Generalised Regression Models

GRM: Generalized Linear Mixed Models (GLMM) **Semester 1, 2022–2023**

1 Introduction

We consider an extension of linear and generalized linear models to include **random effects**. Up to now we have considered all the random variation in the response **y** to result from the 'family' assumed. However, it is often useful to regard some of the unknown parameters in a linear model or generalized linear model as random, e.g. regard some of the unknown parameters as being chosen at random from a larger population. Models containing both **random** and **fixed** parameters are called **mixed models**. The 'lme4' package in R provides software for fitting these models.

Example Variation in the yield of a dyestuff
An experiment was carried out to investigate how much of the variation in yield in the manufacture of a dyestuff was due to the variation between batches in one of the raw materials. Five laboratory determinations of the yield were made for each of six randomly chosen batches of raw material, with the results given below.

	Batch					
	1	2	3	4	5	6
Yields	1545	1540	1595	1445	1595	1520
	1440	1555	1550	1440	1630	1455
	1440	1490	1605	1595	1515	1450
	1520	1560	1510	1465	1635	1480
	1580	1495	1560	1545	1625	1445

For this study, we may not be interested in the particular six batches used; they are merely representatives of a larger population about which we want to make inferences. We can model the effects of the batches as forming a random sample from some distribution, usually a Normal distribution.

If y_{jk} denotes the kth yield measurement for the jth batch then we might assume a model

$$y_{jk} = \mu + a_j + e_{jk} \quad (j = 1, \ldots, g; k = 1, \ldots, m_j), \tag{1}$$

with $g = 6$ and $m_j = 5$. Here μ and a_j denote respectively the expected value of the yield (over batches as well as determinations) and the effect of the jth batch. The a_j are called *random effects*, while μ is a *fixed effect*. [We follow a convention of using Roman and Greek letters for random and fixed effects respectively.] The random variables a_j and e_{jk} might be assumed uncorrelated with expectations zero and variances σ_a^2 and σ^2 respectively. To make inferences about the parameters μ, σ_a^2 and σ^2, we might also take the a_j and e_{jk} to be jointly Normally distributed.

Under this *one-way random-effects* model, we have (conditional on μ, but not on the a_j)

$$\text{var}\left(y_{jk}\right) = \sigma_a^2 + \sigma^2, \quad \text{cov}\left(y_{jk}, y_{jk'}\right) = \sigma_a^2 \quad (k \neq k'), \quad \text{var}\left(\overline{y}_{j\cdot}\right) = \sigma_a^2 + \sigma^2/m_j. \tag{2}$$

2 Which effects should be treated as random?

It is not always clear which of the effects in a statistical model should be treated as random: the decision may depend on the purpose of the analysis as well as the nature of the study. The published advice on when effects should be taken to be random is contradictory. To decide which effects to treat as random, it can be useful to imagine repeating the study: the 'fixed' effects are those whose levels would be the same in the new experiment, and the 'random' effects are those whose levels would be different from before. Thus factors that would be treated as fixed include sex, age groupings, disease types, medical treatments, and measurement times in a repeated-measures experiment. Factors that would usually be treated as random include experimental animals (including humans), years (for studies of annual crops) and batches of raw material used in an experiment. Some that could be treated as either fixed or random are crop varieties and expert assessors.

Some other examples of where random effects might arise are as follows.

1. In educational research, sources of variation in examination performance may include pupils, classes, schools and local authorities. Suitable models would be *hierarchical* or *multi-level* or have *nested* effects, and might include random effects at each level. Fixed effects might include factors relating to teaching methods or class organisation (such as age groups), and covariates (such as pupils' birth dates) could be incorporated at the appropriate level.

2. Similarly, sample surveys may be organised by choosing a sample of local authorities, then parts of the local authority areas, such as post-code sectors, and then individual households. The inferences from such surveys should take account of the possible sources of variation.

3. If measurements are made on related animals, such as litter-mates, we expect that they will be positively correlated within the groups because of the effects of shared genes and shared maternal environment. Conversely, competition for food within litters may lead to a negative correlation between body weights: assuming additive litter effects is then not appropriate.

4. When blocking is used in an experimental design, it may be reasonable to take block effects as random. With incomplete blocks, estimates of treatment differences can be based on the totals over the blocks as well as on differences within blocks. The best combination of these *interblock* and *intrablock* estimates depends on the ratio of the residual and block variances.

3 Linear component of Mixed Models

Linear Models: The usual fixed-effects linear model for an n-vector $\mathbf{y}$ of responses is $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ with $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$ given the design matrix $\mathbf{X}$. This model can be generalized in various ways to a *mixed* model, that is one containing both fixed and random effects. One of the simpler ways is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1 \mathbf{u}_1 + \ldots + \mathbf{Z}_q \mathbf{u}_q + \mathbf{e}, \tag{3}$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_q$ are vectors of random effects with corresponding design matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_q$. Note that the residual vector $\mathbf{e}$ can be included as a vector of random effects (with design matrix $\mathbf{I}_n$); other $\mathbf{u}$-vectors might represent main effects and interaction effects. If $\mathbf{Z}_s$ is $n \times p_s$ ($s =$

$1,\ldots,q$) then $\mathbf{u}_1,\ldots,\mathbf{u}_q$ are assumed to be uncorrelated with each other and with $\mathbf{e}$, and to have zero expectations and variance matrices $\sigma_1^2\mathbf{I}_{p_1},\ldots,\sigma_q^2\mathbf{I}_{p_q}$.

Under (3), the response vector $\mathbf{y}$ has expectation $\mathbf{X}\beta$, but its variance matrix becomes, instead of $\sigma^2\mathbf{I}_n$,

$$\Sigma = \mathrm{var}(\mathbf{y}\,|\,\mathbf{X}) = \sigma_1^2\mathbf{Z}_1\mathbf{Z}_1^T + \ldots + \sigma_q^2\mathbf{Z}_q\mathbf{Z}_q^T + \sigma^2\mathbf{I}_n\,. \tag{4}$$

Model (3) includes the model assumed in the example considered above, as well as more complex factorial models. Also included are regression models in which the slopes and intercepts are random: including an interaction between a random factor and a covariate is interpreted as allowing the regression coefficients to vary between levels of the factor. Such models are used with time as a covariate to analyse repeated-measures data.

Unless the data are balanced, the estimate of a fixed effect depends on whether other effects are treated as fixed or random.

Generalized Linear Models: In the GLM context we have a **linear component**

$$\eta = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{u}_1 + \ldots + \mathbf{Z}_q\mathbf{u}_q\,, \tag{5}$$

with a link function $g(\mu) = \eta$, where the response Y has a distribution that is a member of the exponential family.

4 ML and REML estimation

Under the assumption of Normality, the maximum likelihood (ML) estimates of β, $\sigma_1^2,\ldots,\sigma_q^2$ and σ^2 in (3) maximize the log-likelihood, which is

$$-\frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\Sigma^{-1}(\mathbf{y}-\mathbf{X}\beta)\,. \tag{6}$$

For given σ, the ML estimate $\widehat{\beta}$ of β satisfies the (weighted least-squares) equation

$$\mathbf{X}^T\Sigma^{-1}\mathbf{X}\widehat{\beta} = \mathbf{X}^T\Sigma^{-1}\mathbf{y}\,,$$

but maximizing (6) with respect to $\sigma_1^2,\ldots,\sigma_q^2$ and σ^2 usually requires iterative estimation (including checking for negative variance estimates).

Approximate tests of hypotheses about the *fixed* effects in (3) can be based on differences in the maximized log-likelihood (i.e. on the χ^2 approximation for differences between deviances).

For REML estimation, the likelihood is based not on $\mathbf{y}$ but on a vector of linear functions of $\mathbf{y}$ chosen to have zero expectations under the model. If $\mathbf{X}$ has rank r then an $n\times(n-r)$ matrix $\mathbf{K}$ of full rank can be found satisfying $\mathbf{K}^T\mathbf{X} = \mathbf{O}$. The $(n-r)$-vector $\mathbf{K}^T\mathbf{y}$ then has distribution $N_{n-r}(\mathbf{0},\mathbf{K}^T\Sigma K)$, which does not depend on β. The REML estimates maximize the corresponding likelihood, but do not depend on the particular choice of $\mathbf{K}$.

Under REML, tests of hypotheses about *fixed* effects cannot use the maximized log-likelihood directly because omitting elements of β changes $\mathbf{X}$ and hence changes the matrix $\mathbf{K}$ used to construct the REML likelihood. For the same reason, an AIC criterion based on the REML likelihood cannot be used to compare models having different fixed effects. However, inferences about individual coefficients in a model can be based on the approximate variance matrix of the coefficients: individual estimates divided by their estimated standard errors are compared with $N(0,1)$.

5 The R functions `lmer` and `glmer`

The function `lmer` can be used to estimate fixed effects (including regression coefficients) and variance components in mixed linear models with the form defined in (3). The method of estimation used is REML (by default) or ML (using the option REML=FALSE). Similarly, the function `glmer` can be used to estimate fixed effects (including regression coefficients) in GLMMs with the form defined in (5). These function are provided in the `lme4` package, described at `http://lme4.r-forge.r-project.org/`[1].

The model formulae used in `lmer` include fixed effects as in `lm` or `aov`. A simple random-effect term corresponding to a factor `randfactor`, say, is denoted by `(1 | randfactor)` in an `lmer` formula.

Optional arguments for `lmer` include `data`, `subset` and `na.action`. The result of using `lmer` is an object of class `mer`: methods for `mer` objects include `summary`, `fitted`, `resid`, `coef` and `anova`. Note that the statistics such as `AIC` displayed when `anova` is used are based on ML estimates, even if REML has been used for estimation: see the final paragraph of Section 4.

Example Variation in the yield of a dyestuff (continued)
The file `dyestuff.txt` contains columns with `yield` and `batch`. To fit model (1) we could use

```
dyestuff.mer <- lmer(yield ~ 1 + (1 | as.factor(batch)))
```

The output from `summary(dyestuff.mer)` includes estimates of the variance components σ^2 and σ_a^2, and of the fixed effect, the overall expected yield μ, as follows.

```
Linear mixed model fit by REML
Formula: yield ~ 1 + (1 | as.factor(batch))

   AIC   BIC logLik deviance REMLdev
 325.7 329.9 -159.8    327.4   319.7

Random effects:
 Groups           Name        Variance Std.Dev.
 as.factor(batch) (Intercept) 1764.0   42.001
 Residual                     2451.3   49.510
Number of obs: 30, groups: as.factor(batch), 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1527.50      19.38   78.81
```

[1] R TIP: you can use `install.packages('lme4')` to install the 'lme4' package if not already installed, and then `library(lme4)` before using the functions `lmer`/`glmer`.

Example Contagious bovine pleuropneumonia (CBPP)

Contagious bovine pleuropneumonia (CBPP) is a major disease of cattle in Africa, caused by a mycoplasma. This dataset describes the serological incidence of CBPP in zebu cattle during a follow-up survey implemented in 15 commercial herds located in the Boji district of Ethiopia. The goal of the survey was to study the within-herd spread of CBPP in newly infected herds. Blood samples were quarterly collected from all animals of these herds to determine their CBPP status. These data were used to compute the serological incidence of CBPP (new cases occurring during a given time period). Some data are missing (lost to follow-up).

Below is an analysis with `glmer`.

```
library(lme4)
## response as a matrix
(m1 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
family = binomial, data = cbpp))
## response as a vector of probabilities and usage of argument "weights"
m1p <- glmer(incidence / size ~ period + (1 | herd), weights = size,
family = binomial, data = cbpp)
## Confirm that these are equivalent:
stopifnot(all.equal(fixef(m1), fixef(m1p), tolerance = 1e-5),
all.equal(ranef(m1), ranef(m1p), tolerance = 1e-5))checkConv
13
## GLMM with individual-level variability (accounting for overdispersion)
cbpp$obs <- 1:nrow(cbpp)
(m2 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd) + (1|obs),
family = binomial, data = cbpp))
```

Results of analysis

Output from the `glmer` analysis is given below.

```
> library(lme4)
Loading required package: Matrix

> (m1 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd),
+ family = binomial, data = cbpp))

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]

 Family: binomial  ( logit )

Formula: cbind(incidence, size - incidence) ~ period + (1 | herd)
   Data: cbpp
     AIC      BIC   logLik deviance df.resid
194.0531 204.1799 -92.0266 184.0531       51

Random effects:
 Groups Name        Std.Dev.
 herd   (Intercept) 0.6421
Number of obs: 56, groups:  herd, 15

Fixed Effects:
(Intercept)      period2      period3      period4
    -1.3983      -0.9919      -1.1282      -1.5797
```

```
> m1p <- glmer(incidence / size ~ period + (1 | herd), weights = size,
+ family = binomial, data = cbpp)

> stopifnot(all.equal(fixef(m1), fixef(m1p), tolerance = 1e-5),
+ all.equal(ranef(m1), ranef(m1p), tolerance = 1e-5))checkConv

> 13
[1] 13

> cbpp$obs <- 1:nrow(cbpp)
> (m2 <- glmer(cbind(incidence, size - incidence) ~ period + (1 | herd) +
                  (1|obs+ family = binomial, data = cbpp))

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]

 Family: binomial  ( logit )

Formula: cbind(incidence, size - incidence) ~ period + (1 | herd) + (1 |
    obs)
   Data: cbpp

     AIC      BIC   logLik deviance df.resid
186.6383 198.7904 -87.3192 174.6383       50

Random effects:
 Groups Name        Std.Dev.
 obs    (Intercept) 0.8911
 herd   (Intercept) 0.1840

Number of obs: 56, groups:  obs, 56; herd, 15
Fixed Effects:
(Intercept)      period2      period3      period4
     -1.500       -1.226       -1.329       -1.866
```