provided the assumptions are met.

The theory of generalised linear models is based on large sample distributions of (maximum likelihood) statistics.

Therefore results will be valid only approximately.



Gegeneralised Linear Models

- linear models
  - simple linear regression
  - multiple linear regression
  - one factor ANOVA
  - two factors ANOVA
  - ANCOVA
  - . . .
- logistic regression
- probit analysis
- Poisson regression
- log-linear models for frequency tables
- . . .

Generalised linear models vs. linear models

Assumptions for generalised linear models are less severe in two aspects:

Linear models

- random variables $Y_1, \dots, Y_n$,
- mutually independent
- $Y_i \sim$ Normal
- $\mathbb{E}(Y_i) = \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$
  $\mathbb{V}\mathrm{ar}\,(Y_i) = \sigma^2$ (constant)

Generalised linear models

- random variables $Y_1, \dots, Y_n$,
- mutually independent
- The distribution of Y_i is not necessarily normal but has to be a member of the **Exponential Dispersion Family ($\mathcal{EDF}$)**.
- $g(\,\mathbb{E}(Y_i)) = g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$
  or $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$

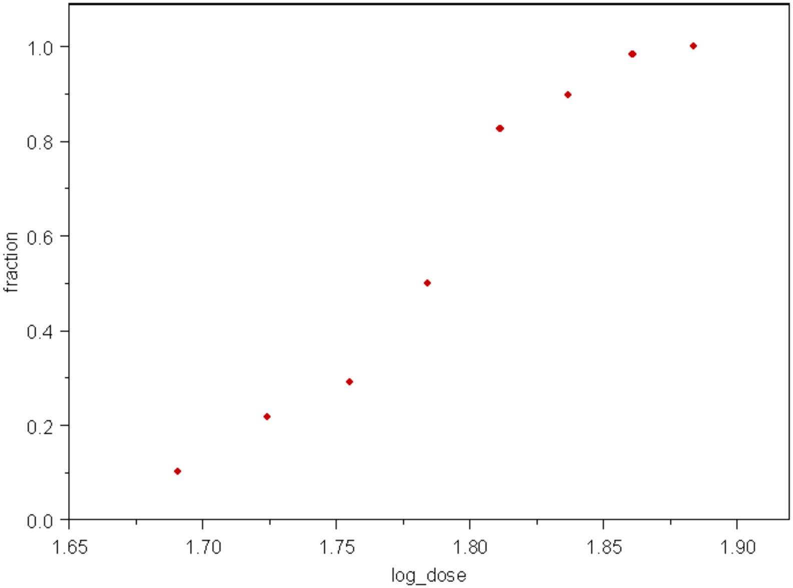
$\mathbb{E}(Y_i)$ is tied to a linear combination $\boldsymbol{x}_i^T \boldsymbol{\beta}$ of the parameters $\boldsymbol{\beta}$ by a monotonous and differentiable function g, not necessarily the identity.

The function g is called the *link function*.
The inverse function g^{-1} is called the *response function*.

Beetles Example

Classical example:

log dose	number of insects	
(x_i)	treated (n_i)	killed
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60



$Y_i \sim$ (Proportional) binomial(n_i, π_i), $i = 1, \ldots, 8$ or $Y_i \sim$ Bernoulli(π_i), $i = 1, \ldots, 481$.

$g(\pi_i) = \ln \dfrac{\pi_i}{1 - \pi_i} = \alpha + \beta x_i$ (link function) or $\pi_i = g^{-1}(\alpha + \beta x_i) = \dfrac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$

$g(\pi_i) = \Phi^{-1}(\pi_i) = \alpha + \beta x_i$ (link function) or $\pi_i = \Phi(\alpha + \beta x_i)$ (response function)

Exponential Dispersion Family

A distribution belongs to the Exponential Dispersion Family ($\mathcal{EDF}$) if its probability mass function or density function can be written as:

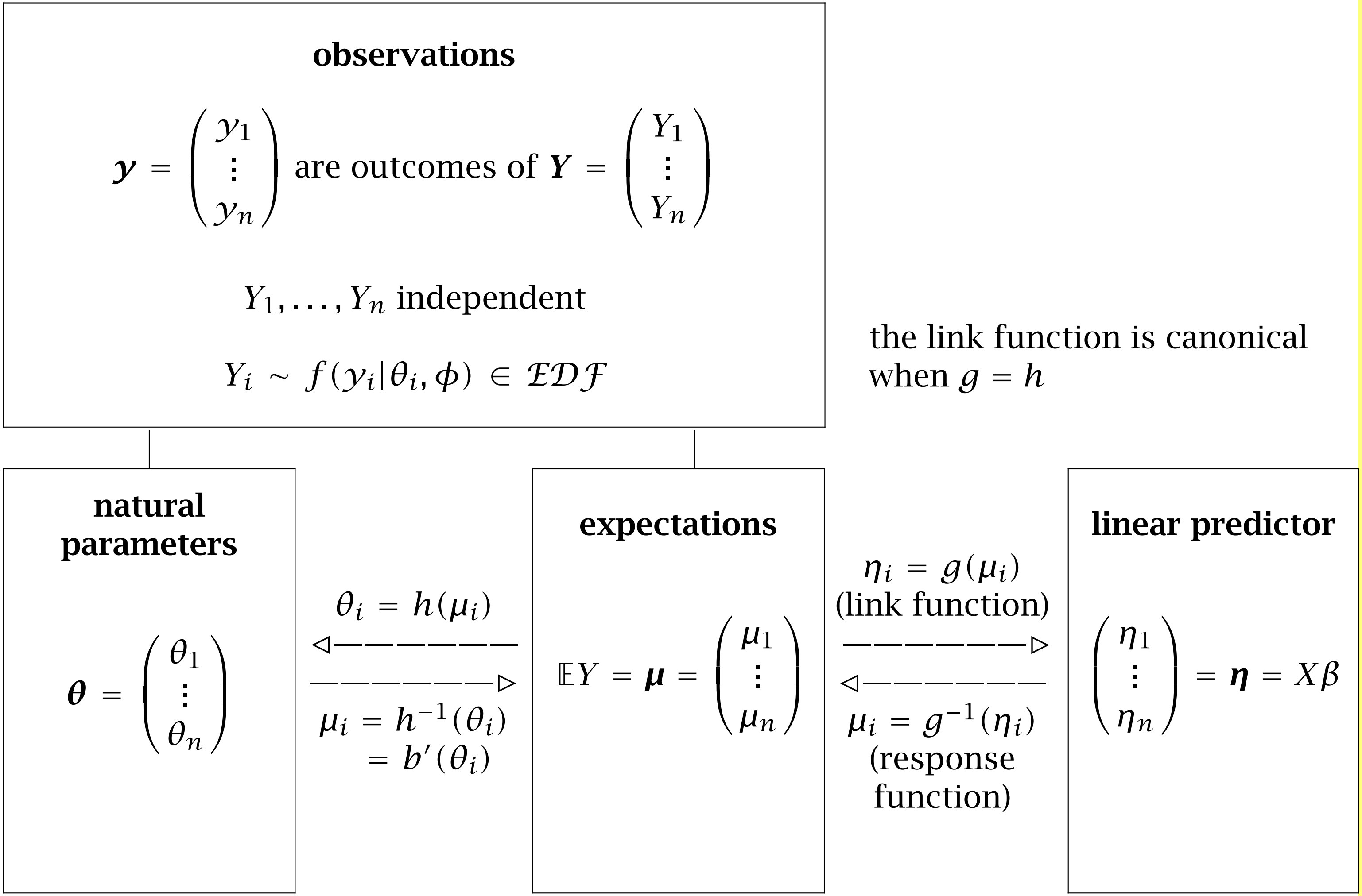
$$f(y_i|\theta_i,\phi) = c_i(y_i,\phi)\exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)}\right].$$

The parameter ϕ is called the dispersion parameter. It is a nuisance parameter not depending on i.

In practice $a_i(\phi)$ has the form: $a_i(\phi) = \frac{\phi}{w_i}$ where the quantities w_i are **weights.**

Distribution of Y_i	θ_i	$a_i(\phi)$	w_i	$b(\theta_i)$	$c_i(y_i,\phi)$	canonical link
Bernoulli(π_i)	$\ln\frac{\pi_i}{1-\pi_i}$	1	1	$\ln(1+e^{\theta_i})$	$I_{\{0,1\}}(y_i)$	logit
Prop.Binom.(n_i,π_i)	$\ln\frac{\pi_i}{1-\pi_i}$	$\frac{1}{n_i}$	n_i	$\ln(1+e^{\theta_i})$	$\binom{n_i}{n_i y_i} I_{\{0,\frac{1}{n_i},\dots,1\}}(y_i)$	logit
Poisson(λ_i)	$\ln\lambda_i$	1	1	e^{θ_i}	$\frac{1}{y_i!} I_{\{0,1,\dots\}}(y_i)$	logarithm
Normal(μ_i,σ^2)	μ_i	σ^2	1	$\frac{1}{2}\theta_i^2$	$\frac{1}{\sqrt{2\pi\phi}} e^{-\frac{y_i^2}{2\phi}}$	identity
Gamma(λ_i,α)	$-\frac{\lambda_i}{\alpha}$	$\frac{1}{\alpha}$	1	$-\ln(-\theta_i)$	$\frac{1}{\Gamma(\frac{1}{\phi})} y_i^{(\frac{1}{\phi}-1)} \phi^{-\frac{1}{\phi}} I_{(0,\infty)}(y_i)$	reciprocal

Basic scheme



observations

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ are outcomes of } \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$Y_1, \dots, Y_n$ independent

$$Y_i \sim f(y_i|\theta_i, \phi) \in \mathcal{EDF}$$

the link function is canonical when $g = h$

natural parameters

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

$\theta_i = h(\mu_i)$

$$\xleftarrow{\hspace{3cm}}$$

$$\xrightarrow{\hspace{3cm}}$$

$\mu_i = h^{-1}(\theta_i)$

$= b'(\theta_i)$

expectations

$$\mathbb{E}Y = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

$\eta_i = g(\mu_i)$

(link function)

$$\xrightarrow{\hspace{3cm}}$$

$$\xleftarrow{\hspace{3cm}}$$

$\mu_i = g^{-1}(\eta_i)$

(response function)

linear predictor

$$\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} = \boldsymbol{\eta} = X\beta$$

Likelihood equations

The parameters $\boldsymbol{\beta}$ are estimated using the maximum likelihood method. The log-likelihood for observations $y_1, \ldots, y_n$ is:

$$\ell(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} \ln c_i(y_i, \phi).$$

The derivative of ℓ with respect to β_j is:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

where

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)} \\ \frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = \frac{\mathbb{V}\mathrm{ar}\,(Y_i)}{a_i(\phi)} \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij}. \end{aligned}$$

This leads to the likelihood equations:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\mathbb{V}\mathrm{ar}\,(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \text{for } j = 1, \ldots, p.$$

Or, in matrix notation:

$$\boldsymbol{X}^T \boldsymbol{V} (\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}$$

where $\boldsymbol{X}$ is a matrix with elements x_{ij} and $\boldsymbol{V}$ a diagonal matrix with $\dfrac{1}{\mathbb{V}\mathrm{ar}\,(Y_i)} \dfrac{\partial \mu_i}{\partial \eta_i}$ as its i-th diagonal element.

These equations can generally not be solved explicitly. Therefore the iterative Newton-Raphson method is used with a modification which gives it the name *Fisher scoring algorithm* or *iterative reweighted least squares (IRWLS) algorithm*.

Remark. In the special case that the link function is canonical ($g = h$ and thus $\theta_i = \eta_i$) and that $a_i(\phi) = \phi$ the likelihood equations remind us of the normal equations in linear models:

$$\boldsymbol{X}^T \hat{\boldsymbol{\mu}} = \boldsymbol{X}^T \boldsymbol{y}.$$

The IRWLS algorithm

In order to find solutions of non-linear equations $\boldsymbol{\psi}(\boldsymbol{\beta}) = \mathbf{0}$ the Newton-Raphson method takes a linear (Taylor series) approximation of $\boldsymbol{\psi}(\boldsymbol{\beta})$ in a neighbourhood of a point $\boldsymbol{\beta}^{(k)}$:

$$\boldsymbol{\psi}(\boldsymbol{\beta}^{(k)}) + \frac{\partial \boldsymbol{\psi}(\boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}).$$

Equating to zero gives us the approximation:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left(\frac{\partial \boldsymbol{\psi}(\boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}^T}\right)^{-1} \boldsymbol{\psi}(\boldsymbol{\beta}^{(k)}).$$

Repeatedly applying this equation starting with $\boldsymbol{\beta}^{(0)}$ gives us a sequence $\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots$ which should converge to a solution $\boldsymbol{\beta}$.

In our case $\boldsymbol{\psi}(\boldsymbol{\beta}) = \dfrac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{y} - \boldsymbol{\mu})$ is the score function and the observed Fisher information $-\dfrac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ is usually replaced by its expectation

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right] \\
&= \mathbb{E}[\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{Y} - \boldsymbol{\mu})\{\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{Y} - \boldsymbol{\mu})\}^T] \\
&= \mathbb{E}[\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{V}\boldsymbol{X}] \\
&= \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X},
\end{aligned}$$

where $\boldsymbol{W} = \boldsymbol{V}(\mathbb{C}\mathrm{ov}(\boldsymbol{Y}))\boldsymbol{V}$ is a diagonal matrix with diagonal elements $W_{ii} = \dfrac{1}{\mathbb{V}\mathrm{ar}(Y_i)}\left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)^2$.

The iteration formula can thus be rewritten as:

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + (\boldsymbol{X}^T\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{(k)}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)}) \\
&= \boldsymbol{S}^{(k)}[\boldsymbol{X}^T\boldsymbol{W}^{(k)}\boldsymbol{X}\boldsymbol{\beta}^{(k)} + \boldsymbol{X}^T\boldsymbol{V}^{(k)}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)})] \\
&= \boldsymbol{S}^{(k)}\boldsymbol{X}^T\boldsymbol{W}^{(k)}[\boldsymbol{X}\boldsymbol{\beta}^{(k)} + (\boldsymbol{W}^{(k)})^{-1}\boldsymbol{V}^{(k)}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)})] \\
&= \boldsymbol{S}^{(k)}\boldsymbol{X}^T\boldsymbol{W}^{(k)}\boldsymbol{z}^{(k)},
\end{aligned}$$

where

$\boldsymbol{S}^{(k)} = (\boldsymbol{X}^T\boldsymbol{W}^{(k)}\boldsymbol{X})^{-1}$,

$\boldsymbol{z}^{(k)} = \boldsymbol{X}\boldsymbol{\beta}^{(k)} + \boldsymbol{U}^{(k)}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)})$ and

$\boldsymbol{U}^{(k)} = (\boldsymbol{W}^{(k)})^{-1}\boldsymbol{V}^{(k)}$ is a diagonal matrix with diagonal elements $U_{ii}^{(k)} = \dfrac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}}$.

As $g(\boldsymbol{y}) \approx g(\boldsymbol{\mu}) + \boldsymbol{U}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{z}$, an obvious choice of starting values is: $z_i^{(0)} = g(y_i)$ and $W_{ii}^{(0)} = 1$ for $i = 1, \ldots, n$.

The IRWLS algorithm

A scheme for the IRWLS algorithm:

starting values:

$$\boldsymbol{X} \leftarrow \text{values } x_{ij}$$

$$\boldsymbol{y} \leftarrow \text{values } y_i$$

$$\boldsymbol{z} \leftarrow g(\boldsymbol{y})$$

$$\boldsymbol{b} \leftarrow (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{z}$$

repeat until $\boldsymbol{b}$ stops changing:

$$\boldsymbol{\eta} \leftarrow \boldsymbol{X}\boldsymbol{b}$$

$$\boldsymbol{\mu} \leftarrow g^{-1}(\boldsymbol{\eta})$$

$$U_{ii} \leftarrow \frac{\partial \eta_i}{\partial \mu_i}$$

$$W_{ii} \leftarrow \frac{1}{\mathbb{V}\text{ar}\,(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

$$\boldsymbol{z} \leftarrow \boldsymbol{X}\boldsymbol{b} + \boldsymbol{U}(\boldsymbol{y} - \boldsymbol{\mu})$$

$$\boldsymbol{S} \leftarrow (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$$

$$\boldsymbol{b} \leftarrow \boldsymbol{S}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}$$

results:

$\boldsymbol{b}$ = estimate for $\boldsymbol{\beta}$

$\hat{\phi}\boldsymbol{S}$ = estimate for $\mathbb{V}\text{ar}\,(\hat{\boldsymbol{\beta}})$

where $\hat{\phi}$ is computed **separately**

The beetles example in S-PLUS:

```
X <- matrix(c(rep(1,8),1.6907,...,1.8839),ncol=2)

n <- c(59,60,62,56,63,59,62,60)

killed <- c(6,13,18,28,52,53,61,60)

y <- killed/n

z <- log((killed+0.5)/(n-killed+0.5))

b <- solve(t(X)%*%X) %*% t(X)%*%z
```

```
eta <- X%*%b

mu <- exp(eta)/(1+exp(eta))

U <- diag(1/(mu*(1-mu)),8)

W <- diag(n*mu*(1-mu),8)

z <- X%*%b+U%*%(y-mu)

S <- solve(t(X)%*%W%*%X)

b <- S%*%t(X)%*%W%*%z
```

Distribution of estimators

Exact distributions of estimators can be derived only in special cases such as in **linear models**.

Approximate distributions are based on the asymptotic theory of maximum likelihood estimators.

Some definitions and properties:

The *score statistic* is $\boldsymbol{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$ with $U_i = \dfrac{\partial \ell}{\partial \beta_i}$ and

$\mathbb{E}(\boldsymbol{U}) = \boldsymbol{0}$.

The expected *Fisher information* is $\mathcal{I} = \mathbb{E}(\boldsymbol{U}\boldsymbol{U}^T)$ with i, j-th element:

$$\mathbb{E}(U_i U_j) = \mathbb{E}\left[\frac{\partial \ell}{\partial \beta_i}\frac{\partial \ell}{\partial \beta_j}\right] = -\,\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \beta_i \partial \beta_j}\right].$$

Under mild conditions convergence in distribution holds:

$$\mathcal{I}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Approximately for finite n we have therefore

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \mathcal{I}^{-1})$$

and for a linear combination $\boldsymbol{A}\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{A}\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\beta}, \boldsymbol{A}\mathcal{I}^{-1}\boldsymbol{A}^T).$$

For example a 95% confidence interval for β_i is approximately: $\hat{\beta}_i \pm 1.96\sqrt{s_{ii}}$, where s_{ii} is the i-th diagonal element of the covariance matrix $\mathcal{I}^{-1}$.

Approximate distributions of $\hat{\theta}_i$, $\hat{\mu}_i$, $\hat{\pi}_i$ or other functions of $\hat{\boldsymbol{\beta}}$ can be found using the delta method.

Goodness of fit

A **saturated model** S is a model that allows for maximum flexibility: its number of parameters equals the number of different linear combinations $\boldsymbol{x}_i^T\boldsymbol{\beta}$.

A **minimal model** is a model with minimum flexibility: it has only one parameter. It assumes that the expected response is constant and does not depend on the explanatory variable(s).

In establishing whether a certain model M fits the data well the likelihood-ratio test statistic is used as a criterion:

$$\lambda_M = 2[\ell_S(\tilde{\boldsymbol{\theta}}) - \ell_M(\hat{\boldsymbol{\theta}})]$$

where $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ are the maximum likelihood estimates of $\boldsymbol{\theta}$ under the models S and M and $\ell_S(\tilde{\boldsymbol{\theta}})$ and $\ell_M(\hat{\boldsymbol{\theta}})$ the maximum log-likelihoods under S and M.

Under certain conditions the distribution of λ_M is asymptotically chi-squared with the number of degrees of freedom equal to the difference in the number of β-parameters under the models S and M.

If $a_i(\phi) = \frac{\phi}{w_i}$ we can write: $\lambda_M = \dfrac{D_M}{\phi}$

where

$$\begin{aligned} D_M &= \sum 2w_i[(y_i\tilde{\theta}_i - b(\tilde{\theta}_i)) - (y_i\hat{\theta}_i - b(\hat{\theta}_i))] \\ &= \sum d_i \end{aligned}$$

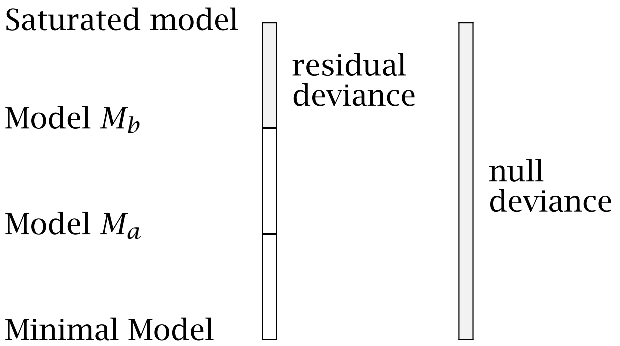
D_M is called the *deviance* and
λ_M the *scaled deviance*.

The *deviance residual* for observation i is defined as $\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i)$.

The *Pearson residual* is $e_i = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mathbb{V}}\text{ar}\,(Y_i)}}$.

Both types of residual can be used to examine where the fit is poor.

Twice the log-likelihood for several models:



Saturated model

Model M_b

Model M_a

Minimal Model

residual deviance

null deviance

Tests of hypotheses

For the testing problem H_0: 'M_a holds' against H_1: 'M_b holds', where M_a and M_b are nested models
(M_a is defined by imposing equality restrictions on the parameters of M_b)
the likelihood-ratio statistic is:

$$\lambda = 2(\ell_b - \ell_a) = 2[(\ell_S - \ell_a) - (\ell_S - \ell_b)] = \frac{D_a - D_b}{\phi}$$

where D_a and D_b are the deviances of the models M_a and M_b.

λ has asymptotically under H_0 a chi-square distribution with df = difference in the number of parameters under the models M_b and M_a.

If the **dispersion parameter** ϕ is unknown then it is estimated by

$$\hat{\phi} = \frac{1}{n - p_b} \sum \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

where $V(\mu_i) = b''(\theta_i)$, if written as a function of μ_i, is the so-called variance function: essentially (apart from the factor ϕ/w_i) it is the variance of Y_i as a function of the mean μ_i.

If we have to choose between a large number of models, nested or not, we have to trade between the goodness of fit and the number p of parameters in the model. A criterion that takes both aspects into account is the Akaike Information Criterion (AIC):

$$\text{AIC} = \frac{D_a}{\hat{\phi}} + 2p$$

(sometimes however $D_a + 2p\hat{\phi}$ is used).

A similar criterion is the Bayesian Information Criterion:

$$\text{BIC} = \frac{D_a}{\hat{\phi}} + p \ln n.$$

For the sake of a fair comparison, all terms in the log-likelihoodfunction have to be included, even if they do not involve parameters. And if two models are compared, the same $\hat{\phi}$ has to be used.

Let Y_i denote the number of insurance claims for a particular make and model of a car. The distribution of Y_i will depend on the number of cars n_i of this type that are insured. An obvious choice for the distribution is therefore a Poisson distribution with parameter $\mu_i = n_i\theta_i$, where θ_i may depend on other characteristics, such as age of the cars and the area where they are used. With the logarithm as the (canonical) link function we get:

$$\ln \mu_i = \ln n_i + \ln \theta_i = \ln n_i + \boldsymbol{x}_i^T \boldsymbol{\beta}$$

where the last term allows for the inclusion of explanatory variables.

This equation differs from the usual specification of the linear component due to the inclusion of the fixed term $\ln n_i$. This term is called an **offset.**

Other examples include epidemiological models where the expected number of persons with a certain disease is proportional to the population size.

Poisson regression and contingency tables

Contingency tables (frequency tables) can originate under different sampling schemes.

Consider three possible setups of a study to investigate whether the chance of a neural tube defect in a foetus is related to the mother's diet.

a. During a certain number of months, for each baby born at the maternity ward of a hospital, we register whether the baby has a neural tube defect or not and we interview the mother about her diet, which could be rated as good, fair or poor in quality. Frequencies y_{ij} are observed in the six possible categories, where $i = 1$ if the baby had a neural tube defect and $i = 2$ if it had not and $j = 1$, 2 or 3 if the diet quality was rated as good, fair or poor, respectively. The counts y_{ij} can be considered as outcomes of independent Poisson distributed variables Y_{ij} with parameters λ_{ij}.

y_{11}	y_{12}	y_{13}
y_{21}	y_{22}	y_{23}

This table can easily be analysed in the GLM framework as the Poisson distribution belongs to the Exponential Dispersion Family.

Poisson regression and contingency tables

b. Suppose we wait until we have gathered a total of n births, where each birth is classified into one of the already mentioned six categories. Then $Y_{11}, \ldots, Y_{23}$ are not independent. Their joint distribution is multinomial with parameters n and $\pi_{11}, \ldots, \pi_{23}$, with $\sum_i \sum_j \pi_{ij} = 1$.

y_{11}	y_{12}	y_{13}	
y_{21}	y_{22}	y_{23}	
			n

c. We interview n_1 mothers who had had a baby with a neural tube defect and n_2 mothers who had not had such a baby about their diet and classify each case into one of the six categories. Now (Y_{11}, Y_{12}, Y_{13}) have a multinomial distribution with parameters n_1 and $(\theta_{11}, \theta_{12}, \theta_{13})$ and, independently from this triplet, (Y_{21}, Y_{22}, Y_{23}) have a multinomial distribution with parameters n_2 and $(\theta_{21}, \theta_{22}, \theta_{23})$, where $\sum_j \theta_{1j} = 1$ and $\sum_j \theta_{2j} = 1$.

y_{11}	y_{12}	y_{13}	n_1
y_{21}	y_{22}	y_{23}	n_2

The tables in **b** and **c** do not fit directly into the GLM framework because the response variables are multidimensional (they have multinomial distributions).

Poisson regression and contingency tables

The **Exponential Dispersion Family** has only one parameter, apart from a possible nuisance parameter ϕ, whereas a multinomial distribution's parameter is multidimensional. We will show that maximum likelihood estimators and their distributions for multinomial models equal those for corresponding Poisson models.

If $Y_1, \ldots, Y_k$ are independent and $Y_i \sim \text{Poisson}(\mu_i)$, then the associated log-likelihoodfunction is

$$\begin{aligned}\ell_1 &= \sum (y_i \ln \mu_i - \mu_i - \ln y_i!) \\ &= \sum (y_i \ln \mu_i) - \mu_+ - \sum \ln y_i!\end{aligned}$$

where a '+' means summation over the indicated subscript.

If $(Y_1, \ldots, Y_k)$ have a multinomial distribution with parameters $n_+, \pi_1, \ldots, \pi_k$, where $\pi_i = \frac{\mu_i}{\mu_+}$, the associated log-likelihoodfunction is

$$\begin{aligned}\ell_2 &= \sum (y_i \ln \pi_i) + \ln y_+! - \sum \ln y_i! \\ &= \sum (y_i \ln \mu_i) - n_+ \ln \mu_+ + \ln n_+! - \sum \ln y_i!\end{aligned}$$

Also define the log-likelihood for a Poisson distribution with parameter μ_+:

$$\ell_3 = y_+ \ln \mu_+ - \mu_+ - \ln y_+!$$

Then the equality $\ell_1 = \ell_2 + \ell_3$ reflects the fact that

$$\mathbb{P}(Y_1 = y_1, \ldots, Y_k = y_k) =$$
$$\mathbb{P}(Y_1 = y_1, \ldots, Y_k = y_k | Y_+ = y_+) \cdot \mathbb{P}(Y_+ = y_+).$$

The left hand side is the product of Poisson probabilities, the first factor on the right hand side is a multinomial probability and the last factor is a Poisson probability because
$Y_+ \sim \text{Poisson}(\mu_+)$.

In a log-linear model we assume that $\ln \mu_i$ is a linear combination of a number of explanatory variables. Usually we include a constant parameter α. Written out explicitly:

$$\ln \mu_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} \quad \text{or} \quad \mu_i = e^{\alpha} \cdot e^{\boldsymbol{x}_i^T \boldsymbol{\beta}},$$

hence $\mu_+ = e^{\alpha} \sum e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}$.

If we substitute $\alpha = \ln \mu_+ - \ln \sum e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}$ and view the log-likelihoods as functions of μ_+ and $\boldsymbol{\beta}$ only, we get: $\ell_1(\mu_+, \boldsymbol{\beta}) = \ell_2(\boldsymbol{\beta}) + \ell_3(\mu_+)$, where ℓ_2 depends on $\boldsymbol{\beta}$ only and ℓ_3 depends on μ_+ only.

All information in ℓ_1 about $\boldsymbol{\beta}$ is in fact contained in ℓ_2. In particular the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ and its covariance matrix $\mathbb{C}\mathrm{ov}\,(\hat{\boldsymbol{\beta}})$ based on ℓ_2 are identical to the corresponding $\hat{\boldsymbol{\beta}}$ and $\mathbb{C}\mathrm{ov}\,(\hat{\boldsymbol{\beta}})$ based on ℓ_1.

Therefore the maximum likelihood estimator for $\boldsymbol{\beta}$ and its distribution are identical in the multinomial and in the Poisson case, provided that a constant α is included in the model.

A similar argument holds for more complicated situations. If we have a three-dimensional contingency table and a product-multinomial distribution for the frequencies Y_{ijk} with marginal totals y_{ij+} fixed by the sampling design, then we can estimate the parameters as if the observations were Poisson distributed. The only condition is that a term ζ_{ij} has to be included in our models so that the marginal totals y_{ij+} for the observed table and the estimated frequency table coincide.

On page 14 and 15 for three different two-dimensional tables we specified models in terms of parameters λ_{ij}, π_{ij} and θ_{ij} which we shall denote by parametrisations $\mathcal{A}$a, $\mathcal{A}$b and $\mathcal{A}$c respectively. In all three cases we could have equivalently used parametrisation $\mathcal{B}$, based on the expected cell frequencies μ_{ij}:

$\mathcal{B}$a. $\mu_{ij} = \lambda_{ij}$ for all i, j.

$\mathcal{B}$b. $\mu_{ij} = n\pi_{ij}$ for all i, j.

$\mathcal{B}$c. $\mu_{ij} = n_i\theta_{ij}$ for all i, j.

A third parametrisation is $\mathcal{C}$, where the saturated model is specified in terms of parameters α, β_i, γ_j and δ_{ij}. The relationship with $\mathcal{B}$ is given by $\ln \mu_{ij} = \alpha + \beta_i + \gamma_j + \delta_{ij}$ where certain restrictions apply for the subscripted parameters.

The model of independence between row- and column factors in cases **a** and **b**, as well as the model of homogeneity in case **c** can now be specified as: $\delta_{ij} = 0$ for all i, j (note that in case **c** the β_i term must be included for reasons explained in the previous column).

Poisson regression and contingency tables

As we did in two-dimensional tables we could analogously parameterise three-dimensional tables by parametrisations of type $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. In the following table five possible models are specified for a three-dimensional table in which only the total is fixed by design ($\mathcal{A}$b applies). Additional sum-to-zero restrictions are assumed under parametrisation $\mathcal{C}$.

Usually we are considering only models which are **comprehensive** (in which all variables are included) and **hierarchical** (if a certain interaction is included, also the corresponding lower order interactions and main effects are included). This enables us to use a shorthand notation which is displayed in the last column.

$\mathcal{A}$ b	$\mathcal{B}b$	$\mathcal{C}b$	shorthand
$\pi_{ijk} =$	$\mu_{ijk} =$	$\ln \mu_{ijk} =$	
π_{ijk}	μ_{ijk}	$\lambda + \alpha_i + \beta_j + \gamma_k + \xi_{jk} + \eta_{ik} + \zeta_{ij} + \delta_{ijk}$	{123} (saturated)
		$\lambda + \alpha_i + \beta_j + \gamma_k + \xi_{jk} + \eta_{ik} + \zeta_{ij}$	{12}, {13}, {23} (partial association)
$\frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}$	$\frac{\mu_{i+k}\mu_{+jk}}{\mu_{++k}}$	$\lambda + \alpha_i + \beta_j + \gamma_k + \xi_{jk} + \eta_{ik}$	{13}, {23} (conditional independence)
$\pi_{i++}\pi_{+jk}$	$\frac{\mu_{i++}\mu_{+jk}}{n}$	$\lambda + \alpha_i + \beta_j + \gamma_k + \xi_{jk}$	{1}, {23} (1 independent of 2 and 3)
$\pi_{i++}\pi_{+j+}\pi_{++k}$	$\frac{\mu_{i++}\mu_{+j+}\mu_{++k}}{n^2}$	$\lambda + \alpha_i + \beta_j + \gamma_k$	{1}, {2}, {3} (complete independence)

If certain marginal totals are fixed by design, the corresponding parameter term has to be included in the model. If for instance the marginal totals y_{ij+} are fixed by design, then the model should at least include the terms ζ_{ij}. The first two models in the table of p. 18 are then allowed, but the last three are not.

It can be shown that a set of minimal sufficient statistics consists of the marginal totals with subscripts corresponding to the shorthand notation. If we take the model ({13}, {23}), then the log-likelihood function is

$$\begin{aligned}\ell &= c + \sum_i \sum_j \sum_k y_{ijk} \ln \mu_{ijk} - \sum_i \sum_j \sum_k \mu_{ijk} \\ &= c + y_{+++}\lambda + \sum_i y_{i++}\alpha_i + \sum_j y_{+j+}\beta_j + \sum_k y_{++k}\gamma_k \\ &\qquad + \sum_i \sum_k y_{i+k}\eta_{ik} + \sum_j \sum_k y_{+jk}\xi_{jk} - \sum_i \sum_j \sum_k \mu_{ijk}.\end{aligned}$$

This is true for the product-Poisson case, and according to the argument on pages 16 – 17 also for the multinomial and the product-multinomial case, provided the parameter terms corresponding to the fixed marginal totals are included.

Now the set $\{y_{+++}, y_{i++}, y_{+j+}, y_{++k}, y_{i+k}, y_{+jk}\}$ is sufficient, where $y_{+++}, y_{i++}, y_{+j+}, y_{++k}$ are simple sums of terms $\{y_{i+k}\}$ and $\{y_{+jk}\}$ which thus form the minimal sufficient statistics.

It can also be proven (see the **remark** on page **7**) that the maximum likelihood estimators $\hat{\mu}_{i+k}$ and $\hat{\mu}_{+jk}$ are equal to y_{i+k} and y_{+jk} respectively.

The maximum likelihood estimators $\hat{\mu}_{ijk}$ are in many cases simple functions of these marginal totals: for the model ({13}, {23}) we have for example

$$\hat{\mu}_{ijk} = \frac{\hat{\mu}_{i+k}\hat{\mu}_{+jk}}{\hat{\mu}_{++k}} = \frac{y_{i+k}y_{+jk}}{y_{++k}}.$$

But in some other cases the maximum likelihood estimators cannot be written as explicit functions of the marginal totals. For three-dimensional models this is the case for the model ({12}, {13}, {23}).

Overdispersion and quasi-likelihood

Some distributions such as the Poisson distribution are described by only one parameter which determines the distribution's mean as well as its variance. It may happen that the estimated variance appears to be much larger than it should be under the assumed distribution. This phenomenon is called overdispersion.

One remedy would be to model the response variable as having a negative binomial distribution, which has an extra parameter. Another solution is the quasi-likelihood approach which bears some resemblance to the Least Squares vs. the maximum likelihood approach in **linear models**.

In ordinary linear models both the maximum likelihood criterion and the least squares criterion lead to the same estimators for the vector $\boldsymbol{\beta}$, although in the latter case the assumptions are much weaker:
no particular distribution of the response variable is assumed, only its mean and variance are specified.

Assumptions and results are summarised in the table on the following page.

A similar approach is possible for generalised linear models: instead of specifying the response variable's distribution as a member of the Exponential Dispersion Family (EDF), we only specify its variance function.

As a result the solutions of the maximum likelihood and the quasi-likelihood equations are identical.

Overdispersion and quasi-likelihood

Assumptions and results for maximum likelihood and least squares estimation in **linear models**.

Maximum Likelihood criterion	Least Squares criterion
Assumptions:	Assumptions:
$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{X\beta}$ $\mathbb{V}\mathrm{ar}\,(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$ $\boldsymbol{Y} \sim \mathrm{Normal}$	$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{X\beta}$ $\mathbb{V}\mathrm{ar}\,(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$
Results:	Results:
$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ $\hat{\sigma}^2 = \frac{1}{n}\lVert\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\rVert^2$ $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ $\mathbb{V}\mathrm{ar}\,(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$	$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ $\mathbb{V}\mathrm{ar}\,(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ under certain conditions: $(\boldsymbol{X}^T\boldsymbol{X})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ hence approximately: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$

Quasi-likelihood

Assumptions and results for maximum likelihood and quasi-likelihood estimation in generalised linear models.

	Maximum Likelihood	Quasi-likelihood
Assumptions	$\mathbb{E}(Y_i) = \mu_i = g^{-1}(\eta_i)$ with $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ $f_{Y_i} \in \mathcal{EDF}$	$\mathbb{E}(Y_i) = \mu_i = g^{-1}(\eta_i)$ with $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ $\mathbb{V}\mathrm{ar}\,(Y_i) = \psi V(\mu_i)$
Score function	$\frac{\partial \ell}{\partial \mu_i} = \frac{Y_i - \mu_i}{\mathbb{V}\mathrm{ar}\,(Y_i)}$	$U_i = \frac{Y_i - \mu_i}{\psi V(\mu_i)}$
Equations	$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathbb{V}\mathrm{ar}\,(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = 0$ or: $\boldsymbol{X}^T \boldsymbol{V}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}$	$\sum_{i=1}^{n} U_i \frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\psi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = 0$ or: $\boldsymbol{X}^T \boldsymbol{V}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}$
Log-likelihood	$\ell(\boldsymbol{\mu}, \boldsymbol{y})$	$Q(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{n} Q(\mu_i, y_i)$ where $Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\psi V(t)} dt$

Using S-PLUS

Generalised Linear Models can be estimated and analysed in S-PLUS by using the graphical user interface or by using the commands window.

Graphical User Interface

GLMs can be applied by selecting:

`Statistics` → `Regression` → `Generalised Linear`.

A dialog window will appear in which the model, possibly the **weights** w_i, and the data can be specified. More information is given in the `User's Guide` (click `Help` → `Online Manuals`) from page 319.

In special cases we have the following alternatives:

Linear regression

`Statistics` → `Regression` → `Linear`

Analysis of Variance

`Statistics` → `ANOVA` → `Fixed effects`

Logistic regression

`Statistics` → `Regression` → `Logistic`

Log-linear model or Poisson regression

`Statistics` → `Regression` → `Log-linear`

Commands window

The general command is `glm` with the possible parameters: `formula`, `family`, `data`, `control`.

`formula` specifies the response variable and the explanatory variables. Example:

`proportion ˜ log.dose`.

`family` specifies the distribution from the class $\mathcal{EDF}$ and the link function if it is not the default link function. Examples:

```
family = binomial
family = binomial(link=probit)
```

`family = gaussian` (normal)

`family = poisson`.

`data` specifies the data-frame that contains the data.

`control` sets parameters that control the iterative estimation process.

S-PLUS – Output of the beetles example

```
Call: glm(formula = proportion ˜ log.dose, family = binomial(link = logit), data = beetles, weights =
treated, na.action = na.exclude, control = list(...))
Deviance Residuals:
       Min         1Q    Median       3Q      Max
 -1.594124 -0.3943968 0.8329153 1.259223 1.593985


Coefficients:
                Value Std. Error   t value
(Intercept) -60.71745   5.179902 -11.72174
   log.dose  34.27032   2.911680  11.76995


(Dispersion Parameter for Binomial family taken to be 1 )
    Null Deviance: 284.2024 on 7 degrees of freedom
Residual Deviance: 11.23223 on 6 degrees of freedom
```

Here $Y_i \sim$ proportional binomial with **weights** $w_i =$ the numbers of insects treated.

In this case the **dispersion parameter** ϕ $(= 1)$ is given, but with some other distributions and in quasi-likelihood models the estimated dispersion parameter $\hat{\phi}$ is given instead.

The **null deviance** is the **deviance** of the **minimal model** in which μ_i = constant. The **residual deviance** is the deviance of the analysed model.

S-PLUS – Coding factors (contrasts)

In a model in which an explanatory variable is a factor, its k levels are associated with parameters $\beta_1, \ldots, \beta_k$. If a constant parameter α is also included, an additional restriction is necessary in order to avoid overparametrisation.

Possible restrictions are:

a. $\sum \beta_i = 0$ (sum-to-zero restriction), type in the S-PLUS commands window:

`> options(contrasts=c("contr.sum","contr.poly"))`. The output gives then the estimated values $\hat{\alpha}$ and $\hat{\beta}_1, \ldots, \hat{\beta}_{k-1}$. The remaining estimate is given by $\hat{\beta}_k = -\hat{\beta}_1 - \cdots - \hat{\beta}_{k-1}$.

b. $\beta_1 = 0$ (corner-point restriction), type in the S-PLUS commands window:

`> options(contrasts=c("contr.treatment","contr.poly"))`.

c. Helmert contrasts (default in S-PLUS).

For more information see `Help` → `Online Manuals` → `Guide to Statistics Volume 1`, Chapter 2 Specifying Models in S-PLUS, page 39 - 43: CONTRASTS: THE CODING OF FACTORS.

S-PLUS – Offset, quasi-likelihood

Offset

If the response vector is `y`, the offset vector is `a` and the only explanatory variable is `x`, then the model is formulated in S-PLUS as

```
y ~ offset(a) + x
```

If the ‘regression line’ has to pass through the origin (no constant parameter should be estimated) the model is formulated as

```
y ~ x - 1
```

If the ‘regression line’ is forced to pass through the offset, and if there is no constant parameter, the model is formulated as

```
y ~ offset(a) + x - 1
```

Quasi-likelihood

Quasi-likelihood models can be estimated through the **menu**:

`Statistics→Regression→Generalised Linear`.

Then specify `quasi` under `Model→Family`.

Several choices are possible for the link function and the variance function.

Through the **Commands window**:

```
> glm(formula, family=quasi(link=..., variance=...),da
```

Possible values for `link` are:

`logit`, `inverse`, `identity`, `1/mu^2`, `log`.

Possible values for `variance` are:

`mu(1-mu)`, `mu^2`, `constant`, `mu^3`, `mu`.

S-PLUS – Adding a distribution

It is possible in S-PLUS to add a distribution from the **Exponential Dispersion Family** and use it in estimating a Generalised Linear Model. Here is a specification for the geometric distribution:

```
geometric <- function()
{ geom.lnk <-list(
       names="log",
         link=function(mu) log(mu),
         inverse=function(eta) exp(eta),
         deriv=function(mu) 1/mu,
       initialize=expression(mu<-y+(y==1)/6))
  geom.var<-list(
         names="mu(mu-1)",
         variance=function(mu) mu*(mu-1),
         deviance=function(mu,y,w,residuals=F)
         {
           devi<-2*((y-1)*log(pmax(1,y-1)*mu/(y*(mu-1)))-log(y/mu))
           if (residuals) sign(y-mu)*sqrt(abs(devi))
           else sum(abs(devi))
    }
  )
 make.family("Geometric",link=geom.lnk,variance=geom.var)
}
```

Literature

Agresti, A. [2002] Categorical Data Analysis, second edition; Wiley, New York.

Azzalini, A. [1996] Statistical inference: based on the likelihood; Chapman & Hall, London.
Chapter 6 is devoted to GLMs.

Dobson, A.J. [2002] An introduction to generalized linear models, 2nd ed.; Chapman & Hall, London.

Fahrmeir, L., Tutz, G. [1994] Multivariate statistical modelling based on generalized linear models; Springer-Verlag, New York.

Gill, J. [2000] Generalized Linear Models: A Unified Approach (Quantitative Applications in the Social Sciences, 134); Sage Publications.

Lindsey, J.K. [1997] Applying Generalized Linear Models; Springer-Verlag, New York.

McCullagh, P., Nelder, J.A. [1989] Generalized Linear Models, 2nd ed.; Chapman & Hall, London.

Myers, R.H., Montgomery, D.C., Vining, G.G. [2002] Generalized Linear Models: With Applications in Engineering and the Sciences; Wiley, New York.

Venables, W.N., Ripley, B.D. [1994] Modern Applied Statistics with S-Plus; Springer-Verlag New York.

In S-PLUS see under `Help→Online Manuals→ Guide to Statistics volume 1`:
Chapter 10: ‘Generalizing the linear model’.

On the World Wide Web see:
`www.statsoft.com/textbook/stathome.html` and select GLZ.

Index