Examples from Multilevel Software Comparative Reviews

Douglas Bates
R Development Core Team
Douglas.Bates@R-project.org

October 3, 2004

Abstract

The Center for Multilevel Modelling at the Institute of Education, London maintains a web site of "Software reviews of multilevel modeling packages". The data sets discussed in the reviews are available at this web site. We have incorporated these data sets in the `lme4` package for R and, in this vignette, provide the results of fitting several models to these data sets.

1 Introduction

2 Two-level normal models

The `Exam` data set is used in fitting examples of two-level normal multilevel models.

```
> data(Exam)
> str(Exam)
`data.frame':        4059 obs. of  10 variables:
 $ school  : Factor w/ 65 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ normexam: num   0.261  0.134 -1.724  0.968  0.544 ...
 $ schgend : Factor w/ 3 levels "mixed","boys",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ schavg  : num  0.166 0.166 0.166 0.166 0.166 ...
 $ vr      : Factor w/ 3 levels "bottom 25%","mi..",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ intake  : Factor w/ 3 levels "bottom 25%","mi..",..: 1 2 3 2 2 1 3 2 2 3 ...
 $ standLRT: num   0.619  0.206 -1.365  0.206  0.371 ...
 $ sex     : Factor w/ 2 levels "F","M": 1 1 2 1 1 2 2 2 1 2 ...
 $ type    : Factor w/ 2 levels "Mxd","Sngl": 1 1 1 1 1 1 1 1 1 1 ...
 $ student : Factor w/ 650 levels "1","2","3","4",..: 143 145 142 141 138 155 158 115 117 113 ...
```

```
> sysgc.time(Em1 <- lme(normexam ~ standLRT + sex + schgend,
+     Exam, ~1 | school), gc = TRUE)
[1] 0.08 0.00 0.09 0.00 0.00
> summary(Em1)
Linear mixed-effects model fit by REML
Fixed: normexam ~ standLRT + sex + schgend
 Data: Exam
      AIC      BIC    logLik
 9361.673 9405.834 -4673.837

Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.085829 0.29297
 Residual             0.562534 0.75002
# of obs: 4059, groups: school, 65

Fixed effects:
                Estimate  Std. Error   DF t value  Pr(>|t|)
(Intercept)  -1.0493e-03  5.5569e-02 4054 -0.0189   0.98494
standLRT      5.5975e-01  1.2450e-02 4054 44.9601 < 2.2e-16
sexM         -1.6739e-01  3.4100e-02 4054 -4.9089 9.519e-07
schgendboys   1.7769e-01  1.1347e-01 4054  1.5659   0.11745
schgendgirls  1.5900e-01  8.9403e-02 4054  1.7784   0.07541

Correlation of Fixed Effects:
            (Intr) stnLRT sexM   schgndb
standLRT    -0.014
sexM        -0.316  0.061
schgendboys -0.395 -0.003 -0.145
schgendgrls -0.622  0.009  0.197  0.245
> sysgc.time(Em2 <- lme(normexam ~ standLRT * sex + schgend,
+     Exam, ~1 | school), gc = TRUE)
[1] 0.05 0.00 0.04 0.00 0.00
> summary(Em2)
Linear mixed-effects model fit by REML
Fixed: normexam ~ standLRT * sex + schgend
 Data: Exam
      AIC      BIC    logLik
 9369.204 9419.673 -4676.602

Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.085856 0.29301
 Residual             0.562666 0.75011
# of obs: 4059, groups: school, 65

Fixed effects:
                  Estimate  Std. Error   DF t value  Pr(>|t|)
(Intercept)   -8.4349e-04  5.5586e-02 4053 -0.0152   0.98789
standLRT       5.5745e-01  1.6662e-02 4053 33.4572 < 2.2e-16
sexM          -1.6733e-01  3.4105e-02 4053 -4.9064 9.638e-07
schgendboys    1.7765e-01  1.1349e-01 4053  1.5653   0.11759
schgendgirls   1.5879e-01  8.9422e-02 4053  1.7757   0.07586
standLRT:sexM  5.1121e-03  2.4584e-02 4053  0.2079   0.83528
```

```
Correlation of Fixed Effects:
            (Intr) stnLRT sexM   schgndb schgndg
standLRT    -0.022
sexM        -0.316  0.040
schgendboys -0.395 -0.001 -0.145
schgendgrls -0.622  0.014  0.196  0.245
stndLRT:sxM  0.018 -0.664  0.008 -0.002  -0.011
> sysgc.time(Em3 <- lme(normexam ~ standLRT * sex + schgend,
+     Exam, ~standLRT | school), gc = TRUE)
[1] 0.06 0.00 0.06 0.00 0.00
> summary(Em3)
Linear mixed-effects model fit by REML
Fixed: normexam ~ standLRT * sex + schgend
 Data: Exam
      AIC      BIC    logLik
 9328.242 9391.329 -4654.121

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 school   (Intercept) 0.083723 0.28935
          standLRT    0.015250 0.12349  0.575
 Residual             0.550374 0.74187
# of obs: 4059, groups: school, 65

Fixed effects:
                 Estimate  Std. Error   DF t value  Pr(>|t|)
(Intercept)   -2.1277e-02  5.3279e-02 4053 -0.3993   0.68966
standLRT       5.5713e-01  2.4349e-02 4053 22.8812 < 2.2e-16
sexM          -1.6859e-01  3.3844e-02 4053 -4.9814 6.576e-07
schgendboys    1.7751e-01  1.0211e-01 4053  1.7384   0.08221
schgendgirls   1.7790e-01  8.2104e-02 4053  2.1668   0.03031
standLRT:sexM -6.8757e-03  2.9540e-02 4053 -0.2328   0.81596

Correlation of Fixed Effects:
            (Intr) stnLRT sexM   schgndb schgndg
standLRT     0.200
sexM        -0.337  0.026
schgendboys -0.354 -0.048 -0.148
schgendgrls -0.600  0.116  0.225  0.218
stndLRT:sxM  0.067 -0.559  0.010  0.094  -0.181
```

There are some interesting aspects of data management that show up in the analysis of these data. The `student` variable is an identifier of the student within the `school`. It would be best to combine the indicators of school and student to get a unique identifier of the student.

```
> Exam$ids <- interaction(Exam$school, Exam$student, drop = TRUE)
> str(Exam)
`data.frame':        4059 obs. of  11 variables:
 $ school  : Factor w/ 65 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ normexam: num   0.261  0.134 -1.724  0.968  0.544 ...
 $ schgend : Factor w/ 3 levels "mixed","boys",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ schavg  : num  0.166 0.166 0.166 0.166 0.166 ...
 $ vr      : Factor w/ 3 levels "bottom 25%","mi..",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
 $ intake  : Factor w/ 3 levels "bottom 25%","mi..",..: 1 2 3 2 2 1 3 2 2 3 ...
 $ standLRT: num   0.619  0.206 -1.365  0.206  0.371 ...
 $ sex     : Factor w/ 2 levels "F","M": 1 1 2 1 1 2 2 2 1 2 ...
 $ type    : Factor w/ 2 levels "Mxd","Sngl": 1 1 1 1 1 1 1 1 1 1 ...
 $ student : Factor w/ 650 levels "1","2","3","4",..: 143 145 142 141 138 155 158 115 117 113 ...
 $ ids     : Factor w/ 4055 levels "1.143","1.145",..: 1 2 3 4 5 6 7 8 9 10 ...
```

Notice that there are 4059 observations but only 4055 unique levels of student within school. We can check the ones that are duplicated

```
> Exam$ids[which(duplicated(Exam$ids))]
[1] 43.86 50.39 52.2  52.21
4055 Levels: 1.143 1.145 1.142 1.141 1.138 1.155 1.158 1.115 1.117 ... 65.56
```

One of these duplicated cases is particularly interesting. One of the students with the duplicated student id 86 in school 43 is the only male student in this mixed school. This is probably a case of a misrecorded school.

3 Three-level Normal Models

Data from the 1997 A-level Chemistry exam are available as `Chem97`.

```
> data(Chem97)
> str(Chem97)
`data.frame':        31022 obs. of  8 variables:
 $ lea      : Factor w/ 131 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ school   : Factor w/ 2410 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ student  : Factor w/ 31022 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ score    : num  4 10 10 10 8 10 6 8 4 10 ...
 $ gender   : Factor w/ 2 levels "M","F": 2 2 2 2 2 2 2 2 2 2 ...
 $ age      : num  3 -3 -4 -2 -1 4 1 4 3 0 ...
 $ gcsescore: num  6.62 7.62 7.25 7.50 6.44 ...
 $ gcsecnt  : num  0.339 1.339 0.964 1.214 0.158 ...
> sysgc.time(mC1 <- lme(score ~ 1, Chem97, ~1 | lea/school))
[1] 0.34 0.02 0.35 0.00 0.00
> summary(mC1)
Linear mixed-effects model fit by REML
Fixed: score ~ 1
 Data: Chem97
      AIC      BIC   logLik
 157881.8 157915.2 -78936.9

Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 2.74872  1.65793
 lea      (Intercept) 0.15349  0.39178
 Residual             8.51609  2.91823
# of obs: 31022, groups: school, 2410; lea, 131

Fixed effects:
              Estimate Std. Error    DF t value  Pr(>|t|)
(Intercept) 5.3190e+00 5.8108e-02 31021  91.536 < 2.2e-16
```

```
> sysgc.time(mC2 <- lme(score ~ gcsecnt, Chem97, ~1 | lea/school))
[1] 0.90 0.01 0.91 0.00 0.00
> summary(mC2)
Linear mixed-effects model fit by REML
Fixed: score ~ gcsecnt
 Data: Chem97
    AIC      BIC   logLik
 141707 141748.7 -70848.5

Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 1.166198 1.07991
 lea      (Intercept) 0.014766 0.12151
 Residual             5.154202 2.27029
# of obs: 31022, groups: school, 2410; lea, 131

Fixed effects:
             Estimate Std. Error    DF t value  Pr(>|t|)
(Intercept) 5.6355e+00 3.1235e-02 31020  180.42 < 2.2e-16
gcsecnt     2.4726e+00 1.6904e-02 31020  146.27 < 2.2e-16

Correlation of Fixed Effects:
        (Intr)
gcsecnt 0.058
```

4 Two-level models for binary data

The data frame `Contraception` provides data from the Bangladesh fertility survey.

```
> data(Contraception)
> str(Contraception)
`data.frame':        1934 obs. of  6 variables:
 $ woman   : Factor w/ 1934 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ district: Factor w/ 60 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ use     : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ livch   : Factor w/ 4 levels "0","1","2","3+": 4 1 3 4 1 1 4 4 2 4 ...
 $ age     : num   18.44  -5.56   1.44   8.44 -13.56 ...
 $ urban   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
> summary(Contraception[, -1])
    district    use      livch         age             urban
 14     : 118   N:1175   0 :530   Min.   :-13.560000   N:1372
 1      : 117   Y: 759   1 :356   1st Qu.: -7.559900   Y: 562
 46     :  86            2 :305   Median : -1.559900
 25     :  67            3+:743   Mean   :  0.002198
 6      :  65                     3rd Qu.:  6.440000
 30     :  61                     Max.   : 19.440000
 (Other):1420
> sysgc.time(mB1 <- GLMM(use ~ urban + age + livch, binomial,
+     Contraception, ~1 | district))
[1] 0.50 0.00 0.54 0.00 0.00
> summary(mB1)
```

```
Generalized Linear Mixed Model

Family: binomial family with logit link
Fixed: use ~ urban + age + livch
Data: Contraception
      AIC      BIC    logLik
 2429.664 2474.203 -1206.832

Random effects:
     Groups        Name    Variance    Std.Dev.
   district (Intercept)     0.21518     0.46387
# of obs: 1934, groups: district, 60

Estimated scale (compare to 1)  0.9844111

Fixed effects:
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -1.6606460  0.1452147 -11.4358 < 2.2e-16
urbanY       0.7193097  0.1183317   6.0788 1.211e-09
age         -0.0261558  0.0078152  -3.3468 0.0008176
livch1       0.0921026  0.1565011   6.9782 2.989e-12
livch2       1.3545533  0.1729641   7.8314 4.824e-15
livch3+      1.3241531  0.1773558   7.4661 8.262e-14

Correlation of Fixed Effects:
        (Intr) urbanY age    livch1 livch2
urbanY  -0.300
age      0.446 -0.046
livch1  -0.589  0.059 -0.211
livch2  -0.631  0.094 -0.378  0.488
livch3+ -0.748  0.098 -0.674  0.539  0.619
> sysgc.time(mB2 <- GLMM(use ~ urban + age + livch, binomial,
+     Contraception, ~1 | district, method = "Laplace"))
Using optimizer nlm
[1] 30.03  0.03 30.07  0.00  0.00
> summary(mB2)
Generalized Linear Mixed Model

Family: binomial family with logit link
Fixed: use ~ urban + age + livch
Data: Contraception
      AIC      BIC    logLik
 2417.616 2428.750 -1206.808

Random effects:
     Groups        Name    Variance    Std.Dev.
   district (Intercept)     0.21239     0.46086
# of obs: 1934, groups: district, 60

Estimated scale (compare to 1)  0.9859618

Fixed effects:
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -1.6897106  0.1459307 -11.5789 < 2.2e-16
urbanY       0.7329914  0.1192200   6.1482 7.835e-10
age         -0.0265950  0.0078772  -3.3762  0.000735
```

```
livch1       1.1091844  0.1576920   7.0339 2.009e-12
livch2       1.3763954  0.1743346   7.8951 2.900e-15
livch3+      1.3452344  0.1787120   7.5274 5.177e-14

Correlation of Fixed Effects:
        (Intr) urbanY age    livch1 livch2
urbanY  -0.301
age      0.448 -0.046
livch1  -0.589  0.059 -0.210
livch2  -0.631  0.094 -0.378  0.487
livch3+ -0.749  0.099 -0.674  0.538  0.618
> sysgc.time(mB3 <- GLMM(use ~ urban + age + livch, family = binomial,
+     data = Contraception, random = ~urban | district))
[1] 0.61 0.00 0.61 0.00 0.00
> summary(mB3)
Generalized Linear Mixed Model

Family: binomial family with logit link
Fixed: use ~ urban + age + livch
Data: Contraception
      AIC      BIC    logLik
 2225.720 2281.394 -1102.860

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 district (Intercept) 0.38774  0.62269
          urbanY      0.66745  0.81698  -0.793
# of obs: 1934, groups: district, 60

Estimated scale (compare to 1)  0.9759564

Fixed effects:
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -1.6665200  0.1572532 -10.5977 < 2.2e-16
urbanY       0.7914232  0.1681257   4.7073 2.510e-06
age         -0.0258502  0.0079082  -3.2688   0.00108
livch1       1.0987723  0.1580051   6.9540 3.550e-12
livch2       1.3342511  0.1745854   7.6424 2.132e-14
livch3+      1.3227367  0.1795440   7.3672 1.743e-13

Correlation of Fixed Effects:
        (Intr) urbanY age    livch1 livch2
urbanY  -0.481
age      0.416 -0.036
livch1  -0.548  0.038 -0.211
livch2  -0.586  0.068 -0.378  0.487
livch3+ -0.695  0.062 -0.674  0.537  0.616
```

5 Growth curve model for repeated measures data

```
> data(Oxboys)
> str(Oxboys)
```

```
Formal class 'groupedData' [package "lme4"] with 6 slots
  ..@ data   :`data.frame':        234 obs. of  4 variables:
  .. ..$ Subject : Ord.factor w/ 26 levels "10"<"26"<"25"<..: 13 13 13 13 13 13 13 13 13 5 ...
  .. ..$ age     : num [1:234] -1.0000 -0.7479 -0.4630 -0.1643 -0.0027 ...
  .. ..$ height  : num [1:234] 140 143 145 147 148 ...
  .. ..$ Occasion: Ord.factor w/ 9 levels "1"<"2"<"3"<"4"<..: 1 2 3 4 5 6 7 8 9 1 ...
  .. ..- attr(*, "FUN")=function (x)
  .. .. ..- attr(*, "source")= chr "function (x) max(x, na.rm = TRUE)"
  ..@ formula:Class 'formula' length 3 height ~ age | Subject
  .. .. ..- attr(*, ".Environment")=length 31 <environment>
  ..@ outer  :Class 'formula' length 2 ~0
  .. .. ..- attr(*, ".Environment")=length 3 <environment>
  ..@ inner  :Class 'formula' length 2 ~0
  .. .. ..- attr(*, ".Environment")=length 3 <environment>
  ..@ labels :List of 2
  .. ..$ y: chr "Height"
  .. ..$ x: chr "Centered age"
  ..@ units  :List of 1
  .. ..$ y: chr "(cm)"
> sysgc.time(mX1 <- lme(height ~ age + I(age^2) + I(age^3) +
+     I(age^4), Oxboys, ~age + I(age^2) | Subject), gc = TRUE)
[1] 0.18 0.00 0.18 0.00 0.00
> summary(mX1)
Linear mixed-effects model fit by REML
Fixed: height ~ age + I(age^2) + I(age^3) + I(age^4)
 Data: Oxboys
      AIC     BIC    logLik
 651.9081 693.372 -313.9541

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 64.03479 8.00217
          age          2.86418 1.69239  0.614
          I(age^2)     0.67430 0.82115  0.215 0.658
 Residual              0.21737 0.46623
# of obs: 234, groups: Subject, 26

Fixed effects:
             Estimate Std. Error  DF t value  Pr(>|t|)
(Intercept) 149.01887    1.57036 229 94.8945 < 2.2e-16
age           6.17418    0.35650 229 17.3187 < 2.2e-16
I(age^2)      1.12823    0.35144 229  3.2103  0.001516
I(age^3)      0.45385    0.16246 229  2.7937  0.005653
I(age^4)     -0.37690    0.30018 229 -1.2556  0.210552

Correlation of Fixed Effects:
         (Intr) age    I(g^2) I(g^3)
age       0.572
I(age^2)  0.076  0.264
I(age^3) -0.001 -0.340  0.025
I(age^4)  0.021  0.016 -0.857 -0.021
> sysgc.time(mX2 <- lme(height ~ poly(age, 4), Oxboys, ~age +
+     I(age^2) | Subject), gc = TRUE)
[1] 0.1 0.0 0.1 0.0 0.0
> summary(mX2)
```

```
Linear mixed-effects model fit by REML
Fixed: height ~ poly(age, 4)
 Data: Oxboys
      AIC      BIC    logLik
 640.8686 682.3324 -308.4343

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 64.03464 8.00216
          age          2.86418 1.69239  0.614
          I(age^2)     0.67429 0.82115  0.215 0.658
 Residual              0.21737 0.46623
# of obs: 234, groups: Subject, 26

Fixed effects:
               Estimate Std. Error  DF t value  Pr(>|t|)
(Intercept)   149.51976    1.59031 229 94.0194 < 2.2e-16
poly(age, 4)1  64.54095    3.32787 229 19.3941 < 2.2e-16
poly(age, 4)2   4.20322    1.02361 229  4.1063 5.597e-05
poly(age, 4)3   1.29077    0.46628 229  2.7682  0.006098
poly(age, 4)4  -0.58547    0.46630 229 -1.2556  0.210552

Correlation of Fixed Effects:
            (Intr) p(,4)1 p(,4)2 p(,4)3
poly(ag,4)1 0.631
poly(ag,4)2 0.230  0.583
poly(ag,4)3 0.000  0.000  0.000
poly(ag,4)4 0.000  0.000  0.000  0.000
```

6 Cross-classification model

```
> data(ScotsSec)
> str(ScotsSec)

`data.frame':        3435 obs. of  6 variables:
 $ verbal : num  11 0 -14 -6 -30 -17 -17 -11 -9 -19 ...
 $ attain : num  10 3 2 3 2 2 4 6 4 2 ...
 $ primary: Factor w/ 148 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ sex    : Factor w/ 2 levels "M","F": 1 2 1 1 2 2 2 1 1 1 ...
 $ social : num  0 0 0 20 0 0 0 0 0 0 ...
 $ second : Factor w/ 19 levels "1","2","3","4",..: 9 9 9 9 9 9 1 1 9 9 ...

> sysgc.time(mS1 <- lme(attain ~ sex, ScotsSec, ~1 | primary +
+     second))

[1] 0.07 0.00 0.07 0.00 0.00

> summary(mS1)

Linear mixed-effects model fit by REML
Fixed: attain ~ sex
 Data: ScotsSec
      AIC      BIC    logLik
 17137.91 17168.62 -8563.956

Random effects:
 Groups   Name        Variance Std.Dev.
 primary  (Intercept) 1.10962  1.0534
 second   (Intercept) 0.36966  0.6080
```

```
 Residual             8.05511  2.8382
# of obs: 3435, groups: primary, 148; second, 19

Fixed effects:
             Estimate Std. Error   DF t value  Pr(>|t|)
(Intercept) 5.2552e+00 1.8432e-01 3433 28.5108 < 2.2e-16
sexF        4.9851e-01 9.8255e-02 3433  5.0737 4.109e-07

Correlation of Fixed Effects:
     (Intr)
sexF -0.264
```