The hglm Package (version 1.2)

Lars Rönnegård
Dalarna University

Xia Shen
Uppsala University

Moudud Alam
Dalarna University

Abstract

This vignette describes the R **hglm** package via a series of applications that may be of interest to applied scientists. The **hglm** package implements the estimation algorithm for hierarchical generalized linear models. The package fits generalized linear models with random effects, where the random effect may come from a conjugate exponential-family distribution (Gaussian, Gamma, Beta or inverse-Gamma). The design matrices both for the fixed and random effects can be explicitly specified, which allows fitting correlated random effects as well as random regression models. Fixed effects may also be modeled in the dispersion parameter. The **hglm** package produces estimates of fixed effects, random effects, variance components as well as their standard errors. Model diagnostics such as deviances and leverages can be visualized. At the end of this vignette, estimates from the **hglm** package are compared to the ones from GenStat for several examples previously published by Lee and Nelder.

Keywords: hierarchical generalized linear models, *h*-likelihood, random effects.

1. Introduction

The **hglm** package implements the estimation algorithm for hierarchical generalized linear model (HGLM; Lee and Nelder 1996). The package fits generalized linear models (GLM; McCullagh and Nelder 1989) with random effects, where the random effect may come from a conjugate exponential-family distribution (normal, gamma, beta or inverse-gamma). The user may explicitly specify the design matrices both for the fixed and random effects, which means that correlated random effects as well as random regression models can be fitted. Fixed effects may also be modeled in the dispersion parameter.

Generalized linear mixed models (GLMM) have previously been implemented in several R (R Development Core Team 2009) function, such as the `glmer()` function in the **lme4** library and in the `glmmPQL()` function in the **MASS** library. In GLMM, the random effects are assumed to be Gaussian whereas the `hglm()` function allow for other distributions for the random effect. The `hglm()` function also extends the fitting algorithm of Gordon Smyth's **dglm** package by including random effects in the linear predictor for the mean. Moreover, the model specification in `hglm()` can be given as a formula or alternatively in terms of `y`, `X`, `Z` and `X.disp`, where `y` is the vector of observed responses, `X` and `Z` are the design matrices for the fixed and random effects, respectively, in the linear predictor for the mean, and `X.disp` is the design matrix for the fixed effects in the dispersion parameter. This enables a more flexible modeling of the random effects than specifying the model by an R formula. Consequently, this option is not as user friendly but gives the user a possibility to fit random regression

Table 1: **hglm** functions.

Function	Description	Reference
`Beta`	Extended usage of the Beta family	
`GLM.MME`	Internal IWLS estimation for `hglm()`	Lee, Nelder, and Pawitan (2006)
`hglm`	Fitting hierarchical generalized linear models	Lee and Nelder (1996)
`inverse.gamma`	Extended usage of the inverse-Gamma family	
- Utilities -		
`plot`	Plot individual deviances and hatvalues for the fitted **hglm** objects	
`print`	Produce basic statistics from **hglm** estimation in a simplified way	
`summary`	Produce standard summary statistics for the fitted **hglm** objects	

models and random effects with known correlation structure.

The **hglm** package produces estimates of fixed effects, random effects, variance components as well as their standard errors. In the output it also produces diagnostics such as deviances and leverages.

New in version 1.2

From version 1.2 of the package it is possible to fit several random effects (see Salamnder example below) and is implemented using sparse matrix techniques in the **Matrix** library.

2. Important implementation details

2.1. Brief overview of the fitting algorithm

The fitting algorithm is described in detail in Lee *et al.* (2006) and we summarize it here. Let n be the number of observations and k be the number of levels in the random effect. The algorithm is then given by:

1. Initialize starting values;
2. Construct an augmented model with response $y_{aug} = \begin{pmatrix} y \\ E(u) \end{pmatrix}$;
3. Use a GLM to estimate β and v given the vector ϕ and the dispersion parameter for the random effect λ. Save the deviances and leverages from the fitted model;
4. Use a gamma GLM to estimate β_d from the first n deviance residuals d and leverages h obtained from the previous model. The response variable and weights for this model are $d/(1-h)$ and $(1-h)/2$, respectively. Update the dispersion parameter by putting ϕ equal to the predicted response values for this model;

5. Use a similar GLM as in Step 4 to estimate λ from the last k deviance residuals and leverages obtained from the GLM in Step 3;

6. Iterate between steps 3-5 until convergence.

2.2. The h-likelihood theory

Let y be the response and u an unobserved random effects. The **hglm** package fits a hierarchical model $y|u \sim f_m(\mu, \phi)$ and $u \sim f_d(\psi, \lambda)$ where f_m and f_d are specified distributions for the mean and dispersion parts of the model.

We follow the notation of Lee and Nelder (1996), which is based on the GLM terminology by McCullagh and Nelder (1989). We also follow the likelihood approach where the model is described in terms of likelihoods. The conditional (log-)likelihood for y given u has the form of a GLM:

$$l(\theta', \phi; y|u) = \frac{y\theta' - b(\theta')}{a(\phi)} + c(y, \phi) \tag{1}$$

where θ' is the canonical parameter, ϕ is the dispersion term, μ' is the conditional mean of y given u where $\eta' = g(\mu')$, i.e. $g(.)$ is a link function for the GLM. The linear predictor μ' is given by $\eta' = \eta + v$ where $\eta = X\beta$ and $v = v(u)$ for some strict monotonic function of u. The hierarchical likelihood (h-likelihood) is defined by:

$$h = l(\theta', \phi; y|u) + l(\alpha; v) \tag{2}$$

where $l(\alpha; v)$ is the log density for v with parameter α. The estimates of β and v are given by $\frac{\partial h}{\partial \beta} = 0$ and $\frac{\partial h}{\partial v} = 0$. The dispersion components are estimated by maximizing the adjusted profile h-likelihood:

$$h_p = \left(h + \frac{1}{2} log|2\pi H^{-1}| \right)_{\beta=\hat{\beta}, v=\hat{v}} \tag{3}$$

where H is the Hessian matrix of the h-likelihood. The dispersion term ϕ can be connected to a liner predictor $X_d\beta_d$ given a link function $g_d(.)$ with $g_d(\phi) = X_d\beta_d$. The adjusted profile likelihoods of l and h may be used for inference of β, v and the dispersion parameters ϕ and λ (pp. 186 in Lee *et al.* 2006).

2.3. Detailed description of the fitting algorithm for a linear mixed model with heteroscedastic residual variance

In this section we describe the fitting algorithm in detail for a linear mixed model where fixed effects are included in the model for the residual variance. The extension to other distributions than Gaussian are described at the end of the section.

Lee and Nelder (1996) showed that linear mixed models can be fitted using a hierarchy of GLM by using an augmented linear model. The linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\mathbf{V} = \mathbf{Z}\mathbf{Z}'\sigma_u^2 + \mathbf{R}\sigma_e^2$$

where $\mathbf{R}$ is a diagonal matrix, and in the first iteration of the HGLM algorithm $\mathbf{R}$ is equal to the identity matrix. The model may be written as an augmented weighted linear model:

$$\mathbf{y}_a = \mathbf{T}_a\boldsymbol{\delta} + \mathbf{e}_a \tag{4}$$

where

$$\mathbf{y}_a = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_q \end{pmatrix}$$

$$\mathbf{T}_a = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}$$

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix}$$

$$\mathbf{e}_a = \begin{pmatrix} \mathbf{e} \\ -\mathbf{u} \end{pmatrix}$$

Here, q is the number of columns in $\mathbf{Z}$, $\mathbf{0}_q$ is a vector of zeros of length q, and $\mathbf{I}_q$ is the identity matrix of size $q \times q$. The variance-covariance matrix of the augmented residual vector is given by

$$V(\mathbf{e}_a) = \begin{pmatrix} \mathbf{R}\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q\sigma_u^2 \end{pmatrix}$$

Given σ_e^2 and σ_u^2, this weighted linear model gives the same estimates of the fixed and random effects ($\boldsymbol{\beta}$ and $\mathbf{u}$ respectively) as Henderson (1976)'s mixed model equations.

The estimates from weighted least squares are given by:

$$\mathbf{T}_a'\mathbf{W}^{-1}\mathbf{T}_a\hat{\boldsymbol{\delta}} = \mathbf{T}_a'\mathbf{W}^{-1}\mathbf{y}_a$$

where $\mathbf{W} \equiv V(\mathbf{e}_a)$.

The two variance components are estimated iteratively by applying a gamma GLM to the residuals e_i^2 and u_i^2 with intercept terms included in the linear predictors. The leverages h_i for these models are calculated from the diagonal elements of the hat matrix:

$$\mathbf{H}_a = \mathbf{T}_a(\mathbf{T}_a'\mathbf{W}^{-1}\mathbf{T}_a)^{-1}\mathbf{T}_a'\mathbf{W}^{-1} \tag{5}$$

A gamma GLM is used to fit the dispersion part of the model with response

$$y_{d,i} = e_i^2/(1-h_i) \tag{6}$$

where $E(y_d) = \mu_d$ and $\mu_d \equiv \phi$ (i.e. σ_e^2 for a Gaussian response). The GLM model for the dispersion parameter is then specified by the link function $g_d(.)$ and the linear predictor $X_d\beta_d$, with prior weights $(1-h_i)/2$, for

$$g_d(\mu_d) = X_d\beta_d \tag{7}$$

Similarly, a gamma GLM is fitted to the dispersion term α (i.e. σ_u^2 for a GLMM) for the random effect v, with

$$y_{\alpha,j} = u_j^2/(1-h_{n+j}) \tag{8}$$

and

$$g_\alpha(\mu_\alpha) = \lambda \tag{9}$$

where the prior weights are $(1-h_{n+j})/2$ and the estimated dispersion term for the random effect is given by $\hat{\alpha} = g_\alpha^{-1}(\hat{\lambda})$.

The algorithm iterates by updating $\mathbf{R} = diag(\hat{\phi})$ and going back to eq. (4).

For a non-Gaussian response variable y, the estimates are obtained simply by fitting a GLM instead of eq. (4) and by replacing e_i^2 and u_j^2 with the deviance residuals from the augmented model (Lee *et al.* 2006).

Based on $\log f_\theta(y|v)$, Lee and Nelder (1996) proposed using the scaled deviance for the goodness-of-fit test, having the estimated degrees of freedom, $d.f. = n - p_D$, where

$$p_D = trace\{(\mathbf{T}_m'\boldsymbol{\Sigma}_m^{-1}\mathbf{T}_m)^{-1}\}\mathbf{T}_m'\boldsymbol{\Sigma}_0^{-1}\mathbf{T}_m$$

and $\boldsymbol{\Sigma}_0^{-1} = \mathbf{W}_{ma}\{diag(\boldsymbol{\Phi}^{-1}, 0)\}$. m represents the mean model. Lee and Nelder (1996) showed that, under the assumed model, degrees of freedom can be estimated as $E(D) \approx n - p_D$. This extends the scaled deviance test for GLMs to HGLMs.

2.4. Distributions and link functions

There are two important classes of models that can be fitted in `hglm`: GLMM and conjugate HGLM. In GLMM we have a Gaussian random effect, whereas the conjugate HGLM has also been commonly used since explicit formulas for the marginal likelihood exist. HGLMs can also be used to fit models in survival analysis (frailty models), where for instance the complementary-log-log link function can be used on binary responses (e.g. Carling, Rönnegård, and Roszbach 2004; Alam and Carling 2008). The gamma distribution plays an important role in modeling responses with a constant coefficient of variation (see Chapter 8 in McCullagh and Nelder 1989), and for such responses with a gamma distributed random effect we have a gamma-gamma model. A summary of the most important models are given in Tables 2.4 and 3.

Table 2: Commonly used distributions and link functions possible to fit with `hglm()`

Model name	$y\|u$ distribution	Link $g(\mu)$	u distribution	Link $v(u)$
Linear mixed model	Gaussian	identity	Gaussian	identity
Binomial conjugate	Binomial	logit	Beta	logit
Binomial GLMM	Binomial	logit	Gaussian	identity
Binomial frailty	Binomial	comp-log-log	Gamma	log
Poisson GLMM	Poisson	log	Gaussian	identity
Poisson conjugate	Poisson	log	Gamma	log
Gamma GLMM	Gamma	log	Gaussian	identity
Gamma conjugate	Gamma	inverse	Inverse-Gamma	inverse
Gamma-Gamma	Gamma	log	Gamma	log

Table 3: `hglm()` code for commonly used models

Model name	`hglm()` code: `family =`	`hglm()` code: `rand.family =`
Linear mixed model	`gaussian(link = identity)`	`gaussian(link = identity)`
Beta-Binomial	`binomial(link = logit)`	`Beta(link = logit)`
Binomial GLMM	`binomial(link = logit)`	`gaussian(link = identity)`
Binomial frailty	`binomial(link = cloglog)`	`Gamma(link = log)`
Poisson GLMM	`poisson(link = log)`	`gaussian(link = identity)`
Poisson frailty	`poisson(link = log)`	`Gamma(link = log)`
Gamma GLMM	`Gamma(link = log)`	`gaussian(link = identity)`
Gamma conjugate	`Gamma(link = inverse)`	`inverse.gamma(link = inverse)`
Gamma-Gamma	`Gamma(link = log)`	`Gamma(link = log)`

2.5. Interacting with the hglm function

The main function is `hglm()` and the input is specified in a similar manner as for `glm()`. For instance, to fit a logit model for y with `week` as fixed effect and `ID` represents the clusters for a normally distributed random intercept, we run

```
R> hglm(fixed = y ~ week, random = ~ 1|ID,
+   family = binomial(link = logit))
```

Given an **hglm** object, the standard generic functions are `print()`, `summary()` and `plot()`. For this example, **hglm** allows an alternative command if the user would like to define the design matrices directly. If the design matrices of `week` and `ID` have been defined as `fixed.design` and `random.design`, respectively, we may run the following command instead.

```
R> hglm(X = fixed.design, y = response, X = random.design,
+   family = binomial(link = logit))
```

Now we analyze the bacteria data available in the **MASS** library using a binomial GLMM. The data consists of observations on the presence of *H.influenzae* at five occasions (at weeks 0, 2, 4, 6 and 11) on 50 individuals. Thirty observations were not reported and there are in total 220 observations. The model diagnostics produced by the `hglm()` function are shown in Figures 1 and 2. In Figure 1 there are two random effects with leverages > 0.7, which correspond to two individuals that only have two observations each. We also see that the assumption of the deviance residuals being gamma distributed is acceptable (Figure 2). The variance of the random individual effect was estimated to 1.559 and the algorithm converged in nine iterations.

```
> library(MASS)
> data(bacteria)
> g1 <- hglm(fixed = y ~ week, random = ~1 | ID, data = bacteria,
+      family = binomial(link = logit))
> summary(g1)

Call:
hglm.formula(family = binomial(link = logit), fixed = y ~ week,
```

```
    random = ~1 | ID, data = bacteria)


----------
MEAN MODEL
----------

Summary of the fixed effects estimates:

             Estimate Std. Error t-value Pr(>|t|)
(Intercept)  2.28703    0.33348   6.858  9.2e-11 ***
week        -0.13407    0.04123  -3.252  0.00135 **
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1
Note: P-values are based on 193 degrees of freedom

Summary of the random effects estimates:

      Estimate Std. Error
IDX01   0.7276     0.9729
IDX02  -0.2731     0.8312
IDX03   0.8392     0.9430
IDX04   0.8392     0.9430
IDX05   0.8392     0.9430
IDX06   0.7276     0.9729
IDX07  -1.4774     0.6997
IDX08  -0.1252     0.8026
IDX09   0.8392     0.9430
IDX10   0.3961     1.0683
IDX11   0.8392     0.9430
IDX12  -0.8526     0.7288
IDX13   0.8392     0.9430
IDX14  -0.7128     0.8851
IDX15   0.8392     0.9430
IDX16  -0.1252     0.8026
IDX17  -0.1252     0.8026
IDX18  -1.4774     0.6997
IDX19  -0.8526     0.7288
IDX20   0.7276     0.9729
IDX21   0.8392     0.9430
IDY01   0.8392     0.9430
IDY02  -1.4774     0.6997
IDY03  -0.8526     0.7288
IDY04  -1.8279     0.7500
IDY05  -1.4179     0.8230
IDY06   0.7535     0.9667
IDY07  -0.8526     0.7288
```

```
IDY08  -0.2409     0.8266
IDY09   0.7735     0.9616
IDY10  -0.1252     0.8026
IDY11   0.8392     0.9430
IDY12   0.3533     1.0811
IDY13  -1.9288     0.7409
IDY14   0.6247     1.0044
IDZ01   0.6247     1.0044
IDZ02   0.8392     0.9430
IDZ03   0.8392     0.9430
IDZ05  -0.8526     0.7288
IDZ06  -0.9623     0.9406
IDZ07  -0.8526     0.7288
IDZ09   0.7535     0.9667
IDZ10   0.8392     0.9430
IDZ11   0.8392     0.9430
IDZ14  -0.1252     0.8026
IDZ15   0.5355     1.0212
IDZ19   0.8392     0.9430
IDZ20  -1.0694     0.7654
IDZ24  -0.8526     0.7288
IDZ26  -0.1252     0.8026

----------------
DISPERSION MODEL
----------------

WARNING: h-likelihood estimates through EQL can be biased.

Dispersion parameter for the mean model:
[1] 0.7611218

Model estimates for the dispersion term:

Link = log

Effects:
  Estimate Std. Error
   -0.2730     0.1017

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

Dispersion parameter for the random effects:
[1] 1.565

Dispersion model for the random effects:
```

```
Link = log

Effects:
IDZ26
  Estimate Std. Error
    0.4478     0.2842

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

EQL estimation converged in 9 iterations.

> plot(g1)
```

2.6. Possible future developments

In the current version of `hglm()` it is possible to a include a single random effect in the mean part of the model. An important development would be to include several random effects in the mean part of the model and also to include random effects in the dispersion parts of the model. The latter class of models are called Double HGLM and have been shown to be a useful tool for modeling heavy tailed distributions Lee and Nelder (2006).

The algorithm of `hglm()` gives true marginal likelihood estimates for conjugate HGLM, whereas for other models the estimates are approximated. Lee and co-workers (see Lee *et al.* 2006, and references therein) have developed higher-order approximations, which give very good estimates. These higher-order approximations are not implemented in the current version of the **hglm** package. For these possible future extensions, we refer to the commercially available GenStat software and also to coming updates of the **hglm** package.

3. Linear mixed model with fixed effects in the residual variance

We consider a normal-normal model with heteroscedastic residual variance. In biology, for instance, this is important if we wish to model a random genetic effect (e.g. Rönnegård and Carlborg 2007) for a trait y and where the residual variance is different between sexes.

For the response y and observation number i we have:

$$y_i|\beta, u, \beta_d \sim N(X_i\beta + Z_i u, \exp(X_{d,i}\beta_d))$$

$$u \sim MVN(0, \mathbf{I}\sigma_u^2)$$

where β and u are the fixed and random effects in the mean part of the model, β_d is the fixed effect in the residual variance part of the model. The variance of the random effect u is given by σ_u^2. The subscript i for the matrices $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{X}_d$ indicate the i:th row. Here, a *log* link function is used for the dispersion term (i.e. the residual variance) and the model for the residual variance is therefore given by $\exp(X_{d,i}\beta_d)$. In the more general GLM notation the dispersion term ϕ is given by the residual variance here and $log(\phi_i) = X_{d,i}\beta_d$.

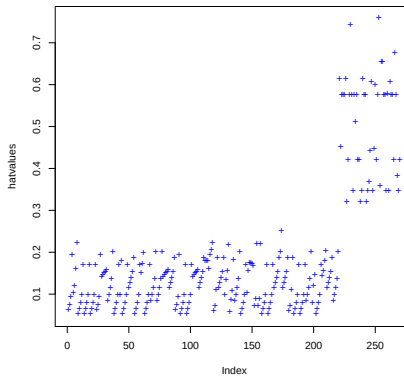

Figure 1: Leverages (i.e. diagonal elements of the augmented hat-matrix) for each observation 1 to 220 , and for each level in the random effect (index 221-282).

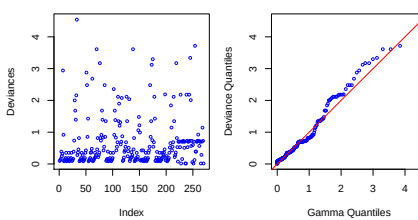

Figure 2: Deviance diagnostics for each observation and level in the random effect.

This model is not possible to fit in the **dglm** package, for instance, because we have random effects in the mean part of the model and it is also an improvement compared to the `glmer()` function since we allow a model for the residual variance.

We simulate data where there are five clusters with 20 observations in each cluster. For the mean part of the model, The simulated intercept value is $\mu = 0$ and the variance for the random effect is $\sigma_u^2 = 0.2$. Given the explanatory variable x_d, the simulated residual variance is 1.0 for $x_d = 0$ and 2.72 for $x_d = 1$. In this example, and the following ones, we show how the input code can be given in terms of the model matrices `y`, `X`, `Z` and `X.disp` instead of using R formula. The output shows that the variance of the random effect is 0.606, and that $\hat{\beta}_d = (0.354, 0.505)$.

```
> n.cluster <- 5
> n.per.cluster <- 20
> sigma2_u <- 0.2
> sigma2_e <- 1
> beta.disp <- 1
> mu <- 0
> n <- n.cluster * n.per.cluster
> set.seed(1234)
> X <- matrix(1, n, 1)
> Z <- diag(n.cluster) %x% rep(1, n.per.cluster)
> a <- rnorm(5, 0, sqrt(sigma2_u))
> X_d <- matrix(1, n, 2)
> X_d[, 2] <- rbinom(n, 1, 0.5)
> e <- rnorm(n, 0, sqrt(sigma2_e * exp(beta.disp * X_d[, 2])))
> y <- mu + Z %*% a + e
> simul1 <- hglm(y = y, X = X, Z = Z, X.disp = X_d)
> summary(simul1)

Call:
hglm.default(X = X, y = y, Z = Z, X.disp = X_d)


----------
MEAN MODEL
----------

Summary of the fixed effects estimates:

     Estimate Std. Error t-value Pr(>|t|)
X.1 -0.004179   0.268339  -0.016    0.988
Note: P-values are based on 96 degrees of freedom

Summary of the random effects estimates:

    Estimate Std. Error
Z.1   0.0454     0.3167
```

```
Z.2   0.0284     0.3187
Z.3   0.4314     0.3177
Z.4  -0.8336     0.3167
Z.5   0.3284     0.3133

----------------
DISPERSION MODEL
----------------

WARNING: h-likelihood estimates through EQL can be biased.

Model estimates for the dispersion term:

Link = log

Effects:
  Estimate Std. Error
1   0.0247     0.1859
2   0.5047     0.2958

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

Dispersion parameter for the random effects:
[1] 0.2982

Dispersion model for the random effects:

Link = log

Effects:
.|Random1
  Estimate Std. Error
   -1.2100     0.7755

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

EQL estimation converged in 4 iterations.
```

4. Poisson model with Gamma distributed random effects

For dependent count data it is common to model a Poisson distributed response with a gamma distributed random effect (Lee *et al.* 2006). If we assume no overdispersion conditional on u and thereby have a fixed dispersion term, this model may be specified as:

$$E(y_i|\beta, u) = \exp(X_i\beta + Z_i v)$$

where a level j in the random effect v is given by $v_j = log(u_j)$ and u_j are iid with gamma distribution having mean and variance: $E(u_j) = 1$, $var(u_j) = \lambda$.

This model is also possible to fit with the **hglm** package and extends other GLMM functions (e.g. `glmer()`) to allow for non-normal distributions for the random effect.

We simulate a Poisson model with random effects and test if there are differences in the dispersion term for an explanatory variable x_d. This example uses the data from the previous example. Hence, the simulated parameters and matrices `mu`, `a` and `Z` are the same. The estimated variance of the random effect is 0.056. The output also gives the estimate and standard error (0.551) of $log(\sigma_u^2)$. (Code continued from the previous example)

```
> eta <- exp(mu + Z %*% a)
> y <- rpois(length(eta), eta)
> simul.pois <- hglm(y = y, X = X, Z = Z, X.disp = X_d, family = poisson(link = log))
> summary(simul.pois)

Call:
hglm.default(X = X, y = y, Z = Z, family = poisson(link = log),
    X.disp = X_d)


----------
MEAN MODEL
----------

Summary of the fixed effects estimates:

    Estimate Std. Error t-value Pr(>|t|)
X.1 -0.07284    0.34560  -0.211    0.834
Note: P-values are based on 95 degrees of freedom

Summary of the random effects estimates:

    Estimate Std. Error
Z.1  -0.7050     0.4203
Z.2   0.3631     0.3762
Z.3   0.8090     0.3655
Z.4  -0.7181     0.4211
Z.5   0.2510     0.3766

----------------
DISPERSION MODEL
----------------

WARNING: h-likelihood estimates through EQL can be biased.

Model estimates for the dispersion term:
```

```
Link = log

Effects:
  Estimate Std. Error
1  -0.0367     0.1859
2   0.3426     0.2963

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

Dispersion parameter for the random effects:
[1] 0.5253

Dispersion model for the random effects:

Link = log

Effects:
.|Random1
  Estimate Std. Error
   -0.6438     0.7511

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

EQL estimation converged in 3 iterations.
```

5. Poisson-Gamma model with an offset for the mean model

The pump failure data of Gaver and O'Muircheartaigh (1987) contains the number of failures y_i and the period of operation t_i recorded for each of 10 pumps, so that the empirical failure rate is y_i/t_i. Gaver and O'Muircheartaigh (1987) mentioned that there are two groups of the pumps, where 4 pumps were operated continuously, and the rest intermittently. We fit the conjugate Poisson-Gamma HGLM with the group effects as fixed effects, the effects of the pumps as random effects, and an offset of $\log t_i$. On a log scale, the dispersion estimate of the random effects is $\log(4.208) = 1.437$.

```
> data(pump)
> offset.model <- hglm(fixed = S ~ factor(Gr), random = ~1 | System,
+     offset = log(t), fix.disp = 1, family = poisson(), rand.family = Gamma(),
+     data = pump)
> print(offset.model)

Call:
hglm.formula(family = poisson(), rand.family = Gamma(), fixed = S ~
    factor(Gr), random = ~1 | System, data = pump, fix.disp = 1,
```

```
    offset = log(t))

---------------------------
Estimates of the mean model
---------------------------

Fixed effects:
(Intercept) factor(Gr)1
 0.07485203 -1.66518573

Random effects:
 as.factor(System)1  as.factor(System)2  as.factor(System)3  as.factor(System)4
          0.2949244           0.1090546           0.4321987           0.5623241
 as.factor(System)5  as.factor(System)6  as.factor(System)7  as.factor(System)8
          0.5988128           2.7105528           0.9377853           0.9377853
 as.factor(System)9 as.factor(System)10
          1.5423544           1.8742075

Dispersion parameter for the mean model: 1

Dispersion parameter for the random effects: 1.047001

Estimation converged in 3 iterations
```

6. Linear mixed model with a correlated random effect

In animal breeding, it is important to estimate variance components prior to ranking of animal performances (Lynch and Walsh 1998). In such models the genetic effect of each animal is modelled as a level in a random effect and the correlation structure $\mathbf{A}$ is a matrix with known elements calculated from the pedigree information. The model is given by

$$y_i|\beta, u \sim N(\, X_i\beta + Z_i u, \sigma_e^2)$$

$$u \sim MVN(0, \mathbf{A}\sigma_u^2)$$

The model may be reformulated as (e.g. Lee *et al.* 2006; Rönnegård and Carlborg 2007)

$$y_i|\beta, u \sim N(\, X_i\beta + Z_i^* u^*, \sigma_e^2)$$

$$u^* \sim MVN(0, \mathbf{I}\sigma_u^2)$$

where $\mathbf{Z}^* = \mathbf{ZL}$ and $\mathbf{L}$ is the Cholesky factorization of $\mathbf{A}$.

The model can be fitted with the `hglm()` function since the input matrix `Z` is user-specified. Now we analyze the data set `QTLMAS` included in the **hglm** package as an example. The data

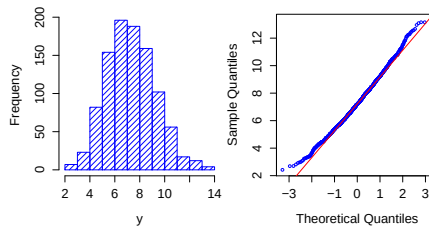

Figure 3: Histogram and qqplot for the analyzed trait.

consists of 2025 individuals from two generations where 1000 individuals have observed trait values y that are approximately normal (Figure 3). The data we analyze was simulated for the QTLMAS 2009 Workshop (Coster, Bastiaansen, Calus, Maliepaard, and Bink 2009)[1].

A longitudinal growth trait was simulated but for simplicity we analyze only the values given on the third occasion at age 265 days. We fitted a model with a fixed intercept and a random animal effect, a, where the correlation structure of a is given by the additive relationhip matrix $\mathbf{A}$ (which is obtained from the available pedigree information). A design matrix $\mathbf{Z}_0$ was constructed giving relating observation id-number in the pedigree. For observation y_i coming from individual j in the ordered pedigree file $\mathbf{Z}_0[i,j] = 1$, and all other elements are 0. Let $\mathbf{L}$ be the Cholesky factorization of $\mathbf{A}$, and $\mathbf{Z} = \mathbf{Z}_0\mathbf{L}$. The design matrix for the fixed effects, $\mathbf{X}$, is a column of ones. The estimated variance components are $\hat{\sigma}^2_e = 2.21$ and $\hat{\sigma}^2_u = 1.50$.

```
> data(QTLMAS)
> y <- QTLMAS[, 1]
> Z <- as.matrix(QTLMAS[, 2:2026])
> X <- matrix(1, 1000, 1)
> test1 <- hglm(y = y, X = X, Z = Z)
> print(test1)

Call:
hglm.default(X = X, y = y, Z = Z)

----------------------------
Estimates of the mean model
```

[1] http://www.qtlmas2009.wur.nl/UK/Dataset

```
---------------------------

Fixed effects:
     X.1
7.279761

Random effects:
          Z1           Z2           Z3           Z4           Z5
-1.1917337052  1.6486047743  1.3194273737 -0.9282585018 -0.4710833169
          Z6           Z7           Z8           Z9          Z10
-1.0583335326  1.0114515633  1.8796419925  0.6117058990 -0.2591250725
          Z11          Z12          Z13          Z14          Z15
-1.4267889421 -0.0051659774 -1.4901201892  0.5553842553 -1.2813772301
.
.
.
        Z1971        Z1972        Z1973        Z1974        Z1975
 0.4247723198 -0.5311791214 -0.5508397034 -0.1570602480  1.0964260666
        Z1976        Z1977        Z1978        Z1979        Z1980
-0.2640223315  0.3479044124 -0.2007819755  0.0842254872  0.1633291648
        Z1981        Z1982        Z1983        Z1984        Z1985
-0.7791366787  0.0628969497 -0.2665420090  0.1179749701  0.3145453021
        Z1986        Z1987        Z1988        Z1989        Z1990
 0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000
        Z1991        Z1992        Z1993        Z1994        Z1995
 0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000
        Z1996        Z1997        Z1998        Z1999        Z2000
 0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000
        Z2001        Z2002        Z2003        Z2004        Z2005
 0.0000000000  0.0000000000  0.0000000000  0.0000000000  0.0000000000
        Z2006        Z2007        Z2008        Z2009        Z2010
-0.1051557809  0.0813358412 -0.5929085599 -0.7107300985  0.4387751595
        Z2011        Z2012        Z2013        Z2014        Z2015
-0.5066362224  0.3862523049  0.8496955195 -0.3028617421  0.7378715231
        Z2016        Z2017        Z2018        Z2019        Z2020
-0.0438317993  0.3284771649  0.5710404831 -0.7629690457 -0.1369179125
        Z2021        Z2022        Z2023        Z2024        Z2025
 0.4938531799 -0.6084524857 -0.1515746280  0.4102069852 -0.2024295270

Dispersion parameter for the mean model: 2.211169

Dispersion parameter for the random effects: 1.502516

Estimation converged in 2 iterations
```

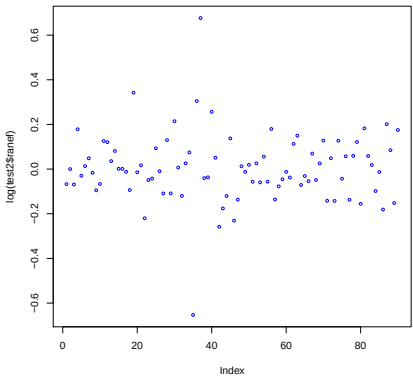

Figure 4: Estimated random effects for the 90 markers on chromosome 1 for trait values at 265 days. The main genetic effect was simulated close to marker number 40.

7. Random regression with a Gamma distributed random effect

The observed trait values y are the same as the previous example. Here we model a Gamma distributed random marker effect on chromosome 1. The incidence matrix $\mathbf{Z}_m$ is 1000×90 since there are 1000 observed phenotypes and 90 genetic markers on chromosome 1 in this data set. An element $\mathbf{Z}_m[i, j]$ is given by the number of copies (0, 1 or 2) for marker j in individual i. The design matrix for the fixed effects, $\mathbf{X}$, is a column of ones. The marker with the highest estimated random effect (Figure 4) was close to the main genetic effect simulated on chromosome 1. (Code continued from the previous example)

```
R> Z.markers <- QTLMAS[,2027:2116]
R> test2 <- hglm(y = y, X = X, Z = Z.markers,
+                rand.family = Gamma(link = log))
R> plot(log(test2$ranef))
```

8. Binomial model with a Beta distributed random effect

The seed germination data presented by Crowder (1978) has previously been analyzed using a binomial GLMM (Breslow and Clayton 1993) and a binomial-beta HGLM (Lee and Nelder 1996). The data consists of 831 observations from 21 germination plates. The effect of seed variety and type of root extract was studied in a 2×2 factorial lay-out. We fit the binomial-beta HGLM used by Lee and Nelder (1996) and setting the convergence criteria in `hglm()` to 0.005 produces the same estimates for the fixed effects as the ones obtained by Lee and Nelder (with differences $< 10^{-2}$). The beta distribution parameter α in Lee and Nelder (1996) was defined as $1/(2a)$ where a is the dispersion term obtained from `hglm()`. The output from the R code given below gives $\hat{a} = 0.026$ and the corresponding estimate given in Lee and Nelder (1996) is $\hat{a} = 1/(2\hat{\alpha}) = 0.023$.

```
> data(seeds)
> germ <- hglm(fixed = r/n ~ extract * I(seed == "073"), weights = n,
+     data = seeds, random = ~1 | plate, family = binomial(), rand.family = Beta(),
+     fix.disp = 1)
> summary(germ)

Call:
hglm.formula(family = binomial(), rand.family = Beta(), fixed = r/n ~
    extract * I(seed == "073"), random = ~1 | plate, data = seeds,
    weights = n, fix.disp = 1)


----------
MEAN MODEL
----------

Summary of the fixed effects estimates:

                                      Estimate Std. Error t-value Pr(>|t|)
(Intercept)                           -0.54234    0.19285  -2.812 0.018401 *
extractCucumber                        1.34005    0.27333   4.903 0.000621 ***
I(seed == "073")TRUE                   0.07519    0.31136   0.241 0.814054
extractCucumber:I(seed == "073")TRUE -0.82704    0.43417  -1.905 0.085927 .
---
Signif. codes:  0 Š***Š 0.001 Š**Š 0.01 Š*Š 0.05 Š.Š 0.1 Š Š 1
Note: P-values are based on 10 degrees of freedom

Summary of the random effects estimates:

                    Estimate Std. Error
as.factor(plate)1    0.4419     0.2510
as.factor(plate)2    0.5021     0.2328
as.factor(plate)3    0.4398     0.2283
as.factor(plate)4    0.5814     0.2386
```

```
as.factor(plate)5    0.5348     0.2468
as.factor(plate)6    0.5334     0.2786
as.factor(plate)7    0.4770     0.2640
as.factor(plate)8    0.4575     0.2667
as.factor(plate)9    0.5678     0.2550
as.factor(plate)10   0.4644     0.3058
as.factor(plate)11   0.5194     0.3007
as.factor(plate)12   0.5194     0.2314
as.factor(plate)13   0.5542     0.2346
as.factor(plate)14   0.4626     0.2404
as.factor(plate)15   0.4239     0.2273
as.factor(plate)16   0.5205     0.2852
as.factor(plate)17   0.4392     0.2857
as.factor(plate)18   0.5128     0.2524
as.factor(plate)19   0.4951     0.2600
as.factor(plate)20   0.5655     0.2493
as.factor(plate)21   0.4875     0.2953

----------------
DISPERSION MODEL
----------------

WARNING: h-likelihood estimates through EQL can be biased.

Dispersion parameter for the mean model:
[1] 1

Model estimates for the dispersion term:

Link = log

Effects:
[1] 1

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

Dispersion parameter for the random effects:
[1] 0.02484

Dispersion model for the random effects:

Link = log

Effects:
as.factor(plate)21
  Estimate Std. Error
   -3.6954     0.5304
```

```
Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

EQL estimation converged in 7 iterations.
```

9. Gamma HGLM with a structured dispersion

In this example, we analyze the semiconductor data taken from Myers, Montgomery, and Vining (2002), which involves a designed experiment in a semiconductor plant. Six factors, lamination temperature, lamination time, lamination pressure, firing temperature, firing cycle time and firing dew point, are employed, and we are interested in the curvature of the substrate devices produced in the plant. The curvature measurement is made four times on each device produced. Each design variable in taken at two levels. The measurement is known to be non-normally distributed, and the measurements taken on the same device are correlated. Myers *et al.* (2002) considered a gamma response model with a log link and used a GEE method assuming an AR(1) working correlation.

We consider a gamma HGLM by adding a random effect for the device in the mean model

$$\log \mu = \beta_0 + x_1\beta_1 + x_3\beta_3 + x_5\beta_5 + x_6\beta_6.$$

And the dispersion model is

$$\log \phi = \gamma_0 + x_2\gamma_2 + x_3\gamma_3$$

The variance λ of random effects represents the between-group variance, while ϕ represents the within-group variance. Results are shown as follows, and residual plots for the mean and dispersion models are in Figures (5, 6, 7). The estimates are the same as those obtained using EQL in GenStat.

```
R> data(semiconductor)
R> gamma.model <- hglm(fixed = y ~ x1 + x3 + x5 + x6,
+                      random = ~ 1|Device,
+                      family = Gamma(link = log),
+                      disp = ~ x2 + x3, data = semiconductor)
R> summary(gamma.model)
R> plot(gamma.model, cex = .6, pch = 1,
+       cex.axis = 1/.6, cex.lab = 1/.6,
+       cex.main = 1/.6, mar = c(3,4.5,0,1.5))

Call:
hglm.formula(fixed = y ~ x1 + x3 + x5 + x6, random = ~1 | Device,
data = semiconductor, family = Gamma(link = log), disp = ~x2 + x3)

DISPERSION MODEL
WARNING: h-likelihood estimates through EQL can be biased.
Model estimates for the dispersion term:
Link = log
Effects:
                        Estimate Std. Error
```

```
(Intercept)  -2.5910     0.1974
x2           -0.6815     0.1973
x3           -0.4989     0.1973

Dispersion = 1 is used in Gamma model on deviances to
calculate the standard error(s).
Dispersion parameter for the random effects
[1] 0.0496

Dispersion model for the random effects:
Link = log
Effects:
Estimate Std. Error
 -3.0037     0.5139

Dispersion = 1 is used in Gamma model on deviances to
calculate the standard error(s).
MEAN MODEL
Summary of the fixed effects estimates
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.71131    0.06819 -69.088  < 2e-16 ***
x1           0.20871    0.06769   3.083  0.00311 **
x3           0.32853    0.06819   4.818 1.05e-05 ***
x5          -0.17288    0.06769  -2.554  0.01326 *
x6          -0.35693    0.06766  -5.276 1.98e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Summary of the random effects estimate
                    Estimate Std. Error
as.factor(Device)1    0.2852     0.1807
as.factor(Device)2    0.0113     0.1807
...
as.factor(Device)16   0.2178     0.1525

EQL estimation converged in 2 iterations.
```

```
> data(semiconductor)
> gamma.model <- hglm(fixed = y ~ x1 + x3 + x5 + x6, random = ~1 |
+     Device, family = Gamma(link = log), disp = ~x2 + x3, data = semiconductor)
> summary(gamma.model)

Call:
hglm.formula(family = Gamma(link = log), fixed = y ~ x1 + x3 +
    x5 + x6, random = ~1 | Device, disp = ~x2 + x3, data = semiconductor)


----------
MEAN MODEL
----------
```

```
Summary of the fixed effects estimates:

            Estimate Std. Error t-value Pr(>|t|)
(Intercept) -4.71076    0.06878 -68.494  < 2e-16 ***
x1           0.20710    0.06829   3.032  0.00381 **
x3           0.32655    0.06878   4.748 1.71e-05 ***
x5          -0.17540    0.06829  -2.568  0.01319 *
x6          -0.35797    0.06826  -5.244 3.04e-06 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1
Note: P-values are based on 51 degrees of freedom

Summary of the random effects estimates:

                     Estimate Std. Error
as.factor(Device)1    0.2853     0.1823
as.factor(Device)2    0.0117     0.1823
as.factor(Device)3   -0.2753     0.1632
as.factor(Device)4   -0.1689     0.1632
as.factor(Device)5   -0.0450     0.1633
as.factor(Device)6   -0.0890     0.1633
as.factor(Device)7    0.1255     0.1539
as.factor(Device)8    0.1786     0.1539
as.factor(Device)9    0.0649     0.1823
as.factor(Device)10  -0.0550     0.1823
as.factor(Device)11   0.0133     0.1632
as.factor(Device)12   0.1240     0.1632
as.factor(Device)13  -0.0469     0.1633
as.factor(Device)14  -0.2201     0.1633
as.factor(Device)15  -0.1217     0.1539
as.factor(Device)16   0.2186     0.1539

----------------
DISPERSION MODEL
----------------

WARNING: h-likelihood estimates through EQL can be biased.

Model estimates for the dispersion term:

Link = log

Effects:
            Estimate Std. Error
(Intercept)  -2.5914     0.1975
x2           -0.6828     0.1974
x3           -0.4992     0.1974
```

```
Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

Dispersion parameter for the random effects:
[1] 0.04969

Dispersion model for the random effects:

Link = log

Effects:
as.factor(Device)16
  Estimate Std. Error
   -3.0019     0.5118

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

EQL estimation converged in 2 iterations.

> plot(gamma.model, cex = 0.6, pch = 1, cex.axis = 1/0.6, cex.lab = 1/0.6,
+     cex.main = 1/0.6, mar = c(3, 4.5, 0, 1.5))
```

10. More than one random effect in the $hglm2$ function

From version 1.2 of the **hglm** package it is possible to fit models having several random effects. The model is specified using the **lme4** formula convension. Here the $hglm2$ function is illustrated using the classical data set from a Salamander mating experiment McCullagh and Nelder (1989). The binary response variable (`Mate`) is mating success or not, and two variance components σ_f^2 and σ_m^2 are estimated for females and males, respectively. The results below show that $\hat{\sigma}_f^2 = 0.72$ and $\hat{\sigma}_m^2 = 0.64$ (the EQL method in Genstat gives $\hat{\sigma}_f^2 = 0.72$ and $\hat{\sigma}_m^2 = 0.62$).

```
> data(salamander)
> hglm.salam <- hglm2(meanmodel = Mate ~ TypeF + TypeM + TypeF *
+     TypeM + (1 | Female) + (1 | Male), family = binomial(), data = salamander,
+     conv = 1e-08, maxit = 40)
> summary(hglm.salam)

Call:
hglm2.formula(meanmodel = Mate ~ TypeF + TypeM + TypeF * TypeM +
    (1 | Female) + (1 | Male), data = salamander, family = binomial(),
    conv = 1e-08, maxit = 40)


----------
```

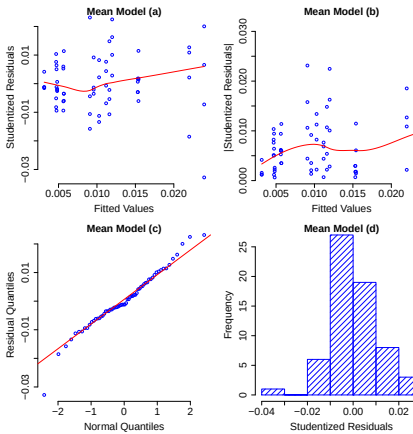

Figure 5: Residual plots of the mean model for the semiconductor data.

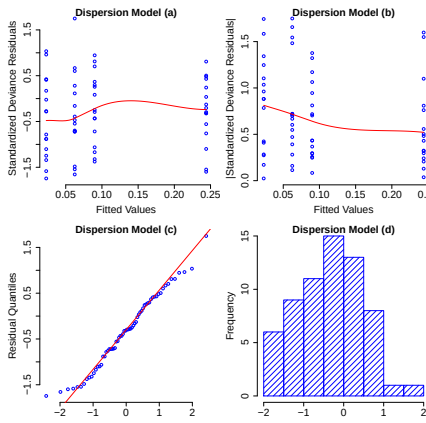

Figure 6: Residual plots of the dispersion model for the semiconductor data.

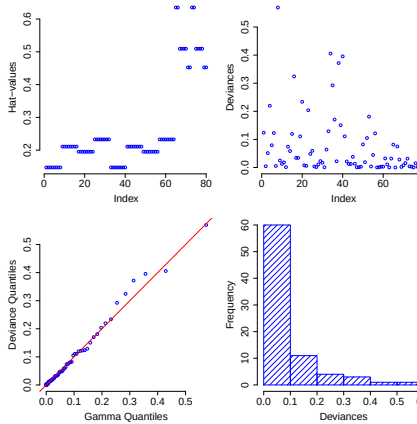

Figure 7: Diagnostics plots of the hat-values and deviances for the semiconductor data.

```
MEAN MODEL
----------

Summary of the fixed effects estimates:

              Estimate Std. Error t-value Pr(>|t|)
(Intercept)     0.7881     0.3210   2.455   0.0146 *
TypeFW         -2.2903     0.4315  -5.308 2.12e-07 ***
TypeMW         -0.5410     0.3890  -1.391   0.1654
TypeFW:TypeMW   2.8208     0.4974   5.671 3.26e-08 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1
Note: P-values are based on 308 degrees of freedom

Summary of the random effects estimates:

                        Estimate Std. Error
(Intercept)| Female:1    0.3982     0.6527
(Intercept)| Female:2    0.7099     0.6902
(Intercept)| Female:3    0.3160     0.6595
(Intercept)| Female:4    0.3957     0.6493
(Intercept)| Female:5    0.7790     0.6812
(Intercept)| Female:6    0.5167     0.6410
(Intercept)| Female:7   -0.9803     0.6256
(Intercept)| Female:8   -0.1591     0.6204
(Intercept)| Female:9    0.1447     0.6252
(Intercept)| Female:10  -0.4881     0.6203
(Intercept)| Female:11   0.6029     0.6719
(Intercept)| Female:12  -0.5651     0.6653
(Intercept)| Female:13  -0.5605     0.6682
(Intercept)| Female:14   0.5665     0.6663
(Intercept)| Female:15  -0.6063     0.6722
(Intercept)| Female:16  -0.8006     0.6627
(Intercept)| Female:17   0.8572     0.6521
(Intercept)| Female:18   0.0633     0.6446
(Intercept)| Female:19   0.7857     0.6571
(Intercept)| Female:20  -0.0257     0.6556
(Intercept)| Female:21  -0.2983     0.6283
(Intercept)| Female:22   0.0533     0.6346
(Intercept)| Female:23   0.5395     0.6480
(Intercept)| Female:24  -0.9501     0.6320
(Intercept)| Female:25  -0.6135     0.6340
(Intercept)| Female:26  -0.2541     0.6178
(Intercept)| Female:27   0.1335     0.6310
(Intercept)| Female:28  -0.5088     0.6200
(Intercept)| Female:29   0.5220     0.6400
(Intercept)| Female:30  -0.1779     0.6223
```

```
(Intercept)| Female:31  -0.2489     0.6744
(Intercept)| Female:32  -0.1533     0.6652
(Intercept)| Female:33   0.2464     0.6566
(Intercept)| Female:34   0.1371     0.6655
(Intercept)| Female:35  -1.0734     0.6837
(Intercept)| Female:36   0.9411     0.6590
(Intercept)| Female:37  -0.1645     0.6583
(Intercept)| Female:38  -1.0692     0.6637
(Intercept)| Female:39   1.2022     0.6563
(Intercept)| Female:40   0.5689     0.6580
(Intercept)| Female:41   0.5804     0.6433
(Intercept)| Female:42  -0.9132     0.6294
(Intercept)| Female:43   0.5328     0.6487
(Intercept)| Female:44  -0.2124     0.6231
(Intercept)| Female:45  -0.1857     0.6319
(Intercept)| Female:46   0.4968     0.6525
(Intercept)| Female:47  -0.9086     0.6359
(Intercept)| Female:48   0.4325     0.6509
(Intercept)| Female:49  -0.2603     0.6279
(Intercept)| Female:50   0.3594     0.6584
(Intercept)| Female:51  -0.2239     0.6581
(Intercept)| Female:52   0.1756     0.6738
(Intercept)| Female:53   0.1811     0.6495
(Intercept)| Female:54   0.0978     0.6444
(Intercept)| Female:55   0.3980     0.6481
(Intercept)| Female:56  -0.7029     0.6709
(Intercept)| Female:57  -0.1518     0.6600
(Intercept)| Female:58  -0.0558     0.6582
(Intercept)| Female:59  -0.2071     0.6689
(Intercept)| Female:60  -0.2149     0.6715

Summary of the random effects estimates:

                      Estimate Std. Error
(Intercept)| Male:1    0.7426     0.6451
(Intercept)| Male:2    0.0169     0.6435
(Intercept)| Male:3    0.7487     0.6439
(Intercept)| Male:4    0.0120     0.6438
(Intercept)| Male:5    0.3625     0.6376
(Intercept)| Male:6   -0.0615     0.6432
(Intercept)| Male:7   -0.6224     0.6429
(Intercept)| Male:8   -0.1213     0.6355
(Intercept)| Male:9    0.1569     0.6235
(Intercept)| Male:10  -0.6016     0.6307
(Intercept)| Male:11  -0.1434     0.6191
(Intercept)| Male:12   0.3657     0.6264
(Intercept)| Male:13   0.7261     0.6516
```

```
(Intercept)| Male:14   0.2652     0.6341
(Intercept)| Male:15  -0.0226     0.6147
(Intercept)| Male:16  -0.0975     0.5973
(Intercept)| Male:17  -0.6542     0.6099
(Intercept)| Male:18   0.5421     0.6252
(Intercept)| Male:19   0.0897     0.6174
(Intercept)| Male:20   0.0191     0.6132
(Intercept)| Male:21   0.3414     0.6217
(Intercept)| Male:22  -0.9736     0.6439
(Intercept)| Male:23   0.8247     0.6369
(Intercept)| Male:24   0.8963     0.6225
(Intercept)| Male:25  -0.1098     0.6319
(Intercept)| Male:26  -0.2067     0.6417
(Intercept)| Male:27  -0.6158     0.6401
(Intercept)| Male:28   0.1946     0.6324
(Intercept)| Male:29  -0.1610     0.6308
(Intercept)| Male:30  -0.5641     0.6463
(Intercept)| Male:31  -0.5385     0.6109
(Intercept)| Male:32  -0.1370     0.6149
(Intercept)| Male:33  -0.0717     0.6268
(Intercept)| Male:34   0.0636     0.6225
(Intercept)| Male:35  -0.1103     0.6088
(Intercept)| Male:36   0.2496     0.6048
(Intercept)| Male:37  -0.7990     0.6079
(Intercept)| Male:38   0.5015     0.6188
(Intercept)| Male:39   0.0504     0.6089
(Intercept)| Male:40   0.1336     0.6074
(Intercept)| Male:41  -0.1889     0.6486
(Intercept)| Male:42   0.2020     0.6411
(Intercept)| Male:43  -0.5403     0.6446
(Intercept)| Male:44   0.1929     0.6405
(Intercept)| Male:45  -0.5436     0.6445
(Intercept)| Male:46  -0.1989     0.6252
(Intercept)| Male:47  -0.2861     0.6363
(Intercept)| Male:48   0.8658     0.6276
(Intercept)| Male:49  -0.6175     0.6440
(Intercept)| Male:50   0.8559     0.6355
(Intercept)| Male:51  -0.9534     0.6000
(Intercept)| Male:52  -0.7414     0.6088
(Intercept)| Male:53   0.9475     0.6327
(Intercept)| Male:54   0.0627     0.6072
(Intercept)| Male:55   0.2102     0.6076
(Intercept)| Male:56   0.6876     0.6521
(Intercept)| Male:57  -0.3009     0.6074
(Intercept)| Male:58   0.1436     0.6120
(Intercept)| Male:59  -0.8807     0.6075
(Intercept)| Male:60   0.3924     0.6237
```

```
----------------
DISPERSION MODEL
----------------

WARNING: h-likelihood estimates through EQL can be biased.

Dispersion parameter for the mean model:
[1] 0.9954023

Model estimates for the dispersion term:

Link = log

Effects:
  Estimate Std. Error
   -0.0046     0.0806

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

Dispersion parameter for the random effects:
[1] 0.7245 0.6402

Dispersion model for the random effects:

Link = log

Effects:
(Intercept)| Female
  Estimate Std. Error
   -0.3222     0.2833

(Intercept)| Male
  Estimate Std. Error
   -0.4460     0.2941

Dispersion = 1 is used in Gamma model on deviances to calculate the standard error(s).

EQL estimation converged in 28 iterations.
```

Summary

The hierarchical generalized linear model approach gives new possibilities to fit generalized linear models with random effects. The **hglm** package extends existing GLMM fitting algorithms to include fixed effects in a model for the residual variance, fit models where the random effect distribution is not necessarily Gaussian and estimate variance components for

correlated random effects. For such models there are important applications in, for instance: genetics (Noh, Yip, Lee, and Pawitan 2006), survival analysis (Ha and Lee 2005), count data (Lee *et al.* 2006) and dichotomous responses (Noh and Lee 2007). We therefore expect that this new package will be of use for applied statisticians in several different fields.

11. Discussions on h-likelihood theory

For the users not previously acquainted with h-likelihood theory, we summarize here the discussions in Lee and Nelder (1996), Lee, Nelder, and Noh (2007), Lee and Nelder (2006) and the collection of discussion papers in Statistical Science vol. 24 no. 3 (Lee and Nelder 2009a; Meng 2009; Louis 2009; Molenberghs, Kenward, and Verbeke 2009; Lee and Nelder 2009b). We try to keep this summary objective to reflect the most important parts of the discussions. Thereafter, we give our view of which parts of the discussion that has not been settled yet and may have consequences on the hglm package.

Summary of discussions by Clayton, Kuha and Firth following Lee and Nelder (1996) and the Discussion section in Lee *et al.* (2007)

In the Discussions following Lee and Nelder (1996), it is claimed that: i) the h-likelihood is not a true likelihood, ii) if there is little information on each of the random effects we can expect the estimates to be biased, and iii) the h-likelihood could be interpreted in a Bayesian viewpoint as a posterior distribution with uniform priors so that we can only expect the estimates to be satisfactory if there is plenty of information on each of the random effects.

In Lee *et al.* (2007) the authors address these criticisms. Specifically they address the following three points:

i) The h-likelihood is not a valid likelihood.

ii) The h-likelihood is not invariant to non-linear transformation of the random effects.

iii) There are problems in the analysis of binary matched pairs data.

The authors claim that the h-likelihood is a valid likelihood. Unlike the classical likelihood, however, which contain the two objects: data (observed random variables) and unknown fixed parameters, the h-likelihood also contains a third object, unobserved random variables. Based on Bjornstad (1996) extended likelihood principle we should expect a joint likelihood, including both fixed and random parameters, to carry all the information about these parameters. (For an excellent review on the extended likelihood principle, see Pawitan (2001)). In Lee and Nelder (2005) the authors explained why they consider the h-likelihood to be a joint likelihood appropriate for statistical inference of HGLMs.

Concerning the h-likelihood and invariance, here it is important to note that the h-likelihood is not a general joint likelihood. In the h-likelihood it is imposed that a proper scale for the random effects is chosen, and consequently the problem of invariance may be an issue for a general joint likelihood but not for the h-likelihood.

One might expect that the h-likelihood has similar problems as the Penalized Quasi-Likelihood (PQL) method (Breslow and Clayton 1993) for analysis of binary matched pairs (i.e. the extreme case of having few observations for each of the random effects). The authors show

however that the estimation method they call HL(2) gives estimates very close to the marginal likelihood estimates obtained by numerical integration. The method HL(2) is a higher order approximation to the h-likelihood. Hence, there does not seem to be a major problem with the h-likelihood itself but rather there is an issue of choosing a computational method that produces correct h-likelihood results.

This is not the only paper where binary outcomes with little information on each of the random effects have been investigated for the h-likelihood. Noh *et al.* (2006) showed that the bias in variance component estimates were very small, compared to PQL, for binary outcomes in related individuals of small families. To obtain good estimates for the h-likelihood an appropriate computational method was developed (similar to the HL(2) method in Noh and Lee (2007)).

Summary of Firth's discussion in Lee and Nelder (2006)

i) The choice of the function for the random effects cannot be made on grounds of additivity, i.e. the transformation $v(u)$ that makes fixed and random effects additive. Two simple examples are given.

ii) The h-likelihood may lead to inconsistent estimators and it is not reasonable for Lee and Nelder to claim that the h-likelihood gives a general method for generalized linear models with random effects.

iii) The problem of bias should not be specific for binary outcomes but rather be a general problem for situations where there is little information on each of the random effects.

The authors reply that it should be noted that although the choice of link function for the random effects is difficult in the examples given by Firth, the adjusted profile h-likelihood still gives satisfactory estimates for the fixed parameters.

Discussion in papers of Statistical Science vol. 24 no. 3

Lee and Nelder (2009a) base their argumentation on the extended likelihood principle (Bjornstad 1996) to explain why the h-likelihood should be used for inference of models with unobservable random effects. The idea of the h-likelihood is that the marginal likelihood should be used for estimating fixed effects, the h-likelihood should be maximized to estimate random effects and the adjusted profile likelihood for the estimation of dispersion parameters. When the marginal likelihood is difficult to compute, the adjusted profile h-likelihood (with profiling over the random effects) is proposed as an alternative.

They criticize the use of estimation methods that are not included in a probabilistic framework, such as GEE and empirical Bayes methods. It is emphasized that the h-likelihood inference can be made without the necessity of inventing priors, as opposed to Bayesian methods.

By using the example of Bayarri, DeGroot, and Kadane (1988), they illustrate the importance of choosing the scale for the random effects so that the maximization of the h-likelihood gives the MLE for the fixed effects. In this example the outcome y follows an exponential distribution conditional on a random effect u, where u is also exponentially distributed. The

point of this example is that useless estimators are derived if the joint likelihood is maximized. However, Lee and Nelder (2009a) claim that the h-likelihood is defined for the specific scale $v = \log u$, which is shown to produce sensible estimates of the fixed parameter and the random effects. They admit that "the choice of the scale in defining the h-likelihood is important to guarantee the meaningfulness of the mode estimation". Furthermore: "The (weak) canonical scale in HGLMs leads to an invariance of a certain extended likelihood. However, in general the validity of such a scale has not been established."

Louis (2009) considers the paper by Lee and Nelder as a piece of "over promotion" and "more of an opinion-piece than a scientific comparison of approaches". He agrees that h-likelihood estimation can be valid and efficient under some settings but it is not globally valid.

Molenberghs *et al.* (2009) review different estimation methods for models with random effects and to them it is clear that no method can "claim uniform superiority over all others". They also emphasis the importance of assessing a suite of models to judge, through sensitivity analysis, how key inferences are vulnerable to model choice, which was neglected in Lee and Nelder (2009a).

Meng (2009) derives the necessary constraints for the scale of the random effects to be used in the h-likelihood. The similarity (and differences) to Bayesian analysis with uniform priors is pointed out, and concludes that the h-likelihood is just an approximation to the Bayesian posterior for the special case of having uniform priors.

In response, Lee and Nelder (2009b) note that both Louis and Meng say that the extended likelihood such as the h-likelihood does carry information about the unobservables but that the Bayesian approach is best suited for the inference. In contrast to the many possible priors in the Bayesian framework, the authors say that "in our system there is only one corresponding prior likelihood", i.e. a uniform prior. Furthermore, the adjusted profile h-likelihood is not merely a Laplace approximation to the marginal posterior distribution in a Bayesian framework, because "it can also eliminate fixed unknowns", and the authors argue that the adjusted profile h-likelihood is quite different from the Bayesian marginal posterior.

Summary of points not settled yet and possible consequences

The conclusion that can be drawn from the studies performed by Lee, Nelder and co-workers over the past 15 years is that we can expect the h-likelihood to give good estimates, but the computational procedure to get good approximations for the maximum h-likelihood estimates might be computationally demanding for data such as binary matched pairs. Uncertainty measures, such as standard errors, have been shown to be possible to derive from the curvature of the h-likelihood for many applications of HGLMs, but the general validity of using the h-likelihood for inference is still being argued by several authors.

In the **hglm** package, we have implemented the Extended Quasi-Likelihood (EQL) procedure, which is the one described in the book by Lee *et al.* (2006), where the estimates are obtained by fitting an interconnected set of GLMs in a very elegant way. The dispersion components are estimated by correcting for the leverages in the mean part of the model, which is a result of applying an adjusted profile likelihood function to the EQL.

The EQL method can give biased results, which we clearly state in the output of the package, but this is not a problem of the h-likelihood itself but rather that the EQL method may sometimes give estimates that are substantially different from the maximum h-likelihood estimates.

Table 4: Comparison of estimates for a binomial-beta model using the seed germination data presented in Lee and Nelder (1996).

	GenStat: HL(0,0)[a]	GenStat: EQL	hglm
Intercept	-0.543	-0.542	-0.542
Seed	0.080	0.077	0.075
Extract	1.337	1.339	1.339
Interaction	-0.822	-0.825	-0.826
Dispersion parameter for the random effects:			
$\log \alpha$	3.096	3.022	3.003

[a] Same estimates as in Lee and Nelder (1996)

Table 5: Comparison of estimates for a Poisson-gamma model using the pump failure data presented in Lee and Nelder (1996).

	GenStat: HL(0,0)[a]	GenStat: EQL	hglm
Intercept	-1.599	-1.590	-1.590
Group(2)	1.668	1.665	1.665
Dispersion parameter for the random effects:			
$\log \alpha$	-0.165	0.046	0.046

[a] Same estimates as in Lee and Nelder (1996)

The most important point from the above discussions is that a proper scale for the random effects is required by the h-likelihood to avoid problems of invariance. For many of the commonly used HGLM this is not an issue, but the user of the hglm package should be aware of this potential problem.

12. Comparison with estimates using GenStat

In the **hglm** package, interconnected GLMs are used to fit HGLMs. The algorithm was implemented using the `glm` function in the R **stats** package and by following Table 7.1 and Table 7.3 in Lee *et al.* (2006). An important part of the implementation was to define a separate GLM function for HGLMs with gamma distributed random effects (as explained in Table 7.1 and Example 6.3 in Lee *et al.* (2006)), and also to compute the deviance components for beta distributed random effects according to Table 7.1 in Lee *et al.* (2006). For a linear mixed model, this algorithm gives REML estimates, and for general HGLMs it gives EQL estimates. The standard errors are the ones produced by the interconnected GLM functions at convergence.

The estimates produced by the hglm function are very similar to the ones using the EQL method implemented in GenStat. For a binomial-beta model using the seed germination data presented in Lee and Nelder (1996) with fixed binomial dispersion parameter, the estimates of fixed effects (Table 4) differed by less than 2×10^{-3} to the EQL estimates in GenStat, and

Table 6: Comparison of estimates for a gamma-normal model using the semiconductor data presented in Lee *et al.* (2006).

	GenStat: HL(0,0)[a]	GenStat: EQL	hglm
Mean model			
Intercept	-4.711	-4.712	-4.711
x_1	0.209	0.210	0.209
x_3	0.328	0.329	0.329
x_5	-1.174	-0.173	-0.173
x_6	-0.357	-0.357	-0.357
Dispersion model			
Intercept	-2.610	-2.588	-2.591
x_2	-0.673	-0.686	-0.682
x_3	-0.492	-0.503	-0.499
Dispersion parameter for the random effects:			
$\log \lambda$	-3.014	-3.028	-3.004

[a] Same estimates as on page 218 of Lee *et al.* (2006)

the dispersion parameter differed by less than 1%.

For a Poisson-gamma model using the pump failure data presented in Lee and Nelder (1996), the estimates (Table 5) differed by less than 1×10^{-3}. By comparing to the HL(0,0) estimates, we can see that here the EQL method seems to give a poor approximation for the dispersion parameter.

For a gamma-normal model using the semiconductor data presented on page 218 of Lee *et al.* (2006), the estimates of fixed effects (Table 6) differed by less than 4×10^{-3} to the EQL estimates in GenStat, and the dispersion parameter differed by less than 1%.

References

Alam M, Carling K (2008). "Computationally feasible estimation of the covariance structure in generalized linear mixed models GLMM." *Journal of Statistical Computation and Simulation*, **78**, 1227–1237.

Bayarri M, DeGroot M, Kadane J (1988). "What is the likelihood function? (with discussion)." *In Ststistical Decision Theory and Related Topics IV. (eds. S. S. Gupta and J. O. Berger) Springer, New York.*, **1**.

Bjornstad J (1996). "On the generalization of the likelihood function and likelihood principle." *Journal of the American Statistical Association*, **91**, 791–806.

Breslow NE, Clayton DG (1993). "Approximate inference in generalized linear mixed models." *Journal of the American Statistical Association*, **88**, 9–25.

Carling K, Rönnegård L, Roszbach K (2004). "An analysis of portfolio credit risk when counterparties are interdependent within industries." *Sveriges Riksbank Working Paper*, **169**.

Coster A, Bastiaansen J, Calus M, Maliepaard C, Bink M (2009). "QTLMAS 2009: Simulated dataset." *Submitted*.

Crowder MJ (1978). "Beta-binomial anova for proportions." *Applied Statistics*, **27**, 34–37.

Gaver DP, O'Muircheartaigh IG (1987). "Robust Empirical Bayes Analyses of Event Rates." *Technometrics*, **29**(1), 1–15.

Ha ID, Lee Y (2005). "Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models." *Biometrika*, **92**, 717–723.

Henderson CR (1976). "A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values." *Biometrics*, **32**(1), 69–83.

Lee Y, Nelder JA (1996). "Hierarchical generalized linear models with Discussion." *J. R. Statist. Soc. B*, **58**, 619–678.

Lee Y, Nelder JA (2005). "Likelihood for random-effect models (with discussion)." *Statistical and Operational Research Transactions*, **29**, 141–182.

Lee Y, Nelder JA (2006). "Double hierarchical generalized linear models with discussion." *Applied Statistics*, **55**, 139–185.

Lee Y, Nelder JA (2009a). "Likelihood inference for models with unobservables: another view." *Statistical Science*, **24**, 255–269.

Lee Y, Nelder JA (2009b). "Rejoinder: Likelihood inference for models with unobservables: another view." *Statistical Science*, **24**, 294–302.

Lee Y, Nelder JA, Noh M (2007). "H-likelihood: problems and solutions." *Statistics and Computing*, **17**, 49–55.

Lee Y, Nelder JA, Pawitan Y (2006). *Generalized linear models with random effects: a unified analysis via h-likelihood*. Chapman & Hall/CRC.

Louis TA (2009). "Discussion of likelihood inference for models with unobservables: another view." *Statistical Science*, **24**, 270–272.

Lynch M, Walsh B (1998). *Genetics and analysis of Quantitative Traits*. Sinauer Associates, Inc. ISBN 087893481.

McCullagh P, Nelder JA (1989). *Generalized linear models*. Chapman & Hall/CRC.

Meng XL (2009). "Decoding the h-likelihood." *Statistical Science*, **24**, 280–293.

Molenberghs G, Kenward MG, Verbeke G (2009). "Discussion of likelihood inference for models with unobservables: another view." *Statistical Science*, **24**, 273–279.

Myers PH, Montgomery DC, Vining GG (2002). *Generalized linear models with applications in engineering and the sciences*. New York: John Wiley and Sons.

Noh M, Lee Y (2007). "REML estimation for binary data in GLMMs." *Journal of Multivariate Analysis*, **98**, 896–915.

Noh M, Yip B, Lee Y, Pawitan Y (2006). "Multicomponent variance estimation for binary traits in family-based studies." *Genetic Epidemiology*, **30**, 37–47.

Pawitan Y (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rönnegård L, Carlborg Ö (2007). "Separation of Base Allele and Sampling Term Effects Gives New Insights in Variance Component QTL Analysis." *BMC Genetics*, **8**(1).

Affiliation:

Lars Rönnegård
School of Technology and Business Studies
Dalarna University
Campus Framtidsdalen, Borlänge, Sweden
E-mail: lrn@du.se
URL: http://users.du.se/~lrn/DUweb/

Xia Shen
Department of Cell and Molecular Biology
Uppsala University
BMC Building, Uppsala, Sweden
and
School of Technology and Business Studies
Dalarna University
Campus Framtidsdalen, Borlänge, Sweden
E-mail: xia.shen@icm.uu.se
URL: http://www.19850911.com

Moudud Alam
School of Technology and Business Studies
Dalarna University
Campus Framtidsdalen, Borlänge, Sweden
E-mail: maa@du.se