Examples from Multilevel Software Comparative Reviews

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

April 23, 2005

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.

```
> 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%","mid 50%",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ intake  : Factor w/ 3 levels "bottom 25%","mid 50%",..: 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 ...
```

```
> system.time(Em1 <- lmer(normexam ~ standLRT + sex + schgend +
+     (1 | school), Exam), gc = TRUE)
[1] 0.15 0.00 0.15 0.00 0.00
> summary(Em1)
Linear mixed-effects model fit by REML
Formula: normexam ~ standLRT + sex + schgend + (1 | school)
   Data: Exam
      AIC      BIC    logLik MLdeviance REMLdeviance
 9361.673 9405.834 -4673.837   9325.501     9347.673
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
```

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 <- with(Exam, school: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%","mid 50%",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ intake  : Factor w/ 3 levels "bottom 25%","mid 50%",..: 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 "1:1","1:4","1:6",..: 48 49 47 46 45 50 51 39 40 38 ...
```

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: 1:1 1:4 1:6 1:7 1:13 1:14 1:16 1:17 1:19 1:22 1:27 ... 65:155
```

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

These results are from the 1997 A-level Chemistry exam. The `school` is nested in `lea` (local education authority) and has unique levels for each of the 2410 schools. It is a good practice to make the nesting explicit by specifying the grouping factors as the 'outer' factor, `lea` in this case, and the interaction of the outer and inner factors, `lea:school` or `school:lea` in this case. This will ensure unique levels for each `school` within `lea` combination.

To fit the model `mC2` we increase the number of EM iterations from its default of 20 to 40. Without this change the current version of the `optim` function in R will declare convergence to an incorrect optimum. By increasing the number of EM iterations we are able to get closer to the optimum before calling `optim` and converge to the correct value. The optim function will be patched so this change will not be needed in future versions of R.

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

```
> system.time(mC1 <- lmer(score ~ 1 + (1 | lea:school) + (1 |
+     lea), Chem97), gc = TRUE)
```

```
[1] 3.94 0.07 4.20 0.00 0.00
```

```
> summary(mC1)
```

```
Linear mixed-effects model fit by REML
Formula: score ~ 1 + (1 | lea:school) + (1 | lea)
   Data: Chem97
      AIC      BIC   logLik MLdeviance REMLdeviance
 157881.8 157915.2 -78936.9   157869.9     157873.8
Random effects:
 Groups     Name        Variance Std.Dev.
 lea:school (Intercept) 2.74981  1.6583
```

```
 lea        (Intercept) 0.15343  0.3917
 Residual               8.51591  2.9182
# of obs: 31022, groups: lea:school, 2410; lea, 131

Fixed effects:
             Estimate Std. Error    DF t value  Pr(>|t|)
(Intercept) 5.3189e+00 5.8108e-02 31021  91.536 < 2.2e-16
> system.time(mC2 <- lmer(score ~ gcsecnt + (1 | school) +
+     (1 | lea), Chem97, control = list(niterEM = 40)), gc = TRUE)
[1] 1.35 0.00 1.35 0.00 0.00
> summary(mC2)
Linear mixed-effects model fit by REML
Formula: score ~ gcsecnt + (1 | school) + (1 | lea)
   Data: Chem97
      AIC      BIC    logLik MLdeviance REMLdeviance
 141707.2 141748.9 -70848.58   141685.8     141697.2
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 1.163183 1.07851
 lea      (Intercept) 0.020849 0.14439
 Residual             5.153861 2.27021
# of obs: 31022, groups: school, 2410; lea, 131

Fixed effects:
             Estimate Std. Error    DF t value  Pr(>|t|)
(Intercept) 5.6377e+00 3.2353e-02 31020  174.26 < 2.2e-16
gcsecnt     2.4726e+00 1.6907e-02 31020  146.25 < 2.2e-16

Correlation of Fixed Effects:
        (Intr)
gcsecnt 0.056
```

4 Two-level models for binary data

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

```
> 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
```

5 Growth curve model for repeated measures data

```
> str(Oxboys)
`data.frame':        234 obs. of  4 variables:
 $ Subject : Factor w/ 26 levels "1","10","11",..: 1 1 1 1 1 1 1 1 1 12 ...
 $ age     : num  -1.0000 -0.7479 -0.4630 -0.1643 -0.0027 ...
 $ height  : num  140 143 145 147 148 ...
 $ Occasion: Factor w/ 9 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 1 ...
 - attr(*, "ginfo")=List of 7
  ..$ formula     :Class 'formula' length 3 height ~ age | Subject
  .. .. ..- attr(*, ".Environment")=length 4 <environment>
  ..$ order.groups: logi TRUE
  ..$ FUN         :function (x)
  .. ..- attr(*, "source")= chr "function (x) max(x, na.rm = TRUE)"
  ..$ outer       : NULL
  ..$ inner       : NULL
  ..$ labels      :List of 2
  .. ..$ age   : chr "Centered age"
  .. ..$ height: chr "Height"
  ..$ units       :List of 1
  .. ..$ height: chr "(cm)"
> system.time(mX1 <- lmer(height ~ age + I(age^2) + I(age^3) +
+     I(age^4) + (age + I(age^2) | Subject), Oxboys), gc = TRUE)
[1] 0.41 0.00 0.42 0.00 0.00
> summary(mX1)
Linear mixed-effects model fit by REML
Formula: height ~ age + I(age^2) + I(age^3) + I(age^4) + (age + I(age^2) |      Subject)
   Data: Oxboys
      AIC     BIC    logLik MLdeviance REMLdeviance
 651.9081 693.372 -313.9541   625.3593     627.9081
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 64.03130 8.00196
          age          2.86408 1.69236  0.614
          I(age^2)     0.67428 0.82115  0.215 0.658
 Residual              0.21738 0.46624
# of obs: 234, groups: Subject, 26

Fixed effects:
             Estimate Std. Error  DF t value  Pr(>|t|)
(Intercept) 149.01887    1.57032 229 94.8971 < 2.2e-16
age           6.17418    0.35650 229 17.3190 < 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  0.210554

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
> system.time(mX2 <- lmer(height ~ poly(age, 4) + (age + I(age^2) |
+     Subject), Oxboys), gc = TRUE)
```

```
[1] 0.39 0.01 0.39 0.00 0.00
```

```
> summary(mX2)
```

```
Linear mixed-effects model fit by REML
Formula: height ~ poly(age, 4) + (age + I(age^2) | Subject)
   Data: Oxboys
      AIC      BIC    logLik MLdeviance REMLdeviance
 640.8686 682.3324 -308.4343   625.3593     616.8686
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 64.03114 8.00195
          age          2.86407 1.69236  0.614
          I(age^2)     0.67428 0.82115  0.215 0.658
 Residual              0.21738 0.46624
# of obs: 234, groups: Subject, 26

Fixed effects:
              Estimate Std. Error  DF t value  Pr(>|t|)
(Intercept)   149.51976    1.59026 229 94.0222 < 2.2e-16
poly(age, 4)1  64.54095    3.32780 229 19.3945 < 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  0.210554

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

```
> 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 ...
```

```
> system.time(mS1 <- lmer(attain ~ sex + (1 | primary) + (1 |
+     second), ScotsSec), gc = TRUE)
```

```
[1] 0.21 0.00 0.21 0.00 0.00
```

```
> summary(mS1)
```

```
Linear mixed-effects model fit by REML
Formula: attain ~ sex + (1 | primary) + (1 | second)
   Data: ScotsSec
      AIC      BIC    logLik MLdeviance REMLdeviance
 17137.91 17168.62 -8563.956   17123.49     17127.91
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 28.5107 < 2.2e-16
sexF        4.9851e-01 9.8255e-02 3433  5.0737 4.109e-07

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