Case Studies for Linear Mixed Models

Keunbaik Lee
Sungkyunkwan University

keunbaik@skku.edu

October 16, 2021

General Guidelines I

- Unlike simple linear regression models, for correlated data we need pay attention to both the mean model and the variance model.
- When the mean model is of primary interest, it may be sufficient to use a simple variance model and use empirical variances to achieve valid inference. Still it might be worthwhile to find an appropriate variance model to improve efficiency.
- When the variance model is also of interest, car must be taken to model it correctly. In addition, the mean model is also critical. When the wrong mean model is used, the variance estimation will not even be consistent.
- Typically the model building process involves the following steps:
  1. Fit an over-elaborated ("saturated") mean model with simple covariance structure (e.g., working independence).

General Guidelines II

2. Use the residuals to explore the variance structure and select a covariance model.
3. Refit the over-elaborated model with the covariance model to see if the goodness-of-fit is adequate.
4. If yes, then try to simplify the mean model. Otherwise repeat the modeling process.

- Keep in mind that modeling is the means not the end. Goodness-of-fit is not the ultimate criterion for selecting models. Simplicity and interpretability are just as important, if not more so. Address the scientific equation of interest.

Motivation I

Recall the orthodontic measurement data, as shown in the following figure. One question of interest is the individual **growth curve.**

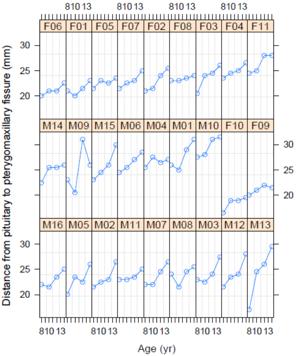


Figure: Orthodontic distance measurements

Motivation II

Let's consider only the girls for the moment and the three modeling strategies:

1. Two-stage analysis: fit a linear regression line to each subject and analyze the resulted slopes as responses in the second stage analysis.
2. Include an indicator variable for subject id in the regression.
3. Linear mixed model.

Two-Stage Analysis

```
> library(nlme)
> data(Orthodont)
>
> # Two Stage Analysis
> OrthFem <- subset(Orthodont,Sex=="Female")
> OrthFem[1:5,]
Grouped Data: distance ~ age | Subject
   distance age Subject    Sex
65     21.0   8     F01 Female
66     20.0  10     F01 Female
67     21.5  12     F01 Female
68     23.0  14     F01 Female
69     21.0   8     F02 Female
>
> of.lis <-lmList(distance~I(age-11),data=OrthFem)
> coef(of.lis)
    (Intercept) I(age - 11)
F10      18.500       0.450
F09      21.125       0.275
F06      21.125       0.375
F01      21.375       0.375
F05      22.625       0.275
F07      23.000       0.550
F02      23.000       0.800
F08      23.375       0.175
F03      23.750       0.850
F04      24.875       0.475
F11      26.375       0.675
> plot(intervals(of.lis))
```

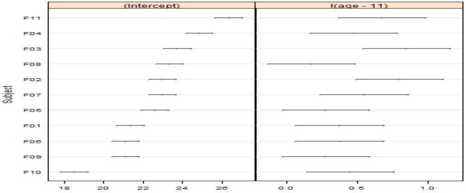


Figure: Confidence intervals (95%) for the coefficients of simple linear models

- There is a lot of variation in the intercepts and slopes are relatively comparable.
- Intuitively, we know this approach is not very efficient, since for every subjects we are estimating two parameters (not counting the standard errors), that are 22 parameters.
- If the data is not balanced, then the individual growth curve parameters are estimated at different precision and we need weight them differently in the subsequent analysis.
- We already know the method (which includes only 2 parameters) that ignores the correlation and uses OLS, is not very efficient.

Fixed Effects

We can include an indicator for subject, thus allow each to have a different intercept.

```
> # Fixed Effects
> of.lm <- lm(distance~factor(Subject,ordered=FALSE)+I(age-11)-1,data=
>
> library(MASS)
> confint(of.lm)
                                         2.5 %     97.5 %
factor(Subject, ordered = FALSE)F10 17.7055622 19.2944378
factor(Subject, ordered = FALSE)F09 20.3305622 21.9194378
factor(Subject, ordered = FALSE)F06 20.3305622 21.9194378
factor(Subject, ordered = FALSE)F01 20.5805622 22.1694378
factor(Subject, ordered = FALSE)F05 21.8305622 23.4194378
factor(Subject, ordered = FALSE)F07 22.2055622 23.7944378
factor(Subject, ordered = FALSE)F02 22.2055622 23.7944378
factor(Subject, ordered = FALSE)F08 22.5805622 24.1694378
factor(Subject, ordered = FALSE)F03 22.9555622 24.5444378
factor(Subject, ordered = FALSE)F04 24.0805622 25.6694378
factor(Subject, ordered = FALSE)F11 25.5805622 27.1694378
I(age - 11)                          0.3724235  0.5866674
> apply(intervals(of.lis)[,,2],2,mean)
    lower      est.     upper
0.1696451 0.4795455 0.7894458
```

- Here we are estimating 12 parameters (11 intercepts and one slope).
- The precision on the slope is substantially better (sd 0.05 vs 0.220).
- The intercepts (and CIs) are similar to those in separate regressions.
- However the intercepts in this model do not have the interpretation as population parameters.

Linear Mixed Model

One solution is to use a random effect for subjects.

```
> # Linear Mixed Model
> of.lme <- lme(distance~I(age-11),random=~1|Subject,data=OrthFem)
> intervals(of.lme)
Approximate 95% confidence intervals

 Fixed effects:
                 lower       est.      upper
(Intercept) 21.3549737 22.6477273 23.9404809
I(age - 11)  0.3724235  0.4795455  0.5866674
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                   lower    est.    upper
sd((Intercept)) 1.313761 2.06847 3.256733

 Within-group standard error:
    lower      est.     upper
0.6105468 0.7800331 0.9965684
> orth.i <-cbind(two.stage=coef(of.lis)[,1],
+                fixed=coef(of.lm)[1:11],
+                random=coef(of.lme)[,1])
> rownames(orth.i) <-NULL
> orth.i
      two.stage  fixed   random
 [1,]    18.500 18.500 18.64240
 [2,]    21.125 21.125 21.17728
 [3,]    21.125 21.125 21.17728
 [4,]    21.375 21.375 21.41869
 [5,]    22.625 22.625 22.62578
 [6,]    23.000 23.000 22.98791
 [7,]    23.000 23.000 22.98791
 [8,]    23.375 23.375 23.35003
 [9,]    23.750 23.750 23.71216
[10,]    24.875 24.875 24.79853
[11,]    26.375 26.375 26.24704
>
```

- The estimate and CI for the slope are very close to the previous model.
- The std. dev. for the random effects (2.07) is slightly smaller than the std. dev. for the intercepts in the previous model (2.10).
- The intercepts are “shrunk” toward the mean.
- At first look, there is one (variance) parameter for the intercepts instead of 11. But there is no free lunch.

Grouped Data Object in nlme

```
> setwd("d:/course/SKKU/Longitudinal_Data_Analysis/2015Fall/R-codes")
>
> library(nlme)
>
> #------------------------------------------#
> ## Grouped data
>
> tracking <- read.table ("tracking.dat", header = TRUE)
> tracking
    Sex Age  Shape Trial1 Trial2 Trial3 Trial4
1     M  31    Box   2.68   4.14   7.22   8.00
2     M  30    Box   7.09   8.55   8.79   9.68
3     M  30    Box   6.05   6.25   7.04   7.80
4     M  27    Box   4.35   6.50   5.17   6.50
5     M  30    Box   4.08   6.00   6.82   6.68

>
> tracklong <- reshape (tracking, direction = "long",
+  varying = 4:7, times = 1:4,
+  split = list (regexp = "l", include = TRUE))
>
> tracklong[1:10,]
     Sex Age Shape time Trial id
1.1     M  31   Box    1  2.68  1
2.1     M  30   Box    1  7.09  2
```

```
3.1    M  30   Box    1  6.05  3
4.1    M  27   Box    1  4.35  4
5.1    M  30   Box    1  4.08  5
6.1    M  28   Box    1  8.22  6
7.1    M  34   Box    1  4.51  7
8.1    M  28   Box    1  7.36  8
9.1    M  28   Box    1  3.34  9
10.1   M  33   Box    1  7.19 10
>
> tracklong <- tracklong[order (tracklong$id, tracklong$time),]
> tracklong <- groupedData (Trial ~ time | id, data = tracklong,
+  outer = ~ Sex * Shape)
>
> gsummary (tracklong)
    Sex Age  Shape time   Trial  id
36    F   6    Box  2.5  0.1475  36
41    F   5    Box  2.5  0.3375  41
42    F  45    Box  2.5  0.4075  42
13    F   7    Box  2.5  0.4550  13
38    F  45    Box  2.5  1.4700  38

> gsummary (tracklong, inv = TRUE, omit = TRUE)
    Sex Age  Shape
36    F   6    Box
41    F   5    Box
42    F  45    Box
```

```
13     F   7     Box
38     F  45     Box

> gapply (tracklong, "Trial", sd)
  36.Trial   41.Trial   42.Trial   13.Trial   38.Trial   60.Trial   47
0.09569918 0.17192537 0.16276261 0.29285947 0.78866977 0.75243494 0.36
...

> plot (tracklong, outer = TRUE, aspect = "fill",
+  xlab = "Trial", ylab = "Contact Time (sec)",
+  auto.key = FALSE, key = NULL)
>
> track.sum <- gsummary(tracklong)
```

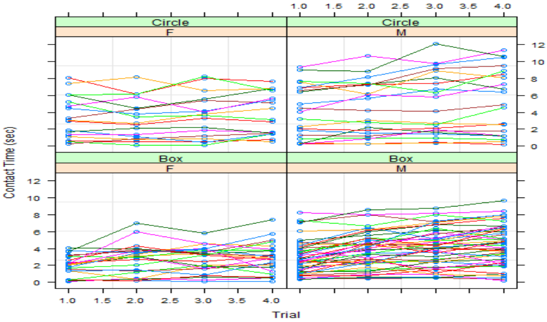


Figure: Tracking data

Orthodontic Data

```
> ## Orthodontic Data
>
> data(Orthodont)
> o10.lm <- lm(distance ~ age * Sex, data = Orthodont)
> summary(o10.lm)

Call:
lm(formula = distance ~ age * Sex, data = Orthodont)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6156 -1.3219 -0.1682  1.3299  5.2469 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    16.3406     1.4162  11.538  < 2e-16 ***
age             0.7844     0.1262   6.217 1.07e-08 ***
SexFemale       1.0321     2.2188   0.465    0.643    
age:SexFemale  -0.3048     0.1977  -1.542    0.126    
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯

Residual standard error: 2.257 on 104 degrees of freedom
Multiple R-squared: 0.4227,     Adjusted R-squared: 0.4061 
F-statistic: 25.39 on 3 and 104 DF,  p-value: 2.108e-12
```

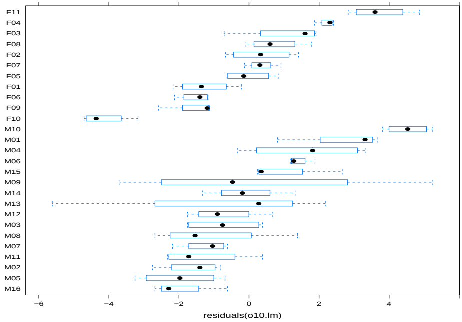
```
>
> anova(o10.lm)
Analysis of Variance Table

Response: distance
           Df Sum Sq Mean Sq F value    Pr(>F)
age         1 235.36 235.356 46.2042 6.884e-10 ***
Sex         1 140.46 140.465 27.5756 8.054e-07 ***
age:Sex     1  12.11  12.114  2.3782    0.1261
Residuals 104 529.76   5.094
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯
>
> drop1(o10.lm, scope = c("Sex", "age:Sex"), test = "F")
Single term deletions

Model:
distance ~ age * Sex
        Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>               529.76 179.75
Sex      1    1.1022 530.86 177.98  0.2164 0.6428
age:Sex  1   12.1142 541.87 180.19  2.3782 0.1261
> o20.lm <- update(o10.lm, ~ . - age:Sex)
> summary(o20.lm)
```

```
Call:
lm(formula = distance ˜ age + Sex, data = Orthodont)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9882 -1.4882 -0.0586  1.1916  5.3711

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.70671    1.11221  15.920  < 2e-16 ***
age          0.66019    0.09776   6.753 8.25e-10 ***
SexFemale   -2.32102    0.44489  -5.217 9.20e-07 ***
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯

Residual standard error: 2.272 on 105 degrees of freedom
Multiple R-squared: 0.4095,     Adjusted R-squared: 0.3983
F-statistic: 36.41 on 2 and 105 DF,  p-value: 9.726e-13
```

There is a gender and age effect.

```
> library(lattice)
> pdf("ortho_lm.pdf", height = 8, width = 8)
> bwplot(getGroups (Orthodont) ˜ residuals (o10.lm))
> dev.off ()
windows
      2
```

```
>
> o10.lis <- lmList(distance ˜ age, data = Orthodont)
> pairs(o10.lis, id = 0.01, adj = -0.5)
> plot(intervals (o10.lis))
```



- The residuals from the same subject tend to have the same sign, indicating some “subject effect”.
- There is not much correlation between the intercept and slope estimates (after re-centering the age).
- A random intercept is perhaps needed.
- The boys seem to have larger intercept. We have not put gender into the mean model yet.

Fitting Linear Mixed Model with lme

The function call has the form:

```
lme(fixed, data, random)
```

A model with both random intercept and slope (The intercept “1” is often omitted from the model formula).

```
> # Fitting linear mixed model with lme
> o10.lme <- lme(distance ~ I(age - 11),
+  data = Orthodont,
+  random = ~ I(age - 11) | Subject)
> summary(o10.lme)
Linear mixed-effects model fit by REML
 Data: Orthodont
      AIC      BIC    logLik
  454.6367 470.6173 -221.3183

Random effects:
 Formula: ~I(age - 11) | Subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2.1343289 (Intr)
I(age - 11) 0.2264278 0.503
Residual    1.3100402
```

```
Fixed effects: distance ~ I(age - 11) 
                Value Std.Error DF  t-value p-value
(Intercept) 24.023148 0.4296601 80 55.91198       0
I(age - 11)  0.660185 0.0712533 80  9.26533       0
 Correlation: 
            (Intr)
I(age - 11) 0.294 

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max 
-3.223106872 -0.493760899  0.007316483  0.472151219  3.916031757 

Number of Observations: 108
Number of Groups: 27 
```

Residuals

For random effects model the residuals can be defined at different levels. The population level (marginal, level 0) residuals are given by

$$r^0 = y - X\beta,$$

and are estimated by

$$\hat{r}^0 = y - X\hat{\beta},$$

- The variance of the population residuals can be estimated, and the residuals can be standardized.
- The effect of estimating the variances of the residuals is small when m is large.
- The mean structure can be investigated using population residuals.

The **subject specific** (conditional, level 1) residuals are given by

$$r^1 = y - X\beta - Zb,$$

and are estimated by

$$\hat{r}^1 = y - X\hat{\beta} - Z\hat{b},$$

where $\hat{b}$ is the BLUP of b.

- When n_i is small, b_i and r^1 will be poorly estimated.

```
> # Residuals
> fitted(o20.lme, level = 0:1)
      fixed   Subject
1   22.61562 24.84572
2   24.18437 26.57649
3   25.75312 28.30725
4   27.32187 30.03802
5   22.61562 21.27478
6   24.18437 22.79641
7   25.75312 24.31803
8   27.32187 25.83966
9   22.61562 22.03311
10  24.18437 23.56449
```

```
11  25.75312 25.09588

> resid(o20.lme, level = 1, type = "p")
         M01          M01          M01          M01          M02
 0.881105017 -1.203387894  0.528797575  0.734311693  0.171916601 -0.22
 ....
attr(,"label")
[1] "Standardized residuals"
>
> newO <- data.frame(Subject = rep (c ("M11", "F03"), each = 3),
+  Sex = rep(c ("Male", "Female"), each = 3),
+  age = rep(16:18, 2))
> predict(o20.lme, newdata = newO)
     M11      M11      M11      F03      F03      F03
26.96809 27.61195 28.25580 26.61357 27.20668 27.79979
attr(,"label")
[1] "Predicted values (mm)"
>
> # Prediction
> predict(o20.lme, newdata = newO, level = 0:1)
  Subject predict.fixed predict.Subject
1     M11      28.89062        26.96809
2     M11      29.67500        27.61195
3     M11      30.45937        28.25580
4     F03      25.04545        26.61357
5     F03      25.52500        27.20668
```

```
6     F03      26.00455       27.79979
> # We see the shrinkage when comparing the predicted random effects w
> # the individual regression coefficients.
>
> o10.lis <- lmList(distance ~ I(age-11), data = Orthodont) # Linear M
> comp0 <- compareFits(coef(o10.lis), coef(o10.lme))
> comp0
># plot(comp0,mark=fixef(o10.lme))
>
>
> plot(comparePred (o10.lis, o10.lme), length.out = 2,
+  lty = 1:2, lwd = 1.5,
+  col = "black", layout = c(8, 4), between = list (y = c(0, 0.5)))
>
> o20.lme <- update(o10.lme, distance ~ I(age - 11) * Sex)
> o20.lmeM <- update(o20.lme, method = "ML")
> plot(compareFits (ranef (o20.lme), ranef (o20.lmeM)),
+ mark = c(0, 0))
>
> o10.lm2 <- lm(distance ~ I(age - 11) * Sex, data = Orthodont)
> anova(o20.lme, o10.lm2)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
o20.lme     1  8 448.5817 469.7368 -216.2908
o10.lm2     2  5 493.5591 506.7811 -241.7796 1 vs 2 50.97746  <.0001
>
> # For groupedData
> o30.lme <- update(o10.lme, random = pdDiag (~ I(age - 11)))
> # otherwise
> o30.lme <- update(o10.lme,
+    random = list (Subject = pdDiag (~ I(age - 11))))
> summary(o30.lme)
Linear mixed-effects model fit by REML
 Data: Orthodont
       AIC     BIC    logLik
  454.9848 468.302 -222.4924

Random effects:
 Formula: ~I(age - 11) | Subject
 Structure: Diagonal
        (Intercept) I(age - 11) Residual
StdDev:    2.134331   0.2264279  1.31004

Fixed effects: distance ~ I(age - 11)
                Value Std.Error DF  t-value p-value
(Intercept) 24.023148 0.4296605 80 55.91193       0
I(age - 11)  0.660185 0.0712533 80  9.26533       0
 Correlation:
            (Intr)
I(age - 11) 0

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
```

```
-2.9027334 -0.4862659  0.0371326  0.4288734  3.9631748

Number of Observations: 108
Number of Groups: 27
> anova(o10.lme, o30.lme)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
o10.lme      1  6 454.6367 470.6173 -221.3183
o30.lme      2  5 454.9848 468.3020 -222.4924 1 vs 2 2.348112  0.1254
>
>
> o40.lme <- update(o10.lme, random = ˜ 1 | Sex / Subject)
> summary(o40.lme)
Linear mixed-effects model fit by REML
 Data: Orthodont
       AIC      BIC    logLik
  452.0344 465.3516 -221.0172

Random effects:
 Formula: ˜1 | Sex
        (Intercept)
StdDev:    1.550378

 Formula: ˜1 | Subject %in% Sex
        (Intercept) Residual
StdDev:    1.807424 1.431592
```

```
Fixed effects: distance ~ I(age - 11)
                Value Std.Error DF  t-value p-value
(Intercept) 23.831367 1.1602756 80 20.53940       0
I(age - 11)  0.660185 0.0616059 80 10.71626       0
 Correlation:
            (Intr)
I(age - 11) 0

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.73925835 -0.54662107 -0.01599557  0.45199558  3.66710262

Number of Observations: 108
Number of Groups:
             Sex Subject %in% Sex
               2               27
> anova(o40.lme, o10.lme)
        Model df      AIC      BIC    logLik   Test   L.Ratio p-value
o40.lme     1  5 452.0344 465.3516 -221.0172
o10.lme     2  6 454.6367 470.6173 -221.3183 1 vs 2 0.6022852  0.4377
> # Simple model is better.
>
> ranef(o40.lme, levels = 1:2)
Level: Sex
       (Intercept)
Male      1.035618
```

```
Female   -1.035618

Level: Subject %in% Sex
            (Intercept)
Male/M16   -1.613865466
Male/M05   -1.613865466
Male/M02   -1.289706704
Male/M11   -1.073600862
Male/M07   -0.965547941
Male/M08   -0.857495021
```

Model Diagnosis

Two important assumptions

1. The within-group errors are iid $N(0, \sigma^2)$ and independent of the random effects.
2. The random effects are normally distributed with mean 0 and a covariance matrix D that does not depend the subject and the random effects are independent (are they identically distributed?) for different subjects.

```
> # Model Diagnosis
> plot(o20.lme, Subject ˜ resid (.), abline = 0)
> plot(o20.lme, resid (., type = "p") ˜ fitted(.) | Sex,id = 0.05)
```

In the general form of linear mixed model:

$$Y_i|b_i \sim N(X_i\beta + Z_ib_i, \sigma^2 B_iC_iB_i),$$

where B is a diagnoal matrix of “weights” to allow heteroscedasticity of the within group errors.

```
> o25.lme <- update(o20.lme, weights = varIdent (form = ˜ 1 | Sex))
> summary(o25.lme)
Linear mixed-effects model fit by REML
 Data: Orthodont
       AIC     BIC    logLik
  429.5225 453.322 -205.7612

Random effects:
 Formula: ˜I(age - 11) | Subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 1.854979 (Intr)
I(age - 11) 0.156517 0.394
Residual    1.629585

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ˜1 | Sex
 Parameter estimates:
     Male    Female
1.0000000 0.4088464
Fixed effects: distance ˜ I(age - 11) + Sex + I(age - 11):Sex
                          Value Std.Error DF  t-value p-value
(Intercept)           24.968750 0.5065098 79 49.29569  0.0000
I(age - 11)            0.784375 0.0991448 79  7.91141  0.0000
SexFemale             -2.321023 0.7612188 25 -3.04909  0.0054
I(age - 11):SexFemale -0.304830 0.1186356 79 -2.56946  0.0121
```

```
 Correlation:
                      (Intr) I(g-11) SexFml
I(age - 11)            0.142
SexFemale             -0.665 -0.095
I(age - 11):SexFemale -0.119 -0.836   0.194

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.89845602 -0.50012102  0.03984999  0.51833974  3.10719509

Number of Observations: 108
Number of Groups: 27
>
> anova(o25.lme, o20.lme)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
o25.lme     1  9 429.5225 453.3220 -205.7612
o20.lme     2  8 448.5817 469.7368 -216.2908 1 vs 2 21.05918  <.0001
> qqnorm(o25.lme, ~ resid (.) | Sex)
```

Checking the Random Effects

We can use Q-Q plots and conditional plots to check the normality and homogeneity of the random effects. However, as we cautioned earlier, these assumptions are harder to check.

```
# Checking the random effects
qqnorm(o20.lme, ~ ranef (.), id = 0.10)
qqnorm(o25.lme, ~ ranef (.), id = 0.10)
pairs(o20.lme, ~ ranef (.) | Sex, id = ~ Subject == "M13")
pairs(o25.lme, ~ ranef (.) | Sex, id = ~ Subject == "M13")
```

- The heteroscedasticity model accommodates the boys' outlying observations with increasing within-group error variance, thus reducing the between-group variance, thus more shrinkage.
- Note here everyone has the same set of covariates so the random effects should be iid. In general it might be necessary to standardize the random effects.

The Variance (Weight) Function

The general variance function for the within-group errors is defined as

$$var(\epsilon_{ij}|b_i) = \sigma^2 g^2(\mu_{ij}, v_{ij}, \delta),$$

for $i = 1, \cdots, m$; $j = 1, \cdots, n_j$, where

$$\mu_{ij} = E(Y_{ij}|b_i),$$

and v_{ij} are covariates and δ are parameters.

- Note that in this general form, ϵ_i and b_i are no longer independent. The assumption is

  $$E(\epsilon_i|b_i) = 0$$

  and it follows that

  $$var(\epsilon_{ij}) = E(var(\epsilon_{ij}|b_i)).$$

- This model introduces some difficulties since integrating out b_i is not always feasible (for nonlinear models). So in nlme, an approxomation is used:

$$var(\epsilon_{ij}|b_i) \approx \sigma^2 g^2(\hat{\mu}_{ij}, v_{ij}, \delta).$$

Variance Functions in nlme

In nlme, the variance functions are provided as varFunc classes. Some examples are:

- Fixed (varFixed): the within-group variance is proportional to some covariates, e.g.,

  $$var(\epsilon_{ij}) = \sigma^2 \text{Age}_{ij} \text{ or } g(\text{Age}_{ij}) = \sqrt{\text{Age}_{ij}}.$$

  It is represented as varFixed($\sim$ Age).
- Different variances per stratum (varIdent): the within-group variances are different for each level of a class variable s:

  $$g(s_{ij}, \delta) = \delta_{s_{ij}},$$

  where by default $\delta_1 = 1$.

- Other possible choices are: varPower, varExp, varConstPower and varComb, the last one being a combination of other functions.
- Note: the variance functions are also available for general linear models fitted with gls (without random effects).

Correlation Functions in nlme

In nlme, correlation structures are specified using the corStruct class. Some examples are:

- Compound symmetry (exchangeable, varCompSymm), e.g.,

  ```
  corCompSymm(~ 1 | Subject)
  ```

  which says with-subject correlation is ρ.
- Autocorrelation of order 1 (AR1): varAR1.

It is often desirable to specify an initial value of the correlation parameter using the value argument of the correlation object constructor.

nlme also provies functions to calculate auto-correlation function (ACF) and the variogram Variogram.

Multicenter AIDS Cohort Study: CD4+ Data

```
> library(nlme)
> CD4 <- read.table("cd4.dat",header=TRUE)
> CD4g <- groupedData(CD4 ~ Time | ID, data = CD4, FUN = median,
+  labels = list (x = "Time since seroconversion",
+  outer = ~ Age,
+  labels = list (y = "CD4+ Cell Number")),
+  units = list (x = "(yr)", y = ""))
> gsummary(CD4g, inv = TRUE, omit = TRUE)[1:10,,drop = FALSE]
        Age
20089  6.31
40445  0.02
20498  4.78
10915  0.32
20014  1.79
41416 -4.30
30048 -3.64
40970 -0.04
20323 -1.76
30827 11.53
> gsummary(CD4g, FUN = function (x) max (x, na.rm = TRUE))[1:10,]
          Time  CD4   Age Packs Drugs Sex Cesd    ID
20089 2.332649  641  6.31     3     1   5    8 20089
40445 4.917180  356  0.02     0     1   0    4 40445
20498 1.806982  823  4.78     0     1   5   17 20498
10915 4.123203  773  0.32     0     1   5   16 10915
```

```
20014 1.872690  913  1.79     1     1  -2   11 20014
41416 3.197810  511 -4.30     0     1  -1    6 41416
30048 4.065709  547 -3.64     0     1   5   24 30048
40970 4.065709  672 -0.04     4     0   0   -1 40970
20323 1.177276 1038 -1.76     0     1   5   37 20323
30827 3.436003  505 11.53     0     1   5   17 30827
> gsummary(CD4g, FUN = function (x) min (x, na.rm = TRUE))[1:10,]
           Time CD4   Age Packs Drugs Sex Cesd    ID
20089 -0.251882  52  6.31     0     0  -2   -5 20089
40445 -0.394251 187  0.02     0     0  -5   -5 40445
20498 -0.273785 123  4.78     0     0  -3    4 20498
10915 -0.758385 139  0.32     0     0  -4   -6 10915
20014 -1.341547 224  1.79     0     1  -4    1 20014
41416 -0.725530  89 -4.30     0     0  -4   -5 41416
30048 -0.249144  39 -3.64     0     0  -4   -5 30048
40970 -0.999316 159 -0.04     3     0  -5   -7 40970
20323 -0.791239  43 -1.76     0     0  -3   -3 20323
30827 -0.249144 101 11.53     0     0  -4   -6 30827
>
> CD4$Time2 <- ifelse(CD4$Time < 0, 0, CD4$Time)
> cd4.lm <- lm(I(sqrt (CD4)) ~ Cesd + Drugs + Sex + Packs +
+  Time2 + I(Time2^2), data = CD4)
> summary(cd4.lm)

Call:
lm(formula = I(sqrt(CD4)) ~ Cesd + Drugs + Sex + Packs + Time2 +
    I(Time2ˆ2), data = CD4)

Residuals:
     Min       1Q   Median       3Q      Max 
-21.5151  -4.0749  -0.4008   3.7172  27.9015 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 28.63913    0.30257  94.654  < 2e-16 ***
Cesd        -0.03455    0.01310  -2.637  0.00842 ** 
Drugs        0.93519    0.29720   3.147  0.00167 ** 
Sex         -0.05574    0.03698  -1.507  0.13186    
Packs        0.97146    0.08753  11.099  < 2e-16 ***
Time2       -4.98658    0.27770 -17.957  < 2e-16 ***
I(Time2ˆ2)   0.75434    0.06654  11.337  < 2e-16 ***
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯

Residual standard error: 6.04 on 2369 degrees of freedom
Multiple R-squared:  0.27,      Adjusted R-squared: 0.2681 
F-statistic:   146 on 6 and 2369 DF,  p-value: < 2.2e-16 

> par(mfrow = c(2, 2))
> plot (cd4.lm)
> temp <- subset(CD4, Time < 4)
> cd4.lmt <- lm(I(sqrt (CD4)) ˜ Time2 + I(Time2ˆ2), data = temp)
> temp$fitted <- fitted(cd4.lmt)^2
> temp <- temp[order (temp$Time),]
> plot(CD4 ~ Time, data = temp, col = "gray50", pch = ".",
+  xlab = "Years since seroconversion", ylab = "CD4+ cell number")
> lines(temp$fitted ~ temp$Time)
>
> CD4.lst <- lmList(I(sqrt (CD4)) ~ Time2 + I(Time2^2)|ID, data = CD4)
> plot(intervals (CD4.lst), layout = c(1, 3))
```

Correlation Structure

Consider a stochastic process $Y(t)$, the autocovariance function is defined as:

$$\begin{aligned}\gamma(t,u) &= cov\left\{Y(t), Y(t-u)\right\} \\ &= E\left\{Y(t)-\mu(t)\right\}\left\{Y(t-u)-\mu(t-u)\right\},\end{aligned}$$

where u is the “lag” and

$$Y(t) = \mu(t) + r(t)$$

and $\mu(t)$ is the trend and $r(t)$ is the residual process such that $E\left\{r(t)\right\} = 0$.

- The process (second-order) stationary if $\gamma(t,u)$ depends only on u (which we take to be positive).
- For stationary process, $\gamma(0)$ is the variance of $Y(t)$ for all t and the autocorrelation function is

$$\rho(u) = \frac{\gamma(u)}{\gamma(0)}.$$

- The autocorrelation function is most useful for equally spaced data. For $t = 1, \ldots, n$, the residuals are

$$r_t = \frac{y_t - \hat{y}_t}{\sqrt{\widehat{var}(Y_t)}}.$$

The empirical autocorrelation function is

$$\hat{\rho}(u) \quad = \quad \widehat{corr}(r_t, r_{t-u}) = \frac{\frac{1}{n-u}\sum_{t=u+1}^{n} r_t r_{t-u}}{\frac{1}{n}\sum_{t=1}^{n} r_t^2}.$$

In Orthodont data,

```
> o10.lme <- lme(distance ~ I(age - 11),
+  data = Orthodont,
+  random = ~ I(age - 11) | Subject)
> o15.lme <- update(o10.lme, distance ~ I(age - 11) + Sex)
> ACF(o15.lme)
  lag          ACF
1   0  1.000000000
2   1 -0.480774256
3   2  0.008214159
4   3 -0.261229105
> plot(ACF (o15.lme), alpha = 0.05)
```

Variogram

Variogram is essier to handle for unequally spaced data. Recall that the variogram is defined as

$$\gamma(u) = \frac{1}{2} E \left\{ Y(t) - Y(t-u) \right\}^2 ,$$

for $u > 0$, and for a stationary process

$$\gamma(u) = \sigma^2 (1 - \rho(u)),$$

where $var(Y) = \sigma^2$.
In CD4+ data,

```
> Variogram(o15.lme)
     variog dist n.pairs
1 1.0528846    1      81
2 0.6734403    2      54
3 0.7141598    3      27
> plot(Variogram (o15.lme))
>
> Serial Correlation for CD4 Data
> CD4.lme <- lme(I(sqrt (CD4)) ˜ Cesd + Drugs + Sex + Packs +
+  Time2 + I(Time2ˆ2), data = CD4,
+  random = ˜ 1 | ID)
> plot(ACF (CD4.lme), alpha = 0.01)
>
> Variogram (CD4.lme)
       variog dist n.pairs
1  0.7505848    1    2007
2  0.8935267    2    1643
3  0.9964452    3    1303
4  1.0822057    4     988
5  1.1458269    5     720
6  1.1963521    6     495
7  1.2022744    7     322
8  1.1616083    8     189
9  1.3729741    9      97
10 1.4406387   10      43
11 0.9941913   11      10
> r <- tapply(resid (CD4.lme), CD4$ID, function (x) x)
> dt <- tapply(CD4$Time, CD4$ID, function (x) {
+  tmp <- outer (x, x, "-")
+  abs (tmp[lower.tri(tmp)])
+ })
> non.singles <- which (sapply (r, length) != 1)
> r <- r[non.singles]
> dt <- dt[non.singles]
> CD4.v <- mapply (function (x, y) Variogram (x, y), r, dt,SIMPLIFY =
> CD4.v <- do.call ("rbind", CD4.v)
> temp <- loess.smooth (x = CD4.v$dist, y = CD4.v$variog,
+  family = "gaussian")
> plot (variog ~ dist, data = CD4.v, ylim = c(0, 100), col = "gray70")
> lines (temp, lty = 1, lwd = 2)
> abline (h = var (unlist (r)), lwd = 2, lty = 2)
>
> # Exponential Correlation
> CD4.lme2 <- lme (I(sqrt (CD4)) ~ Cesd + Drugs + Sex + Packs +
+  Time2 + I(Time2^2), data = CD4,
+  random = ~ 1 | ID,
+  correlation = corExp (form = ~ Time, value = 0.1))
> summary (CD4.lme2)
Linear mixed-effects model fit by REML
 Data: CD4
       AIC      BIC    logLik
  14316.81 14374.52 -7148.407

Random effects:
 Formula: ~1 | ID
        (Intercept) Residual
StdDev:    3.911596 4.674125

Correlation Structure: Exponential spatial correlation
 Formula: ~Time | ID
 Parameter estimate(s):
    range
```

```
0.5057515
Fixed effects: I(sqrt(CD4)) ~ Cesd + Drugs + Sex + Packs + Time2 + I(T
                Value Std.Error   DF   t-value p-value
(Intercept) 29.243415 0.3957184 2001  73.89956  0.0000
Cesd        -0.044401 0.0137543 2001  -3.22812  0.0013
Drugs        0.404451 0.3158650 2001   1.28046  0.2005
Sex          0.050519 0.0380137 2001   1.32897  0.1840
Packs        0.539562 0.1246222 2001   4.32958  0.0000
Time2       -4.686629 0.2698153 2001 -17.36976  0.0000
I(Time2^2)   0.626022 0.0625696 2001  10.00522  0.0000
 Correlation:
           (Intr) Cesd   Drugs  Sex    Packs  Time2
Cesd       -0.061
Drugs      -0.611 -0.019
Sex        -0.050 -0.046 -0.132
Packs      -0.324 -0.025 -0.046 -0.011
Time2      -0.321 -0.009  0.022  0.325  0.025
I(Time2^2)  0.209 -0.003  0.002 -0.238  0.002 -0.930

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.616635203 -0.546364399  0.005226372  0.563678216  4.387458255

Number of Observations: 2376
Number of Groups: 369
>
```

```
> anova (CD4.lme2, CD4.lme)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
CD4.lme2     1 10 14316.81 14374.52 -7148.407
CD4.lme      2  9 14458.52 14510.45 -7220.261 1 vs 2 143.7077  <.0001
> intervals (CD4.lme2)
Approximate 95\% confidence intervals

 Fixed effects:
                  lower        est.       upper
(Intercept) 28.46735160 29.24341477 30.01947795
Cesd        -0.07137503 -0.04440069 -0.01742635
Drugs       -0.21500758  0.40445104  1.02390966
Sex         -0.02403141  0.05051916  0.12506973
Packs        0.29515938  0.53956233  0.78396528
Time2       -5.21577709 -4.68662868 -4.15748028
I(Time2^2)   0.50331403  0.62602233  0.74873063
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: ID
                   lower     est.    upper
sd((Intercept)) 3.519681 3.911596 4.347151

 Correlation structure:
          lower      est.     upper
```

```
range 0.4303557 0.5057515 0.5943562
attr(,"label")
[1] "Correlation structure:"

 Within-group standard error:
   lower     est.    upper
4.480871 4.674125 4.875713
> summary (CD4.lme)
Linear mixed-effects model fit by REML
 Data: CD4
      AIC      BIC    logLik
  14458.52 14510.45 -7220.261

Random effects:
 Formula: ˜1 | ID
        (Intercept) Residual
StdDev:    4.237735 4.349483

Fixed effects: I(sqrt(CD4)) ˜ Cesd + Drugs + Sex + Packs + Time2 + I(T
                Value Std.Error   DF   t-value p-value
(Intercept) 29.194329 0.3904274 2001  74.77530  0.0000
Cesd        -0.050820 0.0140008 2001  -3.62980  0.0003
Drugs        0.359554 0.3180973 2001   1.13033  0.2585
Sex          0.074187 0.0371871 2001   1.99497  0.0462
Packs        0.611049 0.1230691 2001   4.96509  0.0000
Time2       -4.527128 0.2235104 2001 -20.25466  0.0000
```

```
I(Time2^2)   0.600922 0.0517042 2001  11.62231  0.0000
 Correlation:
           (Intr) Cesd   Drugs  Sex    Packs  Time2
Cesd       -0.055
Drugs      -0.634 -0.015
Sex        -0.034 -0.051 -0.142
Packs      -0.331 -0.048 -0.033 -0.018
Time2      -0.286 -0.012  0.035  0.382  0.032
I(Time2^2)  0.188 -0.005 -0.001 -0.283 -0.001 -0.933

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.20671697 -0.56388389  0.00678614  0.56113461  4.50474956

Number of Observations: 2376
Number of Groups: 369
```