Introduction to General and Generalized Linear Models
Hierarchical models

Henrik Madsen
Poul Thyregod

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Kgs. Lyngby

January 2011
This lecture

- Introduction, approaches to modelling of overdispersion
- Hierarchical Poisson Gamma model
- Conjugate prior distributions
- The generalized one-way random effects model
- The Binomial Beta model
- Normal distributions with random variance
- Hierarchical generalized linear models
Introduction

A characteristic property of the generalized linear models is that the variance, \( \text{Var}[Y] \) in the distribution of the response is a known function, \( V(\mu) \), that only depends on the mean value \( \mu \)

\[
\text{Var}[Y_i] = \lambda_i V(\mu) = \frac{\sigma^2}{w_i} V(\mu)
\]

where \( w_i \) denotes a known weight, associated with the \( i \)'th observation, and where \( \sigma^2 \) denotes a dispersion parameter common to all observations, irrespective of their mean.

The dispersion parameter \( \sigma^2 \) does serve to express overdispersion in situations where the residual deviance is larger than what can be attributed to the variance function \( V(\mu) \) and known weights \( w_i \).

We shall describe an alternative method for modeling overdispersion, viz. by hierarchical models analogous to the mixed effects models for the normally distributed observations.
A starting point in a hierarchical modeling is an assumption that the distribution of the random “noise” may be modeled by an exponential dispersion family (Binomial, Poisson, etc.), and then it is a matter of choosing a suitable (prior) distribution of the mean-value parameter $\mu$.

It seems natural to choose a distribution with a support that coincides with the mean value space $\mathcal{M}$ rather than using a normal distribution (with a support constituting all of the real axis $\mathbb{R}$).

In some applications an approach with a normal distribution of the canonical parameter is used. Such an approach is sometimes called \textit{generalized linear mixed models} (GLMMs).
Introduction

- Although such an approach is consistent with a formal requirement of equivalence between mean values space and support for the distribution of $\mu$ in the binomial and the Poisson distribution case, the resulting marginal distribution of the observation is seldom tractable, and the likelihood of such a model will involve an integral which cannot in general be computed explicitly.

- Also, the canonical parameter does not have a simple physical interpretation and, therefore, an additive “true value” + error, with a normally distributed “error” on the canonical parameter to describe variation between subgroups, is not very transparent.

- Instead, we shall describe an approach based on the so-called standard conjugated distribution for the mean parameter of the within group distribution for exponential families.

- These distributions combine with the exponential families in a simple way, and lead to marginal distributions that may be expressed in a closed form suited for likelihood calculations.
Hierarchical Poisson Gamma model - example

The table shows the distribution of the number of daily episodes of thunderstorms at Cape Kennedy, Florida, during the months of June, July and August for the 10-year period 1957–1966, total 920 days.

<table>
<thead>
<tr>
<th>Number of episodes, $z_i$</th>
<th>Number of days, $#i$</th>
<th>Poisson expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>803</td>
<td>791.85</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>118.78</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>8.91</td>
</tr>
<tr>
<td>3+</td>
<td>3</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table: The distribution of days with 0, 1, 2 or more episodes of thunderstorm at Cape Kennedy.

All observational periods are $n_i = 1$ day.
Hierarchical Poisson Gamma model - example

- The data represents *counts* of events (episodes of thunderstorms) distributed in time.

- A completely random distribution of the events would result in a Poisson distribution of the number of daily events.

- The variance function for the Poisson distribution is $V(\mu) = \mu$; therefore, a Poisson distribution of the daily number of events would result in the variance in the distribution of the daily number of events being equal to the mean, $\hat{\mu} = \bar{y}_+ = 0.15$ thunderstorms per day.

- The empirical variance is $s^2 = 0.1769$, which is somewhat larger than the average. We further note that the observed distribution has heavier tails than the Poisson distribution. Thus, one might be suspicious of overdispersion.
Hierarchical Poisson Gamma model - example

- Poisson exp.
- Neg. Bin. exp.
- Observed

<table>
<thead>
<tr>
<th>Number of episodes</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

Number of days

- 10^3
- 10^2
- 10^1
- 10^0
- 10^{-1}
Formulation of hierarchical model

Theorem (Compound Poisson Gamma model)

Consider a hierarchical model for $Y$ specified by

$$Y | \mu \sim \text{Pois}(\mu),$$

$$\mu \sim \text{G}(\alpha, \beta),$$

i.e. a two stage model.

In the first stage a random mean value $\mu$ is selected according to a Gamma distribution. The $Y$ is generated according to a Poisson distribution with that value as mean value. Then the the marginal distribution of $Y$ is a negative binomial distribution, $Y \sim \text{NB}(\alpha, 1/(1 + \beta))$
Theorem (Compound Poisson Gamma model, continued)

The probability function for $Y$ is

$$P[Y = y] = g_Y(y; \alpha, \beta)$$

$$= \frac{\Gamma(y + \alpha)}{y!\Gamma(\alpha)} \frac{\beta^y}{(\beta + 1)^{y+\alpha}}$$

$$= \binom{y + \alpha - 1}{y} \frac{1}{(\beta + 1)^\alpha} \left(\frac{\beta}{\beta + 1}\right)^y$$

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

where we have used the convention

$$\binom{z}{y} = \frac{\Gamma(z + 1)}{\Gamma(z + 1 - y) y!}$$

for $z$ real and $y$ integer values.
For integer values of $\alpha$ the negative binomial distribution is known as the distribution of the number of “failures” until the $\alpha$’th success in a sequence of independent Bernoulli trials where the probability of success in each trial is $p = 1/(1 + \beta)$.

For $\alpha = 1$ the distribution is known as the *geometric distribution*. 
Formulation of hierarchical model

Decomposition of the marginal variance, signal/noise ratio

If \( \mu \sim G(\alpha, \beta) \) then \( \mathbb{E}[\mu] = \alpha \beta \) and \( \text{Var}[\mu] = \alpha \beta^2 \).

Then we have the decomposition

\[
\text{Var}[Y] = \mathbb{E}[\text{Var}[Y|\mu]] + \text{Var}[\mathbb{E}[Y|\mu]] = \mathbb{E}[\mu] + \text{Var}[\mu] = \alpha \beta + \alpha \beta^2
\]

of the total variation in variation within groups and between groups, respectively.

We may now introduce a signal/noise ratio as

\[
\gamma = \frac{\text{Var}[\mathbb{E}[Y|\mu]]}{\mathbb{E}[\text{Var}[Y|\mu]]} = \frac{\alpha \beta^2}{\alpha \beta} = \beta
\]
Inference on individual group means

Theorem (Conditional distribution of $\mu$)

Consider the hierarchical Poisson-Gamma model and assume that a value $Y = y$ has been observed.

Then the conditional distribution of $\mu$ for given $Y = y$ is a Gamma distribution,

$$\mu \mid Y = y \sim G(\alpha + y, \beta/(\beta + 1))$$

with mean

$$E[\mu \mid Y = y] = \frac{\alpha + y}{(1/\beta + 1)}$$
Inference on individual group means

- In a Bayesian framework, we would identify the distribution of $\mu$ as the prior distribution, the distribution of $Y | \mu$ as the sampling distribution, and the conditional distribution of $\mu$ for given $Y = y$ as the posterior distribution.

- When the posterior distribution belongs to the same distribution family as the prior one, we say that the prior distribution is conjugate with respect to that sampling distribution.

- Using conjugate priors simplifies the modeling. To derive the posterior distribution, it is not necessary to perform the integration, as the posterior distribution is simply obtained by updating the parameters of the prior one.
Reparameterization of the Gamma distribution

Instead of the usual parameterization of the gamma distribution of $\mu$ by its shape parameter $\alpha$ and scale parameter $\beta$, we may choose a parameterization by the mean value, $m = \alpha \beta$, and the signal/noise ratio $\gamma = \beta$

\[ \gamma = \beta \]
\[ m = \alpha \beta \]

The parameterization by $m$ and $\gamma$ implies that the degenerate one-point distribution of $\mu$ in a value $m_0$ may be obtained as limiting distribution for Gamma distributions with mean $m_0$ and signal/noise ratios $\gamma \to 0$. Moreover, under that limiting process the corresponding marginal distribution of $Y$ (negative binomial) will converge towards a Poisson distribution with mean $m_0$. 
Conjugate prior distributions

Definition (Standard conjugate distribution for an exponential dispersion family)

Consider an exponential dispersion family $ED(\mu, V(\mu)/\lambda)$ for $\theta \in \Omega$. Let $\mathcal{M} = \tau(\Omega)$ denote the mean value space for this family. Let $m \in \mathcal{M}$ and consider

$$g_\theta(\theta; m, \gamma) = \frac{1}{C(m, \gamma)} \exp\left(\frac{\theta m - \kappa(\theta)}{\gamma}\right)$$

with

$$C(m, \gamma) = \int_{\Omega} \exp\left(\frac{\theta m - \kappa(\theta)}{\gamma}\right) \, d\theta$$

for all (positive) values of $\gamma$ for which the integral converges. This distribution is called the standard conjugate distribution for $\theta$. The concept has its roots in the context of Bayesian parametric inference to describe a family of distributions whose densities have the structure of the likelihood kernel.
Conjugate prior distributions

- When the variance function, $V(\mu)$ is at most quadratic, the parameters $m$ and $\gamma$ have a simple interpretation in terms of the mean value parameter, $\mu = \tau(\theta)$, viz.

$$m = E[\mu]$$

$$\gamma = \frac{Var[\mu]}{E[Var(\mu)]}$$

with $\mu = E[Y|\theta]$, and with $Var(\mu)$ denoting the variance function

- The use of the symbol $\gamma$ is in agreement with our introduction of $\gamma$ as signal to noise ratio for normally distributed observations and for the Poisson-Gamma hierarchical model.
When the variance function for the exponential dispersion family is at most quadratic, the standard conjugate distribution for $\mu$ coincides with the standard conjugate distribution for $\theta$.

However, for the Inverse Gaussian distribution, the standard conjugate distribution for $\mu$ is improper.

The parameterization of the natural conjugate distribution for $\mu$ by the parameters $m$ and $\gamma$ has the advantage that location and spread are described by separate parameters. Thus, letting $\gamma \to 0$, the distribution of $\mu$ will converge towards a degenerate distribution with all its mass in $m$. 
The generalized one-way random effects model

Definition (The generalized one-way random effects model)

Now consider a hierarchical model with \( k \) randomly selected groups, \( i = 1, 2, \ldots, k \), and measurements \( Y_{ij}, j = 1, \ldots, n_i \) from subgroup \( i \).

1. Conditional on the group mean, \( \mu_i \), the measurements, \( Y_{ij} \) are independent and distributed according to an exponential dispersion model with mean \( \mu_i \), variance function, \( V(\mu) \), and precision parameter \( \lambda \) (the sampling distribution).

2. The group means, \( \mu_i \) are independent random variables distributed according to the natural conjugate distribution of the sampling distribution.

The model may be thought of as a two-stage process: In the first stage, a group is selected, the group mean value \( \mu_i \) is selected from the specified distribution. Then, with that value of \( \mu_i \) a set of \( n_{ij} \) independent observations \( Y_{ij}, j = 1, \ldots, n_i \) are generated according to the exponential dispersion model with that value, \( \mu_i \).
Theorem (Marginal distributions in the generalized one-way random effects model)

Consider a generalized one-way random effects model. The moments in the marginal distribution of $Y_{ij}$ are

$$
E[Y_{ij}] = E[E[Y_i | \mu]] = E[\mu] = m
$$

$$
\text{Cov}[Y_{ij}, Y_{hl}] = \begin{cases} 
E[V(\mu)]\{\gamma + 1/\lambda\} & \text{for } (i, j) = (h, l) \\
\text{Var}[\mu] & \text{for } i = h, j \neq l \\
0 & \text{for } i \neq h
\end{cases}
$$

Thus, the parameter, $\gamma$, in the distribution of the group means reflects the usual decomposition of total variation into variation within the groups and variation between the groups. As the variation, $V(\mu)$, within a specific group depends on the group mean, $\mu$, the within-group variation is understood as an “average”, $E[V(\mu)]$. 
The generalized one-way random effects model

Marginal and simultaneous distributions

The marginal distribution of the average in a group, $\overline{Y}_i$, has mean and variance

$$E[\overline{Y}_i] = E[E[\overline{Y}_i|\mu]] = E[\mu] = m$$

$$\text{Var}[\overline{Y}_i] = E[V(\mu)] \left( \gamma + \frac{1}{\lambda n_i} \right)$$

Thus, whenever $\text{Var}[\mu] > 0$, there is overdispersion in the sense that the variance in the marginal distribution exceeds the average variance in the within group distributions.
The generalized one-way random effects model

Marginal and simultaneous distributions

Measurements, $Y_{ij}$ and $Y_{ik}$ in the same group are correlated with *intraclass correlation*

$$
\rho = \frac{\text{Cov}[Y_{ij}, Y_{ik}]}{\text{Var}[Y_{ij}]} = \frac{\text{Var}[\mu]}{\text{E}[V(\mu)]/\lambda + \text{Var}[\mu]} = \frac{\gamma}{1/\lambda + \gamma}
$$

and hence,

$$
\gamma = \frac{\rho}{1 - \rho} \frac{1}{\lambda}
$$
The likelihood function corresponding to a set of observations, $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$ is constructed from the marginal probabilities for the group means, or group totals as a function of the parameters $m$ and $\gamma$. 
Consider the between-group sum of squares,

$$SSB = \sum_{i=1}^{N} n_i \left( \bar{Y}_i - \bar{Y} \right)^2$$

with $\bar{Y}$ denoting the usual (weighted) average of the subgroup averages. It may be shown that

$$E[SSB] = (k - 1) E[V(\mu)] \left( \frac{1}{\lambda} + n_0 \gamma \right),$$

with $n_0$ denoting the *weighted average sample size*. Thus, a natural candidate for a method of moments estimate of the between-group dispersion, $\gamma$, is

$$S^2_2 = \frac{1}{N - 1} \sum_{i=1}^{N} n_i (Y_i - \bar{Y})^2$$

with expected value

$$E[S^2_2] = E[V(\mu)] \left( \frac{1}{\lambda} + n_0 \gamma \right).$$
Inference of individual group means, $\mu_i$

Theorem (Conditional distribution of group mean $\mu_i$ after observation of $\bar{y}_i$)

Consider a generalized one-way random effects model and assume that the sample from the $i$'th group resulted in the values $(y_{i1}, y_{i2}, \ldots, y_{in_i})$. Then the conditional distribution of the canonical parameter $\theta$ (the posterior distribution) given this sample result depends only on $n$ and $\bar{y}_i$.

The probability density function in the conditional distribution of $\theta$ is

$$w(\theta | \bar{Y}_i = \bar{y}_i) = \frac{1}{C(m_1, \gamma_1)} \exp \left( \frac{\theta m_1 - \kappa(\theta)}{\gamma_1} \right)$$

with

$$m_1 = m_{post} = \frac{m/\gamma + n\lambda \bar{y}}{1/\gamma + n\lambda}$$

$$\gamma_1 = \gamma_{post} = \frac{1}{1/\gamma + n\lambda}.$$
Theorem (Conditional distribution of group mean $\mu_i$ after observation of $\bar{y}_i$, continued)

Thus, the posterior distribution for $\mu$ is also of the same form as the prior distribution. We have just updated the parameters $m$ and $\gamma$. The updating may be expressed

$$
\frac{1}{\gamma_{post}} = \frac{1}{\gamma_{prior}} + n\lambda
$$

$$
\frac{m_{post}}{\gamma_{post}} = \frac{m_{prior}}{\gamma_{prior}} + n\lambda \bar{y}
$$

Recall that $1/\gamma = \frac{E[V(\mu)]}{\text{Var}[\mu]}$ is a measure of the relative precision in the distribution of $\mu$ relative to the precision in the sampling distribution of the $Y$’s, i.e.

$$
\gamma_{post} = \frac{\text{Var}[\mu | \bar{y}]}{E[V(\mu) | \bar{y}]}.
$$
Hierarchical Binomial-Beta distribution model

The natural conjugate distribution to the binomial is a Beta-distribution.

Theorem

Consider the generalized one-way random effects model for \( Z_1, Z_2, \ldots, Z_k \) given by

\[
Z_i | p_i \sim B(n, p_i) \\
p_i \sim \text{Beta}(\alpha, \beta)
\]

i.e. the conditional distribution of \( Z_i \) given \( p_i \) is a Binomial distribution, and the distribution of the mean value \( p_i \) is a Beta distribution. Then the marginal distribution of \( Z_i \) is a Polya distribution with probability function

\[
P[Z = z] = g_Z(z) = \binom{n}{z} \frac{\Gamma(\alpha + x)}{\Gamma(\alpha)} \frac{\Gamma(\beta + n - z)}{\Gamma(\beta)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)}
\]

for \( z = 0, 1, 2, \ldots, n \).
Hierarchical Binomial-Beta distribution model

- The Polya distribution is named after the Hungarian mathematician G. Polya, who first described this distribution – although in another context.

- Instead of $m$, we shall use the symbol $\pi$ to denote $E[p]$. Thus, we shall use a parameterization given by

\[
\pi = \frac{\alpha}{\alpha + \beta}, \quad \gamma = \frac{1}{\alpha + \beta}
\]

as

\[
E[p] = \pi = \frac{\alpha}{\alpha + \beta}
\]

\[
\text{Var}[p] = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{\pi(1 - \pi)}{\alpha + \beta + 1}
\]

\[
E[V(p)] = E[p(1 - p)] = \frac{\pi(1 - \pi)}{1 + \gamma}
\]
Hierarchical Binomial-Beta distribution model

- The moments in the marginal distribution of $Z$ are

\[
E[Z] = n\pi \\
\text{Var}[Z] = n\frac{\pi(1 - \pi)}{1 + \gamma} [1 + n\gamma]
\]

and the moments for the fraction $Y = Z/n$ are

\[
E[Y] = \pi \\
\text{Var}[Y] = \frac{\pi(1 - \pi)}{1 + \gamma} \left(\gamma + \frac{1}{n}\right).
\]
The Binomial Beta model

Estimation of individual group means

Theorem (Conditional distribution of $p$ for given $z$)

Consider the hierarchical binomial beta model in and assume that a value $Z = z$ has been observed. Then the conditional distribution of $p$ is a Beta distribution

$$p|Z = z \sim \text{Beta}(\alpha + z, \beta + n - z)$$

with mean

$$E[p|Z = z] = \frac{\alpha + z}{\alpha + \beta + n}.$$
The Binomial Beta model

Estimation of individual group means

The conditional mean as a weighted average

We find the updating formulae

\[
\frac{1}{\gamma_{post}} = \frac{1}{\gamma_{prior}} + n
\]

\[
\pi_{post} = \frac{\alpha + z}{\alpha + \beta + n} = w\pi_{prior} + \left(1 - w\right)\frac{z}{n}
\]

with

\[
w = \frac{\alpha + \beta}{\alpha + \beta + n} = \frac{1/\gamma_{prior}}{1/\gamma_{prior} + n}.
\]
Normal distributions with random variance

As a non-trivial example of a hierarchical distribution we consider the hierarchical normal distribution model with random variance:

**Theorem**

Consider a generalized one-way random effects model specified by

\[
Y_i | \sigma_i^2 \sim N(\mu, \sigma_i^2)
\]

\[
1/\sigma_i^2 \sim G(\alpha, 1/\beta)
\]

where \(\sigma_i^2\) are mutually independent for \(i = 1, \ldots, k\).

The marginal distribution of \(Y_i\) under this model is

\[
\frac{Y_i - \mu}{\sqrt{\beta/\alpha}} \sim t(2\alpha)
\]

where \(t(2\alpha)\) is a \(t\)-distribution with \(2\alpha\) degrees of freedom, i.e. a distribution with heavier tails than the normal distribution.
Definition (Hierarchical generalized linear model)

Consider a set of observations $Y = (Y_1, Y_2, \ldots, Y_k)^T$ such that for a given value of a parameter $\theta$ the distribution of $Y_i$ is given by an exponential dispersion model with canonical parameter space $\Omega$ (for $\theta$), mean value $\mu = \kappa'(\theta)$, mean value space $\mathcal{M}$ (for $\mu$) and canonical link $\theta = g(\mu)$. The variables in a hierarchical generalized linear model are

1. the observed responses $y_1, y_2, \ldots, y_k$ ($\in \mathcal{M}$)
2. the unobserved state variables $u_1, u_2, \ldots, u_q$ ($\in \mathcal{M}$)
3. and the corresponding unobserved canonical variables $v_i = g(u_i)$ ($\in \Omega$)
Hierarchical generalized linear models

Definition (Hierarchical generalized linear model, continued)

The linear predictor is of the form

\[ \theta = g(\mu|v) = X\beta + Zv \]

The distribution of \( V \in \Omega \) is a conjugated distribution to the canonical parameter \( \theta \). The derived distribution of \( U \in \mathcal{M} \) is the corresponding conjugated distribution to the mean value parameter \( \mu \) such that \( E[U] = \psi \).
Hierarchical generalized linear models

Estimation in a hierarchical generalized linear model

The estimation can be performed by an extended generalized linear model for $y$ and $\psi$.

The mean value parameters $\beta$ and $\nu$ may be estimated by an iterative procedure solving

$$
\begin{pmatrix}
  X'\Sigma_0^{-1}X & X'\Sigma_0^{-1}Z \\
  Z'\Sigma_0^{-1}X & Z'\Sigma_0^{-1}Z + \Sigma_1^{-1}
\end{pmatrix}
\begin{pmatrix}
  \hat{\beta} \\
  \hat{\nu}
\end{pmatrix}
= 
\begin{pmatrix}
  X'\Sigma_0^{-1}z_0 \\
  Z'\Sigma_0^{-1}z_0 + \Sigma_1^{-1}z_1
\end{pmatrix}
$$

where $z_0 = \eta_0 + (y - \mu_0)(\partial\eta_0/\partial\mu_0)$ and $z_1 = \nu_1 + (\psi_1 - u)(dv_1/du)$ are the adjusted dependent variables for the distribution of $y$ given $\nu$ and of $\nu$, respectively, in analogy with the estimation in generalized linear models.

In the one-way random effect models we have considered in the examples the hierarchical likelihood estimates of group means are the empirical Bayes estimates derived in the examples.