The hglm Package

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. it also produces diagnostics such as deviances and leverages. 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.

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; McGullagh 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 models and random effects with known correlation structure.

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	

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.

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 McGullagh 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

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

and $\mathbf{\Sigma}_0^{-1} = \mathbf{W}_{ma}\{diag(\mathbf{\Phi}^{-1}, 0)\}$. 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 McGullagh 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

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

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)

```
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.

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

```
Call:
hglm.formula(fixed = y ~ week, random = ~1 | ID, data = bacteria,
family = binomial(link = logit))
```

```
DISPERSION MODEL
WARNING: h-likelihood estimates through EQL can be biased.
Model estimates for the dispersion term:[1] 0.761503

Model estimates for the dispersion term:
Link = log
Effects:
Estimate Std. Error
 -0.2725     0.1018

Dispersion = 1 is used in Gamma model on deviances to
calculate the standard error(s).
Dispersion parameter for the random effects
[1] 1.559

Dispersion model for the random effects:
Link = log
Effects:
Estimate Std. Error
  0.4440     0.2838

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)  2.27761    0.33280   6.844  7.7e-11 ***
week        -0.13343    0.04114  -3.243  0.00137 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Summary of the random effects estimate
          Estimate Std. Error
IDX01   0.7271     0.9701
IDX02  -0.2728     0.8311
...
IDZ26  -0.1250     0.8023

HL estimation converged in 9 iterations.
```

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 ver-

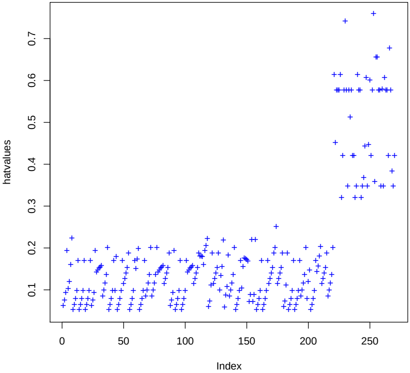

Figure 1: Leverages (i.e. diagonal elements of the augmented hat-matrix) for each observation 1 to 220 , and for each level in the random effect (index 221-282).

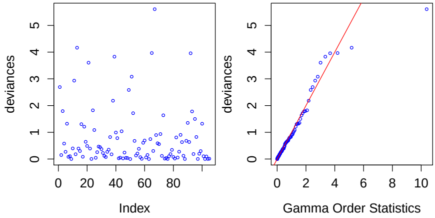

Figure 2: Deviance diagnostics for each observation and level in the random effect.

sion of the **hglm** package. Furthermore, in the current version of `hglm()`, we have not included the possibility to specify models with offsets. 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 $log(\phi_i) = X_{d,i}\beta_d$.

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 $\hat{\beta}_d = (0.354, 0.505)$.

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

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

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:
Estimate Std. Error
 -1.2100     0.7755

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|)
X.1 -0.004131   0.268339  -0.015    0.988

Summary of the random effects estimate
    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

HL 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)

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

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

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] 0.5254

Dispersion model for the random effects:
Link = log
Effects:
Estimate Std. Error
 -0.6436     0.7513

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|)
X.1 -0.07421    0.34566  -0.215     0.83

Summary of the random effects estimate
    Estimate Std. Error
Z.1  -0.7049     0.4208
Z.2   0.3631     0.3762
Z.3   0.8089     0.3655
Z.4  -0.7180     0.4216
Z.5   0.2509     0.3767

HL 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 $\log(4.208) = 1.437$.

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

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

```
Fixed effects:
(Intercept) factor(Gr)1
 0.08801297 -1.63909609
Random effects:
as.factor(System)1  as.factor(System)2  as.factor(System)3 ...
        0.25013564          0.05840759          0.37517720
...

Dispersion parameter for the mean model:[1] 0.001927418
Dispersion parameter for the random effects:[1] 4.208412

Estimation converged in 5 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 **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 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

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

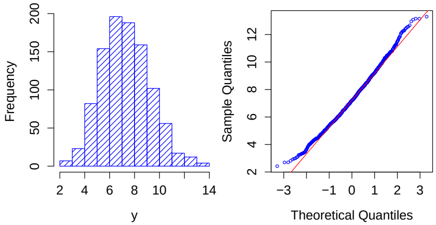

Figure 3: Histogram and qqplot for the analyzed trait.

are 0. Let **L** be the Cholesky factorization of **A**, and $\mathbf{Z} = \mathbf{Z}_0\mathbf{L}$. The design matrix for the fixed effects, **X**, is a column of ones. The estimated variance components are $\hat{\sigma}_e^2 = 2.21$ and $\hat{\sigma}_u^2 = 1.50$.

```
R> data(QTLMAS)
R> y <- QTLMAS[,1]
R> Z <- QTLMAS[,2:2026]
R> X <- matrix(1, 1000, 1)
R> test0 <- hglm(y = y, X = X)
R> test1 <- hglm(y = y, X = X,Z = Z)
R> print(test1)
```

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

Fixed effects:
     X.1
7.279766

Random effects:
          V1            V2            V3            V4
-1.191733707  1.648604776  1.319427376 -0.928258503
...

Variance of common error term:[1] 2.211169
Variance of the random effects:[1] 1.502516

Estimation converged in 2 iterations
```

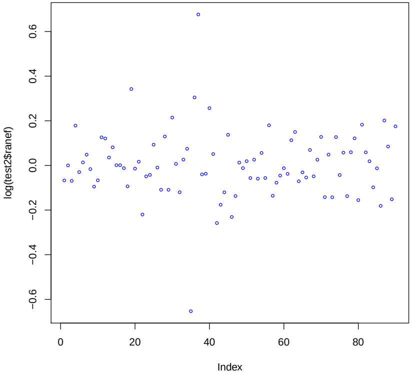

Figure 4: Estimated random effects for the 90 markers on chromosome 1 for trait values at 265 days. The main genetic effect was simulated close to marker number 40.

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 $\hat{a} = 1/(2\hat{\alpha}) = 0.023$. We conclude that the **hglm** package can replicate the results given by Lee and Nelder (1996).

```
R> data(seeds)
R> germ <- hglm(fixed = r/n ~ extract*I(seed=="O73"),
+               weights = n, data = seeds,
+               random = ~1|plate, family = binomial(),
+               rand.family = Beta(), conv = 0.005)
R> summary(germ)
```

```
Call:
hglm.formula(fixed = r/n ~ extract * I(seed == "O73"), random = ~1 |
plate, data = seeds, family = binomial(), rand.family = Beta(),
conv = 0.005, weights = n)

DISPERSION MODEL
WARNING: h-likelihood estimates through EQL can be biased.
Model estimates for the dispersion term:[1] 1.166024

Model estimates for the dispersion term:
Link = log
Effects:
Estimate Std. Error
0.1536     0.4558

Dispersion = 1 is used in Gamma model on deviances to
calculate the standard error(s).
Dispersion parameter for the random effects
[1] 0.02577

Dispersion model for the random effects:
Link = log
Effects:
Estimate Std. Error
-3.6585     0.5209

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)                -0.54202    0.20541  -2.639 0.017243 *
extractCucumber             1.34061    0.29110   4.605 0.000252 ***
I(seed == "O73")TRUE        0.07213    0.33057   0.218 0.829870
ex...:I(seed == "O73")TRUE -0.82877    0.46109  -1.797 0.090057 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Summary of the random effects estimate
                   Estimate Std. Error
as.factor(plate)1    0.4397     0.2685
as.factor(plate)2    0.5022     0.2486
...
as.factor(plate)21   0.4869     0.3179

HL estimation converged in 5 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

HL estimation converged in 2 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 al-

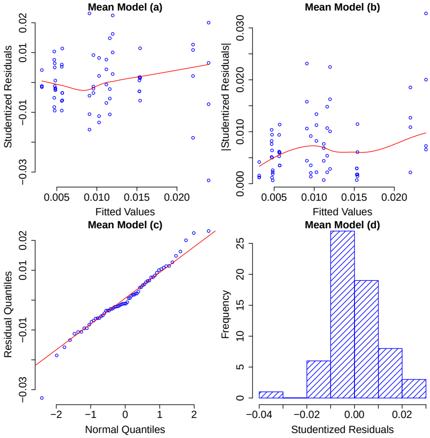

Figure 5: Residual plots of the mean model for the semiconductor data.

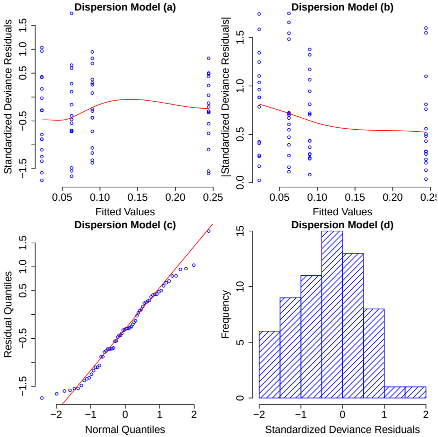

Figure 6: Residual plots of the dispersion model for the semiconductor data.

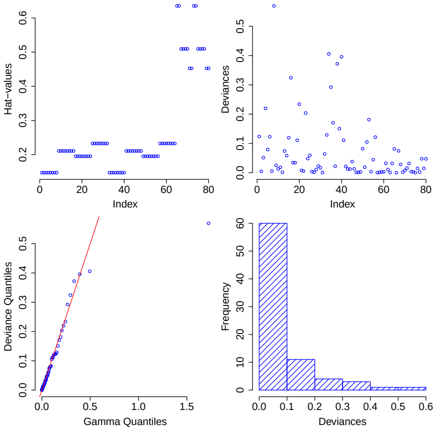

Figure 7: Diagnostics plots of the hat-values and deviances for the semiconductor data.

gorithms 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.

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.

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

Lee Y, Nelder JA, Pawitan Y (2006). *Generalized linear models with random effects*. Chapman & Hall/CRC.

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

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

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.

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
The Linnaeus Centre for Bioinformatics
Uppsala University
BMC Building, Uppsala, Sweden
and
School of Technology and Business Studies
Dalarna University
Campus Framtidsdalen, Borlänge, Sweden
E-mail: xia.shen@lcb.uu.se
URL: http://www.anst.uu.se/xiash103/

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