Contents

1 Generalized Linear Mixed Models
   1 Introduction
   2 Motivation
   3 Definition
   4 Estimation and algorithm
   5 Analysis of Deviance
   6 Examples

2 Overdispersion
   1 Motivation
   2 Models
   3 Examples
Agricultural Science - different types of data (responses):
- continuous: weight, height, diameter
- discrete: count, proportion

Model choice - important part of the research: search for a simple model which explains well the data.

All models envolve:
- a systematic component - related to the explanatory variables (regression model, analysis of variance model, analysis of covariance model);
- a random component - related to the distributions followed by the response variables;
- a link between systematic and random components.
Motivating example – Carnation meristem culture

<table>
<thead>
<tr>
<th></th>
<th>0.0</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>1.0</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>l</td>
<td>v</td>
<td>l</td>
<td>v</td>
<td>l</td>
<td>v</td>
</tr>
<tr>
<td>1</td>
<td>2.5</td>
<td>0</td>
<td>3</td>
<td>5.5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
<td>0</td>
<td>2</td>
<td>4.3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>0</td>
<td>4</td>
<td>2.7</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2.5</td>
<td>1</td>
<td>3</td>
<td>4.3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4.0</td>
<td>0</td>
<td>5</td>
<td>5.4</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>4.0</td>
<td>0</td>
<td>3</td>
<td>3.8</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>3.0</td>
<td>0</td>
<td>3</td>
<td>4.3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>3.0</td>
<td>0</td>
<td>4</td>
<td>6.0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>5.0</td>
<td>0</td>
<td>3</td>
<td>5.0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>4.0</td>
<td>0</td>
<td>2</td>
<td>5.0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>2.0</td>
<td>1</td>
<td>3</td>
<td>4.5</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>4.0</td>
<td>0</td>
<td>3</td>
<td>4.0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>3.0</td>
<td>0</td>
<td>4</td>
<td>3.3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>3.5</td>
<td>1</td>
<td>3</td>
<td>4.3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>3.0</td>
<td>0</td>
<td>3</td>
<td>4.5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>2.0</td>
<td>1</td>
<td>3</td>
<td>3.8</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>5.5</td>
<td>0</td>
<td>3</td>
<td>4.7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>3.0</td>
<td>0</td>
<td>4</td>
<td>2.2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>2.5</td>
<td>0</td>
<td>2</td>
<td>3.8</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>2.0</td>
<td>0</td>
<td>3</td>
<td>5.0</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

- **Response variables**: number of shoots (s), average length of shoots (l), vitrification (v)
- **Distributions**: ??
- **Systematic component**: regression model, completely randomized experiment.
- **Links**: ??
GLM Overdispersion

![Graphs showing the relationship between BAP dose and number of shoots, average length of shoots, and vitrification.](image)
Motivating example – Rotenon toxicity

<table>
<thead>
<tr>
<th>Dose ($d_i$)</th>
<th>$m_i$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>2.6</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>3.8</td>
<td>48</td>
<td>16</td>
</tr>
<tr>
<td>5.1</td>
<td>46</td>
<td>24</td>
</tr>
<tr>
<td>7.7</td>
<td>49</td>
<td>42</td>
</tr>
<tr>
<td>10.2</td>
<td>50</td>
<td>44</td>
</tr>
</tbody>
</table>

- Response variable: $Y_i$ – number of dead insects out of $m_i$ insects (Martin, 1942).
- Distribution: Binomial.
- Systematic component: regression model, completely randomized experiment.
- Aim: Lethal doses.
The curve of the observed proportions against doses has an S shape.
Bioassay or biological assay – the measurement of the potency of a stimulus by means of the reactions it produces in living matter.

The stimulus may be physiological, chemical, biological, etc.
- insects exposed to chemical stimuli (insecticides or biological, e.g., a virus, stimuli)
- other agricultural bioassays may involve large animals, fungi, plants or plant parts such as leaves.

The response here is quantal or binary – just 2 possible responses (success or failure).
- an insect dies or survives
- a seed germinates or fails to germinate
- a cutting roots or fails to root

The rotenone data example
The tolerance of an individual (e.g., an insect) is the dose or concentration or intensity of the stimulus above which the individual responds (e.g., dies) and below which it does not respond.

An individual dies if his tolerance is smaller than a given dose.

Tolerance varies between individuals – it is a random variable.

Suppose

- \( z \): tolerance of a randomly chosen individual
- \( f_Z(z) \): the probability density function of \( Z \)
- \( x \): dose of the stimulus
Then, for an individual chosen at random from the population

\[
P(\text{death}|x) = P(Z < x) = \int_{-\infty}^{x} f(z)dz
\]

It is often reasonable to assume that the tolerance \( Z \) has a normal distribution, that is, \( Z \sim N(\mu, \sigma^2) \). Then, the cumulative normal distribution is

\[
\pi = P(\text{death}|x) = P(Z < x) = P \left( \frac{Z - \mu}{\sigma} < \frac{x - \mu}{\sigma} \right) = \Phi \left( \frac{x - \mu}{\sigma} \right)
\]

The inverse of the cumulative normal function, \( \text{probit}(\pi) = \Phi^{-1}(\pi) = \alpha + \beta x \), where \( \alpha = -\frac{\mu}{\sigma} \) and \( \beta = \frac{1}{\sigma} \), is called the \textbf{probit} transformation.
Alternatively, we can assume that tolerance $Z$ has a logistic distribution with mean $\mathbb{E}(Z) = \mu$, variance $\text{Var}(Z) = \pi^2 \tau^2 / 3$ and pdf

$$f_Z(z; \mu, \tau) = \frac{\exp \left( \frac{z - \mu}{\tau} \right)}{\tau} \left[ 1 + \exp \left( \frac{z - \mu}{\tau} \right) \right]^{-2}, \quad \mu \in \mathbb{R}, \quad \tau > 0,$$

Then, the cumulative logistic distribution is

$$\pi = P(Z \leq x) = F(x) = \int_{-\infty}^{x} \frac{\beta e^{\alpha + \beta z}}{(1 + e^{\alpha + \beta z})^2} dz = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

where $\alpha = -\mu / \tau$ and $\beta = 1 / \tau$.

The inverse of the cumulative distribution of the logistic distribution is called the \textbf{logit} transformation

$$\text{logit}(\pi) = \log \frac{\pi}{1 - \pi} = \alpha + \beta x$$

The normal and logistic distributions are symmetrical around the mean.
A different distribution, which is asymmetrical, for the tolerance $Z$ is the extreme value (Gumbel) distribution with mean $\mathbb{E}(Z) = \alpha + \gamma \tau$, variance $\text{Var}(Z) = \pi^2 \tau^2 / 6$, where $\gamma \approx 0, 577216$ is the Euler number defined by $\gamma = -\psi(1) = \lim_{n \to \infty} \left( \sum_{i=1}^{n} i^{-1} - \log n \right)$, $\psi(p) = \frac{d \log \Gamma(p)}{dp}$ is the digamma function, and pdf

$$f_Z(z; \alpha, \tau) = \frac{1}{\tau} \exp \left( \frac{z - \alpha}{\tau} \right) \exp \left[ - \exp \left( \frac{z - \alpha}{\tau} \right) \right], \quad \alpha \in \mathbb{R}, \quad \tau > 0,$$

Then, the cumulative Gumbel distribution is

$$\pi = F(x) = \int_{-\infty}^{x} \beta \exp \left( \alpha + \beta z - e^{\alpha + \beta z} \right) dz = 1 - \exp \left[ - \exp \left( \alpha + \beta x \right) \right]$$

where $\alpha = -\mu / \tau$ and $\beta = 1 / \tau$.

The inverse of the cumulative Gumbel distribution is called complementary log-log transformation

$$\text{c-loglog}(\pi) = \log(-\log(1 - \pi)) = \alpha + \beta x$$

The exposed theory shows one type of biological justification for the link function.
The three components of a generalized linear model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989, Cordeiro and Demétrio, 2011) are:

- independent random variables $Y_i$, $i = 1, \ldots, n$, from an exponential family distribution with means $\mu_i$ and constant dispersion (scale) parameter $\phi$,

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

where $\mu = \mathbb{E}[Y] = b'(\theta)$ and $\text{Var}(Y) = \phi b''(\theta) = \phi V(\mu)$, $V(\mu)$ called variance function.

- a linear predictor vector $\eta$ given by

$$\eta = X\beta$$

where $\beta$ is a vector of $p$ unknown parameters and $X = [x_1, \ldots, x_n]^T$ is the $n \times p$ design matrix;

- a link function $g(\cdot)$ relating the mean to the linear predictor, i.e.

$$g(\mu_i) = \eta_i = x_i^T \beta$$
Normal Models

\( Y_i, \ i = 1, \ldots, n, \) a continuous response variable,
\( Y_i \sim N(\mu_i, \sigma^2) \) with mean \( \mu_i \) and constant variance \( \sigma^2 \)
We model the mean \( \mu_i \) in terms of the explanatory variables \( x_i \).
As a glm

- **Random component:**

\[
f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_i - \mu_i)^2}{2\sigma^2} \right] = \exp \left[ \frac{1}{\sigma^2} \left( y_i \mu_i - \frac{\mu_i^2}{2} \right) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2} \right]
\]

where

\[
\theta_i = \mu_i, \ \phi = \sigma^2, \ \ b(\theta_i) = \frac{\mu_i^2}{2} = \frac{\theta_i^2}{2}, \ \ c(y_i, \phi) = -\frac{1}{2} \left[ \frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2) \right], \ \ V(\mu_i) = 1
\]

- **Systematic component:** \( \eta_i = x_i^T \beta \)

- **Link function:** \( \eta_i = \mu_i \) (identity link, canonical link)
Binomial regression model

$Y_i$, count of successes out of a sample of size $m_i$, $i = 1, \ldots, n$

$Y_i \sim \text{Bin}(m_i, \pi_i)$ with mean $\mathbb{E}[Y_i] = \mu_i = m_i\pi_i$, and

$\text{Var}(Y_i) = m_i\pi_i(1 - \pi_i)$

We model the expected proportions $\pi_i \in [0, 1]$ in terms of the explanatory variables $x_i$. As a glm

- **Random component:**

  \[
  f(y_i; \pi_i) = \left(\frac{m_i}{y_i}\right)\pi_i^{y_i}(1 - \pi_i)^{m_i - y_i} = \exp \left[ y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + m_i \log(1 - \pi_i) + \log \left( \frac{m_i}{y_i} \right) \right],
  \]

  where $y_i = 0, 1, \ldots, m_i$,

  \[
  \phi = 1, \quad \theta_i = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \log \left( \frac{\mu_i}{m_i - \mu_i} \right) \Rightarrow \mu_i = \frac{m_e^{\theta_i}}{(1 + e^{\theta_i})},
  \]

  \[
  b(\theta) = -m_i \log(1 - \pi_i) = m_i \log (1 + e^{\theta_i}), \quad c(y_i, \phi) = \log \left( \frac{m_i}{y_i} \right), \quad V(\mu_i) = \mu_i(m_i - \mu_i)/m_i
  \]

- **Systematic component:**

  \[
  \eta_i = \beta'x_i,
  \]
**Link function:**

The canonical link function is the logit

$$\eta_i = g(\pi_i) = g\left(\frac{\mu_i}{m_i}\right) = \log\left(\frac{\mu_i}{m_i - \mu_i}\right) = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

Other common choices are

- probit $\eta_i = g(\pi_i) = g\left(\frac{\mu_i}{m_i}\right) = \Phi^{-1}(\mu_i/m_i) = \Phi^{-1}(\pi_i)$
- complementary log-log link

$$\eta_i = g(\pi_i) = g\left(\frac{\mu_i}{m_i}\right) = \log\{-\log(1 - \pi_i)\}.$$
Poisson regression models

$Y_i, \ i = 1, \ldots, n$, are counts with means $\mu_i$
$Y_i \sim \text{Pois} (\mu_i)$ with mean $\mu_i$ and variance $\text{Var}(Y_i) = \mu_i$

As a glm

- **Random component:**

$$f(y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \exp[y_i \log(\mu_i) - \mu_i - \log(y_i!)]$$

where

$\phi = 1, \ \theta_i = \log(\mu_i) \Rightarrow \mu_i = \exp(\theta_i), \ b(\theta) = \mu_i = \exp(\theta_i), \ c(y_i, \phi) = \log(y_i!), \ \text{V}(\mu_i) = \mu_i$

- **Systematic component:**

$$\eta_i = \beta' x_i,$$

- **Link function:**

The canonical link function is the log

$$\eta_i = g(\mu_i) = \log(\mu_i)$$
For different observation periods/areas/volumes:

\[ Y_i \sim \text{Pois}(t_i \lambda_i) \]

Taking a log-linear model for the rates,

\[ \log(\lambda_i) = x_i^T \beta \]

results in the following log-linear model for the Poisson means

\[ \log(\mu_i) = \log(t_i \lambda_i) = \log(t_i) + x_i^T \beta, \]

where the \( \log(t_i) \) is included as a fixed term, or *offset*, in the model.
Components of some distributions from exponential family

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\phi$</th>
<th>$\theta$</th>
<th>$b(\theta)$</th>
<th>$c(y, \phi)$</th>
<th>$\mu(\theta)$</th>
<th>$V(\mu)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal: $N(\mu, \sigma^2)$</td>
<td>$\sigma^2$</td>
<td>$\mu$</td>
<td>$\frac{\theta^2}{2}$</td>
<td>$-\frac{1}{2} \left[ \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right]$</td>
<td>$\theta$</td>
<td>1</td>
</tr>
<tr>
<td>Poisson: $P(\mu)$</td>
<td>1</td>
<td>$\log(\mu)$</td>
<td>$e^\theta m \log(1 + e^\theta)$</td>
<td>$-\log(y!)$</td>
<td>$e^\theta \frac{m e^\theta}{1 + e^\theta}$</td>
<td>$\frac{\mu}{m} (m - \mu)$</td>
</tr>
<tr>
<td>Binomial: $B(m, \pi)$</td>
<td>1</td>
<td>$\log \left( \frac{\mu}{m - \mu} \right)$</td>
<td>$m \log(1 + e^\theta)$</td>
<td>$\log \left( \frac{m}{y} \right)$</td>
<td>$\log \left[ \frac{\Gamma(k + y)}{\Gamma(k)y!} \right]$</td>
<td>$\frac{\mu}{m} (k \pi + 1)$</td>
</tr>
<tr>
<td>Negative Binomial: $BN(\mu, k)$</td>
<td>1</td>
<td>$\log \left( \frac{\mu}{\mu + k} \right)$</td>
<td>$-k \log(1 + e^\theta)$</td>
<td>$-\frac{1}{\theta}$</td>
<td>$\frac{1}{1 - e^\theta}$</td>
<td>$\frac{\mu}{k}^2$</td>
</tr>
<tr>
<td>Gamma: $G(\mu, \nu)$</td>
<td>$\nu^{-1}$</td>
<td>$\frac{1}{\nu}$</td>
<td>$-\log(-\theta)$</td>
<td>$\nu \log(\nu y) - \log(y) - \log \Gamma(\nu)$</td>
<td>$-\frac{1}{\theta}$</td>
<td>$\frac{\mu^2}{\nu}$</td>
</tr>
<tr>
<td>Inverse Gaussian: $IG(\mu, \sigma^2)$</td>
<td>$\sigma^2$</td>
<td>$\frac{1}{2\mu^2}$</td>
<td>$-(-2\theta)^{1/2}$</td>
<td>$-\frac{1}{2} \left[ \log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y} \right]$</td>
<td>$(-2\theta)^{-1/2}$</td>
<td>$\mu^3$</td>
</tr>
<tr>
<td>Distribution</td>
<td>Canonical link functions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>Identity: $\eta = \mu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>Logarithmic: $\eta = \log(\mu)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binomial</td>
<td>Logistic: $\eta = \log\left(\frac{\pi}{1 - \pi}\right) = \log\left(\frac{\mu}{m - \mu}\right)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td>Reciprocal: $\eta = \frac{1}{\mu}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>Reciprocal squared: $\eta = \frac{1}{\mu^2}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Maximum likelihood estimation

\[ \ell(\beta) = \sum_{i=1}^{n} \ell_i(\theta_i, \phi; y_i) = \phi^{-1} \sum_{i=1}^{n} [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi), \]

\[ U_r = \frac{\partial \ell(\beta)}{\partial \beta_r} = \sum_{i=1}^{n} \frac{d \ell_i}{d \theta_i} \frac{d \mu_i}{d \eta_i} \frac{d \eta_i}{d \beta_r} = \phi^{-1} \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{V_i} d\mu_i x_{ir}, \text{ because} \]

\[ \ell(\beta) = f(\theta_1, \ldots, \theta_i, \ldots, \theta_n) \]

\[ \theta_i = \int V_i^{-1} d\mu_i = q(\mu_i) \]

\[ \mu_i = g^{-1}(\eta_i) = h(\eta_i) \]

\[ \eta_i = \sum_{r=1}^{p} x_{ir} \beta_r \]

where \( \mu_i = b'(\theta_i) \) and \( d\mu_i / d\theta_i = V_i \)
The equations $U_r = 0$ are usually non-linear and have no analytic solution. Therefore, we rely on numerical methods (Newton-Raphson, Fisher-scoring, etc) to solve them.

The Fisher-scoring method gives

$$\beta^{(m+1)} = \beta^{(m)} + (K^{(m)})^{-1}U^{(m)},$$

where the Fisher information matrix $K$ has elements

$$\kappa_{r,s} = -E \left[ \frac{\partial^2 \ell(\beta)}{\partial \beta_r \partial \beta_s} \right] = E \left[ \frac{\partial \ell(\beta)}{\partial \beta_r} \frac{\partial \ell(\beta)}{\partial \beta_s} \right] = \phi^{-2} \sum_{i=1}^{n} E(Y_i - \mu_i)^2 \frac{1}{V_i^2} \left( \frac{d\mu_i}{d\eta_i} \right)^2 x_{ir}x_{is}$$
Estimation algorithm (Nelder and Wedderburn, 1972) – Iteratively weighted least squares (IWLS)

\[ X^T W^{(m)} X \beta^{(m+1)} = X^T W^{(m)} z^{(m)} \]

where

- \( X = [x_1, x_2, \ldots, x_n]^T \) is a design matrix \( n \times p \),
- \( W = \text{diag}\{W_i\} \) – depends on the known (prior) weights \( (w_i) \), variance function (distribution) and link function

\[
W_i = \frac{w_i}{V(\mu_i)} \left( \frac{d \mu_i}{d \eta_i} \right)^2
\]

- \( \beta \) – parameter vector \( p \times 1 \)
- \( z \) – a vector \( n \times 1 \) (adjusted response variable) – depends on \( y \) and link function

\[
z_i = \eta_i + (y_i - \mu_i) \frac{d \eta_i}{d \mu_i}
\]
Therefore, the Fisher scoring algorithm proceeds as follows.

1. Choose an initial estimate $\beta^{(m)}$ for $\hat{\beta}$ at $m = 0$.
2. Evaluate $\eta^{(m)}$, $W^{(m)}$ and $z^{(m)}$ at $\beta^{(m)}$.
3. Calculate $\beta^{(m+1)} = [X^T W^{(m)} X]^{-1} X^T W^{(m)} z^{(m)}$.
4. If $||\beta^{(m+1)} - \beta^{(m)}|| >$ some prespecified (small) tolerance then set $m \rightarrow m + 1$ and go to 2.
5. Use $\beta^{(m+1)}$ as the solution for $\hat{\beta}$. 

Clarice G.B. Demétrio
Generalized Linear Models and Extensions
When the dispersion parameter is unknown, it may be estimated by the Pearson Estimator

\[ \hat{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \]

where \( \hat{\mu}_i = g^{-1}(\hat{\beta}' x_i) \) is the \( i \)th fitted value.

Some computer packages estimate \( \phi \) by the deviance estimator \( D(\hat{\beta})/(n - p) \); but it cannot be recommended because of problems with bias and inconsistency in the case of a non-constant variance function.

For positive data, the deviance may also be sensitive to rounding errors for small values of \( y_i \).

The asymptotic variance of \( \hat{\beta} \) is estimated by the inverse (Fisher) information matrix, giving

\[ \text{Var}(\hat{\beta}) = K = \phi(X^T W X)^{-1}, \]

where \( W \) is calculated from \( \hat{\beta} \).

The standard error \( \text{se}(\hat{\beta}_j) \) is calculated as the square-root of the \( j \)th diagonal element of this matrix, for \( j = 1, \ldots, p \).
When $\phi$ is known, a $1 - \alpha$ confidence interval for $\beta_j$ is defined by the endpoints

$$\hat{\beta}_j \pm se(\hat{\beta}_j)z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ standard normal quantile.

For $\phi$ unknown, we replace $\phi$ by $\hat{\phi}$ in $K$ and a $1 - \alpha$ confidence interval for $\beta_j$ is defined by the endpoints

$$\hat{\beta}_j \pm se(\hat{\beta}_j)t_{(1-\alpha/2)}(n-p)$$

where $t_{(1-\alpha/2)}(n-p)$ is the $1 - \alpha/2$ quantile of Student’s t distribution with $n - p$ degrees of freedom.
Analysis of deviance is the method of parameter inference for generalized linear models based on the deviance, generalizing ideas from ANOVA, and first introduced by Nelder and Wedderburn (1972).

The situation is similar to regression analysis, in the sense that model terms must be eliminated sequentially, and the significance of a term may depend on which other terms are in the model.

The deviance $D$ measures the distance between $y$ and $\hat{\mu}$, given by

$$S = \frac{D(\hat{\beta})}{\phi} = -2[\log L(\hat{\mu}, y) - \log L(y, y)] = 2\phi^{-1} \sum_{i=1}^{n} w_i [y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)]$$

where $L(\hat{\mu}, y)$ and $L(y, y)$ are the likelihood function values for the current and saturated models, $\tilde{\theta}_i = \theta(y_i)$, $\hat{\theta}_i = \theta(\hat{\mu}_i)$ and $D(\hat{\beta}) = \sum_{i=1}^{n} w_i d(y_i; \hat{\mu}_i)$. 

### Deviance for some models

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$D_p = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$</td>
</tr>
<tr>
<td>Binomial</td>
<td>$D_p = 2 \sum_{i=1}^{n} \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (m_i - y_i) \log \left( \frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right]$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$D_p = 2 \sum_{i=1}^{n} \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (\hat{\mu}_i - y_i) \right]$</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>$D_p = 2 \sum_{i=1}^{n} \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (y_i + k) \log \left( \frac{\hat{\mu}_i + k}{y_i + k} \right) \right]$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$D_p = 2 \sum_{i=1}^{n} \left[ \log \left( \frac{\hat{\mu}_i}{y_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$D_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}$</td>
</tr>
</tbody>
</table>
We consider separately the cases where $\phi$ is known and unknown, but first we introduce some notation.

Let $M_1$ denote a model with $p$ parameters, and let $D_1 = D(\hat{\beta})$ denote the minimized deviance under $M_1$.

Let $M_2$ denote a sub-model of $M_1$ with $q < p$ parameters, and let $D_2$ denote the corresponding minimized deviance, where $D_2 \geq D_1$. 
Mainly relevant for discrete data, for which, in general, $\phi = 1$.

The deviance $D_1$ is a measure of goodness-of-fit of the model $M_1$; and is also known as the $G^2$ statistic in discrete data analysis.

A more traditional goodness-of-fit statistic is Pearson’s $X^2$ statistic

$$X^2 = \sum \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Asymptotically, for large $w$ the statistics $D_1$ and $X^2$ are equivalent and distributed as $\chi^2(n - p)$ under $M_1$.

Various numerical and analytical investigations have shown that the limiting $\chi^2$ distribution is approached faster for the $X^2$ statistic than for $D_1$, at least for discrete data.

A formal level $\alpha$ goodness-of-fit test for $M_1$ is obtained by rejecting $M_1$ if $X^2 > \chi^2_{(1-\alpha)}(n - p)$

This test may be interpreted as a test for overdispersion.

The fit of a model is a complex question, cannot be summarized in a single number – supplement with an inspection of residuals.
To test the sub-model $M_2$ with $q < p$ we use the log likelihood ratio statistic

$$D_2 - D_1 \sim \chi^2(p - q)$$

$M_2$ is rejected at level $\alpha$ if $D_2 - D_1 > \chi^2(1-\alpha)(p - q)$

In the case where $\phi \neq 1$ we use the scaled deviance $D/\phi$ instead of $D$; and the scaled Pearson statistic $X^2/\phi$ instead of $X^2$ and so on.
Unknown dispersion $\phi$ parameter

- The dispersion parameter is usually unknown for continuous data.
- In the discrete case we may prefer to work with unknown dispersion parameter, if evidence of overdispersion has been found in the data.
- There is no formal goodness-of-fit test available based on $X^2$ – the fit of the model $M_1$ to the data must be checked by residual analysis.
- $X^2$ is used to estimate the dispersion parameter

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

where $\hat{\mu}_i = g^{-1}(\hat{\beta}' x_i)$ is the $i$th fitted value.

- To test the sub-model $M_2$ with $q < p$ parameters inference may be based on $F$–statistic,

$$F = \frac{(D_2 - D_1)/(p-q)}{\hat{\phi}} \sim F(p-q, n-p)$$

- We reject $M_2$ at level $\alpha$ if $F > F_{1-\alpha}(p-q, n-p)$
Suppose a completely randomized experiment with two factors $A$ (with $a$ levels) and $B$ (with $b$ levels) and $r$ replications

<table>
<thead>
<tr>
<th>Model</th>
<th>DF</th>
<th>Deviance</th>
<th>Deviance Diff.</th>
<th>DF Diff.</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>$rab - 1$</td>
<td>$D_1$</td>
<td>$D_1 - D_A$</td>
<td>$a - 1$</td>
<td>$A$ ignoring $B$</td>
</tr>
<tr>
<td>$A$</td>
<td>$a(rb - 1)$</td>
<td>$D_A$</td>
<td>$D_A - D_{A+B}$</td>
<td>$b - 1$</td>
<td>$B$ including $A$</td>
</tr>
<tr>
<td>$A+B$</td>
<td>$a(rb - 1) - (b - 1)$</td>
<td>$D_{A+B}$</td>
<td>$D_{A+B} - D_{A*B}$</td>
<td>$(a - 1)(b - 1)$</td>
<td>Interaction $AB$ included $A$ and $B$</td>
</tr>
<tr>
<td>$A+B+A.B$</td>
<td>$ab(r - 1)$</td>
<td>$D_{A*B}$</td>
<td>$D_{A*B}$</td>
<td>$ab(r - 1)$</td>
<td>Residual</td>
</tr>
<tr>
<td>Saturated</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Residual analysis

- **Pearson residual**
  \[ r_{Pi} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \]
  reflect the skewness of the underlying distribution.

- **Deviance residual**
  \[ r_{Di} = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d(y_i; \hat{\mu}_i)} \]
  which is much closer to being normal than the Pearson residual, but has a bias (Jorgensen, 2011).

- **Modified deviance residual** (Jorgensen, 1997)
  \[ r_{Di}^* = r_{Di} + \phi \frac{r_{Wi}}{r_{Di}} \log \frac{r_{Wi}}{r_{Di}} \]
  where \( r_{Wi} \) is the Wald residual defined by
  \[ r_{Wi} = [g_0(y_i) - g_0(\mu_i)] \sqrt{V(y_i)} \]
  where \( g_0 \) is the canonical link.
All those residuals have approximately mean zero and variance
\( \phi(1 - h_i) \), where \( h_i \) is the \( i \)th diagonal element of the matrix
\[
H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}.
\]

Use standardized residuals such as \( r_{Di}^*(1 - h_i)^{1/2} \), which are nearly normal with variance \( \phi \).

Plot residuals against the fitted values – to check the proposed variance function.

Normal Q-Q plot (or normal Q-Q plot with simulated envelopes) for the residuals – to check the correctness of the distributional assumption.
Checking for the link function

Suppose the link function \( g_0(\mu) = g(\mu, \lambda_0) = X\beta \), nested in a parametric family \( g(\mu, \lambda) \), indexed by the parameter \( \lambda \), for example,

\[
g(\mu, \lambda) = \begin{cases} 
\frac{\mu^\lambda - 1}{\lambda} & \lambda \neq 0 \\
\log(\mu) & \lambda = 0 
\end{cases}
\]

which includes the identity, logarithmic links, or the Aranda-Ordaz family,

\[
\mu = \begin{cases} 
1 - (1 + \lambda e^\eta)^{-\frac{1}{\lambda}} & \lambda e^\eta > -1 \\
1 & \text{c.c.}
\end{cases}
\]

which includes the identity, complementary log-log links.

The Taylor expansion for \( g(\mu, \lambda) \) around \( \lambda_0 \), gives

\[
g(\mu, \lambda) \simeq g(\mu, \lambda_0) + (\lambda - \lambda_0)u(\lambda_0) = X\beta + \gamma u(\lambda_0)
\]

em que \( u(\lambda_0) = \frac{\partial g(\mu, \lambda)}{\partial \lambda} \bigg|_{\lambda=\lambda_0} \).

In general it is used \( u(\lambda_0) = \hat{\eta}^2 \).
Suppose the link function used was $\eta = g(\mu)$ and that the true link is $g^*(\mu)$. Then,

$$g(\mu) = g[g^{-1}(\eta)] = h(\eta)$$

The null hypothesis is $H_0 : h(\eta) = \eta$ and $H_a : h(\eta) = \text{ não linear}$. Using Taylor expansion for $g(\mu)$ we have:

$$g(\mu) \approx h(0) + \eta h'(0) + \eta^2 \frac{h''(0)}{2}$$

then, the added variable is $\hat{\eta}^2$, assuming that the model has terms for which the mean is marginal.

**Formal tests**
- Likelihood ratio test
- Score Test
- Wald Test

**Graphics**
- Added variable plot
- Partial residual plot
R commands for GLM

```r
glm(resp ~ linear predictor + offset(of), weights = w,
family=familyname(link ="linkname" ))
```

The `resp` is the response variable `y`. For a binomial regression model it is necessary to create:

```r
resp<-cbind(y,n-y)
```

The possible families ("canonical link") are:

```r
binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

The default family is the gaussian family and default links are the canonical links (don’t need to be declared). Other possible links are "probit", "cloglog", "cauchit", "sqrt", etc. To see more, type

```r
? glm
```
**Example**

Average daily fat yields (kg/day) from milk from a single cow for each of 35 weeks (McCulloch, 2001)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.31</td>
<td>0.39</td>
<td>0.50</td>
<td>0.58</td>
<td>0.59</td>
</tr>
<tr>
<td>0.68</td>
<td>0.66</td>
<td>0.67</td>
<td>0.70</td>
<td>0.72</td>
</tr>
<tr>
<td>0.65</td>
<td>0.64</td>
<td>0.57</td>
<td>0.48</td>
<td>0.46</td>
</tr>
<tr>
<td>0.31</td>
<td>0.33</td>
<td>0.36</td>
<td>0.30</td>
<td>0.26</td>
</tr>
<tr>
<td>0.29</td>
<td>0.31</td>
<td>0.29</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>0.11</td>
<td>0.07</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
A typical model:
Fat yield \( \approx \alpha t^\beta e^{\gamma t} \) where \( t=\)week

Transform
\[
Y_i = \alpha t_i^\beta e^{\gamma t_i} e^{\epsilon_i}
\]
\[
\log(Y_i) = \mu_i + \epsilon_i = \log \alpha + \beta \log(t_i) + \gamma_i t + \epsilon_i
\]
\[
\log(Y_i) \sim N(\log \alpha + \beta \log(t_i) + \gamma_i t, \sigma^2)
\]
\[
E[\log(Y_i)] = \log \alpha + \beta \log(t_i) + \gamma_i t
\]

Link
\[
Y_i = \mu_i + \xi_i = \alpha t_i^\beta e^{\gamma t_i} + \xi_i
\]
\[
Y_i \sim N(\alpha t_i^\beta e^{\gamma t_i}, \tau^2)
\]
\[
E[Y_i] = \alpha t_i^\beta e^{\gamma t_i}
\]
\[
\log(E[Y_i]) = \log \alpha + \beta \log(t_i) + \gamma_i t
\]
Plot of fat yield for each week – Observed values and fitted curve

周恩来 (kg/day)

* Observed
* Log transformation
* Log link

Clarice G.B. Demétrio
Generalized Linear Models and Extensions
# Average daily fat yields (kg/day) from milk
# from a single cow for each of 35 weeks
# McCulloch (2001)
# Ruppert, Cressie, Carroll (1989), Biometrics, 45:637–656

fatyield.dat<-scan(what=list(yield=0))
0.31 0.39 0.50 0.58 0.59 0.64
0.68 0.66 0.67 0.70 0.72 0.68
0.65 0.64 0.57 0.48 0.46 0.45
0.31 0.33 0.36 0.30 0.26 0.34
0.29 0.31 0.29 0.20 0.15 0.18
0.11 0.07 0.06 0.01 0.01

fatyield.dat$week=1:35
attach(fatyield.dat)
lweek<-log(week)

## Normal model for log(fat)
lyield<-log(yield)
mod1<-lm(lyield~week+lweek)
summary(mod1)
anova(mod1)
## Normal model with log link

```r
mod2 <- glm(yield ~ week + lweek, family = gaussian(link = "log"))
fit2 <- fitted(mod2)
summary(mod2)
anova(mod2)
```

# Plotting

```r
plot(c(0, 35), c(0, 0.9), type = "n", xlab = "Weeks", ylab = "Fat yield (kg/day)"
points(week, yield, pch = "*

w <- seq(1, 35, 0.1)
lines(w, predict(mod2, data.frame(week = w, lweek = log(w)), type = "response"), col = "green", lty = 1)
lines(w, exp(predict(mod1, data.frame(week = w, lweek = log(w)), type = "response")), col = "red", lty = 2)
legend(20, 0.9, c("Observed", "Log transformation", "Log link"), lty = c(-1, 2, 1), pch = c("*", ", " ), col = c("black", "red", "green"), cex = .6)
title(sub = "Figure 1. Fat yield (kg/day) for each week")
```
Binomial regression model

Example

Batches of 20 pyrethroid-resistant moths (*Heliothis virescens*), a cotton crop pest) of each sex were exposed to a range of doses of cypermethrin two days after emergence from pupation. The number of moths which were either knocked down or dead was recorded after 72h.

<table>
<thead>
<tr>
<th>Doses ($d_i$)</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2.0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4.0</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>8.0</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>16.0</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>32.0</td>
<td>20</td>
<td>16</td>
</tr>
</tbody>
</table>

- Response variable: $Y_i$ – number of dead insects out of 20 insects.
- Distribution: Binomial.
- Systematic component: regression model, completely randomized experiment.
- Aim: Lethal doses.
### Cypermethrin toxicity example

#### Deviance residuals

<table>
<thead>
<tr>
<th>Terms fitted in model</th>
<th>d.f.</th>
<th>Deviance</th>
<th>$X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>11</td>
<td>124.9</td>
<td>101.4</td>
</tr>
<tr>
<td>Sex</td>
<td>10</td>
<td>118.8</td>
<td>97.4</td>
</tr>
<tr>
<td>Dose</td>
<td>6</td>
<td>15.2</td>
<td>12.9</td>
</tr>
<tr>
<td>Sex + Dose</td>
<td>5</td>
<td>5.0</td>
<td>3.7</td>
</tr>
</tbody>
</table>

#### Analysis of deviance

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Deviance</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>6.1</td>
<td>0.0144</td>
</tr>
<tr>
<td>Sex</td>
<td>Dose</td>
<td>1</td>
<td>10.2</td>
</tr>
<tr>
<td>Dose</td>
<td>5</td>
<td>109.7</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Dose</td>
<td>Sex</td>
<td>5</td>
<td>113.8</td>
</tr>
<tr>
<td>Residual</td>
<td>5</td>
<td>5.0</td>
<td>0.5841</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>124.9</td>
<td></td>
</tr>
</tbody>
</table>
Cypermethrin toxicity example

Models

\[ \text{logit}(p) = \alpha_j + \beta_j \log_2(\text{dose}) \] – different logistic regression lines

\[ \text{logit}(p) = \alpha_j + \beta \log_2(\text{dose}) \] – common slope

\[ \text{logit}(p) = \alpha + \beta_j \log_2(\text{dose}) \] – common intercept

\[ \text{logit}(p) = \alpha + \beta \log_2(\text{dose}) \] – same logistic regression line.
Cypermethrin toxicity example

Residual Deviances

<table>
<thead>
<tr>
<th>Terms fitted in model</th>
<th>d.f.</th>
<th>Deviance</th>
<th>$X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>11</td>
<td>124.9</td>
<td>101.4</td>
</tr>
<tr>
<td>Sex + Sex log$_2$(dose)</td>
<td>8</td>
<td>4.99</td>
<td>3.51</td>
</tr>
<tr>
<td>Sex + log$_2$(dose)</td>
<td>9</td>
<td>6.75</td>
<td>5.31</td>
</tr>
<tr>
<td>Const. + Sex log$_2$(dose)</td>
<td>9</td>
<td>5.04</td>
<td>3.50</td>
</tr>
<tr>
<td>Const. + log$_2$(dose)</td>
<td>10</td>
<td>16.98</td>
<td>14.76</td>
</tr>
</tbody>
</table>

Analysis of deviance

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Deviance</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>6.1</td>
<td>0.0144</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>1</td>
<td>112.0</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>6.8</td>
<td>0.7473</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>124.9</td>
<td></td>
</tr>
</tbody>
</table>
Regression equations
males
\[
\log \frac{\hat{p}}{1 - \hat{p}} = -2.372 + 1.535 \log_2(\text{dose})
\]
females
\[
\log \frac{\hat{p}}{1 - \hat{p}} = -3.473 + 1.535 \log_2(\text{dose})
\]

Lethal Doses
males
\[
\log_2(\hat{LD}_{50}) = \frac{2.372}{1.535} = 1.55 \Rightarrow \hat{LD}_{50} = 4.69
\]
females
\[
\log_2(\hat{LD}_{50}) = \frac{3.473}{1.535} = 2.26 \Rightarrow \hat{LD}_{50} = 9.61
\]
Cypermethrin toxicity example

Plot of observed proportions and fitted curves
y <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex<-factor(rep(c("M","F"), c(6,6)))
ldose<-rep(0:5,2)
dose<-2**ldose
dose<-factor(dose)
Cyper.dat <- data.frame(sex, dose, ldose, y)
attach(Cyper.dat)

plot(ldose,y/20, pch=c(rep("*",6),rep("+",6)),
col=c(rep("green",6), rep("red",6)),
xlab="log(dose)" , ylab="Proportion killed")

resp<-cbind(y,20-y)

mod1<-glm(resp~1, family=binomial)
mod2<-glm(resp~dose, family=binomial)
mod3<-glm(resp~sex, family=binomial)
mod4<-glm(resp~dose+sex, family=binomial)
anova(mod1, mod2, mod4, test="Chisq")
anova(mod1, mod3, mod4, test="Chisq")
mod5<-glm(resp~ldose, family=binomial)
mod6<-glm(resp~sex+ldose-1, family=binomial)
mod7<-glm(resp~ldose/sex, family=binomial)
mod8<-glm(resp~ldose*sex, family=binomial)
anova(mod1, mod5, mod6, mod8, test="Chisq")
anova(mod1, mod5, mod6, mod7, mod8, test="Chisq")
summary(mod6)

## Plotting
plot(c(1,32), c(0,1), type="n", xlab="log(dose)", ylab="Proportions", log="x")
points(2**ldose,y/20, pch=c(rep("*",6),rep("+",6)),
col=c(rep("green",6),rep("red",6)))
ld<-seq(0,5,0.1)
lines(2**ld, predict(mod6,data.frame(ldose=ld,
sex=factor(rep("M",length(ld)),levels=levels(sex))),
type="response"), col="green")
lines(2**ld, predict(mod6,data.frame(ldose=ld,
sex=factor(rep("F",length(ld)),levels=levels(sex))),
type="response"), col="red")
CS2 toxicity

Mortality of adult beetles after five hours’ exposure to gaseous carbon disulphid (Bliss, 1935).

<table>
<thead>
<tr>
<th>Log dosage</th>
<th>Number exposed</th>
<th>Number killed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6907</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>1.7242</td>
<td>60</td>
<td>13</td>
</tr>
<tr>
<td>1.7552</td>
<td>62</td>
<td>18</td>
</tr>
<tr>
<td>1.7842</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>1.8113</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>1.8369</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>1.8610</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>1.8839</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

Considerations

- **Response variable**: $Y_i$ – number of killed beetles out of $m_i$ exposed beetles (Pregibon, 1980).
- **Distribution**: Binomial.
- **Systematic component**: regression model, completely randomized experiment.
- **Aim**: Lethal doses.
Binomial logistic model with $\eta = \beta_0 + \beta_1 \log(dose)$ gives deviance 11.23 on 6 d.f. $\Rightarrow$ lack of fit

Deviance for link test 8.037 on 1 d.f. $\Rightarrow$ need for different link

Added variable plot for $U = \hat{\eta}^2$ $\Rightarrow$ evidence that $\gamma \neq 0$, spread throughout data

Figure: CS2 - Added variable plot

- C-loglog link with a linear lp gives good fit
- Logistic link with a quadratic lp gives good fit
Figure: Beetle mortality - Observed proportions and fitted curves
# *** Mortality of adult beetle Example ***

y <- c(6, 13, 18, 28, 52, 53, 61, 60)
m <- c(59, 60, 62, 56, 63, 59, 62, 60)
ldose <- c(1.6907, 1.7242, 1.7552, 1.7842, 1.8113, 1.8369, 1.8610, 1.8839)
beetle.dat <- data.frame(ldose, m, y)
attach(beetle.dat)
resp<-cbind(y,m-y)

## Logistic link function
mod1<-glm(resp ~ poly(I(ldose))+poly(I(ldose),2), family=binomial(link="logit"))
anova(mod1, test="Chisq")

mod1l<-glm(resp ~ 1, family=binomial(link="logit"))
1-pchisq(deviance(mod1l), df.residual(mod1l))
print(sum(residuals(mod1l, 'pearson')^2))

mod2l<-glm(resp ~ I(ldose), family=binomial(link="logit"))
1-pchisq(deviance(mod2l), df.residual(mod2l))
print(sum(residuals(mod2l, 'pearson')^2))

## Link function test
LP2l <- (predict(mod2l,type = c("link")))^2
mod3l<-update(mod2l , .~. +LP2l, family=binomial(link="logit"))
anova(mod3l, test="Chisq")

mod4l<-glm(resp ~ poly(I(ldose),2), family=binomial(link="logit"))
1-pchisq(deviance(mod4l), df.residual(mod4l))
print(sum(residuals(mod4l, 'pearson')^2))
## Link function test
LP2q <- (predict(mod4l,type = c("link")))^2/mod5l<-update(mod4l , .~. +LP2q, family=binomial(link="logit"))
anova(mod5l, test="Chisq")

## Added variable plot
rp <- residuals(mod2l, 'pearson')
W <- mod2l$fitted*(1- mod2l$fitted)*m
U <- LP2l
mod1n <- glm(U~I(ldose), family=gaussian, weights=W)
rU <- (U - mod1n$fitted)
plot(rU,rp, xlab="Ordinary residuals", ylab="Pearson residuals")

# Partial residual plot
resparc <- residuals(mod3l, 'pearson') + mod3l$coef[3]*U
plot(U, resparc, xlab="U", ylab="Residual + component")

summary(mod6l<-glm(resp ~ I(ldose)+I(ldose^2), family=binomial(link="logit")))
anova(mod6l, test="Chisq")
library(MASS)
# variance-covariance matrix of estimated parameters
vcov(mod6l)
## Complementary log-log link function

```r
modc<-glm(resp ~ poly(I(ldose))+poly(I(ldose),2), family=binomial(link="cloglog"))
anova(modc, test="Chisq")
```

```r
mod1c<-glm(resp ~ 1, family=binomial(link="cloglog"))
1-pchisq(deviance(mod1c), df.residual(mod1c))
print(sum(residuals(mod1c, 'pearson')^2))
```

```r
mod2c<-glm(resp ~ I(ldose), family=binomial(link="cloglog"))
1-pchisq(deviance(mod2c), df.residual(mod2c))
print(sum(residuals(mod2c, 'pearson')^2))
```

## Link function test

```r
LP2l <- (predict(mod2c,type = c("link")))^2
mod3c<-update(mod2c , .~. +LP2l, family=binomial(link="cloglog"))
anova(mod3c, test="Chisq")
```

## Added variable plot

```r
rp <- residuals(mod2c, 'pearson')
W <- mod2c$fitted*(1- mod2c$fitted)*m
U <- LP2l
mod1n <- glm(U~I(ldose), family=gaussian, weights=W)
rU <- (U - mod1n$fitted)
plot(rU,rp, xlab="Ordinary residuals", ylab="Pearson residuals")
```

# Partial residual plot

```r
resparc <- residuals(mod3c, 'pearson') + mod3c$coef[3]*U
plot(U,resparc, xlab="U", ylab="Residual + component")
```
R program

```r
summary(mod5c <- glm(resp ~ I(ldose), family = binomial(link = "cloglog")))
library(MASS)
# variance-covariance matrix of estimated parameters
vcov(mod5c)

# Doses LD50 and LD90
dose.p(mod5c); dose.p(mod5c, p = 0.9)

# Doses LD25, LD50, LD75
dose.p(mod5c, p = 1:3/4)

# Plotting
plot(c(1.69, 1.89), c(0, 1), type = "n", xlab = "Log(dose)",
ylab = "Proportion of beetles killed")
points(ldose, y/m, pch = "*")
ld <- seq(1.69, 1.89, 0.01)
lines(ld, predict(mod4l, data.frame(ldose = ld), type = "response"), col = "blue", lty = 1)
lines(ld, predict(mod2c, data.frame(ldose = ld), type = "response"), col = "red", lty = 2)
legend(1.7, 1.0, c("Observed", "Logit", "Cloglog"), lty = c(-1, 1, 2),
pch = c("*", " ", " "), col = c("black", "blue", "red"))
```
Germination of Orobanche seed

Example

<table>
<thead>
<tr>
<th></th>
<th>O. aegyptiaca 75</th>
<th>O. aegyptiaca 73</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bean</td>
<td>Cucumber</td>
</tr>
<tr>
<td>10/39</td>
<td>5/6</td>
<td>8/16</td>
</tr>
<tr>
<td>23/62</td>
<td>53/74</td>
<td>10/30</td>
</tr>
<tr>
<td>23/81</td>
<td>55/72</td>
<td>8/28</td>
</tr>
<tr>
<td>26/51</td>
<td>32/51</td>
<td>23/45</td>
</tr>
<tr>
<td>17/39</td>
<td>46/79</td>
<td>0/4</td>
</tr>
<tr>
<td></td>
<td>10/13</td>
<td></td>
</tr>
</tbody>
</table>

- Response variable: $Y_i$ – number of germinated seeds out of $m_i$ seeds.
- Distribution: Binomial.
- Systematic component: factorial $2 \times 2$ (2 species, 2 extracts), completely randomized experiment (Crowder, 1978).
- Aim: to see how germination is affected by species and extracts.
- Problem: overdispersion.
Germination of Orobanche seed

**Simple Binomial Model**
- No of seeds germinating $y_i$ as y-variable
- Binomial logit model
- Species * Extract interaction model

**Analysis of deviance**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Deviance</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>1</td>
<td>55.97</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>S</td>
<td>1</td>
<td>56.49</td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>2.54</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>E</td>
<td>1</td>
<td>3.06</td>
</tr>
<tr>
<td>S.E</td>
<td>1</td>
<td>6.41</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>17</td>
<td>33.28</td>
<td>0.0096</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>98.72</td>
<td></td>
</tr>
</tbody>
</table>

- Interaction significant
- Some evidence of overdispersion?
Germination of Orobanche seed

Normal plot with simulated envelope

![Normal plot with simulated envelope](image-url)
Species <- factor(c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2))
Extract <- factor(c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2))
y <- c(10, 23, 23, 26, 17, 5, 53, 55, 32, 46, 10, 8, 10, 8, 23, 0, 3, 22, 15, 32, 3)
m <- c(39, 62, 81, 51, 39, 6, 74, 72, 51, 79, 13, 16, 30, 28, 45, 4, 12, 41, 30, 51, 7)
orobanch <- data.frame(Species, Extract, m, y)
attach(orobanch)
resp<-cbind(y,m-y)
p<-y/m

# Binomial fit
orobanchB.fit<-glm(resp~Species*Extract, family=binomial)
summary(orobanchB.fit)
anova(orobanchB.fit, test="Chisq")

orobanchB.fit<-glm(resp~Extract*Species, family=binomial)
summary(orobanchB.fit)
anova(orobanchB.fit, test="Chisq")
Poisson regression model

Example

Storing of micro-organisms

Bacterial concentrations (counts per fixed area) measured at initial freezing (−70°C) and at 1, 2, 6, 12 months.

<table>
<thead>
<tr>
<th>Time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>6</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>31</td>
<td>26</td>
<td>19</td>
<td>15</td>
<td>20</td>
</tr>
</tbody>
</table>

Hypothesized that count decays over time, i.e.

\[
\text{average count } \propto \frac{1}{(\text{Time})^\gamma}
\]

Model

\[
\text{Count } \sim \text{ Pois}(\mu)
\]

\[
\log \mu = \beta_0 + \beta_1 \log(\text{Time})
\]

Avoid problems at time 0 by using

\[
\log \mu = \beta_0 + \beta_1 \log(\text{Time} + 0.1)
\]
Residual deviances

<table>
<thead>
<tr>
<th>Model</th>
<th>d.f.</th>
<th>Deviance</th>
<th>$X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>4</td>
<td>7.0672</td>
<td>7.1532</td>
</tr>
<tr>
<td>log(Time)</td>
<td>3</td>
<td>1.8338</td>
<td>1.8203</td>
</tr>
</tbody>
</table>

Analysis of deviance

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Deviance</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>1</td>
<td>5.2334</td>
<td>0.0222</td>
</tr>
<tr>
<td>Error</td>
<td>3</td>
<td>1.8338</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>7.0672</td>
<td></td>
</tr>
</tbody>
</table>

$log(\hat{\mu}) = 3.149 - 0.1261 \ log(\text{Time})$
Storing of micro-organisms example

Plot of bacterial concentration: observed values and fitted curve
R program

ltim <- log((tim <- c(0, 1, 2, 6, 12)) + 0.1)
lcount <- log(count <- c(31, 26, 19, 15, 20))
bacteria.dat <- data.frame(tim, count, ltim, lcount)

with(bacteria.dat, {
  par(mfrow=c(1,2))
  plot(tim, count, xlab="Time in months", ylab="Counts")
  plot(ltim, lcount, xlab="Log(time in months)", ylab="Log(counts)"
  par(mfrow=c(1,1))
})

mod1 <- glm(count ~ tim, family=poisson, bacteria.dat)
anova(mod1, test="Chisq")

mod2 <- glm(count ~ ltim, family=poisson)
anova(mod2, test="Chisq")

plot(c(0,12), c(15,31), type="n", xlab="Time in months", ylab="Counts")
points(tim, count, pch="*

x <- seq(0, 12, 0.1)
lp <- predict(mod2, data.frame(ltim=log(x+0.1)), type="response")
lines(x, lp, lty=1)
title(sub="Figure 1. Data and Fitted curve")
Exercise Relative potency - Toxicity of insecticide to flour beetles

Considerations

- Response variable: $Y_i$ – number of dead insects out of $m_i$ insects.
- Distribution: Binomial.
- Systematic component: ANOVA with regression model, completely randomized experiment.
- Aim: Lethal doses and comparison of insecticides.
Exercise

Count of the number of plant species on plots that have different biomass (a continuous explanatory variable with three levels: high, mid and low) (Venables and Ripley, 1994)

**Tabla:** Number of plant species (Y), quantity of biomass (X) and levels of pH of the soil.

<table>
<thead>
<tr>
<th>pH level</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>18</td>
<td>0.1008</td>
<td>15</td>
<td>2.6292</td>
<td>13</td>
<td>0.6526</td>
<td>8</td>
<td>3.6787</td>
<td>9</td>
<td>1.5079</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0.1385</td>
<td>9</td>
<td>3.2522</td>
<td>9</td>
<td>1.5553</td>
<td>2</td>
<td>4.8315</td>
<td>8</td>
<td>2.3259</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.8635</td>
<td>3</td>
<td>4.4172</td>
<td>8</td>
<td>1.6716</td>
<td>17</td>
<td>0.2897</td>
<td>12</td>
<td>2.9957</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>1.2929</td>
<td>2</td>
<td>4.7808</td>
<td>14</td>
<td>2.8700</td>
<td>14</td>
<td>0.0775</td>
<td>14</td>
<td>3.5381</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>2.4691</td>
<td>18</td>
<td>0.0501</td>
<td>13</td>
<td>2.5107</td>
<td>15</td>
<td>1.4290</td>
<td>7</td>
<td>4.3645</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2.3665</td>
<td>19</td>
<td>0.4828</td>
<td>4</td>
<td>3.4976</td>
<td>17</td>
<td>1.1207</td>
<td>3</td>
<td>4.8705</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>0.1757</td>
<td>30</td>
<td>1.3767</td>
<td>21</td>
<td>2.5510</td>
<td>18</td>
<td>3.0002</td>
<td>13</td>
<td>4.9056</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>5.3433</td>
<td>9</td>
<td>7.7000</td>
<td>24</td>
<td>0.5536</td>
<td>26</td>
<td>1.9902</td>
<td>26</td>
<td>2.9126</td>
</tr>
<tr>
<td>Mid</td>
<td>20</td>
<td>3.2164</td>
<td>21</td>
<td>4.9798</td>
<td>15</td>
<td>5.6587</td>
<td>8</td>
<td>8.1000</td>
<td>31</td>
<td>0.7395</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>1.5269</td>
<td>18</td>
<td>2.2321</td>
<td>16</td>
<td>3.8852</td>
<td>19</td>
<td>4.6265</td>
<td>20</td>
<td>5.1209</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>8.3000</td>
<td>25</td>
<td>0.5112</td>
<td>23</td>
<td>1.4782</td>
<td>25</td>
<td>2.9345</td>
<td>22</td>
<td>3.5054</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>4.6179</td>
<td>11</td>
<td>5.6969</td>
<td>17</td>
<td>6.0930</td>
<td>24</td>
<td>0.7300</td>
<td>27</td>
<td>1.1580</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.4692</td>
<td>39</td>
<td>1.7308</td>
<td>44</td>
<td>2.0897</td>
<td>35</td>
<td>3.9257</td>
<td>25</td>
<td>4.2667</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>5.4819</td>
<td>23</td>
<td>6.6846</td>
<td>18</td>
<td>7.5116</td>
<td>19</td>
<td>8.1322</td>
<td>12</td>
<td>9.5721</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>0.0866</td>
<td>35</td>
<td>1.2369</td>
<td>30</td>
<td>2.5320</td>
<td>30</td>
<td>3.4079</td>
<td>33</td>
<td>4.6050</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5.3677</td>
<td>26</td>
<td>6.5608</td>
<td>36</td>
<td>7.2420</td>
<td>18</td>
<td>8.5036</td>
<td>7</td>
<td>9.3909</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>0.7648</td>
<td>39</td>
<td>1.1764</td>
<td>34</td>
<td>2.3251</td>
<td>31</td>
<td>3.2228</td>
<td>24</td>
<td>4.1361</td>
</tr>
</tbody>
</table>
Overdispersion in glms (Hinde and Demétrio, 1998a,b)

Residual Deviance ≈ Residual d.f.

What if Residual Deviance $\gg$ Residual d.f.?

1. Badly fitting model
   - omitted terms/variables
   - incorrect relationship (link)
   - outliers

2. variation greater than predicted by model: $\Rightarrow$ Overdispersion
   - count data: $\text{Var}(Y) > \mu$
   - counted proportion data:
     \[ \text{Var}(Y) > n\pi(1 - \pi) \]
Overdispersion in glms (Hinde and Demétrio, 1998a,b)

Causes of Overdispersion

- variability of experimental material
  - individual level variability
- correlation between individual responses
  e.g. litters of rats
- cluster sampling
  e.g. areas; schools; classes; children
- aggregate level data
- omitted unobserved variables
- excess zero counts (structural and sampling zeros)

Consequences
With correct mean model we have consistent estimates of $\beta$ but:

- incorrect standard errors
- selection of overly complex models
**Worldwide Airline Fatalities, 1976-85**

<table>
<thead>
<tr>
<th>Year</th>
<th>Fatal accidents</th>
<th>Passenger deaths</th>
<th>Passenger miles (100 million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>24</td>
<td>734</td>
<td>3863</td>
</tr>
<tr>
<td>1977</td>
<td>25</td>
<td>516</td>
<td>4300</td>
</tr>
<tr>
<td>1978</td>
<td>31</td>
<td>754</td>
<td>5027</td>
</tr>
<tr>
<td>1979</td>
<td>31</td>
<td>877</td>
<td>5481</td>
</tr>
<tr>
<td>1980</td>
<td>22</td>
<td>814</td>
<td>5814</td>
</tr>
<tr>
<td>1981</td>
<td>21</td>
<td>362</td>
<td>6033</td>
</tr>
<tr>
<td>1982</td>
<td>26</td>
<td>764</td>
<td>5877</td>
</tr>
<tr>
<td>1983</td>
<td>20</td>
<td>809</td>
<td>6223</td>
</tr>
<tr>
<td>1984</td>
<td>16</td>
<td>223</td>
<td>7433</td>
</tr>
<tr>
<td>1985</td>
<td>22</td>
<td>1066</td>
<td>7107</td>
</tr>
</tbody>
</table>
Worldwide Airline Fatalities, 1976-85

Simple Models

- Passenger miles \((m_i)\) as exposure variable
- Poisson log-linear model
- Linear time trend

\[
Y_i \sim \text{Pois}(m_i \lambda_i) \quad (1)
\]

\[
\log \lambda_i = \beta_0 + \beta_1 \text{year}_i
\]

Fatal accidents:
Deviance(time trend) = 20.68
Residual Deviance = 5.46 on 8 d.f.

Passenger deaths:
Deviance(time trend) = 202.1
Residual Deviance = 1051.5 on 8 d.f.

⇒ compounding with aircraft size
Two broad categories

- assume some more general form for the variance function, possibly with additional parameters.

- assume a two-stage model for the response with the response model parameter following some distribution.
  Maximum likelihood estimation (conjugate distribution models) or approximate methods (e.g. using first two moments as above)
  Full hierarchical model – Bayesian methods
Overdispersed Proportion Data

$Y_i$ successes out of $m_i$ trials, $i = 1, \ldots, n$.

Model expected proportions $\pi_i$ with link function $g$ and

$$g(\pi_i) = \beta' x_i$$

- Constant overdispersion

$$\text{Var}(Y_i) = \phi m_i \pi_i (1 - \pi_i)$$

- A general variance function:

Overdispersion allowed to depend upon both $m_i$ and $\pi_i$.

$$\text{Var}(Y_i) = m_i \pi_i (1 - \pi_i) \times \left[ 1 + \phi (m_i - 1)^{\delta_1} \{\pi_i (1 - \pi_i)\}^{\delta_2} \right]$$
Mean-variance Models (Hinde and Demétrio, 1998a,b)

Count data
Random variables $Y_i$ represent counts with means $\mu_i$.

- **Constant overdispersion**

\[
\text{Var}(Y_i) = \phi \mu_i
\]

can arise through a simple compounding process.
Suppose that $N_i \sim \text{Pois}(\mu N)$ and $T = \sum_{i=1}^{N_i} X_i$, $X_i$ are iid random variables.

\[
\mathbb{E}[T] = \mu_T = \mathbb{E}_{N}(\mathbb{E}[T|N]) = \mu_N \mu_X
\]

\[
\text{Var}(T) = \mathbb{E}_{N}[\text{Var}(T|N)] + \text{Var}_{N}(\mathbb{E}[T|N])
\]

\[
= \mu_T \left( \frac{\sigma_X^2}{\mu_X} + \mu_X \right) = \mu_T \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]}
\]

- **A general variance function**

\[
\text{Var}(Y_i) = \mu_i \left\{1 + \phi \mu_i^\delta\right\}
\]
Beta-Binomial

\[ Y_i|P_i \sim \text{Bin}(m_i, P_i) \]
\[ E(P_i) = \pi_i \quad \text{Var}(P_i) = \phi \pi_i (1 - \pi_i) \]

Unconditionally, \( E(Y_i) = m_i \pi_i \) and
\[ \text{Var}(Y_i) = m_i \pi_i (1 - \pi_i) [1 + (m_i - 1) \phi] \]

Taking \( P_i \sim \text{Beta}(\alpha_i, \beta_i) \), with \( \alpha_i + \beta_i \) fixed, gives beta-binomial distribution for \( Y_i \) with the same variance function.
The same variance function results from assuming that individual binary responses are not independent but have a constant correlation. Writing \( Y_i = \sum_{j=1}^{m_i} R_{ij} \), where \( R_{ij} \) are Bernoulli random variables with

\[
\mathbb{E}[R_{ij}] = \pi_i \text{ and } \text{Var}(R_{ij}) = \pi_i(1 - \pi_i)
\]

then, assuming a constant correlation \( \rho \) between the \( R_{ij} \)'s for \( j \neq k \), we have

\[
\text{Cov}(R_{ij}, R_{ik}) = \rho \pi_i(1 - \pi_i)
\]

and

\[
\begin{align*}
\mathbb{E}[Y_i] &= m_i \pi_i \\
\text{Var}(Y_i) &= \sum_{j=1}^{m_i} \text{Var}(R_{ij}) + \sum_{j=1}^{m_i} \sum_{k \neq j} \text{Cov}(R_{ij}, R_{ik}) \\
&= m_i \pi_i(1 - \pi_i) + m_i(m_i - 1)[\rho \pi_i(1 - \pi_i)] \\
&= m_i \pi_i(1 - \pi_i)[1 + \rho (m_i - 1)],
\end{align*}
\]
Logistic-normal and related models
Random effect in the linear predictor

\[ \eta_i = \beta' x_i + \sigma z_i \]

- assume \( z_i \sim N(0, 1) \)
  - Probit-normal model - a convenient interpretation as a threshold model for a normally distributed latent variable (McCulloch, 1994).
  - Logistic-normal using EM algorithm with Gaussian quadrature.
  - Approximate approach using a Williams type III model with

\[
\text{Var}(Y_i) = m_i \pi_i (1 - \pi_i) [1 + \phi(m_i - 1) \pi_i (1 - \pi_i)]
\]

- make no specific distributional assumption about \( z \) - estimate a discrete mixing distribution by non-parametric maximum likelihood (NPML).
Considered as the two-stage model, the logit($P_i$) have a normal distribution with variance $\sigma^2$, i.e. \( \text{logit}(P_i) \sim N(x_i^T \beta, \sigma^2) \). Writing

$$U_i = \text{logit}(P_i) = \log \frac{P_i}{1 - P_i} \Rightarrow P_i = \frac{e^{U_i}}{1 + e^{U_i}}$$

and using Taylor series for $P_i$, around $U_i = \mathbb{E}[U_i] = x_i^T \beta$, we have

$$P_i = \frac{e^{x_i^T b \beta}}{(1 + e^{x_i^T \beta})} + \frac{e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})^2} (U_i - x_i^T \beta) + o(U_i - x_i^T \beta).$$

Then

$$\mathbb{E}(P_i) \approx \frac{e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})} := \pi_i$$

and

$$\text{Var}(P_i) \approx \left[ \frac{e^{x_i^T \beta}}{(1 + e^{x_i^T \beta})^2} \right]^2 \text{Var}(U_i) = \sigma^2 \pi_i^2 (1 - \pi_i)^2$$

Consequently the variance function for the logistic-normal model can be approximated by

$$\text{Var}(Y_i) \approx m_i \pi_i (1 - \pi_i) [1 + \sigma^2 (m_i - 1) \pi_i (1 - \pi_i)]$$

which Williams (1982) refers to as a type III variance function.
Two-stage models – Count data (Hinde and Demétrio, 1998a,b)

Negative Binomial Type Variance

- Variation in Poisson rate parameter:

\[ Y_i | \theta_i \sim \text{Pois}(\theta_i), \quad \theta_i \sim \text{Gamma}(k, \lambda_i) \]

leads to negative binomial distribution with

\[ E[Y_i] = \mu_i = \frac{k}{\lambda_i} \]

and

\[ \text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{k} \]

For known \( k \), in the 1-parameter exponential family so still in glm framework.

- Different assumptions for the \( \Gamma \)-distribution lead to different parameterizations with different overdispersed variance functions, e.g. \( \theta_i \sim \Gamma(k_i, \lambda) \) gives

\[ \text{Var}(Y_i) = \mu_i \left(1 + \frac{1}{\lambda}\right) = \phi\mu_i \]
Two-stage models – Count data (Hinde and Demétrio, 1998a,b)

**Negative binomial distribution**

\[ Y_i | \theta_i \sim \text{Pois}(\theta_i) \]
\[ \theta_i \sim \text{Gamma}(k, \lambda_i), \ i = 1, \ldots, n \]

This leads to a **negative binomial distribution** for the \( Y_i \) with

\[
 f_{Y_i}(y_i; \mu_i, k) = \frac{\Gamma(k + y_i)}{\Gamma(k)y_i!} \frac{\mu_i^{y_i}k^k}{(\mu_i + k)^{k+y_i}}, \quad y_i = 0, 1, \ldots
\]

and

\[
 \mathbb{E}(Y_i) = \frac{k}{\lambda_i} = \mu_i
\]

\[
 \text{Var}(Y_i) = \mathbb{E}_{\theta_i}[\text{Var}(Y_i | \theta_i)] + \text{Var}_{\theta_i}(\mathbb{E}[Y_i | \theta_i])
\]

\[
 = \mathbb{E}[\theta_i] + \text{Var}(\theta_i) = \frac{k}{\lambda_i} + \frac{k}{\lambda_i^2}
\]

\[
 \text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{k}
\]
Two-stage models – Count data (Hinde and Demétrio, 1998a,b)

**Poisson-normal and related models**

Individual level random effect in the linear predictor

\[ \eta_i = \beta' x_i + \sigma Z_i \]

- assume \( Z_i \sim N(0, 1) \), so

\[ Y_i | Z_i \sim \text{Pois}(\lambda_i) \quad \text{with} \quad \log \lambda_i = x_i^T \beta + \sigma Z_i \]

where \( Z_i \sim N(0, 1) \), which gives

\[
\begin{align*}
E[Y_i] &= E_Z \left( E[Y_i | Z_i] \right) = E_Z [e^{x_i^T \beta + \sigma Z_i}] \\
&= e^{x_i^T \beta + \frac{1}{2} \sigma^2} := \mu_i \\
\text{Var}(Y_i) &= E_Z \left[ \text{Var}(Y_i | Z_i) \right] + \text{Var}_Z \left( E[Y_i | Z_i] \right) \\
&= e^{x_i^T \beta + \frac{1}{2} \sigma^2} + \text{Var}_Z(e^{x_i^T \beta + \sigma Z_i}) \\
&= e^{x_i^T \beta + \frac{1}{2} \sigma^2} + e^{2x_i^T \beta + \sigma^2} (e^{\sigma^2} - 1).
\end{align*}
\]

i.e. a variance function of the form

\[ \text{Var}(Y_i) = \mu_i + k' \mu_i^2 \]
Germination of Orobanche seed

Example

<table>
<thead>
<tr>
<th>O. aegyptiaca 75</th>
<th>O. aegyptiaca 73</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bean</td>
<td>Cucumber</td>
</tr>
<tr>
<td>10/39</td>
<td>5/6</td>
</tr>
<tr>
<td>23/62</td>
<td>53/74</td>
</tr>
<tr>
<td>23/81</td>
<td>55/72</td>
</tr>
<tr>
<td>26/51</td>
<td>32/51</td>
</tr>
<tr>
<td>17/39</td>
<td>46/79</td>
</tr>
<tr>
<td></td>
<td>0/4</td>
</tr>
<tr>
<td></td>
<td>10/13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bean</th>
<th>Cucumber</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/16</td>
<td>3/12</td>
</tr>
<tr>
<td>10/30</td>
<td>22/41</td>
</tr>
<tr>
<td>8/28</td>
<td>15/30</td>
</tr>
<tr>
<td>23/45</td>
<td>32/51</td>
</tr>
<tr>
<td>0/4</td>
<td>3/7</td>
</tr>
</tbody>
</table>

- Response variable: $Y_i$ – number of germinated seeds out of $m_i$ seeds.
- Distribution: Binomial.
- Systematic component: factorial $2 \times 2$ (2 species, 2 extracts), completely randomized experiment (Crowder, 1978).
- Aim: to see how germination is affected by species and extracts.
- Problem: overdispersion.
Models for Orobanche Data (Hinde and Demétrio, 1998a,b)

Binomial:

\[ \text{Var}(Y_i) = m_i \pi_i (1 - \pi_i) \]

- residual deviance for the interaction model is 33.28 on 17 df. – overdispersion

Quasi-likelihood:

\[ \text{Var}(Y_i) = m_i \pi_i (1 - \pi_i) \]

- constant overdispersion \( \tilde{\phi} = 1.862 \)
- only marginal evidence of interaction
- extract only important factor

Williams:

\[ \text{Var}(Y_i) = m_i \pi_i (1 - \pi_i) [1 + \phi (m_i - 1)] \]

- moment estimate \( \tilde{\phi} = 0.0249 \)
- only marginal evidence of interaction
- extract only important factor
- note similarity to QL, even though \( m_i \) not equal.
**Table:** Deviances with overdispersion estimated from maximal model.

<table>
<thead>
<tr>
<th></th>
<th>Bin ML</th>
<th>Const QL</th>
<th>Beta-Binomial ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>E — S</td>
<td>56.49</td>
<td>30.34</td>
<td>32.69</td>
</tr>
<tr>
<td>S — E</td>
<td>3.06</td>
<td>1.64</td>
<td>2.88</td>
</tr>
<tr>
<td>S.Et</td>
<td>6.41</td>
<td>3.44</td>
<td>4.45</td>
</tr>
<tr>
<td>( \hat{\phi} )</td>
<td>1.862</td>
<td>0.012</td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Deviances with overdispersion re-estimated for each model.

<table>
<thead>
<tr>
<th>Source</th>
<th>Beta-Binomial ML</th>
<th>Logistic-Normal ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>S</td>
<td>15.44</td>
</tr>
<tr>
<td>S</td>
<td>E</td>
<td>2.73</td>
</tr>
<tr>
<td>S.E</td>
<td>4.13</td>
<td>4.17</td>
</tr>
</tbody>
</table>
Half-normal plots with simulation envelopes (Hinde and Demétrio, 1998a,b)
```R
Species <- factor(c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2))
Extract <- factor(c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2))
y <- c(10, 23, 23, 26, 17, 5, 53, 55, 32, 46, 10, 8, 10, 8, 23, 0, 3, 22, 15, 32, 3)
m <- c(39, 62, 81, 51, 39, 6, 74, 72, 51, 79, 13, 16, 30, 28, 45, 4, 12, 41, 30, 51, 7)
orobanch <- data.frame(Species, Extract, m, y)
rm(Species, Extract, m, y)
attach(orobanch)
resp<-cbind(y,m-y)

# Binomial fit
oro.Bin<-glm(resp~Species*Extract, family=binomial)
anova(oro.Bin, test="Chisq")
summary(oro.Bin, cor=FALSE)

oro.Bin<-glm(resp~Extract*Species, family=binomial)
anova(oro.Bin, test="Chisq")
summary(oro.Bin, cor=FALSE)

## Pearson estimate of phi
(X2<-sum(residuals(oro.Bin, 'pearson')^2))
(phi<-X2/df.residual(oro.Bin))
```
R program (cont)

# Quasilikelihhod fit
oro.QL<-glm(resp~Species*Extract, family=quasibinomial)
summary(oro.QL, cor=FALSE)
summary(oro.QL)$dispersion
anova(orobanchQL.fit, test="F")

### Beta-binomial model
library(aod)
# re-estimating phi
oro.BBin4 <- betabin(cbind(y,m-y) ~Extract,~1, data=orobanch)
oro.BBin3 <- betabin(cbind(y,m-y) ~Species,~1, data=orobanch)
oro.BBin2 <- betabin(cbind(y,m-y) ~Extract+Species,~1, data=orobanch)
oro.BBin1 <- betabin(cbind(y,m-y) ~Extract+Species+Extract:Species,~1, data=orobanch)
anova(oro.BBin4,oro.BBin2, oro.BBin1)
anova(oro.BBin3,oro.BBin2, oro.BBin1)
summary(oro.BBin2)

# fixing phi estimated by the maximum model
f1 <- betabin(cbind(y, m - y) ~ Extract+Species+Extract:Species, ~ 1, data=orobanch, 
fixpar = list(5, 0.01238))
f2 <- betabin(cbind(y, m - y) ~ Extract+Species, ~ 1, data=orobanch, 
fixpar = list(4, 0.01238))
f3 <- betabin(cbind(y, m - y) ~ Species, ~ 1, data=orobanch, 
fixpar = list(3, 0.01238))
f4 <- betabin(cbind(y, m - y) ~ Extract, ~ 1, data=orobanch, 
fixpar = list(3, 0.01238))

anova(f4,f2,f1)
anova(f3,f2,f1)
# logistic normal
library(lme4)
ind <- (1:length(y))
# re-estimating phi
oro.LN1<-glmer(resp~Extract*Species + (1|ind), family=binomial(link="logit"))
oro.LN1
summary(oro.LN1)@coefs
summary(oro.LN1)@REmat
oro.LN2<-glmer(resp~Extract+Species + (1|ind), family=binomial(link="logit"))
oro.LN2
oro.LN4<-glmer(resp~Extract + (1|ind), family=binomial(link="logit"))
oro.LN4
oro.LN3<-glmer(resp~Species + (1|ind), family=binomial(link="logit"))
oro.LN3
anova(oro.LN4,oro.LN2,oro.LN1)
anova(oro.LN3,oro.LN2,oro.LN1)
### Worldwide Airline Fatalities, 1976-85

<table>
<thead>
<tr>
<th>Year</th>
<th>Fatal accidents</th>
<th>Passenger deaths</th>
<th>Passenger miles (100 million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>24</td>
<td>734</td>
<td>3863</td>
</tr>
<tr>
<td>1977</td>
<td>25</td>
<td>516</td>
<td>4300</td>
</tr>
<tr>
<td>1978</td>
<td>31</td>
<td>754</td>
<td>5027</td>
</tr>
<tr>
<td>1979</td>
<td>31</td>
<td>877</td>
<td>5481</td>
</tr>
<tr>
<td>1980</td>
<td>22</td>
<td>814</td>
<td>5814</td>
</tr>
<tr>
<td>1981</td>
<td>21</td>
<td>362</td>
<td>6033</td>
</tr>
<tr>
<td>1982</td>
<td>26</td>
<td>764</td>
<td>5877</td>
</tr>
<tr>
<td>1983</td>
<td>20</td>
<td>809</td>
<td>6223</td>
</tr>
<tr>
<td>1984</td>
<td>16</td>
<td>223</td>
<td>7433</td>
</tr>
<tr>
<td>1985</td>
<td>22</td>
<td>1066</td>
<td>7107</td>
</tr>
</tbody>
</table>
Simple Models

- Passenger miles \( (m_i) \) as exposure variable
- Poisson log-linear model
- Linear time trend

\[ Y_i \sim \text{Pois}(m_i \lambda_i) \quad (2) \]
\[ \log \lambda_i = \beta_0 + \beta_1 \text{year}_i \]

**Fatal accidents:**
Deviance(time trend) = 20.68
Residual Deviance = 5.46 on 8 d.f.

**Passenger deaths:**
Deviance(time trend) = 202.1
Residual Deviance = 1051.5 on 8 d.f.

⇒ compounding with aircraft size
### Confidence Intervals – airline data

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>Observed</td>
</tr>
<tr>
<td>1978</td>
<td>Poisson</td>
</tr>
<tr>
<td>1980</td>
<td>Negative binomial</td>
</tr>
<tr>
<td>1982</td>
<td>Quasi-Poisson</td>
</tr>
<tr>
<td>1984</td>
<td></td>
</tr>
</tbody>
</table>

- *Observed*
- Poisson
- Negative binomial
- Quasi-Poisson

---

**Clarice G.B. Demétrio**

*Generalized Linear Models and Extensions*
```r
R program

```
R program (cont)

# Number of deaths
# Poisson fit
plot(year, death, main="Number of Fatal Accidents vs Year")

death1.fit <- glm(death ~ year + offset(log(miles)), family = poisson)
summary(death1.fit)
anova(death1.fit, test = "Chisq")

inflim <- with(predict(death1.fit, se = TRUE, type = "response"),
               fit - pnorm(0.975) * se.fit)
suplim <- with(predict(death1.fit, se = TRUE, type = "response"),
               fit + pnorm(0.975) * se.fit)

plot(c(1976, 1985), c(350, 1200), type="n", xlab="Year", ylab="Number of deaths")
points(year, death)
x <- seq(1976, 1985, 1)
fv1 <- predict(death1.fit, data.frame(year = x), type = "response")
lines(x, fv1, lty = 1)
lines(x, inflim, lty = 2)
lines(x, suplim, lty = 2)

# Negative binomial fit
library(MASS)
death2.fit <- glm.nb(death ~ year + offset(log(miles)), link = log)
summary(death2.fit)
anova(death2.fit, test = "F")
inflim <- with(predict(death2.fit, se = TRUE, type = "response"),
               fit - qt(0.975, 8) * se.fit)
suplim <- with(predict(death2.fit, se = TRUE, type = "response"),
               fit + qt(0.975, 8) * se.fit)
fv2 <- predict(death2.fit, data.frame(year = x), type = "response")
# lines(x, fv2, lty = 1)
# lines(x, fv1, lty = 1, col = "blue")
lines(x, inflim, lty = 3, col = "blue")
lines(x, suplim, lty = 3, col = "blue")

# Quasilikelihhod fit
dead3.fit <- glm(death ~ year + offset(log(miles)), family = quasipoisson)
summary(dead3.fit)
anova(dead3.fit, test = "F")
inflim <- with(predict(dead3.fit, se = TRUE, type = "response"),
fit - qt(0.975, 8) * se.fit)
suplim <- with(predict(dead3.fit, se = TRUE, type = "response"),
fit + qt(0.975, 8) * se.fit)

fv3 <- predict(dead3.fit, data.frame(year = x), type = "response")
# lines(x, fv3, lty = 1, col = "red")
lines(x, inflim, lty = 4, col = "red")
lines(x, suplim, lty = 4, col = "red")
legend(1976, 1200.0, c("Observed", "Poisson", "Negative binomial", "Quasi-Poisson"), lty = c(-1, 2, 3, 4), pch = c("*", " ", " ", " "), col = c("black", "black", "blue", "red"))

# Poisson log-normal
library(lme4)
ind <- (1:length(death))
death.LN1 <- glmer(death ~ year + offset(log(miles)) + (1 | ind),
family = poisson(link = "log"))
death.LN1
summary(death.LN1)@coefs
summary(death.LN1)@REmat


