

Chapter 7. Random Effects Models

MAST90139 Statistical Modelling for Data Science Slides

Guoqi Qian

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

- 1 §7.1 Linear Random Effects Models for Normal Data
 - §7.1.1 Two-stage random effects models
 - §7.1.2 Statistical Inference for Random Effects Models
 - §7.1.3 Fitting random-effects or mixed-effects model in R `lme4`
- 2 §7.2 Random Effects in GLM
- 3 §7.3. Estimation Based on Posterior Modes
 - §7.3.1 Known variance-covariance components
 - §7.3.2 Unknown variance-covariance components
- 4 §7.4. Estimation by Integration Techniques
- 5 §7.6. Bayesian Mixed Models
- 6 §7.7. Marginal estimation approach to random effects models

Chapter 7 Outline (1)

- This chapter is concerned with random effects for analysing both normal and non-normal response data that are assumed to be clustered or correlated.
- In each case the data consist of repeated observations $(y_{it}, \mathbf{x}_{it})$, $t = 1, \dots, T_i$, for each individual or unit $i = 1, \dots, n$, where y_{it} is the response variable and \mathbf{x}_{it} a covariate vector.
- The T_i data points for each unit i comprise a cluster.
- Responses in different clusters are usually assumed to be independent of each other.

Chapter 7 Outline (2)

- Most of the models considered so far specify the effects of covariates on the response to be constant across clusters (hence they are population-averaged effects). This may become inappropriate if one is interested in the effects on individual responses, because they may exist unobservable heterogeneity in the effects across clusters; namely, the covariate effects may be cluster- or subject-specific.
- To account for this heterogeneity, one approach is to treat some or all covariate effects to be random.
- This gives rise to the **random effects** or **mixed effects** models.

Linear random effects models for normal data

For clustered Gaussian data, linear random effects models provide an efficient tool for analysing cluster-specific intercepts and/or covariate effects.

Two-stage random effects models (1)

- **At the first stage**, the normal response y_{it} are assumed to depend linearly on unknown population-specific effects β and on unknown cluster-specific effects \mathbf{b}_i

$$y_{it} = \mathbf{z}_{it}^T \beta + \mathbf{w}_{it}^T \mathbf{b}_i + \varepsilon_{it} \quad (1)$$

where \mathbf{z}_{it} and \mathbf{w}_{it} represent design vectors, often $\mathbf{w}_{it} \subset \mathbf{z}_{it}$, and ε_{it} 's are uncorrelated normal random variables with $E(\varepsilon_{it}) = 0$ and $\text{Var}(\varepsilon_{it}) = \sigma_\varepsilon^2$.

- The design vector \mathbf{z}_{it} and thus \mathbf{w}_{it} may depend on deterministic or stochastic covariates and on past responses:

$$\mathbf{z}_{it} = \mathbf{z}_{it}(\mathbf{x}_{it}, y_{i,t-1}^*), \quad \text{with } y_{i,t-1}^* = (y_{i,t-1}, \dots, y_{i1}).$$

- Furthermore, it is assumed that \mathbf{z}_{it} contains the intercept term “1”.

Two-stage random effects models (2)

- **At the second stage** the effects \mathbf{b}_i are assumed to vary independently from one cluster to another according to a normal mixing distribution with mean $E(\mathbf{b}_i) = \mathbf{0}$ and covariance matrix $\text{Cov}(\mathbf{b}_i) = Q$. Namely,

$$\mathbf{b}_i \sim \text{MVN}(\mathbf{0}, Q), \quad Q > 0 \text{ positive definite.} \quad (2)$$

We also assume the sequences $\{\varepsilon_{it}\}$ and $\{\mathbf{b}_i\}$ are mutually uncorrelated.

- In matrix notation, model (1) takes the form

$$\mathbf{y}_i = Z_i\beta + W_i\mathbf{b}_i + \varepsilon_i \quad (3)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})^T$, $Z_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT_i})^T$, $W_i = (\mathbf{w}_{i1}, \dots, \mathbf{w}_{iT_i})^T$, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT_i})^T$, and $\varepsilon_i \sim \text{MVN}(0, \sigma_\varepsilon^2 I)$.

Two-stage random effects models (3)

- Model (3) can be rewritten as a multivariate heteroscedastic linear regression model

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^* \quad (4)$$

where the distribution vector $\boldsymbol{\varepsilon}_i^* = (\varepsilon_{i1}^*, \dots, \varepsilon_{iT_i}^*)^T$ with components $\varepsilon_{it}^* = \mathbf{w}_{it}^T \mathbf{b}_i + \varepsilon_{it}$ or $\varepsilon_i^* = \mathbf{W}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i$, are independent and normally distributed:

$$\boldsymbol{\varepsilon}_i^* \sim \text{MVN}(\mathbf{0}, V_i), \quad \text{with } V_i = \sigma_\varepsilon^2 I + \mathbf{W}_i \mathbf{Q} \mathbf{W}_i^T \quad (5)$$

- Equation (4), together with the covariance structure (5), represents the **marginal version** of the linear random effects model, where the conditioning on \mathbf{b}_i is dropped.
- From (5) we see responses y_{it} in the same cluster are correlated.

Some special linear random effects models

In the following we introduce several special linear random effects models:

- 1 Varying intercepts
- 2 (partially) varying slopes or varying covariate effects
- 3 cluster-specific effects being hierarchically nested

Random intercept models (1)

- In many situations, covariates that have cluster-specific effects on the response variable may have not been collected due to various reasons. To account for the cluster-specific effects of such covariates on the response, a **linear model with random intercepts** τ_i may be appropriate:

$$y_{it} = \tau_i + \mathbf{x}_{it}^T \boldsymbol{\gamma} + \varepsilon_{it} \quad (6)$$

where the slope coefficient $\boldsymbol{\gamma}$ is a fixed vector constant and intercepts τ_i 's are assumed to be i.i.d. with $E(\tau_i) = \tau$ and $\text{Var}(\tau_i) = \sigma^2$.

- The unobserved deviation $\tau_i - \tau$ may be interpreted as the cluster-specific effects of the individuals in the sample.
- Assuming normality gives a linear random effects model of the form (1) and (2) with

$$\boldsymbol{\beta}^T = (\tau, \boldsymbol{\gamma}^T), \quad \mathbf{z}_{it}^T = (1, \mathbf{x}_{it}^T), \quad \mathbf{w}_{it} = 1, \quad \mathbf{b}_i = (\tau_i - \tau) \sim N(0, \sigma^2).$$

Random intercept models (2)

- The **random intercept model** also takes into account intra-cluster correlation of the Gaussian outcomes. From (5)

$$\text{corr}(y_{it}, y_{is}) = \rho_{ts} = \rho = \frac{\sigma^2}{\sigma_{\varepsilon}^2 + \sigma^2}, \quad t \neq s.$$

- Random intercept models are also called **error components** or **variance components models**.
- The primary objective is to analyse the variance components σ_{ε}^2 and σ^2 , which stand for variability *within* and *between*, respectively, the clusters.

Random slope models

- Varying slope coefficients arise in particular in longitudinal studies, where intercept and slope coefficients are specific to each time series (which can be regarded as a cluster). The corresponding random effects model has the form

$$y_{it} = \tau_i + \mathbf{x}_{it}^T \boldsymbol{\gamma}_i + \varepsilon_{it}$$

where $\boldsymbol{\beta}_i^T = (\tau_i, \boldsymbol{\gamma}_i^T)$ vary independently across clusters according to a normal density with $E(\boldsymbol{\beta}_i) = \boldsymbol{\beta}$ and $\text{Cov}(\boldsymbol{\beta}_i) = \mathbf{Q} > 0$, i.e.

$$\boldsymbol{\beta}_i \sim \text{MVN}(\boldsymbol{\beta}, \mathbf{Q}).$$

- The mean $\boldsymbol{\beta}$ is the population-averaged effect of $\mathbf{z}_{it} = (1, \mathbf{x}_{it}^T)^T$. The variance matrix \mathbf{Q} gives $\boldsymbol{\beta}_i$'s heterogeneity and correlation.
- Rewrite $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{b}_i$. Then $\mathbf{b}_i \sim \text{MVN}(\mathbf{0}, \mathbf{Q})$. The model becomes a special case of (1) with $\mathbf{z}_{it}^T = \mathbf{w}_{it}^T = (1, \mathbf{x}_{it}^T)$.
- Random slope model is also called **random coefficient regression model**.

Multilevel models

- The linear random effects model just introduced are based on a single level cluster structure. In some applications, clustering occurs on more than one levels and the clusters are hierarchically nested.
- For example, in studying students' performance in a subject, each student's two test marks at the beginning and end of the semester are used as the response variable observations. Each student is regarded as a cluster consisting of the two test marks. Since there are several tutors in this subject and a student's test mark may be related to who his/her tutor is, it sounds reasonable to determined the clusters according to whose tutorial class a student sits in and which Id number he/she has. Here, the ID levels are nested within the tutor levels.
- The linear random effects model (1) and (2) can be used to account for multilevel cluster-specific effects.

§7.1.2 Statistical Inference for Random Effects Models (1)

- The main objective is to estimate β , σ_ε^2 and Q , and the random effects \mathbf{b}_i .
- Estimation is often based on a frequentist approach, where β , σ_ε^2 and Q are treated as **fixed** parameters.
- Alternatively, Bayesian methods can be applied, where β , σ_ε^2 and Q are treated as **random** parameters with some prior distribution.
- In the following we start with the frequentist approach including **empirical Bayes** methods.

§7.1.2 Statistical Inference for Random Effects Models (2)

Situation 1. Known variance-covariance components

- Suppose σ_ε^2 and Q are known. Then estimating β is usually based on the marginal model defined by (4) and (5). It is easy to show that the MLE of β is equal to the weighted least squares solution

$$\hat{\beta} = \left(\sum_{i=1}^n Z_i^T V_i^{-1} Z_i \right)^{-1} \sum_{i=1}^n Z_i^T V_i^{-1} \mathbf{y}_i. \quad (7)$$

The MLE $\hat{\beta}$ can also be shown to be BLUE.

- Estimation of the unobserved random effects \mathbf{b}_i is based on the posterior density of \mathbf{b}_i , given the data $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n)$.
- Due to the normality and linearity assumptions, the posterior pdf of \mathbf{b}_i is also normal and does not depend on ε_{it} . It can be shown that the optimal point estimator of \mathbf{b}_i is

$$\hat{\mathbf{b}}_i = E(\mathbf{b}_i | \mathbf{y}_i) = QW_i^T V_i^{-1} (\mathbf{y}_i - Z_i \beta). \quad (8)$$

$\hat{\mathbf{b}}_i$ is BLUE, and also maximises the log-posterior density.

§7.1.2 Statistical Inference for Random Effects Models (3)

Situation 2. unknown variance-covariance components

- Let $\theta = (\sigma_\varepsilon^2, Q)$. If θ is unknown, it is estimated together with β .
- Two approaches are available: **MLE & RMLE**.
- **Maximum likelihood estimation (MLE)**
 - This is based on maximising the marginal log-likelihood of the marginal model (4) and (5). This is equivalent to the minimisation of

$$\ell(\beta, \theta) = \sum_{i=1}^n [\log |V_i| + (\mathbf{y}_i - Z_i\beta)^T V_i^{-1} (\mathbf{y}_i - Z_i\beta)].$$

- The criticism of the MLE for θ is that this estimator does not take into account the loss in degrees of freedom resulting from the estimation of β . The larger the dimension of β is, the larger is the bias of the MLE for θ .

§7.1.2 Statistical Inference for Random Effects Models (4)

Restricted maximum likelihood estimation (RMLE)

- RMLEs generally yield a smaller bias for θ .
- The idea of an RMLE is to construct a likelihood that depends only on θ but not β . Such a likelihood can be derived using a Bayesian formulation of linear random effects model, where β are considered random variables having a vague or totally flat prior distribution, e.g. $\beta \sim \text{MVN}(\beta^*, \Gamma)$ with $\Gamma \rightarrow \infty$ or $\Gamma^{-1} \rightarrow 0$, so that the prior distribution of β is just a constant.
- Then the marginal likelihood of θ can be obtained by integrating the joint likelihood of (β, θ) over β . Maximising the marginal likelihood would give the RMLE of θ .
- However, the numerical calculation of the RMLE for θ is very complicated, where Newton-Raphson or Fisher scoring algorithms, and also EM algorithm are usually used.

§7.1.2 Statistical Inference for Random Effects Models (5)

EM algorithm for estimating $\delta = (\beta^T, \mathbf{b}_1^T, \dots, \mathbf{b}_n^T)^T$ and $\theta = (\sigma_\varepsilon^2, Q)$:

- 1 Choose starting values $\theta^{(0)} = (\sigma_\varepsilon^{2(0)}, Q^{(0)})$. For $p = 0, 1, 2, \dots$,
- 2 Compute $\hat{\delta}^{(p)} = (\hat{\beta}^{(p)T}, \hat{\mathbf{b}}_1^{(p)T}, \dots, \hat{\mathbf{b}}_n^{(p)T})^T$ from (7) and (8) using current $\hat{\theta}^{(p)} = (\sigma_\varepsilon^{2(p)}, Q^{(p)})$. Also compute the current residuals $\mathbf{e}_i^{(p)} = \mathbf{y}_i - Z_i \hat{\beta}^{(p)} - W_i \hat{\mathbf{b}}_i^{(p)}$, $i = 1, \dots, n$ and compute the posterior covariance matrices $\text{Cov}(\mathbf{b}_i | \mathbf{y}_i, \hat{\theta}^{(p)})$ and $\text{Cov}(\varepsilon_i | \mathbf{y}_i, \hat{\theta}^{(p)})$.
- 3 **EM-step** Compute $\hat{\theta}^{(p+1)} = (\sigma_\varepsilon^{2(p+1)}, Q^{(p+1)})$ by

$$\sigma_\varepsilon^{2(p+1)} = \frac{1}{T_1 + \dots + T_n} \sum_{i=1}^n \left[\mathbf{e}_i^{(p)T} \mathbf{e}_i^{(p)} + \text{tr} \left(\text{Cov}(\varepsilon_i | \mathbf{y}_i, \hat{\theta}^{(p)}) \right) \right] \quad (9)$$

$$Q^{(p+1)} = \frac{1}{n} \sum_{i=1}^n \left[\hat{\mathbf{b}}_i^{(p)} \hat{\mathbf{b}}_i^{(p)T} + \text{Cov}(\mathbf{b}_i | \mathbf{y}_i, \hat{\theta}^{(p)}) \right] \quad (10)$$

The posterior covariance matrices can be obtained from the joint normal density defined by the model.

- 4 RMLEs may have larger MSE than MLEs.

§7.1.3 Fitting random-/mixed-effects model in R lme4

- The lme4 package contains functions for fitting linear mixed models, generalized linear mixed models and nonlinear mixed models.
- The lmer() function in lme4 is used to fit linear and generalized linear models. **Basic syntax:**

```
lmer(formula, data, family=NULL, REML=TRUE)
```

- formula: Symbolic description of the model to be fitted
- data: Optional dataframe
- family: Description of the error distribution and link function, if NULL a linear mixed model is fitted
- REML: Logical, if TRUE estimate using REML (provides a consistent estimate of the variance components); if FALSE estimate using ML
- lmer() returns a mer object; see ?"mer-class" for an explanation of the slots of mer

§7.1.3 Formula `lmer()`

- A random-effects term in `lmer()` is specified by a linear model term and a grouping factor separated by '|'; i.e. a random effect is a linear model term conditional on the level of the grouping factor.
- The entire random-effects expression should be enclosed in parentheses since the precedence of '|' as an operator is lower than most other operators used in linear model formulas.
- For example
 - **Random intercept,**
`lmer(Reaction ~ Days + (1 | Subject), data=sleepstudy)`
 - **Random intercept and slope,**
`lmer(Reaction ~ Days + (Days | Subject), data=sleepstudy)`
- See the vignettes for how to fit nested random effects,
`vignette(package="lme4")`

§7.1.3 R functions for inference after `lmer()`

- Functions used for inference and prediction,

<code>summary()</code>	Summarize model results
<code>anova()</code>	Sequential tests of fixed effects and model comparison
<code>VarCorr()</code>	Extract variance components
<code>ranef()</code>	Predict random effects
<code>residuals()</code>	Extract residuals
- See `?"mer-class"` for a complete list of available methods

§7.1.3 Example: A sleep deprivation study (1)

- **Description:** The average reaction time per day for subjects in a sleep deprivation study. On day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night. The observations represent the average reaction time on a series of tests given each day to each subject.
- **Format:** A data frame with 180 observations on the following 3 variables.
 - **Reaction:** Average reaction time (ms)
 - **Days:** Number of days of sleep deprivation
 - **Subject:** Subject number on which the observation was made.
- The **aim** is to see how average reaction time is affected by the number of days of sleep deprivation, and whether the effect is cluster-specific.

§7.1.3 Example: A sleep deprivation study (2)

```
library(lme4)
data(sleepstudy); head(sleepstudy)

  Reaction Days Subject
1 249.5600    0    308
2 258.7047    1    308
3 250.8006    2    308
4 321.4398    3    308
5 356.8519    4    308
6 414.6901    5    308

str(sleepstudy)

'data.frame':   180 obs. of  3 variables:
 $ Reaction: num  250 259 251 321 357 ...
 $ Days    : num   0  1  2  3  4  5  6  7  8  9 ...
 $ Subject : Factor w/ 18 levels "308","309","310",...: 1 1 1 1 1 1 1 1 1 1 ...

require(lattice)
xyplot(Reaction ~ Days | Subject, data=sleepstudy,
       type = c("g", "p", "r"),
       xlab = "Days of sleep deprivation",
       ylab = "Average reaction time (ms)", aspect = "xy")

# Trellis plot of data
# Clear that there are different intercepts and slopes for each subject
# type = c("g", "p", "r"), plots grid lines, points, and regression line
```

§7.1.3 Example: A sleep deprivation study (3)

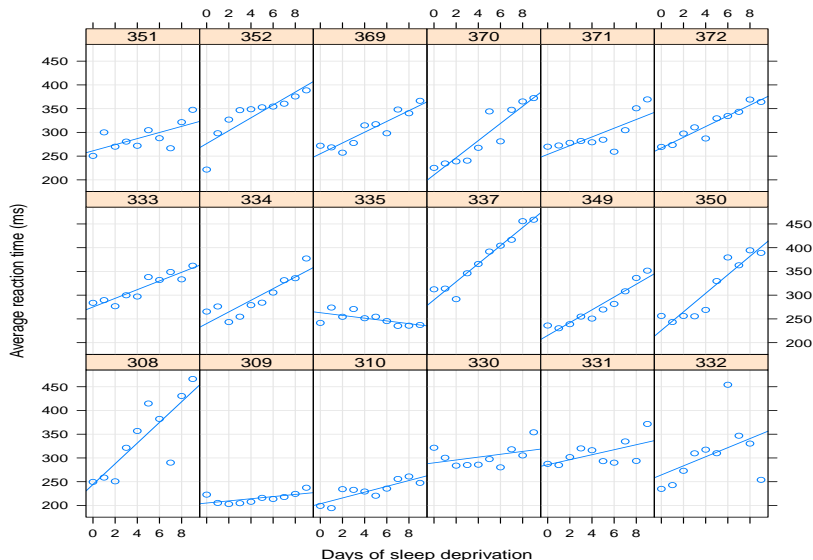


Figure 1: There are different intercepts and slopes for each subject

§7.1.3 Example: A sleep deprivation study (4)

```
#Random intercept model
fit1 <- lmer(Reaction ~ Days + (1 | Subject), data=sleepstudy); summary(fit1)
```

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 | Subject); REML criterion at convergence: 1786.5

Scaled residuals:

	Min	1Q	Median	3Q	Max
	-3.2257	-0.5529	0.0109	0.5188	4.2506

Random effects:

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	1378.2	37.12
Residual		960.5	30.99

Number of obs: 180, groups: Subject, 18

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	251.4051	9.7467	25.79
Days	10.4673	0.8042	13.02

Correlation of Fixed Effects:

(Intr)	
Days	-0.371

§7.1.3 Example: A sleep deprivation study (5)

The fitted random intercept model is

$$\hat{y}_{it} = b_{0i} + 251.4051 + 10.4673 \cdot \text{Days} + \varepsilon_{it}$$

with $t = \text{Days} = 0, 1, \dots, 9$ and $i = 1, \dots, 18$, where

$$b_{0i} \stackrel{iid}{\sim} N(0, 1378.2) \quad \text{and independent of} \quad \varepsilon_{it} \stackrel{iid}{\sim} N(0, 960.5)$$

```
VarCorr(fit1)
```

Groups	Name	Std.Dev.
Subject	(Intercept)	37.124
Residual		30.991

```
# Model Diagnostics
```

```
y.hat <- fitted(fit1) # Fitted values
```

```
int.hat <- ranef(fit1)[[1]][[1]] # Predicted intercepts
```

```
res.hat <- residuals(fit1) # Estimated residuals
```

```
qqnorm(int.hat, main="Random Intercepts"); qqline(int.hat)
```

```
qqnorm(res.hat, main="Residuals"); qqline(res.hat)
```

```
plot(y.hat, res.hat, xlab="Fitted Values", ylab="Residuals")
```

```
abline(h=0, lty=2)
```

§7.1.3 Example: A sleep deprivation study (6)

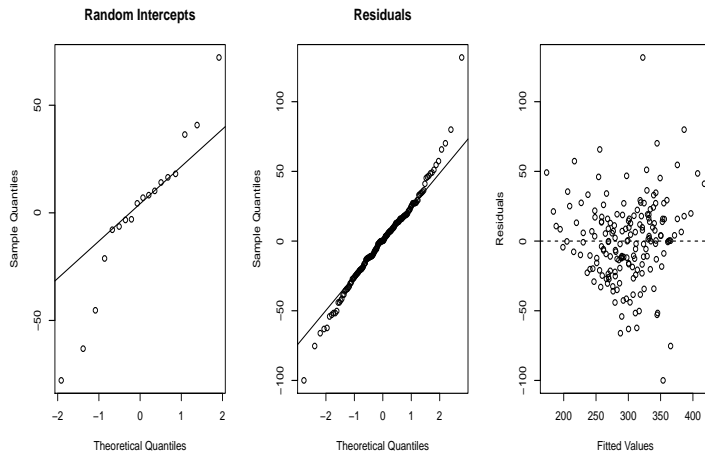


Figure 2: Model diagnostics plots for model `fit1`

§7.1.3 Example: A sleep deprivation study (7)

```
# Random intercept and slope model
fit2 <- lmer(Reaction ~ Days + (Days | Subject), data=sleepstudy); summary(fit2)
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject); REML criterion at convergence: 1743.6
```

Scaled residuals:

	Min	1Q	Median	3Q	Max
	-3.9536	-0.4634	0.0231	0.4633	5.1793

Random effects:

Groups	Name	Variance	Std.Dev.	Corr	#Cov=0.07*24.737*5.923=10.256
Subject	(Intercept)	611.90	24.737		
	Days	35.08	5.923	0.07	
Residual		654.94	25.592		

Number of obs: 180, groups: Subject, 18

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	251.405	6.824	36.843
Days	10.467	1.546	6.771

Correlation of Fixed Effects:
(Intr)

Days -0.138

§7.1.3 Example: A sleep deprivation study (8)

The fitted random intercept and slope model is

$$\hat{y}_{it} = b_{0i} + 251.405 + (10.467 + b_{1i})\text{Days} + \varepsilon_{it}$$

with $t = \text{Days} = 0, 1, \dots, 9$ and $i = 1, \dots, 18$, where

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \stackrel{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 611.90 & 10.256 \\ 10.256 & 35.08 \end{pmatrix}\right) \text{ indep. of } \varepsilon_{it} \stackrel{iid}{\sim} N(0, 654.94)$$

```
VarCorr(fit2)
```

Groups	Name	Std.Dev.	Corr
Subject	(Intercept)	24.7366	
	Days	5.9229	0.066
Residual		25.5918	

```
#Model diagnostics
```

```
y.hat2 <- fitted(fit2) # Fitted values
```

```
int.hat2 <- ranef(fit2)[[1]][[1]] # Predicted intercepts
```

```
res.hat2 <- residuals(fit2) # Estimated residuals
```

```
qqnorm(int.hat2, main="Random Intercepts"); qqline(int.hat2)
```

```
qqnorm(res.hat2, main="Residuals"); qqline(res.hat2)
```

```
plot(y.hat2, res.hat2, xlab="Fitted Values", ylab="Residuals")
```

```
abline(h=0, lty=2)
```

§7.1.3 Example: A sleep deprivation study (9)

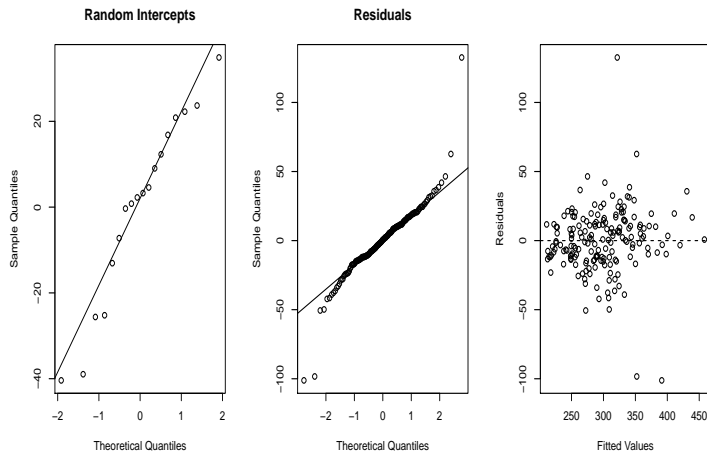


Figure 3: Model diagnostics plots for model `fit2`

§7.1.3 Example: A sleep deprivation study (10)

Compare model `fit2` with model `fit1`:

```
anova(fit1, fit2)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: sleepstudy
```

```
Models:
```

```
fit1: Reaction ~ Days + (1 | Subject)
```

```
fit2: Reaction ~ Days + (Days | Subject)
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi	Df	Pr(>Chisq)
fit1	4	1802.1	1814.8	-897.04	1794.1				
fit2	6	1763.9	1783.1	-875.97	1751.9	42.139		2	7.072e-10 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GLM with random effects (1)

For non-normal response data one cannot expect to get a closed form like $\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*$ as in (4) for normal data.

GLM with random effects: The model is constructed also in two stages.

- **Stage 1.** The conditional density $f(y_{it}|\mathbf{b}_i)$ is of the simple uni- or multi-variate exponential family type with conditional mean:

$$\mu_{it} = E(y_{it}|\mathbf{b}_i) = h(\eta_{it}), \quad \text{with} \quad \eta_{it} = \mathbf{z}_{it}^T\boldsymbol{\beta} + \mathbf{w}_{it}^T\mathbf{b}_i \quad (11)$$

where $h(\cdot)$ is a response function, and η_{it} is the linear predictor.

GLM with random effects (2)

- **Stage 2.** Cluster-specific effects \mathbf{b}_i are assumed to be independent and normally distributed with $E(\mathbf{b}_i) = \mathbf{0}$ and unknown $\text{Cov}(\mathbf{b}_i) = Q > 0$; i.e. $\mathbf{b}_i \stackrel{i.i.d.}{\sim} \text{MVN}(\mathbf{0}, Q)$, $i = 1, \dots, n$; or follows a more general pdf $p(\mathbf{b}_i, Q)$.

As an additional assumption, conditional independence of observations within and between clusters is required, i.e.

$$f(Y|B; \beta) = \prod_{i=1}^n f(\mathbf{y}_i | \mathbf{b}_i; \beta), \quad \text{with} \quad f(\mathbf{y}_i | \mathbf{b}_i; \beta) = \prod_{t=1}^{T_i} f(y_{it} | \mathbf{b}_i; \beta) \quad (12)$$

where $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ and $B = (\mathbf{b}_1, \dots, \mathbf{b}_n)$.

- When the conditioning on \mathbf{b}_i is dropped, observations within clusters are dependent.
- Observations in different clusters are both conditionally and marginally independent.

Example 1 of random effects GLM

Example 1. (binary logistic random intercept model)

Response y_{it} is binary. Then the model for y_{it} is

$$\pi_{it} = P(y_{it} = 1 | b_i) = \frac{\exp(\eta_{it})}{1 + \exp(\eta_{it})} \quad (13)$$

with $\eta_{it} = \mathbf{z}_{it}^T \boldsymbol{\beta} + b_i$, $b_i \stackrel{iid}{\sim} N(0, \sigma^2)$, $t = 1, \dots, T_i$; $i = 1, \dots, n$, where the design vector \mathbf{z}_{it} is assumed to contain the intercept term “1”.

If model (13) is the true one, the covariate effects measured by the population-averaged model with $\eta_{it} = \mathbf{z}_{it}^T \boldsymbol{\beta}$ are closer to zero than the underlying subject-specific model (13).

Example 2 of random effects GLM (1)

Example 2. (multinomial logistic random effects model) (1)

- Suppose response \mathbf{y}_{it} is multi-categorical with $\mathbf{y}_{it} = (y_{it1}, \dots, y_{itk})^T$ where $y_{itj} = 1$ if category j is observed for individual i at time t , and $y_{itj} = 0$ otherwise, $j = 1, \dots, k$.
- Then multinomial random effects models for nominal or ordered responses are completely specified by conditional probabilities $\boldsymbol{\pi}_{it} = (\pi_{it1}, \dots, \pi_{itq})^T$, $q = k - 1$, with

$$\boldsymbol{\pi}_{it} = \mathbf{h}(\boldsymbol{\eta}_{it}), \quad \boldsymbol{\eta}_{it} = \mathbf{Z}_{it}\boldsymbol{\beta} + \mathbf{W}_{it}\mathbf{b}_i.$$

- Specifically, a multinomial logistic random effects model accounting for cluster- and category-specific effects is given by

$$\pi_{itj} = \frac{\exp(\eta_{itj})}{1 + \sum_{m=1}^q \exp(\eta_{itm})}, \quad \text{with } \eta_{itj} = \alpha_{ij} + \mathbf{x}_{it}^T \boldsymbol{\gamma}_j, \quad j = 1, \dots, q.$$

Example 2 of random effects GLM (2)

Example 2. (multinomial logistic random effects model) (2)

- Assume $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iq})^T \stackrel{i.i.d.}{\sim} \text{MVN}(\alpha = (\alpha_1, \dots, \alpha_q)^T, Q)$, the linear predictors can be rewritten as

$$\eta_{itj} = \alpha_j + b_{ij} + \mathbf{x}_{it}^T \gamma_j, \quad \text{where } b_{ij} = \alpha_{ij} - \alpha_j.$$

- Defining $\beta^T = (\alpha_1, \gamma_1^T, \dots, \alpha_q, \gamma_q^T)$ and $\{\mathbf{b}_i^T = (b_{i1}, \dots, b_{iq}), i = 1, \dots, n\} \stackrel{i.i.d.}{\sim} \text{MVN}(\mathbf{0}, Q)$, one obtains a conditional multinomial model of form (11) with

$$Z_{it} = \begin{bmatrix} 1 & \mathbf{x}_{it}^T & & \\ & 1 & \mathbf{x}_{it}^T & \\ & & \ddots & \\ & & & 1 & \mathbf{x}_{it}^T \end{bmatrix}, \quad W_{it} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}_{q \times q}$$

Example 3 of random effects GLM (1)

Example 3. (ordered response random effects model) (1)

- The ordinal response models in §6.3 and §6.4 can be extended to include random effects for handling the subject-specific heterogeneity of the covariate effects.
- Let $F(\cdot)$ be a known cdf, e.g. the logistic cdf. Conditional response probabilities may be specified by

$$\pi_{it1} = F(\theta_{i1} + \mathbf{x}_{it}^T \boldsymbol{\gamma}), \quad \sum_{j=1}^r \pi_{itj} = F(\theta_{ir} + \mathbf{x}_{it}^T \boldsymbol{\gamma}), \quad r = 2, \dots, q,$$

with **cluster-specific thresholds** being ordered

$$-\infty = \theta_{i0} < \theta_{i1} < \dots < \theta_{iq} < \theta_{ik} = \infty \quad (14)$$

and covariate effects $\boldsymbol{\gamma}$ being constant over clusters.

Example 3 of random effects GLM (2)

Example 3. (ordered response random effects model) (2)

- The simplest random effects model is given by **cluster-specific shifting** of thresholds where

$$\theta_{ir} = \theta_r + b_i, \quad b_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2).$$

- Then $\eta_{itr} = \theta_r + b_i + \mathbf{x}_{it}^T \boldsymbol{\gamma}$
- The matrices in the linear form $\boldsymbol{\eta}_{it} = \mathbf{Z}_{it} \boldsymbol{\beta} + \mathbf{W}_{it} \mathbf{b}_i$ are given by

$$\mathbf{Z}_{it} = \begin{bmatrix} 1 & 0 & \cdots & \mathbf{x}_{it}^T \\ 0 & 1 & \cdots & \mathbf{x}_{it}^T \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & \mathbf{x}_{it}^T \end{bmatrix}, \quad \mathbf{W}_{it} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \text{where } \boldsymbol{\beta}^T = (\theta_1, \dots, \theta_q, \boldsymbol{\gamma}^T)$$

Example 3 of random effects GLM (3)

Example 3. (ordered response random effects model) (3)

- Distribution of θ_{ir} 's can be extended to

$$\boldsymbol{\theta}_i^T = (\theta_{i1}, \dots, \theta_{iq}) \stackrel{i.i.d.}{\sim} \text{MVN}(\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^T, Q)$$

- If the threshold means are not well separated and σ^2 is not small enough, the ordering (14) may be violated and numeric problems will occur in the estimating procedure.
- To overcome this problem, θ_{ir} 's may be reparameterised as seen in §6.3.3 using

$$\alpha_{i1} = \theta_{i1}, \quad \alpha_{ij} = \log(\theta_{ij} - \theta_{i,j-1}), \quad j = 2, \dots, q$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{iq})^T \in \mathbb{R}^q$ is unrestricted, and $\boldsymbol{\alpha}_i \stackrel{i.i.d.}{\sim} \text{MVN}(\boldsymbol{\alpha}, Q)$ is assumed.

Example 3 of random effects GLM (4)

Example 3. (ordered response random effects model) (4)

- The conditional response probabilities are now

$$\pi_{it1} = F(\alpha_{i1} + \mathbf{x}_{it}^T \boldsymbol{\gamma}), \quad \sum_{m=1}^j \pi_{itm} = F\left(\alpha_{i1} + \sum_{m=2}^j e^{\alpha_{im}} + \mathbf{x}_{it}^T \boldsymbol{\gamma}\right), \quad j=2, \dots, q.$$

- Rewriting the reparameterised threshold as $\alpha_i = \boldsymbol{\alpha} + \mathbf{b}_i$, with $\mathbf{b}_i \stackrel{i.i.d.}{\sim} \text{MVN}(\mathbf{0}, Q)$, and defining $\boldsymbol{\beta} = (\boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T)^T$ yields a **cumulative logistic random effects model**, where

$$Z_{it} = \begin{bmatrix} 1 & 0 & \cdots & 0 & \mathbf{x}_{it}^T \\ 0 & 1 & \cdots & 0 & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \mathbf{0} \end{bmatrix}, \quad W_{it} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

§7.2 Example: Analyzing Ohio children data by GLMM (1)

Recall the Ohio children data introduced in Chapter 6. Note that each child had 4 binary values of the response (infection history) in the study, comprising a cluster. It seems that the response variable may be dependent on some unobserved cluster-specific attributes. Therefore, a random intercept logistic regression model can also be fitted to the data, producing the following R output.

```
library{lme4}  
fitOC <- glmer(resp ~ age + smoke + (1|id), family=binomial, data=ohio)  
summary(fitOC)
```

§7.2 Example: Analyzing Ohio children data by GLMM (2)

```
Generalized linear mixed model fit by max likelihood (Laplace Approx.) [glmerMod]
Family: binomial ( logit ); Formula: resp ~ age + smoke + (1 | id); Data: ohio
      AIC      BIC    logLik deviance df.resid
1597.9   1620.6   -794.9   1589.9     2144
```

Scaled residuals:

	Min	1Q	Median	3Q	Max
	-1.403	-0.180	-0.158	-0.132	2.518

Random effects:

Groups	Name	Variance	Std.Dev.
id	(Intercept)	5.49	2.34

Number of obs: 2148, groups: id, 537

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.374	0.275	-12.27	<2e-16 ***
age	-0.177	0.068	-2.60	0.0093 **
smoke	0.415	0.287	1.44	0.1485

Correlation of Fixed Effects:

	(Intr)	age
age	0.227	
smoke	-0.419	-0.010

§7.2 Example: Analyzing Ohio children data by GLMM (3)

- 1 Let y_{it} be the response value resp of child i at age t . Write down the model involved in the analysis, including the probability distributions of the response and the random effects, if any, and any other necessary information of the model. Give the estimates of the parameters appearing in the model.
- 2 Estimate the odds ratio of wheezing for a child for every one year older in age.
- 3 Estimate the odds ratio of wheezing for a 10-year old child with a smoking mum versus a 9-year old child with nonsmoking mum. Also calculate an approximate standard error of your odds ratio estimate.

§7.2 Example: Analyzing Ohio children data by GLMM (4)

- 1 Give the estimates of the parameters appearing in the model.
 - Let $\pi_{it} = P(y_{it} = 1 | \text{age}, \text{smoke})$, $t = 1, \dots, 4$; $i = 1, \dots, 537$. Then the logistic random intercept model is used here

$$\log \frac{\pi_{it}}{1 - \pi_{it}} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{smoke} + b_i, \quad b_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$$

Here $\hat{\beta}_0 = -3.374$, $\hat{\beta}_1 = -0.177$, $\hat{\beta}_2 = 0.415$ and $\hat{\sigma}^2 = 5.49$.

- 2 Estimate the OR of wheezing for a child for every 1-yr older in age.

- $\widehat{\text{OR}} = e^{\hat{\beta}_1} = e^{-0.177} = 0.838$

- 3 Estimate the OR of wheezing for a 10-yr old

- $\widehat{\text{OR}} = e^{\hat{\beta}_1 + \hat{\beta}_2} = e^{-0.177 + 0.415} = e^{0.238} = 1.269$.

$$\begin{aligned} \text{s.e.}(\hat{\beta}_1 + \hat{\beta}_2) &= \sqrt{\text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) + 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)} \\ &= \sqrt{0.068^2 + 0.287^2 - 2 \times 0.010 \times 0.068 \times 0.287} = 0.294. \end{aligned}$$

- Hence $\text{s.e.}(\widehat{\text{OR}}) \approx e^{\hat{\beta}_1 + \hat{\beta}_2} \cdot \text{s.e.}(\hat{\beta}_1 + \hat{\beta}_2) = 1.269 \times 0.294 = 0.373$.

§7.3. Estimation based on posterior modes

- The intention here is to estimate β and $B = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ based on posterior modes rather than posterior means of the posterior distribution, thereby avoiding computing multivariate integrals which can be extremely difficult.
- Write $\delta^T = (\beta, \mathbf{b}_1^T, \dots, \mathbf{b}_n^T)$. The posterior density of δ can be obtained from the joint pdf of $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ and δ , where $\mathbf{b}_i \stackrel{i.i.d.}{\sim} \text{MVN}(\mathbf{0}, Q)$ is assumed for \mathbf{b}_i and a vague or flat prior density is assumed for β .
- Denote $p(\delta; Q)$ as the joint prior density of δ . It is easy to see that

$$p(\delta; Q) \propto \prod_{i=1}^n p(\mathbf{b}_i; Q).$$

§7.3.1 Known variance-covariance components (1)

- Suppose Q is known. Applying Bayes theorem and the independence assumption (12), the posterior pdf of δ is

$$f(\delta|Y; Q) = \frac{\prod_{i=1}^n f(\mathbf{y}_i|\mathbf{b}_i, \beta) \cdot \prod_{i=1}^n p(\mathbf{b}_i; Q)}{\int \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{b}_i, \beta) p(\mathbf{b}_i; Q) d\mathbf{b}_1 \cdots d\mathbf{b}_n d\beta} \quad (15)$$

- An optimal point estimator for δ is the posterior mean of (15), but is difficult to calculate due to intractability of the denominator in (15).
- Therefore, **posterior mode estimator** of δ , also called **maximum a posteriori (MAP) estimator** of δ , which is the minimizer of (15) is used.

§7.3.1 Known variance-covariance components (2)

- Since from (15), $f(\boldsymbol{\delta}|Y; Q) \propto \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}) \cdot \prod_{i=1}^n p(\mathbf{b}_i; Q)$, maximizing (15) is equivalent to maximizing the log-posterior

$$\sum_{i=1}^n \log f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\beta}) + \sum_{i=1}^n \log p(\mathbf{b}_i; Q) \quad \text{with respect to } \boldsymbol{\delta}. \quad (16)$$

- Maximizing (16) can be done using Fisher scoring. In the case $\mathbf{b}_i \stackrel{i.i.d}{\sim} \text{MVN}(\mathbf{0}, Q)$, (16) becomes, up to an irrelevant additive constant,

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^n \log f(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\beta}) - \frac{1}{2} \sum_{i=1}^n \mathbf{b}_i^T Q^{-1} \mathbf{b}_i \quad (17)$$

- Maximizing (17) becomes maximising penalised log-likelihood.
- In each Fisher scoring step, efficient use should be made of the block structure of the second derivative of $\ell(\boldsymbol{\delta})$, cf. §7.3.3.

§7.3.2 Unknown variance-covariance components (1)

- Generally, the variance-covariance components of Q are unknown.
- Estimating Q can be based on an EM-type algorithm, which can be deduced from an EM algorithm that indirectly maximizes the marginal log-likelihood

$$\ell(Q) = \log \int \prod_{i=1}^n f(\mathbf{y}_i | \mathbf{b}_i, \beta) p(\mathbf{b}_i; Q) d\mathbf{b}_1 \cdots d\mathbf{b}_n d\beta \quad (18)$$

- Direct maximization of (18) is difficult due to high-dimensional intergration involved.

§7.3.2 Unknown variance-covariance components (2)

- Indirect maximization of (18) by EM starts from the joint log-density

$$\log f(Y, \delta; Q) = \log f(Y|\delta) + \log p(\delta; Q)$$

In the $(p+1)$ th cycle of the EM, the E-step consists of computing

$$M(Q|Q^{(p)}) = E\{\log f(Y, \delta; Q|Y; Q^{(p)})\}$$

which is the conditional expectation of the complete-data log-density given Y and the previous $Q^{(p)}$. Because now $Q^{(p)}$ is known, we have essentially the case of known variance-covariance components considered in §7.3.1.

- To avoid integrations necessary for computing the conditional expectations of random effects appearing in $\log p(\delta; Q)$, **the conditional expectations are replaced by the conditional modes** obtained from maximizing (17) for $Q = Q^{(p)}$. The resulting EM-type algorithm is described in detail in §7.3.3, not to be pursued here.

§7.4. Estimation by Integration Techniques

- Instead of using posterior models to approximate the posterior means, one can approximate the integrations involving δ and Q by numerical methods or Monte Carlo methods.
- Numerical methods include **univariate Gauss-Hermite integration** and **multivariate Gauss-Hermite integration**.
- Monte Carlo methods include **importance sampling**, **rejection sampling**, **Gibbs sampling** and **Metropolis-Hastings algorithm** etc.
- Examples can be found in §7.5.
- Details are not pursued here.

§7.6. Bayesian Mixed Models

§7.3 and §7.4 described empirical Bayes approaches for mixed models, with unknown fixed effects parameters and variance parameters, but a mixing prior distribution for random effects components.

In this section, a pure Bayes approach is used, where all parameters are treated as random.

§7.7. Marginal estimation approach to random effects models

We have already seen this in §7.1 (cf. (4) and (5)).

Marginal estimation approach can also be applied to GLMM, but with increased mathematical and computational complications.