Chapter 19
Mixed Effects Modelling Applied on American Foulbrood Affecting Honey Bees Larvae

A.F. Zuur, L.B. Gende, E.N. Ieno, N.J. Fernández, M.J. Eguaras, R. Fritz, N.J. Walker, A.A. Saveliev, and G.M. Smith

19.1 Introduction

In this chapter, we apply mixed modelling to honeybee data. The data are considered nested because multiple observations were taken from the same hive. A total of 24 hives were sampled.

American Foulbrood (AFB) is an infectious disease affecting the larval stage of honeybees (*Apis mellifera*) and is the most widespread and destructive of the brood diseases (Shimanuki, 1997). The causative agent is *Paenibacillus larvae* (Genersch et al., 2006) and the spore forming bacterium infects queen, drone, and worker larvae. Only the spore stage of the bacterium (Fig. 19.1) is infectious to honey bee larvae. The spores germinate into the vegetative stage soon after they enter the larval gut and continue to multiply until larval death. The spores are extremely infective and resilient, and one dead larva may contain billions of spores (Hansen and Brødsgaard, 1999).

Although adult bees are not directly affected by AFB, some of the tasks carried out by workers might have an impact on the transmission of AFB spores within the colony and on the transmission of spores between colonies. When a bee hatches from its cell, its first task is to clean the surrounding cells, and its next task is tending and feeding of larvae. Here, the risk of transmitting AFB spores is particularly great if larvae that succumbed to AFB are cleaned prior to feeding susceptible larvae (Lindstrom, 2006).

Because AFB is extremely contagious, hard to cure, and lethal at the colony level, it is of importance to detect outbreaks, before they spread and become difficult to control (Lindstrom, 2006). Reliable detection methods are also important for studies of pathogen transmission within and between colonies. Of the available methods, sampling adult bees has been shown the most effective (Nordström et al., 2002). Hornitzky and Karlovskis (1989) introduced the method of culturing adult honey bees for AFB, and demonstrated that spores can be detected from colonies without

A.F. Zuur (✉)
Highland Statistics Ltd., Newburgh, AB41 6FN, United Kingdom

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_19,

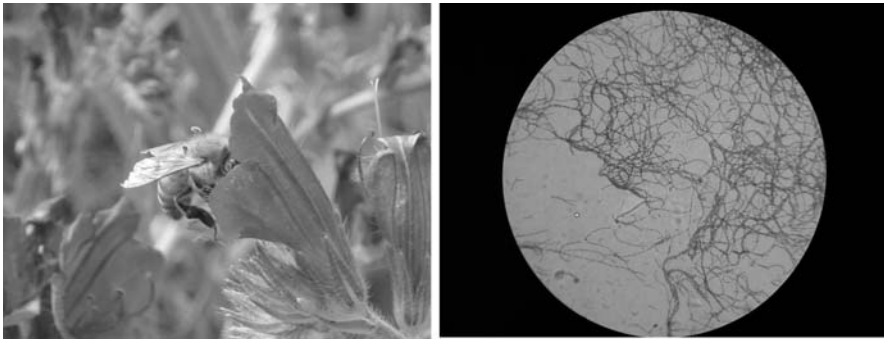

Fig. 19.1 *Left*: Honeybee. *Right*: Vegetative stage of the bacteria at microscopic level

clinical symptoms. Recently, culturing of *P. larvae* from adult honey bee samples has been shown to be a more sensitive tool for AFB screening compared to culturing of honey samples (Nordström et al., 2002). When samples of adult bees are used, the detection level of *P. larvae* is closely linked to the distribution of spores among the bees. For this reason, we will model the density of *P. larvae* with the potential explanatory variables as number of bees in the hive, presence or absence of AFB, and hive identity. Technical details on how spores were counted can be found in Hornitzky and Karlovskis (1989).

19.2 Data Exploration

There are three observations per hive, with a total of 24 hives. Figure 19.2A shows a Cleveland dotplot for the spores (density) conditional on hives. Recall from Chapter 2 that this graph groups the observations from the same hive along the vertical axis, and the values of the spores can be read from the horizontal axis. Two hives have considerably higher values than the others, indicating that serious problems with homogeneity can be expected if linear regression or mixed effects modelling is applied. One option is to use different variances per hive (Chapter 4), but this would result in 24 extra variances. This might make the estimation process for multiple variances with generalised least squares (GLS) unstable. We therefore prefer to transform the data using a logarithmic transformation. A square root transformation was also tried, but was considered too weak to ensure homogeneity. Because some observations have the value of 0, a $\log_{10}(Y_{ij} + 1)$ transformation was applied, where Y_{ij} is the density of spores in observation j in hive i, with $j = 1, \ldots, 3$, and $i = 1, \ldots, 24$. The transformed data are shown in Fig. 19.2B. The R code to access the data, transform the spore data, and make the two Cleveland dotplots, is given below. The first two commands are used to access the data. The `par` command sets up the graphical window and the `mar` option controls the amount of white space around the individual panels. The `dotchart` command was discussed in Chapter 2.

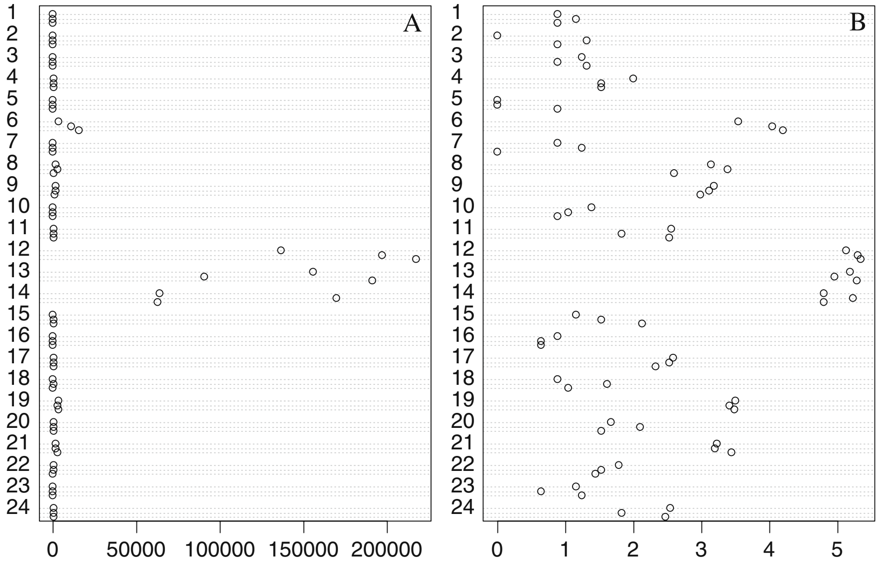

Fig. 19.2 **A**: Cleveland dotplot for the untransformed spores (densities) data. The data are grouped by hives. **B**: Cleveland dotplot for the log10-transformed data. The vertical axes show the three observations per hive and the horizontal axes the values of the spores data

```
> library(AED); data(Bees)
> Bees$fhive <- factor(Bees$hive)
> Bees$Lspobee <- log10(Bees$spobee + 1)
> op<- par(mfrow = c(1, 2), mar = c(3, 4, 1, 1))
> dotchart(Bees$spobee, groups = Bees$fhive)
> dotchart(Bees$Lspobee, groups = Bees$fhive)
> par(op)
```

Instead of using the Cleveland dotplot, we could have used a conditional boxplot. However, with only three values per hive, this would have been less useful.

The explanatory variable `Infection` quantifies the degree of infection (AFB), with values 0 (none), 1 (minor), 2 (moderate), and 3 (major). Although mixed effects modelling can cope with a certain degree of unbalanced data, in this case it may be better to convert the variable `Infection` in 0 (no infection) and 1 (infection is present) as there are only a few observations that have the value 2 or 3 for this variable. The R code to do this is

```
> Bees$Infection01 <- Bees$Infection
> Bees$Infection01[Bees$Infection01 > 0] <- 1
> Bees$fInfection01 <- factor(Bees$Infection01)
```

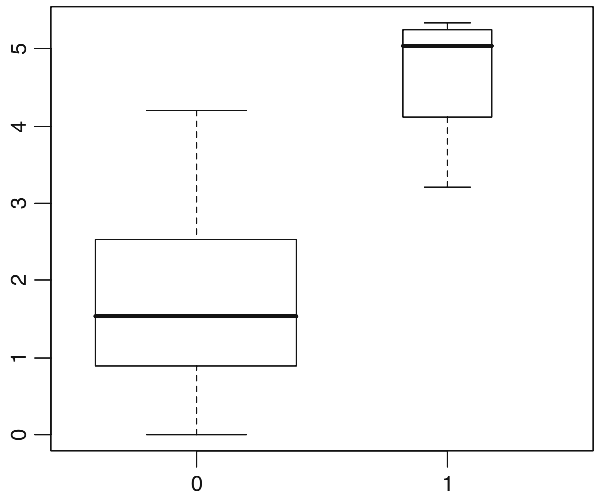

Fig. 19.3 Boxplot of log-transformed spores densities conditional on the variable fInfection01 (AFB). Note that there are considerably more observations with fInfection01 equal to 0. The width of a boxplot is proportional to sample size

All observations for the variable Infection that are larger than 0 are set to 1. After this transformation, 17% of its values are equal to 1 and 73% are 0.

A boxplot of spores conditional on Infection01 shows clear differences between the two levels (Fig. 19.3). The boxplot was made with the command

```
> boxplot(LSpobee ~ fInfection01, data = Bees,
          varwidth = TRUE)
```

Other graphical validation tools were also applied, for example, the coplot and xyplot, but no clear patterns were found. These graphs and R code are not presented here.

19.3 Analysis of the Data

The response variable is the log-transformed density of spores and the explanatory variables are infection (nominal with two classes) and number of bees. To investigate whether there is a hive effect, we first applied a linear regression model on the data. As explanatory variables we used infection and number of bees together with their interaction. The standardised residuals from this model were plotted against hive (Fig. 19.4) and show a clear pattern. In this graph, we want to see residuals that are scattered around zero, but in this case, we have various hives where all three residuals are above the zero line or all are below the zero line. This indicates there is within-hive correlation.

An option is to include hive as an explanatory variable. However, if we do this as a fixed term, paying the price of losing 23 degrees of freedom is rather high! And on top of that, the resulting model would only hold for these 24 hives. A logical

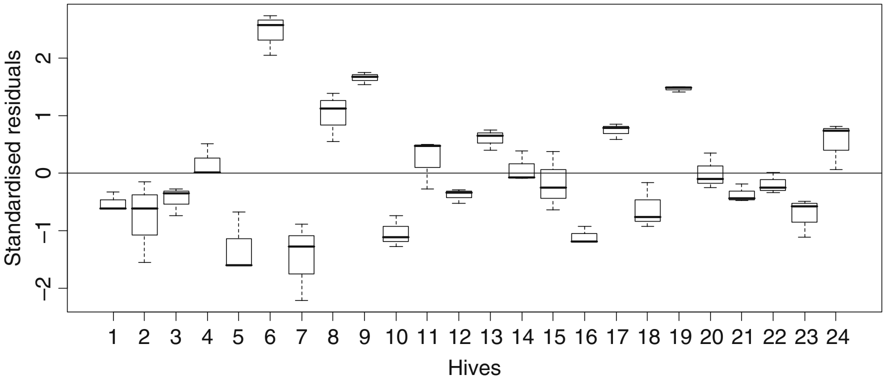

Fig. 19.4 Standardised residuals from the linear regression model where log transformed spores are modelled as a function of infection, number of bees, and their interaction

solution is to proceed with a random intercept model (Chapter 5). The advantages of such an approach are (i) it only requires one extra parameter (the variance for the random intercept), compared to the linear regression model that required 23 extra parameters; (ii) we can make a statement for hives in general and not only these 24; and (iii) as an extra bonus, it introduces a correlation structure between observations of the same hive.

The following R code was used to apply the linear regression model, extract the normalised residuals and produce the boxplot in Fig. 19.4. The `abline` command adds the horizontal line at zero.

```
> M1 <- lm(LSpobee ~ fInfection01 * BeesN, data = Bees)
> E1 <- rstandard(M1)
> plot(E1 ~ Bees$fHive, xlab = "Hives",
        ylab = "Standardised residuals")
> abline(0, 0)
```

Recall from Chapters 4 and 5 that the selection approach for linear mixed effects models should broadly follow a protocol consisting of 10 steps. In step 1, we start with a model that has as many explanatory variables as possible (in the fixed part of the model), then we find the optimal random structure (steps 2–6), the optimal fixed structure (steps 7–8), present the results of the optimal model using REML estimation (step 9), and finally, give an interpretation (step 10). We follow these same steps here.

Step 1 of the Protocol

Earlier in this chapter, we started with a model that contained all the explanatory variables and their interaction in the fixed part of the model. In this case, there are only two fixed explanatory variables.

Steps 2–6 of the Protocol

Starting with a random intercept model, we have

$$\begin{aligned}\text{LSpobee}_{ij} &= \alpha + \beta_1 \times \text{BeesN}_{ij} + \beta_2 \times \text{fInfection01}_{ij} \\ &\quad + \beta_3 \times \text{BeesN}_{ij} \times \text{fInfection01}_{ij} + a_i + \varepsilon_{ij}\end{aligned}$$

In words, the log-transformed spores are modelled as an intercept (α), plus a linear 'number of bees per hive' effect (BeesN), an infection effect (fInfection01), the interaction between these two terms, a random intercept a_i that is assumed to be normally distributed with mean 0 and variance $\sigma_a{}^2$, and something that is 'real' noise (ε_{ij}). The index i refers to hives ($i = 1, \ldots, 24$) and j to the observation within a hive ($j = 1, \ldots, 3$). The term ε_{ij} is the within-hive variation, and is assumed to be independently normally distributed with mean 0 and variance σ^2.

We use the function lme from the R package nlme to fit the random intercept model in Equation (19.1). To assess whether the mixed effects model is better than the ordinary linear regression model, we need to refit the latter one using the gls function without the random intercept. The anova function can then be used to compare AICs or apply a likelihood ratio test. The required R code and output of the anova command are given below.

```
> library(nlme)
> M2<-gls(LSpobee ~ fInfection01 * BeesN, data = Bees)
> M3<-lme(LSpobee ~ fInfection01 * BeesN,
          random =~ 1 | fHive, data = Bees)
> anova(M2,M3)
```

```
   Model df      AIC      BIC     logLik   Test  L.Ratio p-value
M2     1  5 251.5938 262.6914 -120.79692
M3     2  6 175.0129 188.3299  -81.50643 1 vs 2 78.58097  <.0001
```

We can either use the AIC to select the optimal model or apply the likelihood ratio test. The AIC values indicate that the mixed model is preferred. The problem with the likelihood ratio test is that we are testing on the boundary (Chapter 5). The correct p-value is obtained by typing

```
> 0.5 * (1 - pchisq(78.58097, 1))
```

This is still smaller than 0.001; so both approaches favour the mixed model.

There are a few ways to extend the random part of the model. We can try a random intercept and slope model, and we can try using multiple variances. As to the first option, the BeesN effect may be different per hive and the same may hold for the fInfection01 effect. However, both options gave higher AICs. The R code for these models and model comparisons are given below.

```
> M4 <- lme(LSpobee ~ fInfection01 * BeesN,
         random  =~ 1 + BeesN  | fHive, data = Bees)
```

```
> M5 <- lme(LSpobee ~ fInfection01 * BeesN,
        random  =~ 1  + fInfection01  | fHive, data = Bees)
> anova(M2, M3, M4, M5)
```

```
   Model df      AIC      BIC      logLik   Test  L.Ratio  p-value
M2     1  5 251.5938 262.6914  -120.79692
M3     2  6 175.0129 188.3299   -81.50643 1 vs 2 78.58097   <.0001
M4     3  8 178.8460 196.6020   -81.42299 2 vs 3  0.16689   0.9199
M5     4  8 177.7606 195.5167   -80.88032
```

As extending the model with random slopes gives no improvement, we can look at an alternative of adding multiple variances for the residuals ε_{ij}. One option is to fit the model with and without multiple variances and compare them with the AIC or the likelihood ratio test. Another option is to plot the residuals of the model that is the best so far, the random intercept model in Equation (19.1), and see whether anything is wrong. We chose the second approach. The command `plot(M3, col = 1)` produces a plot of the residuals against fitted values for the random intercept model (Fig. 19.5). Note that there is some evidence of heterogeneity as the residual spread is slightly smaller for larger fitted values. These are actually the observations for which `Infection01` is equal to 1 (this can be seen by using colours or different symbols), which suggests extending the random intercept model in Equation (19.1) from $\varepsilon_{ij} \sim \text{N}(0, \sigma^2)$ to $\varepsilon_{ij} \sim \text{N}(0, {\sigma_k}^2)$, where $k = 1, 2$.

This means that we use a variance for the observations that have no infection and a different variance for the observations that have `Infection01` = 1. Technically, the `varIdent` variance structure is used for this; see also Chapter 4. The AIC of this model (171.65) is slightly better than the random intercept model (175.01) in

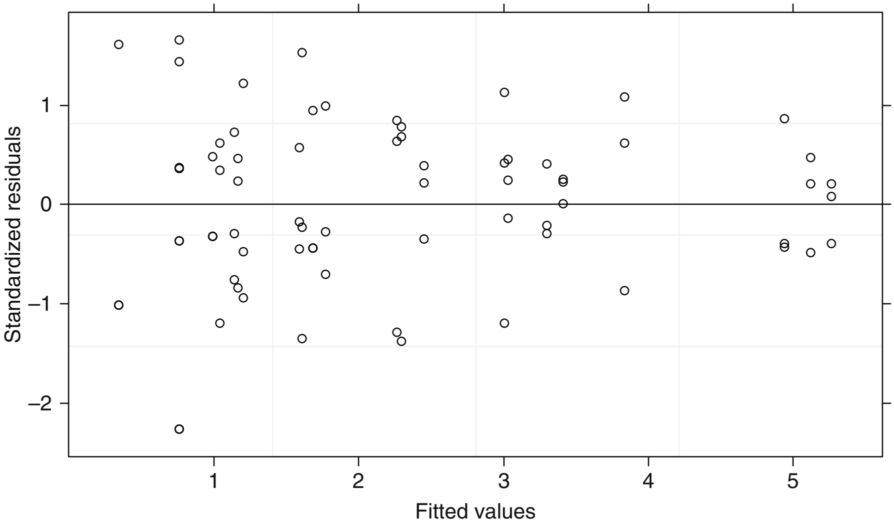

Fig. 19.5 Residuals versus fitted values for the mixed model

Equation (19.1), and the likelihood ratio test gives a *p*-value of 0.02, indicating that we have weak evidence to reject the null hypothesis that both variances are the same. The normalised residuals (not shown here) now look better.

The R code below fits the new model, and compares it with the random intercept model.

```
> M6 <- lme(LSpobee ~ fInfection01 * BeesN,
            random  =~ 1  | fHive, data = Bees,
            weights = varIdent(form  =~ 1  | fInfection01))
> anova(M3, M6)
```

```
  Model df      AIC      BIC    logLik   Test L.Ratio  p-value
M3    1  6 175.0129 188.3299  -81.50643
M6    2  7 171.6587 187.1952  -78.82933 1 vs 2  5.3542  0.0207
```

Steps 7 and 8 of the Protocol

We now continue with the seventh and eighth step of the protocol to find the optimal fixed structure for the selected random structure. This means that using our optimal random structure (random intercept plus two variances for ε_{ij}), we need to look at the optimal fixed structure. As discussed in Chapters 4 and 5, we can either do this using the *t*-statistics from the `summary` command, sequential *F*-tests using the `anova` command, or likelihood ratio tests of nested models. The first two approaches require REML estimation with the third approach needing ML estimation. We will use the last approach as the first two approaches can easily be carried out by the reader, and there is a higher degree for 'confusion' with the third approach.

In the first step, we need to apply the model with all terms and a model without the interaction. Note that we cannot drop any of the main terms yet. The `update` command is used to fit the model without the interaction term; see also Chapters 4 and 5.

```
> M7full  <- lme(LSpobee  ~ fInfection01 * BeesN,
          random =~ 1  | fHive, method = "ML", data = Bees
          weights = varIdent(form  =~ 1  | fInfection01))
> M7sub  <- update(M7full, .~.  -fInfection01 : BeesN)
> anova(M7full, M7sub)
```

```
     Model df      AIC      BIC    logLik   Test  L.Ratio   p-value
M7full   1  7 129.8792 145.8159  -57.93962
M7sub    2  6 128.4452 142.1052  -58.22262 1 vs 2 0.5660039   0.4519
```

The `anova` command gives $L = 0.56$ ($df = 1$) with $p = 0.45$, allowing us to drop the interaction term to give a model with two main terms. We can now either switch to approach one and use the *t*-statistics to assess the significance of these two main terms or we can be consistent and go on with the likelihood ratio testing approach. We prefer consistency. The following code reapplies the model, drops each of the main terms in turn, and then applies the likelihood ratio test.

```
> M8full <- lme(LSpobee  ~ fInfection01  + BeesN,
       random =~ 1  | fHive, method =  "ML", data = Bees,
       weights = varIdent(form  =~ 1  | fInfection01))
> M8sub1 <- update(M8full, .~.  -fInfection01)
> M8sub2 <- update(M8full, .~.  -BeesN)
> anova(M8full, M8sub1)
```

```
      Model df      AIC      BIC     logLik   Test  L.Ratio  p-value
M8full    1  6 128.4452 142.1052  -58.22262
M8sub1    2  5 144.6700 156.0533  -67.33497 1 vs 2 18.22471   <.0001
```

```
> anova (M8full,M8sub2)
```

```
      Model df      AIC      BIC     logLik   Test  L.Ratio p-value
M8full    1  6 128.4452 142.1052  -58.22262
M8sub2    2  5 129.3882 140.7715  -59.69408 1 vs 2 2.942923  0.0863
```

The two anova commands give $p < 0.001$ and $p = 0.08$, making the term beesN the least significant, and we continue without it. This leaves us with one final model comparison of the models with and without the term fInfection01. The following R code is used:

```
> M9full  <- lme(LSpobee ~ fInfection01,
       random  =~ 1  | fHive, method =  "ML", data = Bees,
       weights = varIdent(form  =~ 1  | fInfection01))
> M9sub1 <- update(M9full, .~.  -fInfection01)
> anova(M9full, M9sub1)
```

```
      Model df      AIC      BIC     logLik   Test  L.Ratio  p-value
M9full    1  5 129.3882 140.7715  -59.69408
M9sub1    2  4 147.0532 156.1599  -69.52661 1 vs 2 19.66507   <.0001
```

The last anova command gives $L = 19.66$ ($df = 1$, $p < 0.0001$), indicating that infection is highly significant. So after a considerably amount of R coding, we end up with a model where only one fixed explanatory variable, infection, is highly significant.

Step 9 of the Protocol

In the last two steps of the protocol (9 and 10), we have to refit the model with REML, further validate and present the results, and then explain what it all means. The last part is the difficult bit and will be done in the discussion. The first part is easy:

```
> Mfinal <- lme(LSpobee ~ fInfection01,
          random =~ 1 |fHive, data = Bees, method="REML",
          weights = varIdent(form  =~ 1  | fInfection01))
> summary(Mfinal)
```

```
Linear mixed-effects model fit by REML
 Data: Bees
       AIC      BIC    logLik
  130.1747 141.4171  -60.08733

Random effects:
 Formula: ~1  | fHive
        (Intercept)  Residual
 StdDev:  0.9892908 0.3615819

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1  | fInfection01
 Parameter estimates:
       0        1
1.000000 0.473795

Fixed effects: LSpobee  ~ fInfection01
                  Value Std.Error DF  t-value p-value
(Intercept)    1.757273 0.2260837 48 7.772666       0
fInfection011 2.902090 0.5461078 22 5.314135       0

 Correlation:
               (Intr)
fInfection011  -0.414

Standardized Within-Group Residuals:
       Min          Q1         Med          Q3         Max
-2.1548732  -0.6068385  0.2019003  0.5621671  1.6855583

Number of Observations: 72
Number of Groups: 24
```

Let us to summarise all this information. The optimal model is given by

$$\text{LSpobee}_{ij} = 1.75 + 2.90 \times \text{fInfection01}_{ij} + a_i + \varepsilon_{ij}$$

where $a_i \sim \mathrm{N}(0, 0.98^2)$. For the within-hive residuals, we have $\varepsilon_{ij} \sim \mathrm{N}(0, 0.36^2)$ if the observation has no disease (`Infection01` = 0) and $\varepsilon_{ij} \sim \mathrm{N}(0, 0.36^2 \times 0.47^2)$ if it has a disease (`Infection01` = 1). If an observation has no diseases, then the expected density of spores is 1.75 on the logarithmic scale. If it has a disease, then the expected density is 1.75 + 2.90 = 4.65. Depending on the hive, there is a random variation on both expected values. This is due to the random intercept, and 95% of its values are between –1.96 × 0.36 and 1.96 × 0.36.

Finally, we inspect the residuals of the optimal model. This should actually be done in steps 7 and 8, but because we want to do this for the REML estimates, we do it here. We need to inspect the optimal model for homogeneity of the residuals ε_{ij}.

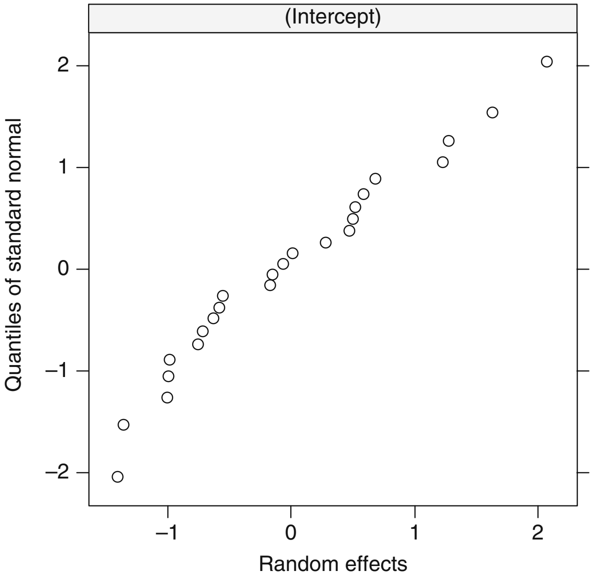

Fig. 19.6 QQ-plot of the mixed effects model `MFinal`

We have already discussed how to do this using the command `plot(Mfinal)`. Results are not presented here, but we can safely say they indicate homogeneity. We can also assume normality of these residuals. This can be verified with `qqnorm(Mfinal)`. It produces a QQ-plot of the normalised residuals. Results are not presented here, but normality is a reasonable conclusion in this case. Finally, we need to verify the normality assumption for the random effects. Use the R command `qqnorm(Mfinal, ~ranef (.),col = 1)`, and again, normality seems a reasonable conclusion (Fig. 19.6).

Another useful command is `intervals(Mfinal)`. It shows the approximate 95% confidence bands of the parameters and random variances.

```
Approximate 95% confidence intervals

 Fixed effects:
                  lower     est.    upper
(Intercept)   1.302701 1.757273 2.211845
fInfection011 1.769532 2.902090 4.034648
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: fHive
                      lower      est.    upper
sd((Intercept))  0.7259948 0.9892908 1.348076
```

```
 Variance function:
      lower     est.     upper
1 0.2770579 0.473795 0.8102339
attr(,"label")
[1] "Variance function:"

 Within-group standard error:
    lower      est.     upper
0.2904009 0.3615819 0.4502102
```

We have now finished steps 1–9 of the protocol and we discuss the interpretation of the model in the next section.

19.4 Discussion

In this chapter, we applied linear mixed effects modelling because the data are nested (three observations per hive). The model showed that there is a significant disease effect on the spore density data. The intraclass correlation is $0.98^2/(0.98^2 + 0.36^2) = 0.88$ if a hive has no disease and $0.98^2/(0.98^2 + 0.36^2 \times 0.47^2) = 0.97$ if a hive has the disease. This is rather high, and means that the effective sample size is considerably smaller than $3 \times 24 = 72$ (Chapter 5). We might as well take one sample per hive and sample more hives.

If the number of spores are analysed instead of density, we can use generalised estimation equations with a Poisson distribution (Chapter 9) or generalised linear mixed modelling with a Poisson distribution (Chapter 13).

19.5 What to Write in a Paper

A paper based on the results presented in this chapter should include a short description of the problem (introduction) and the set up of the experiment (methods). It will need to justify the use of the logarithmic transformation on spores densities and the use of mixed effects modelling. You should also outline the protocol for model selection, and in the results section, mention how you got to the final model. There is no need to present all the R code or results of intermediate models. You may want to include one graph showing homogeneity of the residuals. You should also present the estimated parameters, standard errors, *t*-values, and *p*-values of the optimal model. Warn the reader that the data are unbalanced (not many observations with a disease); so care is needed with the interpretation.

Acknowledgments We would like to thank Fernando Rodriguez, beekeeper from Buenos Aires Province and Sergio Ruffinengo for his collaboration in this project and also MalenaSabatino for the honey bee photography.