Mixed models in R using the lme4 package
Part 7: Generalized linear mixed models

Douglas Bates

University of Wisconsin - Madison
and R Development Core Team
<Douglas.Bates@R-project.org>

Merck, Sharp & Dohme; Rahway, NJ
Sept 24, 2010

Outline

1. Generalized Linear Mixed Models

Outline

1. Generalized Linear Mixed Models

2. Specific distributions and links

Outline

Outline

Outline

Outline

Generalized Linear Mixed Models

- When using linear mixed models (LMMs) we assume that the response being modeled is on a continuous scale.

- Sometimes we can bend this assumption a bit if the response is an ordinal response with a moderate to large number of levels. For example, the Scottish secondary school test results in the mlmRev package are integer values on the scale of 1 to 10 but we analyze them on a continuous scale.

- However, an LMM is not suitable for modeling a binary response, an ordinal response with few levels or a response that represents a count. For these we use generalized linear mixed models (GLMMs).

- To describe GLMMs we return to the representation of the response as an n-dimensional, vector-valued, random variable, $\mathcal{Y}$, and the random effects as a q-dimensional, vector-valued, random variable, $\mathcal{B}$.

Parts of LMMs carried over to GLMMs

- Random variables
    - $\mathcal{Y}$ the response variable
    - $\mathcal{B}$ the (possibly correlated) random effects
    - $\mathcal{U}$ the orthogonal random effects, such that $\mathcal{B} = \Lambda_\theta \mathcal{U}$
- Parameters
    - β - fixed-effects coefficients
    - σ - the common scale parameter (not always used)
    - θ - parameters that determine $\mathrm{Var}(\mathcal{B}) = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T}$
- Some matrices
    - X the $n \times p$ model matrix for β
    - Z the $n \times q$ model matrix for b
    - P fill-reducing $q \times q$ permutation (from Z)
    - Λ_θ relative covariance factor, s.t. $\mathrm{Var}(\mathcal{B}) = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T}$

The conditional distribution, $\mathcal{Y}|\mathcal{U}$

- For GLMMs, the marginal distribution, $\mathcal{B} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right)$ is the same as in LMMs except that σ^2 is omitted. We define $\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$ such that $\mathcal{B} = \boldsymbol{\Lambda}_\theta \mathcal{U}$.
- For GLMMs we retain some of the properties of the conditional distribution for a LMM

  $$(\mathcal{Y}|\mathcal{U} = \boldsymbol{u}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}, \sigma^2 \boldsymbol{I}\right) \text{ where } \boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{u}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{u}$$

  Specifically

  - The conditional distribution, $\mathcal{Y}|\mathcal{U} = \boldsymbol{u}$, depends on $\boldsymbol{u}$ only through the conditional mean, $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{u})$.
  - Elements of $\mathcal{Y}$ are *conditionally independent*. That is, the distribution, $\mathcal{Y}|\mathcal{U} = \boldsymbol{u}$, is completely specified by the univariate, conditional distributions, $\mathcal{Y}_i|\mathcal{U}, i = 1, \ldots, n$.
  - These univariate, conditional distributions all have the same form. They differ only in their means.
- GLMMs differ from LMMs in the form of the univariate, conditional distributions and in how $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{u})$ depends on $\boldsymbol{u}$.

Some choices of univariate conditional distributions

- Typical choices of univariate conditional distributions are:
    - The *Bernoulli* distribution for binary (0/1) data, which has probability mass function

$$p(y|\mu) = \mu^y(1-\mu)^{1-y}, \quad 0 < \mu < 1, \quad y = 0, 1$$

    - Several independent binary responses can be represented as a *binomial* response, but only if all the Bernoulli distributions have the same mean.
    - The *Poisson* distribution for count $(0, 1, \dots)$ data, which has probability mass function

$$p(y|\mu) = e^{-\mu} \frac{\mu^y}{y!}, \quad 0 < \mu, \quad y = 0, 1, 2, \dots$$

- All of these distributions are completely specified by the conditional mean. This is different from the conditional normal (or Gaussian) distribution, which also requires the common scale parameter, σ.

The link function, g

- When the univariate conditional distributions have constraints on μ, such as $0 < \mu < 1$ (Bernoulli) or $0 < \mu$ (Poisson), we cannot define the conditional mean, $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}$, to be equal to the linear predictor, $\boldsymbol{X\beta} + \boldsymbol{X\Lambda}_\theta \boldsymbol{u}$, which is unbounded.

- We choose an invertible, univariate *link function*, g, such that $\eta = g(\mu)$ is unconstrained. The vector-valued link function, $\boldsymbol{g}$, is defined by applying g component-wise.

$$\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu}) \quad \text{where} \quad \eta_i = g(\mu_i), \quad i = 1, \dots, n$$

- We require that g be invertible so that $\mu = g^{-1}(\eta)$ is defined for $-\infty < \eta < \infty$ and is in the appropriate range ($0 < \mu < 1$ for the Bernoulli or $0 < \mu$ for the Poisson). The vector-valued inverse link, $\boldsymbol{g}^{-1}$, is defined component-wise.

"Canonical" link functions

- There are many choices of invertible scalar link functions, g, that we could use for a given set of constraints.

- For the Bernoulli and Poisson distributions, however, one link function arises naturally from the definition of the probability mass function. (The same is true for a few other, related but less frequently used, distributions, such as the gamma distribution.)

- To derive the canonical link, we consider the logarithm of the probability mass function (or, for continuous distributions, the probability density function).

- For distributions in this "exponential" family, the logarithm of the probability mass or density can be written as a sum of terms, some of which depend on the response, y, only and some of which depend on the mean, μ, only. However, only one term depends on **both** y and μ, and this term has the form $y \cdot g(\mu)$, where g is the canonical link.

The canonical link for the Bernoulli distribution

- The logarithm of the probability mass function is

$$\log(p(y|\mu)) = \log(1 - \mu) + y \log\left(\frac{\mu}{1 - \mu}\right), \ 0 < \mu < 1, \ y = 0, 1.$$
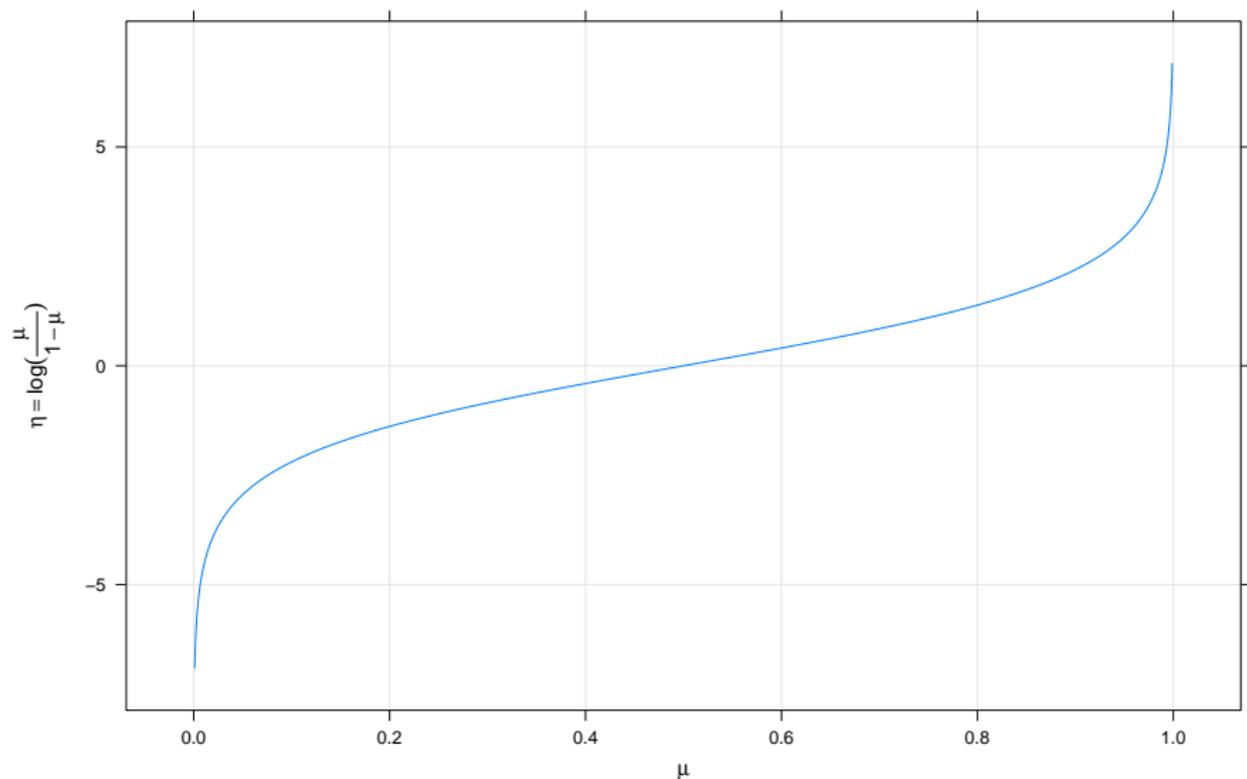
- Thus, the canonical link function is the *logit* link

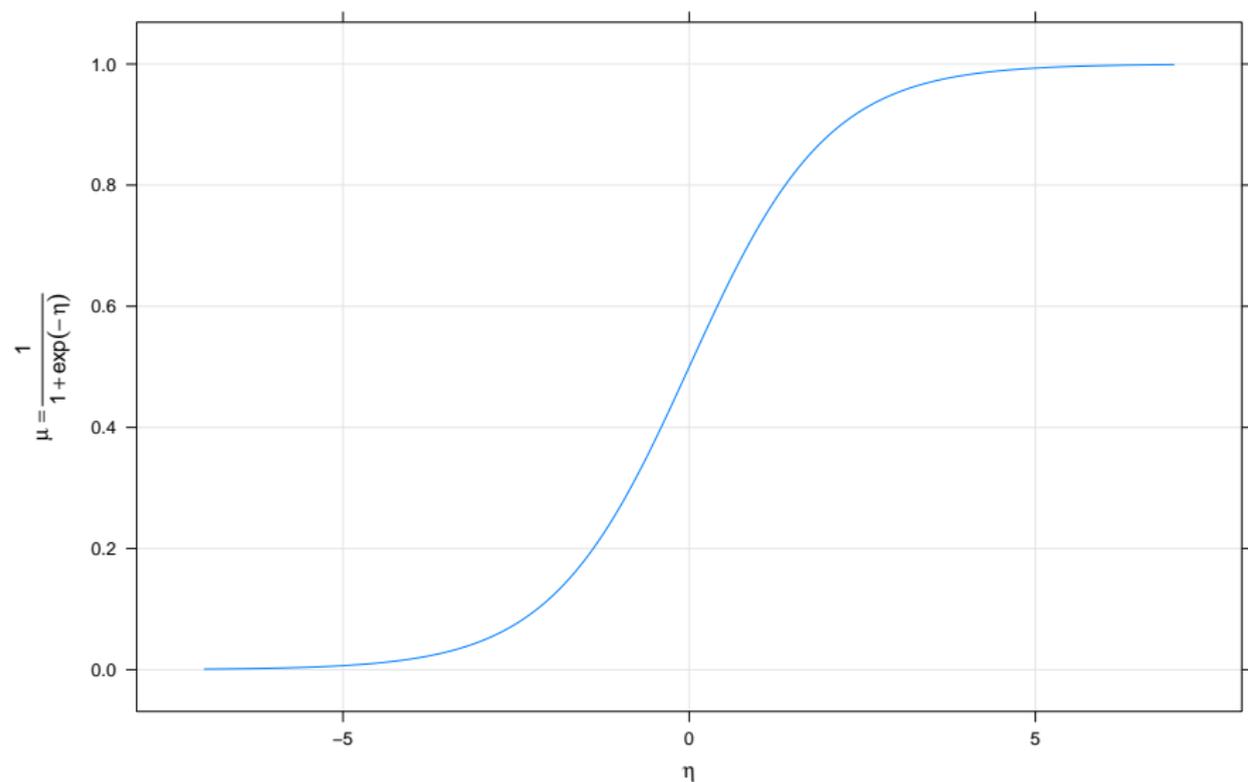$$\eta = g(\mu) = \log\left(\frac{\mu}{1 - \mu}\right).$$

- Because $\mu = P[\mathcal{Y} = 1]$, the quantity $\mu/(1 - \mu)$ is the odds ratio (in the range $(0, \infty)$) and g is the logarithm of the odds ratio, sometimes called "log odds".

- The inverse link is

$$\mu = g^{-1}(\eta) = \frac{e^\eta}{1 + e^\eta} = \frac{1}{1 + e^{-\eta}}$$

Plot of canonical link for the Bernoulli distribution

Plot of inverse canonical link for the Bernoulli distribution

The canonical link for the Poisson distribution

- The logarithm of the probability mass is

$$\log(p(y|\mu)) = \log(y!) - \mu + y\log(\mu)$$

- Thus, the canonical link function for the Poisson is the *log* link

$$\eta = g(\mu) = \log(\mu)$$

- The inverse link is

$$\mu = g^{-1}(\eta) = e^{\eta}$$

The canonical link related to the variance

- For the canonical link function, the derivative of its inverse is the variance of the response.

- For the Bernoulli, the canonical link is the logit and the inverse link is $\mu = g^{-1}(\eta) = 1/(1 + e^{-\eta})$. Then

$$\frac{d\mu}{d\eta} = \frac{e^{-\eta}}{(1 + e^{-\eta})^2} = \frac{1}{1 + e^{-\eta}} \frac{e^{-\eta}}{1 + e^{-\eta}} = \mu(1 - \mu) = \mathrm{Var}(\mathcal{Y})$$

- For the Poisson, the canonical link is the log and the inverse link is $\mu = g^{-1}(\eta) = e^{\eta}$. Then

$$\frac{d\mu}{d\eta} = e^{\eta} = \mu = \mathrm{Var}(\mathcal{Y})$$

The unscaled conditional density of $\mathcal{U}|\mathcal{Y} = y$

- As in LMMs we evaluate the likelihood of the parameters, given the data, as

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) = \int_{\mathbb{R}^q} [\mathcal{Y}|\mathcal{U}](\boldsymbol{y}|\boldsymbol{u}) \, [\mathcal{U}](\boldsymbol{u}) \, d\boldsymbol{u},$$

- The product $[\mathcal{Y}|\mathcal{U}](\boldsymbol{y}|\boldsymbol{u})[\mathcal{U}](\boldsymbol{u})$ is the unscaled (or *unnormalized*) density of the conditional distribution $\mathcal{U}|\mathcal{Y}$.

- The density $[\mathcal{U}](\boldsymbol{u})$ is a spherical Gaussian density $\frac{1}{(2\pi)^{q/2}} e^{-\|\boldsymbol{u}\|^2/2}$.

- The expression $[\mathcal{Y}|\mathcal{U}](\boldsymbol{y}|\boldsymbol{u})$ is the value of a probability mass function or a probability density function, depending on whether $\mathcal{Y}_i|\mathcal{U}$ is discrete or continuous.

- The linear predictor is $\boldsymbol{g}(\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}$. Alternatively, we can write the conditional mean of $\mathcal{Y}$, given $\mathcal{U}$, as

$$\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{u}) = \boldsymbol{g}^{-1}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\right)$$

The conditional mode of $\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}} = \boldsymbol{y}$

- In general the likelihood, $L(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y})$ does not have a closed form. To approximate this value, we first determine the *conditional mode*

$$\tilde{\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\beta}) = \arg \max_{\boldsymbol{u}} [\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}](\boldsymbol{y}|\boldsymbol{u}) \, [\boldsymbol{\mathcal{U}}](\boldsymbol{u})$$

  using a quadratic approximation to the logarithm of the unscaled conditional density.

- This optimization problem is (relatively) easy because the quadratic approximation to the logarithm of the unscaled conditional density can be written as a penalized, weighted residual sum of squares,

$$\tilde{\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\beta}) = \arg \min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{W}^{1/2}(\boldsymbol{\mu}) \left(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{u}) \right) \\ -\boldsymbol{u} \end{bmatrix} \right\|^2$$

  where $\boldsymbol{W}(\boldsymbol{\mu})$ is the diagonal weights matrix. The weights are the inverses of the variances of the $\mathcal{Y}_i$.

The PIRLS algorithm

- Parameter estimates for generalized linear models (without random effects) are usually determined by iteratively reweighted least squares (IRLS), an incredibly efficient algorithm. PIRLS is the penalized version. It is iteratively reweighted in the sense that parameter estimates are determined for a fixed weights matrix W then the weights are updated to the current estimates and the process repeated.

- For fixed weights we solve

$$\min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{W}^{1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{u}) \right) \\ -\boldsymbol{u} \end{bmatrix} \right\|^2$$

as a nonlinear least squares problem with update, $\boldsymbol{\delta_u}$, given by

$$\boldsymbol{P} \left(\boldsymbol{\Lambda}_\theta^\mathsf{T} \boldsymbol{Z}^\mathsf{T} \boldsymbol{M} \boldsymbol{W} \boldsymbol{M} \boldsymbol{Z} \boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q \right) \boldsymbol{P}^\mathsf{T} \boldsymbol{\delta_u} = \boldsymbol{\Lambda}_\theta^\mathsf{T} \boldsymbol{Z}^\mathsf{T} \boldsymbol{M} \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{u}$$

where $\boldsymbol{M} = d\boldsymbol{\mu}/d\boldsymbol{\eta}$ is the (diagonal) Jacobian matrix. Recall that for the canonical link, $\boldsymbol{M} = \mathsf{Var}(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}) = \boldsymbol{W}^{-1}$.

The Laplace approximation to the deviance

- At convergence, the sparse Cholesky factor, L, used to evaluate the update is

$$LL^\mathsf{T} = P\left(\Lambda_\theta^\mathsf{T} Z^\mathsf{T} M W M Z \Lambda_\theta + I_q\right) P^\mathsf{T}$$

or

$$LL^\mathsf{T} = P\left(\Lambda_\theta^\mathsf{T} Z^\mathsf{T} M Z \Lambda_\theta + I_q\right) P^\mathsf{T}$$

if we are using the canonical link.

- The integrand of the likelihood is approximately a constant times the density of the $\mathcal{N}(\tilde{u}, LL^\mathsf{T})$ distribution.

- On the deviance scale (negative twice the log-likelihood) this corresponds to

$$d(\boldsymbol{\beta}, \boldsymbol{\theta} | \boldsymbol{y}) = d_g(\boldsymbol{y}, \boldsymbol{\mu}(\tilde{u})) + \|\tilde{u}\|^2 + \log(|L|^2)$$

where $d_g(\boldsymbol{y}, \boldsymbol{\mu}(\tilde{u}))$ is the GLM deviance for $\boldsymbol{y}$ and $\boldsymbol{\mu}$.

Modifications to the algorithm

- Notice that this deviance depends on the fixed-effects parameters, $\boldsymbol{\beta}$, as well as the variance-component parameters, $\boldsymbol{\theta}$. This is because $\log(|\boldsymbol{L}|^2)$ depends on $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}$ and, hence, on $\boldsymbol{\beta}$. For LMMs $\log(|\boldsymbol{L}|^2)$ depends only on $\boldsymbol{\theta}$.

- It is likely that modifying the PIRLS algorithm to optimize simultaneously on $\boldsymbol{u}$ and $\boldsymbol{\beta}$ would result in a value that is very close to the deviance profiled over $\boldsymbol{\beta}$.

- Another approach is adaptive Gauss-Hermite quadrature (AGQ). This has a similar structure to the Laplace approximation but is based on more evaluations of the unscaled conditional density near the conditional modes. It is only appropriate for models in which the random effects are associated with only one grouping factor

Contraception data

- One of the data sets in the "mlmRev" package, derived from data files available on the multilevel modelling web site, is from a fertility survey of women in Bangladesh.
- One of the (binary) responses recorded is whether or not the woman currently uses artificial contraception.
- Covariates included the woman's age (on a centered scale), the number of live children she had, whether she lived in an urban or rural setting, and the district in which she lived.
- Instead of plotting such data as points, we use the $0/1$ response to generate scatterplot smoother curves versus age for the different groups.

Contraception use versus age by urban and livch
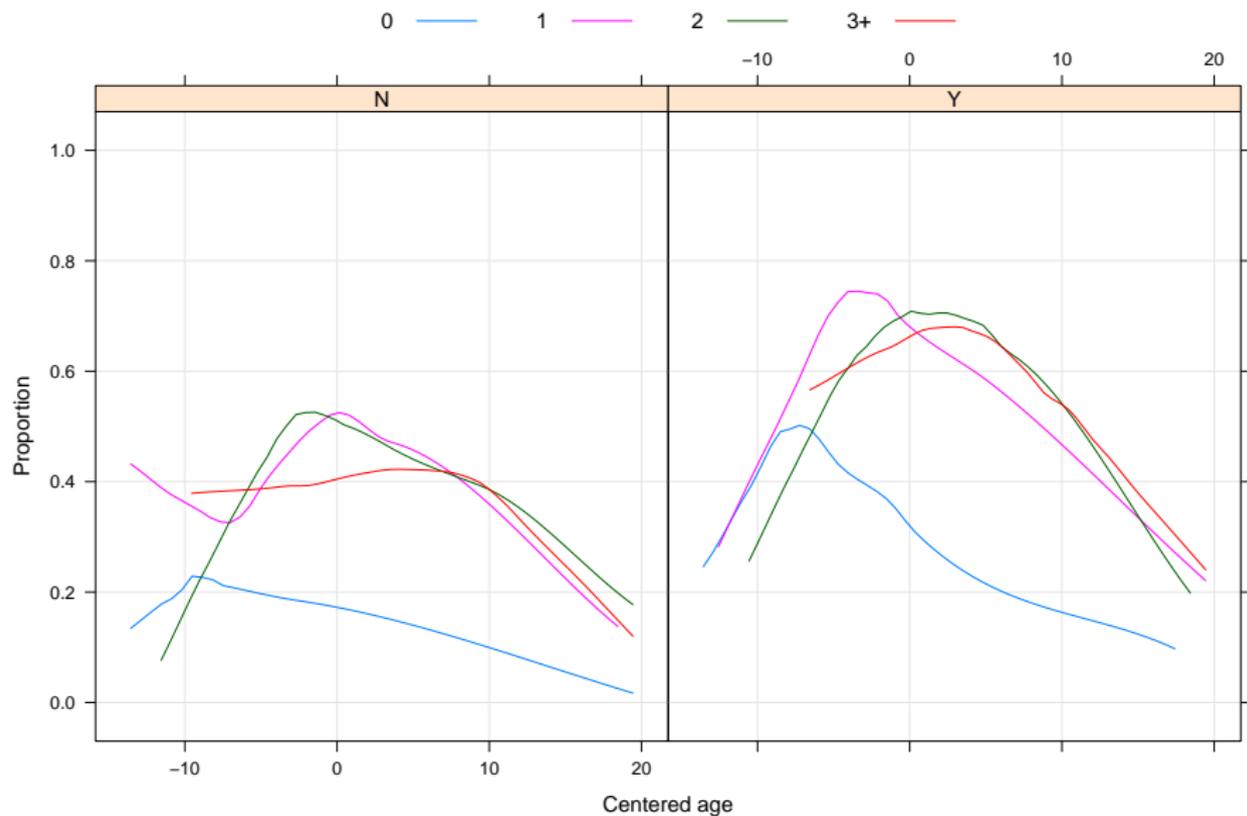
Comments on the data plot

- These observational data are unbalanced (some districts have only 2 observations, some have nearly 120). They are not longitudinal (no "time" variable).

- Binary responses have low per-observation information content (exactly one bit per observation). Districts with few observations will not contribute strongly to estimates of random effects.

- Within-district plots will be too imprecise so we only examine the global effects in plots.

- The comparisons on the multilevel modelling site are for fits of a model that is linear in age, which is clearly inappropriate.

- The form of the curves suggests at least a quadratic in age.

- The urban versus rural differences may be additive.

- It appears that the livch factor could be dichotomized into "0" versus "1 or more".

Preliminary model using Laplacian approximation

```
Generalized linear mixed model fit by maximum likelihood ['merMod']
 Family: binomial
Formula: use ~ age + I(age^2) + urban + livch + (1 | district)
   Data: Contraception
     AIC       BIC    logLik  deviance
 2388.774  2433.313 -1186.387  2372.774
Random effects:
 Groups   Name        Variance Std.Dev.
 district (Intercept) 0.2253   0.4747
Number of obs: 1934, groups: district, 60

Fixed effects:
             Estimate Std. Error z value
(Intercept) -1.019342   0.174010  -5.858
age          0.003516   0.009212   0.382
I(age^2)    -0.004487   0.000723  -6.205
urbanY       0.684625   0.119691   5.720
livch1       0.801901   0.161899   4.953
livch2       0.901037   0.184801   4.876
livch3+      0.899413   0.185435   4.850
```

Comments on the model fit

- This model was fit using the Laplacian approximation to the deviance.
- There is a highly significant quadratic term in age.
- The linear term in age is not significant but we retain it because the age scale has been centered at an arbitrary value (which, unfortunately, is not provided with the data).
- The urban factor is highly significant (as indicated by the plot).
- Levels of livch greater than 0 are significantly different from 0 but may not be different from each other.

Reduced model with dichotomized livch

```
Generalized linear mixed model fit by maximum likelihood ['merMod']
 Family: binomial
Formula: use ~ age + I(age^2) + urban + ch + (1 | district)
   Data: Contraception
      AIC       BIC    logLik  deviance
 2385.231  2418.635 -1186.615  2373.231
Random effects:
 Groups   Name        Variance Std.Dev.
 district (Intercept) 0.2242   0.4734
Number of obs: 1934, groups: district, 60

Fixed effects:
             Estimate Std. Error z value
(Intercept) -0.9913326  0.1675636  -5.916
age          0.0061720  0.0078273   0.789
I(age^2)    -0.0045584  0.0007142  -6.382
urbanY       0.6804429  0.1194818   5.695
chY          0.8462275  0.1470584   5.754
```

Comparing the model fits

- A likelihood ratio test can be used to compare these nested models.

```
> anova ( cm2 , cm1 )

Data: Contraception
Models:
cm2: use ~ age + I(age^2) + urban + ch + (1 | district)
cm1: use ~ age + I(age^2) + urban + livch + (1 | district)
    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
cm2  6 2385.2 2418.6 -1186.6   2373.2
cm1  8 2388.8 2433.3 -1186.4   2372.8 0.4567      2     0.7958
```
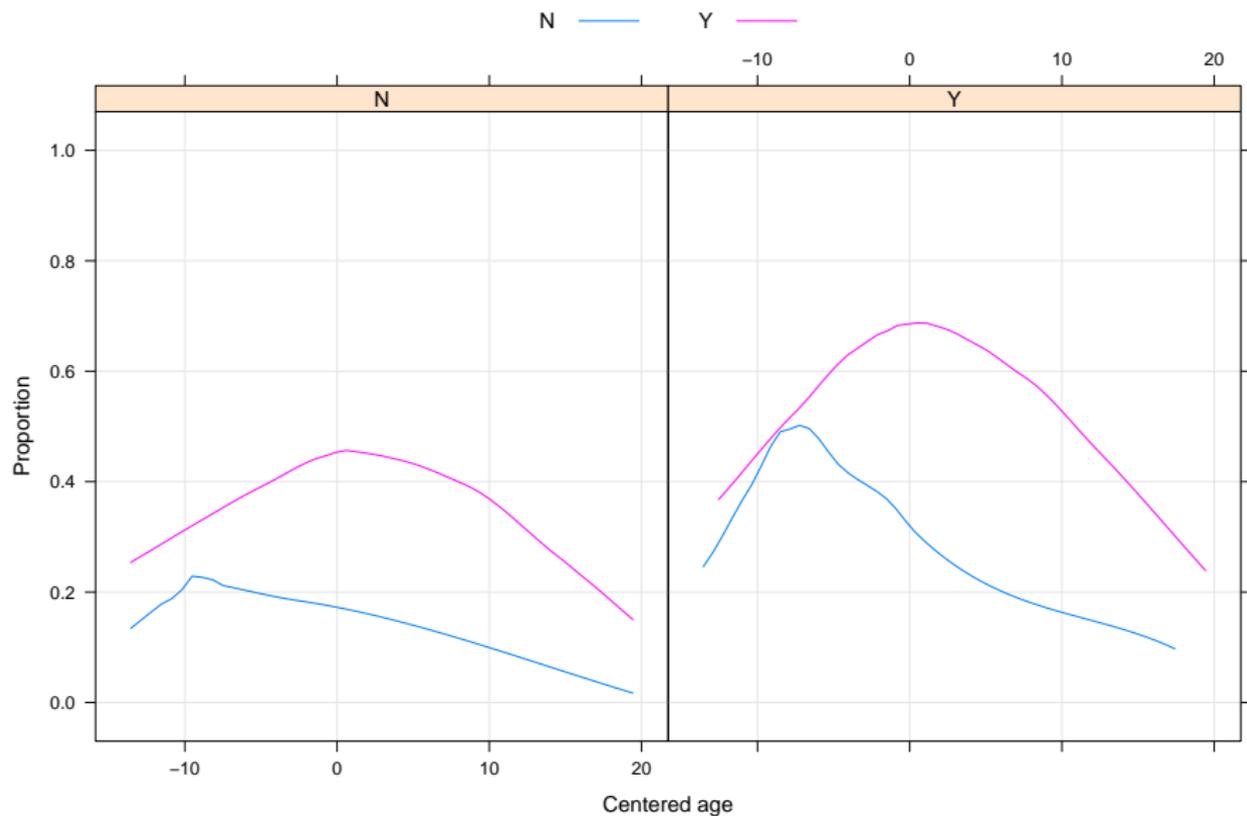
- The large p-value indicates that we would not reject cm2 in favor of cm1 hence we prefer the more parsimonious cm2.
- The plot of the scatterplot smoothers according to live children or none indicates that there may be a difference in the age pattern between these two groups.

Contraception use versus age by urban and ch

Allowing age pattern to vary with ch

```
Generalized linear mixed model fit by maximum likelihood ['merMod']
 Family: binomial
Formula: use ~ age * ch + I(age^2) + urban + (1 | district)
   Data: Contraception
      AIC       BIC    logLik  deviance
 2379.227  2418.198 -1182.613  2365.227
Random effects:
 Groups   Name        Variance Std.Dev.
 district (Intercept) 0.2225   0.4717
Number of obs: 1934, groups: district, 60

Fixed effects:
             Estimate Std. Error z value
(Intercept) -1.304583   0.213473  -6.111
age         -0.046526   0.021701  -2.144
chY          1.190985   0.205945   5.783
I(age^2)    -0.005660   0.000833  -6.795
urbanY       0.700970   0.120055   5.839
age:chY      0.067224   0.025293   2.658
```
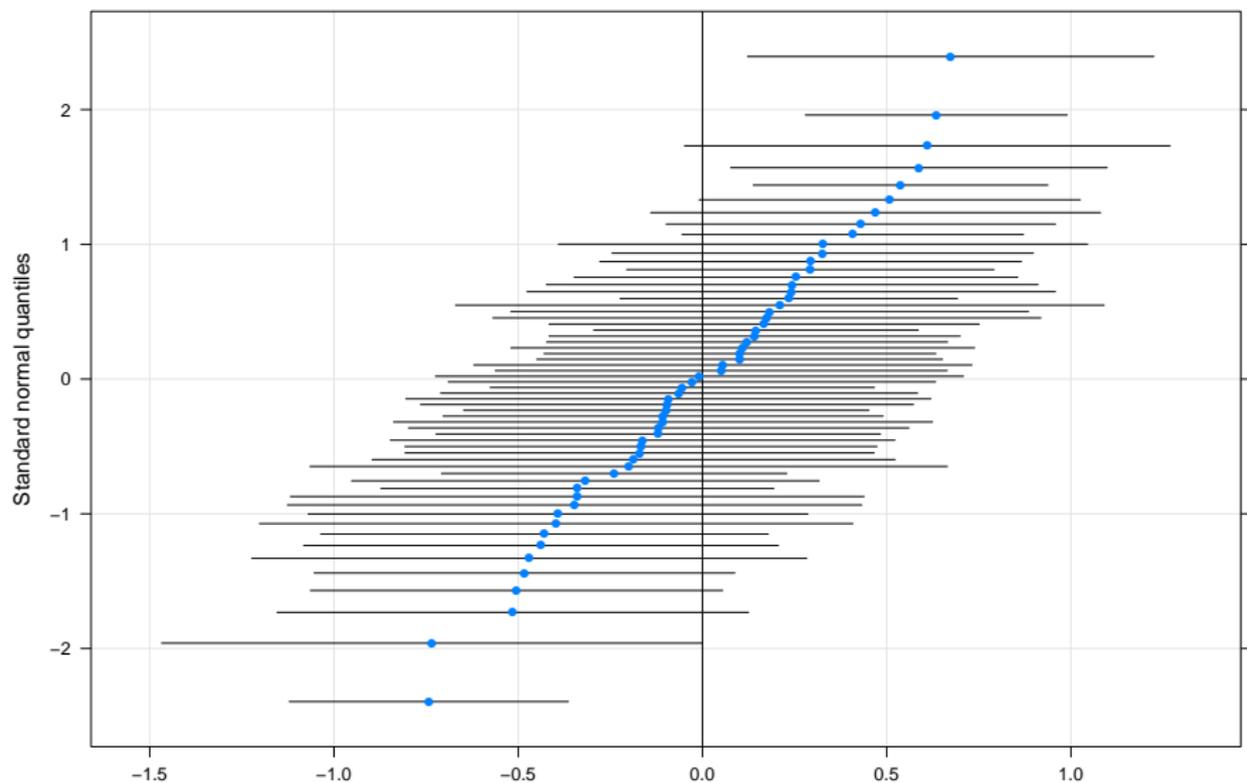
Prediction intervals on the random effects

Extending the random effects

- We may want to consider allowing a random effect for urban/rural by district. This is complicated by the fact the many districts only have rural women in the study

```
      district
urban  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
    N 54  20   0  19  37  58  18  35  20  13  21  23  16  17  14  18
    Y 63   0   2  11   2   7   0   2   3   0   0   6   8 101   8   2
      district
urban 17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32
    N 24  33  22  15  10  20  15  14  49  13  39  45  25  45  27  24
```

Including a random effect for urban by district

```
Generalized linear mixed model fit by maximum likelihood ['merMod']
 Family: binomial
Formula: use ~ age * ch + I(age^2) + urban + (urban | district)
   Data: Contraception
      AIC       BIC    logLik  deviance
 2371.636  2421.742 -1176.818  2353.636
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 district (Intercept) 0.3772   0.6142
          urbanY      0.5271   0.7260   -0.794
Number of obs: 1934, groups: district, 60

Fixed effects:
            Estimate Std. Error z value
(Intercept) -1.308648  0.221297  -5.914
age         -0.045071  0.021729  -2.074
chY          1.182726  0.206610   5.724
I(age^2)    -0.005509  0.000839  -6.567
urbanY       0.766746  0.159863   4.796
age:chY      0.064845  0.025348   2.558
```

Significance of the additional random effect

```
> anova(cm4,cm3)

Data: Contraception
Models:
cm3: use ~ age * ch + I(age^2) + urban + (1 | district)
cm4: use ~ age * ch + I(age^2) + urban + (urban | district)
    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
cm3  7 2379.2 2418.2 -1182.6   2365.2
cm4  9 2371.6 2421.7 -1176.8   2353.6 11.591      2   0.003042
```
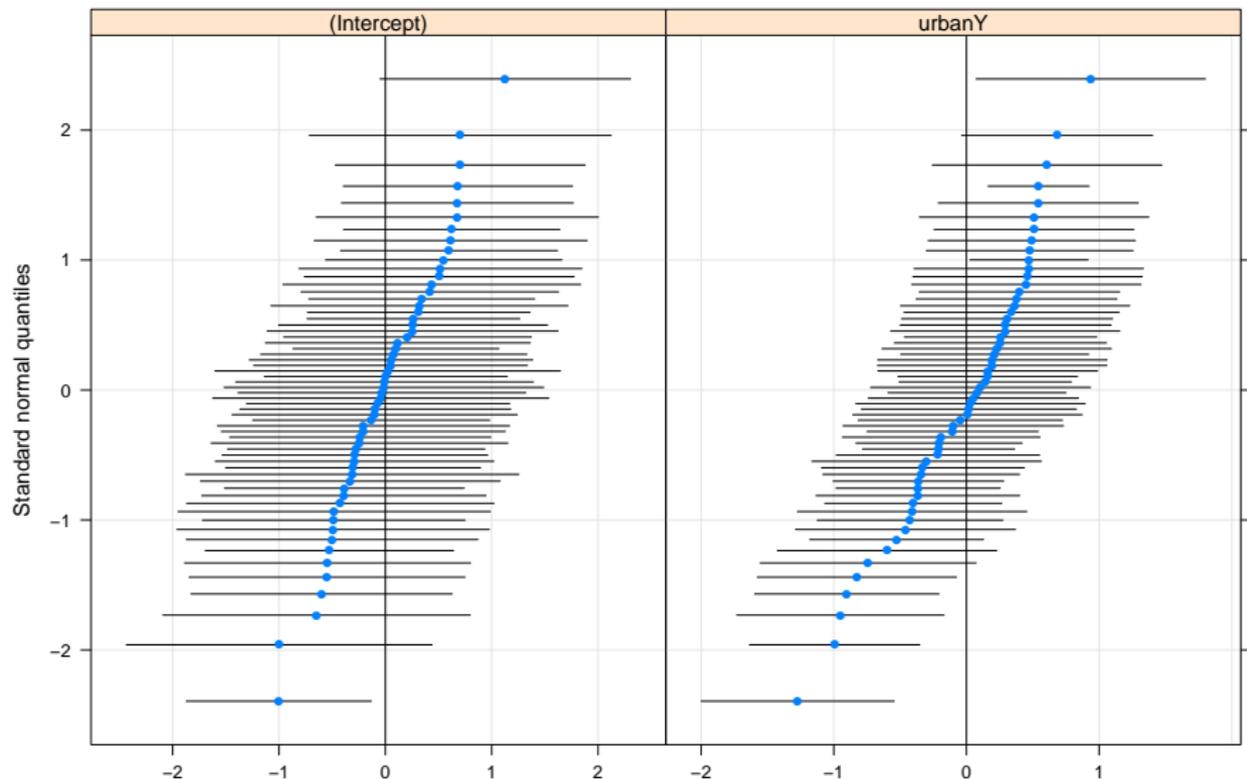
- The additional random effect is highly significant in this test.
- Most of the prediction intervals still overlap zero.
- A scatterplot of the random effects shows several random effects vectors falling along a straight line. These are the districts with all rural women or all urban women.

Prediction intervals for the bivariate random effects

Scatter plot of the BLUPs
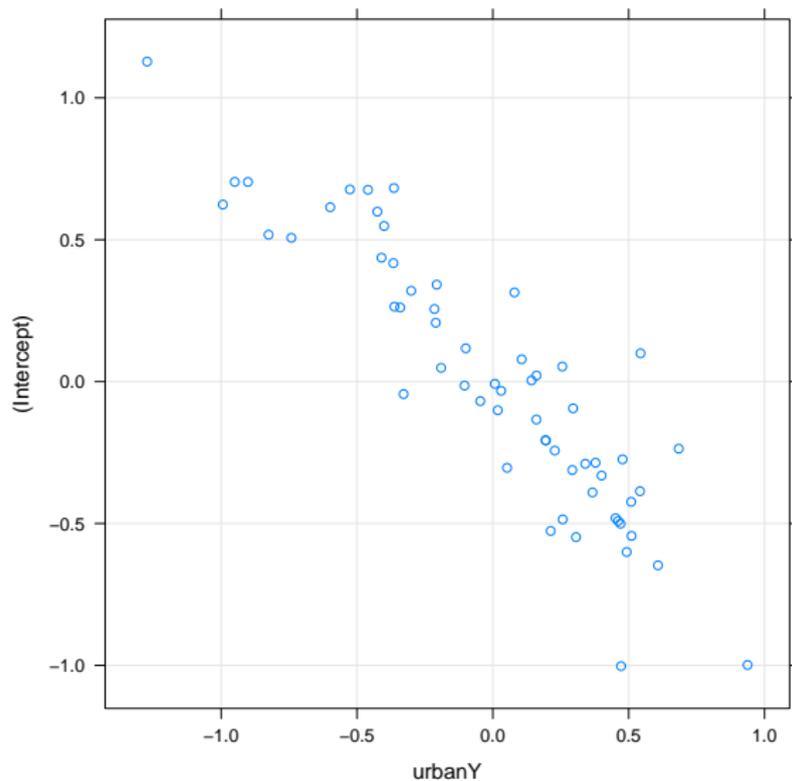
Nested simple, scalar random effects versus vector-valued

```
Generalized linear mixed model fit by maximum likelihood ['merMod']
 Family: binomial
Formula: use ~ age * ch + I(age^2) + urban + (1 | urban:district) + (1 |
   Data: Contraception
      AIC      BIC   logLik deviance
 2371.675  2416.214 -1177.838  2355.675
Random effects:
 Groups          Name        Variance  Std.Dev.
 urban:district (Intercept) 0.3086027 0.55552
 district       (Intercept) 0.0006808 0.02609
Number of obs: 1934, groups: urban:district, 102; district, 60

Fixed effects:
             Estimate Std. Error z value
(Intercept) -1.3042507  0.2188531  -5.959
age         -0.0448232  0.0217800  -2.058
chY          1.1810768  0.2070578   5.704
I(age^2)    -0.0054804  0.0008384  -6.537
urbanY       0.7613874  0.1683946   4.521
age:chY      0.0646389  0.0253854   2.546
```

Using the interaction term only

```
Generalized linear mixed model fit by maximum likelihood ['merMod']
 Family: binomial
Formula: use ~ age * ch + I(age^2) + urban + (1 | urban:district)
   Data: Contraception
      AIC       BIC    logLik  deviance
 2368.542  2407.514  -1177.271  2354.542
Random effects:
 Groups         Name         Variance  Std.Dev.
 urban:district (Intercept)  0.3218    0.5673
Number of obs: 1934, groups: urban:district, 102

Fixed effects:
             Estimate  Std. Error  z value
(Intercept) -1.3094339  0.2196450   -5.962
age         -0.0448437  0.0218185   -2.055
chY          1.1814382  0.2073624    5.697
I(age^2)    -0.0055277  0.0008409   -6.574
urbanY       0.7643321  0.1702667    4.489
age:chY      0.0646063  0.0254321    2.540
```

Comparing models with random effects for interactions

```
> anova(cm6,cm5,cm4)

Data: Contraception
Models:
cm6: use ~ age * ch + I(age^2) + urban + (1 | urban:district)
cm5: use ~ age * ch + I(age^2) + urban + (1 | urban:district) + (1 |
cm5:     district)
cm4: use ~ age * ch + I(age^2) + urban + (urban | district)
    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
cm6  7 2368.5 2407.5 -1177.3   2354.5
cm5  8 2371.7 2416.2 -1177.8   2355.7 0.0000      1     1.0000
cm4  9 2371.6 2421.7 -1176.8   2353.6 2.0393      1     0.1533
```

- The random effects seem to best be represented by a separate random effect for urban and for rural women in each district.
- The districts with only urban women in the survey or with only rural women in the survey are naturally represented in this model.

Conclusions from the example

- Again, carefully plotting the data is enormously helpful in formulating the model.
- Observational data tend to be unbalanced and have many more covariates than data from a designed experiment. Formulating a model is typically more difficult than in a designed experiment.
- A generalized linear model is fit with the function `glmer()` which requires a `family` argument. Typical values are `binomial` or `poisson`
- Profiling is not provided for GLMMs at present but will be added.
- We use likelihood-ratio tests and z-tests in the model building.

A word about overdispersion

- In many application areas using "pseudo" distribution families, such as `quasibinomial` and `quasipoisson`, is a popular and well-accepted technique for accomodating variability that is apparently larger than would be expected from a binomial or a Poisson distribution.

- This amounts to adding an extra parameter, like σ, the common scale parameter in a LMM, to the distribution of the response.

- It is possible to form an estimate of such a quantity during the IRLS algorithm but it is an artificial construct. There is no probability distribution with such a parameter.

- I find it difficult to define maximum likelihood estimates without a probability model. It is not clear how this "distribution which is not a distribution" could be incorporated into a GLMM. This, of course, does not stop people from doing it but I don't know what the estimates from such a model would mean.

Summary

- GLMMs allow for the conditional distribution, $\mathcal{Y}|\boldsymbol{B} = \boldsymbol{b}$, to be other than a Gaussian. A Bernoulli (or, more generally, a binomial) distribution is used to model binary or binomial responses. A Poisson distribution is used to model responses that are counts.

- The conditional mean depends upon the linear predictor, $\boldsymbol{X\beta} + \boldsymbol{Zb}$, through the inverse link function, $\boldsymbol{g}^{-1}$.

- The conditional mode of the random effects, given the observed data, $\boldsymbol{y}$, is determined through penalized iteratively reweighted least squares (PIRLS).

- We optimize the Laplace approximation at the conditional mode to determine the mle's of the parameters. In some simple cases, a more accurate approximation, adaptive Gauss-Hermite quadrature (AGQ), can be used instead, at the expense of greater computational complexity.