Linear Mixed Model: Case Studies

General Guidelines

- 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, care 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).
  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 question of interest.

Fitting Linear Mixed Effects Model

Grouped Data Object in `nlme`

The "tracking" data:

```
 library(nlme) # groupedData
 library(lattice) # histogram
> tracking <- read.table ("tracking.dat", header = TRUE)

> tracking[1:4,]
    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

> tracklong <- reshape (tracking, direction = "long",
+                       varying = 4:7, times = 1:4,
+                       split = list (regexp = "l", include = TRUE))

> tracklong <- tracklong[order (tracklong$id, tracklong$time),]

> tracklong[1:4,]
    Sex Age Shape time Trial id
1.1   M  31   Box    1  2.68  1
1.2   M  31   Box    2  4.14  1
```

```
1.3   M  31   Box    3  7.22  1
1.4   M  31   Box    4  8.00  1
```

```
> 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
......

> 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
......
```

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

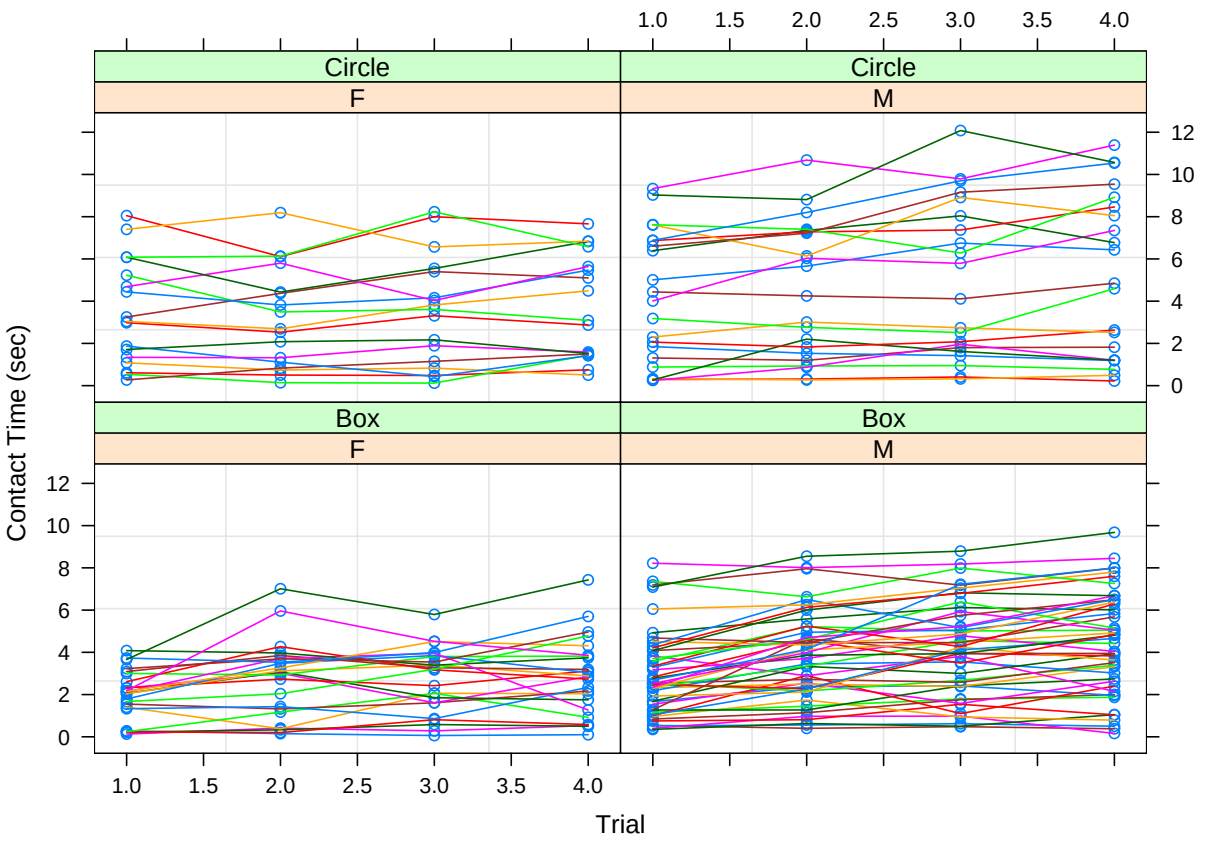


Figure 1: Tracking data

```
> track.sum <- gsummary (tracklong)
> histogram (~ Trial | Sex * Shape, data = track.sum, xlab="Seconds")
```

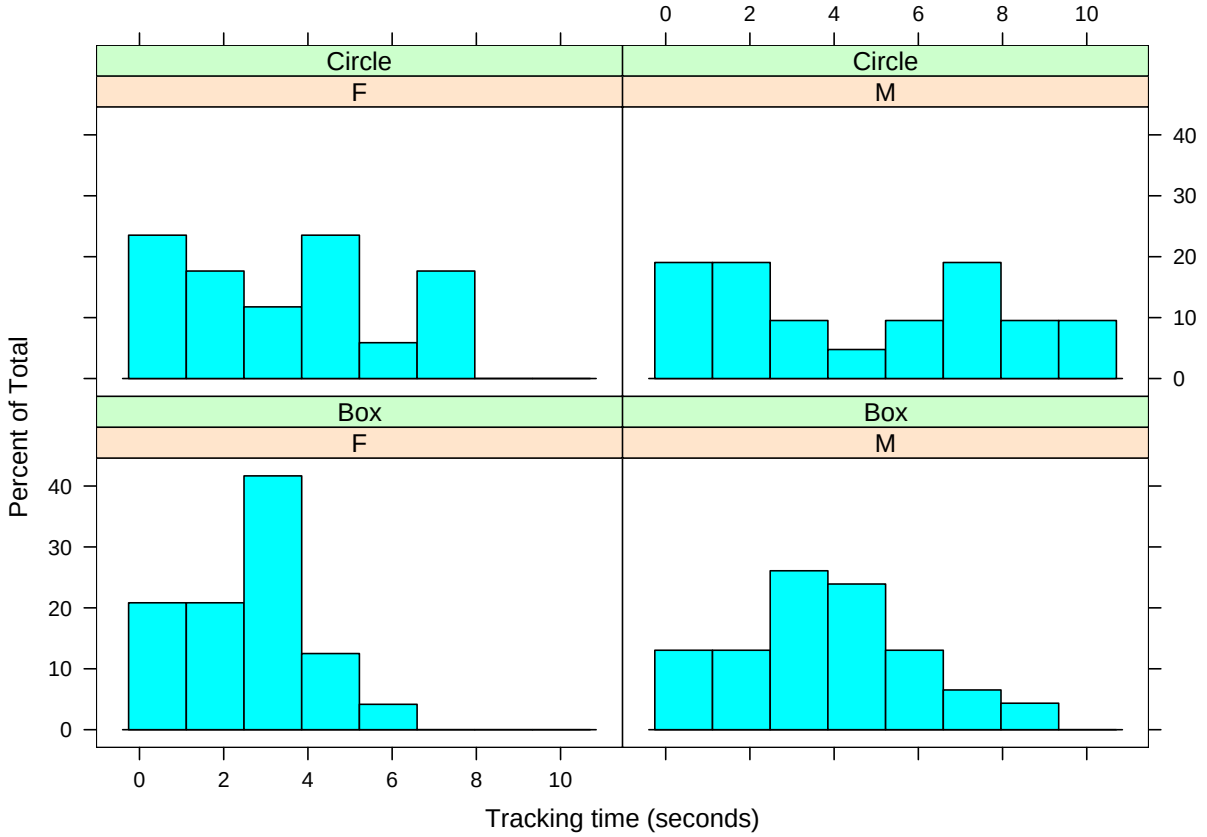


Figure 2: Histogram of mean contact time for each subject

Fitting Linear Models with `lm` and `lmList`

A brief review of the standard linear modeling functions in R with the orthodontic data.

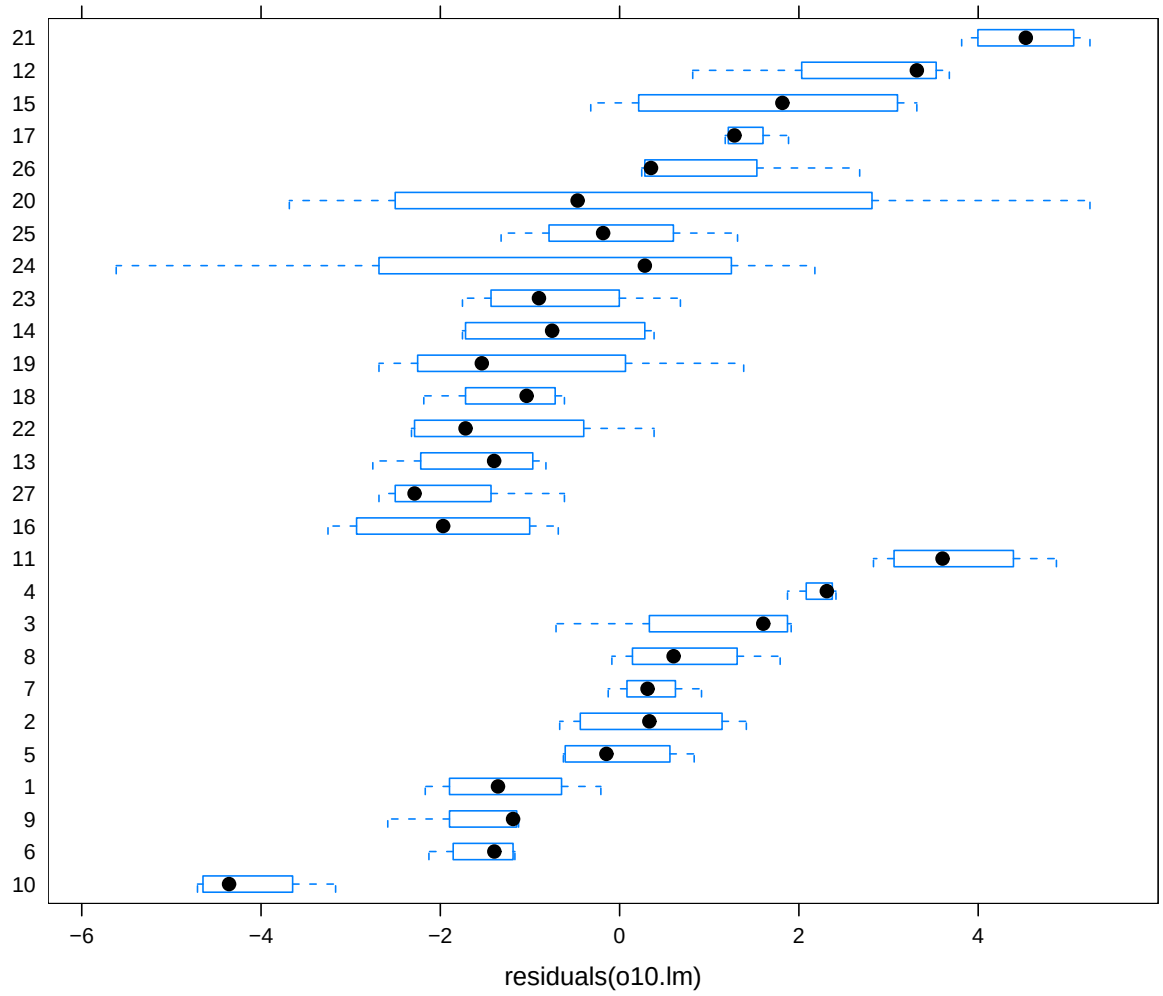
```
> Orth.new <- groupedData (distance ~ age | child,data = as.data.frame(Orthodont),
                           FUN = mean,outer=~male)
> o10.lm <- lm (distance ~ age * male, data = Orth.new)
> summary (o10.lm)

Call:
lm(formula = distance ~ age * male, data = Orth.new)


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)  17.3727     1.7080  10.171  < 2e-16 ***
age           0.4795     0.1522   3.152  0.00212 **
male         -1.0321     2.2188  -0.465  0.64279
age:male      0.3048     0.1977   1.542  0.12608
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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  235.36 46.2042 6.884e-10 ***
male        1 140.46  140.46 27.5756 8.054e-07 ***
age:male    1  12.11   12.11  2.3782    0.1261
Residuals 104 529.76    5.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> drop1 (o10.lm, scope = c("age:male"), test = "F")
Single term deletions

Model:
distance ~ age * male
         Df Sum of Sq    RSS    AIC F value  Pr(F)
<none>                529.76 179.75
age:male  1     12.11 541.87 180.19  2.3782 0.1261

> o20.lm <- update (o10.lm, ~ . - age:male)
```

```
> summary (o20.lm)
Call:
lm(formula = distance ~ age + male, data = Orth.new)

Residuals:
     Min       1Q   Median       3Q      Max
-5.98819 -1.48819 -0.05856  1.19160  5.37106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.38569    1.12857  13.633  < 2e-16 ***
age          0.66019    0.09776   6.753 8.25e-10 ***
male         2.32102    0.44489   5.217 9.20e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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
```

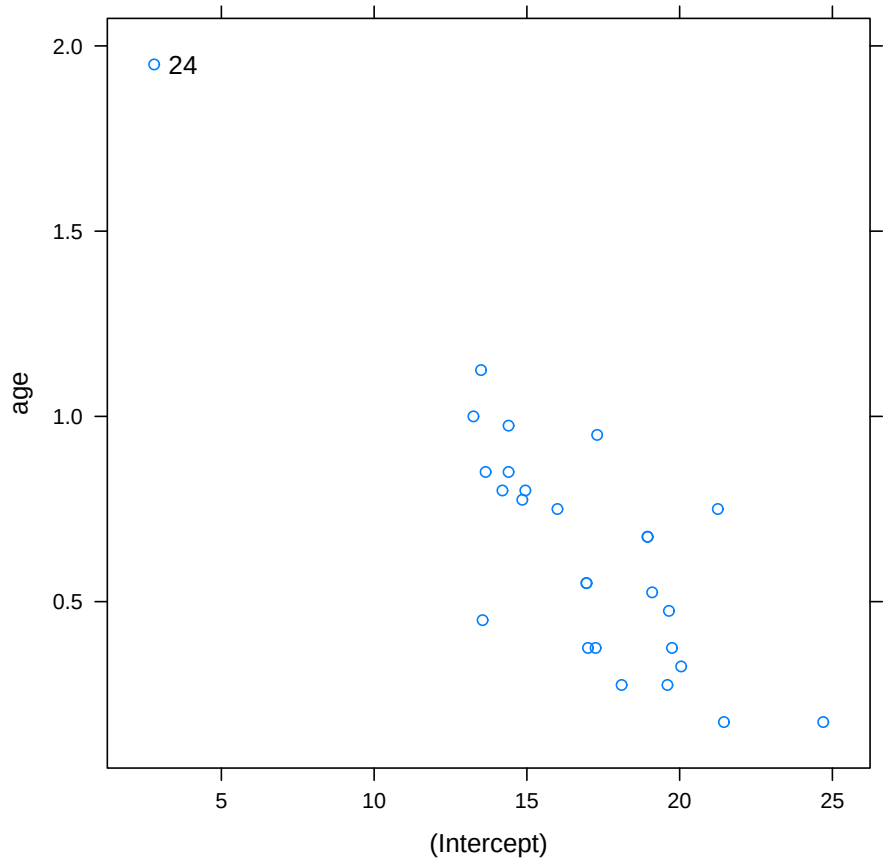
- Based on the lm models, there is a gender effect but not on the growth (i.e. no age*sex interaction).

```
> library(lattice)
> bwplot (getGroups (Orthodont) ~ residuals (o10.lm))
```



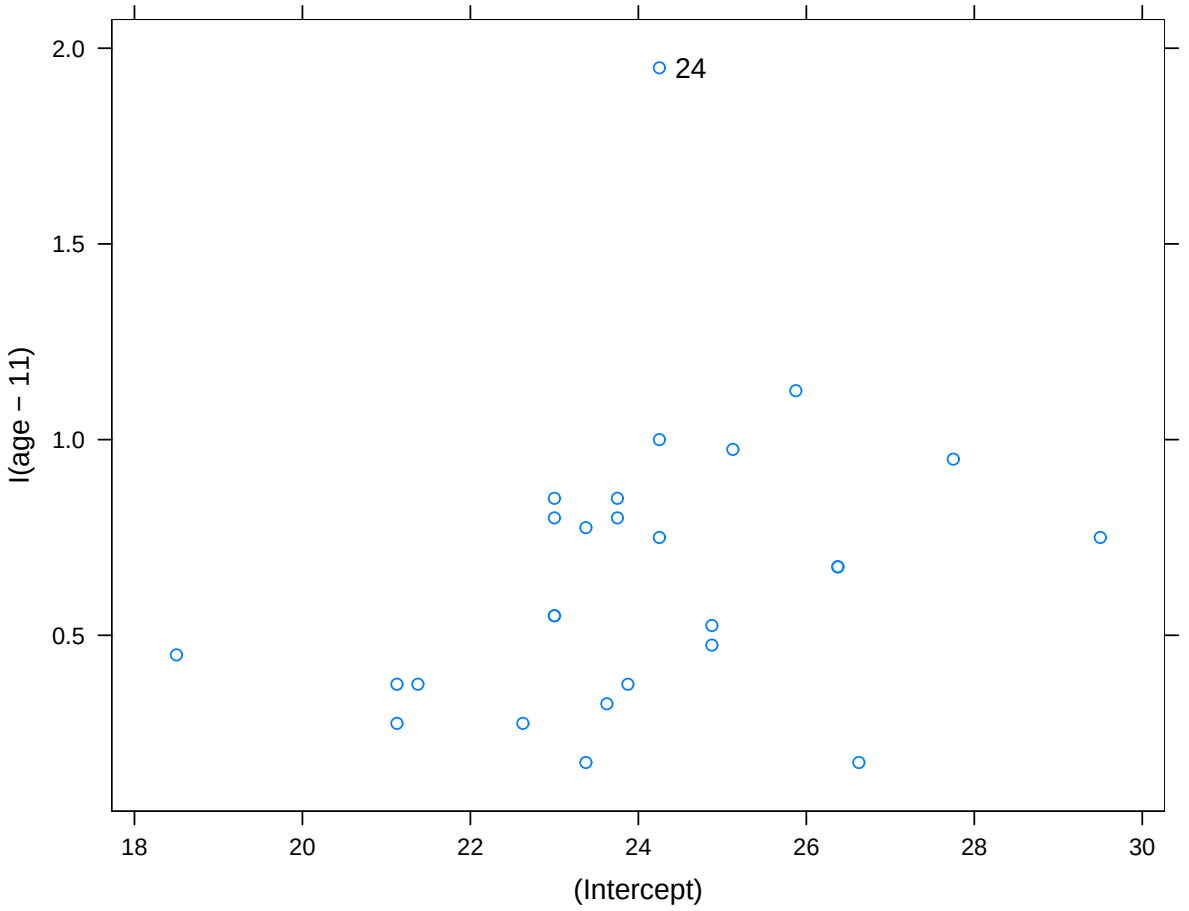
- The residuals from the same subject tend to have the same sign, indicating some "subject effect".

```
> o10.lis <- lmList (distance ~ age , data = Orth.new)
> pairs (o10.lis, id = 0.01, adj = -0.5)
```



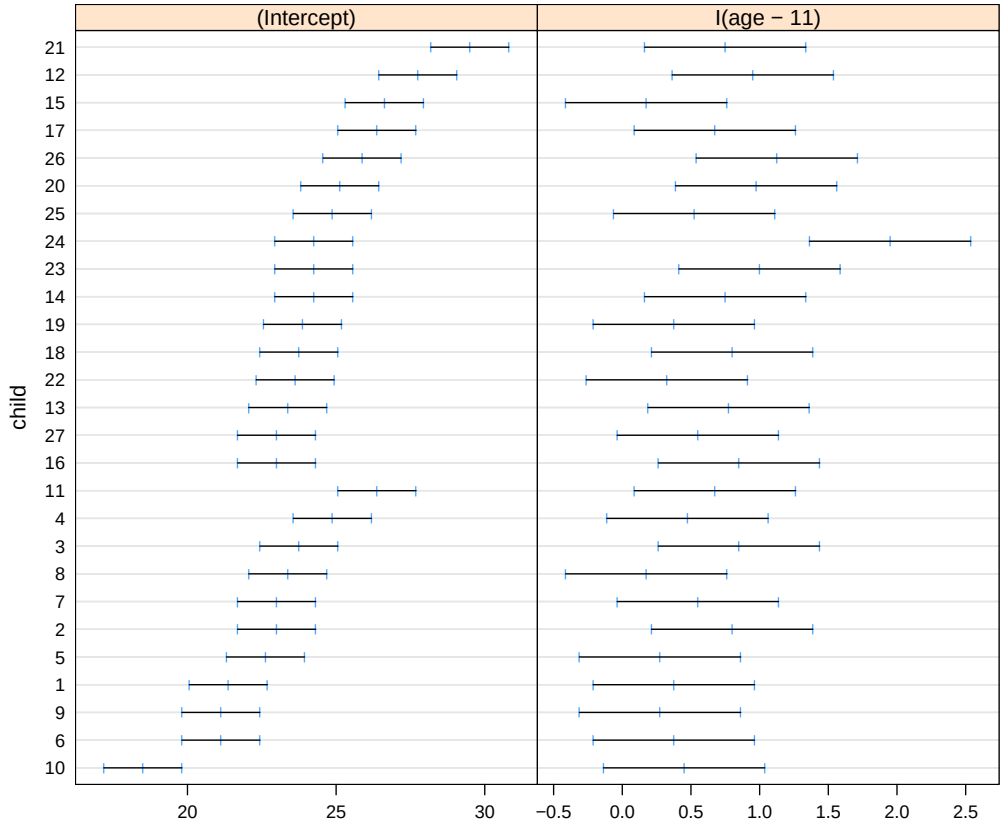
- There is negative correlation between the intercept and slope estimates and an outlier with an unusually low intercept, compensated by a large slope.

```
> o10.lis <- lmList (distance ~ I(age - 11), data = Orth.new)
> pairs (o10.lis, id = 0.01, adj = -0.5)
```



- There is not much correlation between the intercept and slope estimates after centering the age.

```
> plot (intervals (o10.lis))
```



- Note: the 95% CI's from **`lmList`** function are wrong!
- A random intercept is perhaps needed.
- The boys seem to have larger intercept. Note that we haven't 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).

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

Random effects:
 Formula: ~I(age - 11) | child
 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
-3.223106881 -0.493760898  0.007316482  0.472151220  3.916031750

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

Fit the "saturated model":

```
> o20.lme <- update (o10.lme, distance ~ I(age - 11) * male)
> summary (o20.lme)
Linear mixed-effects model fit by REML
 Data: Ortho.new
       AIC      BIC    logLik
  448.5817 469.7368 -216.2908

Random effects:
 Formula: ~I(age - 11) | child
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
```

```
(Intercept) 1.8303268 (Intr)
I(age - 11) 0.1803454 0.206
Residual    1.3100396

Fixed effects: distance ~ I(age - 11) + male + I(age - 11):male
                    Value Std.Error DF  t-value p-value
(Intercept)      22.647727 0.5861389 79 38.63884  0.0000
I(age - 11)       0.479545 0.1037193 79  4.62349  0.0000
male              2.321023 0.7614168 25  3.04829  0.0054
I(age - 11):male  0.304830 0.1347353 79  2.26243  .0264
 Correlation:
                  (Intr) I(g-11) male
I(age - 11)       0.102
male             -0.770 -0.078
I(age - 11):male -0.078 -0.770   0.102

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.168078306 -0.385939100  0.007103934  0.445154631  3.849463339

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:

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

and are estimated by:

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

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

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

and are estimated by:

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

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

- When n_i is small, $\boldsymbol{b}_i$ and $\boldsymbol{r}^1$ will be poorly estimated.
- The previous "raw" residuals can be standardized. The "standardized" (or Pearson) residuals correspond to the raw residuals divided by the estimated standard deviation.

The fitted values and residuals:

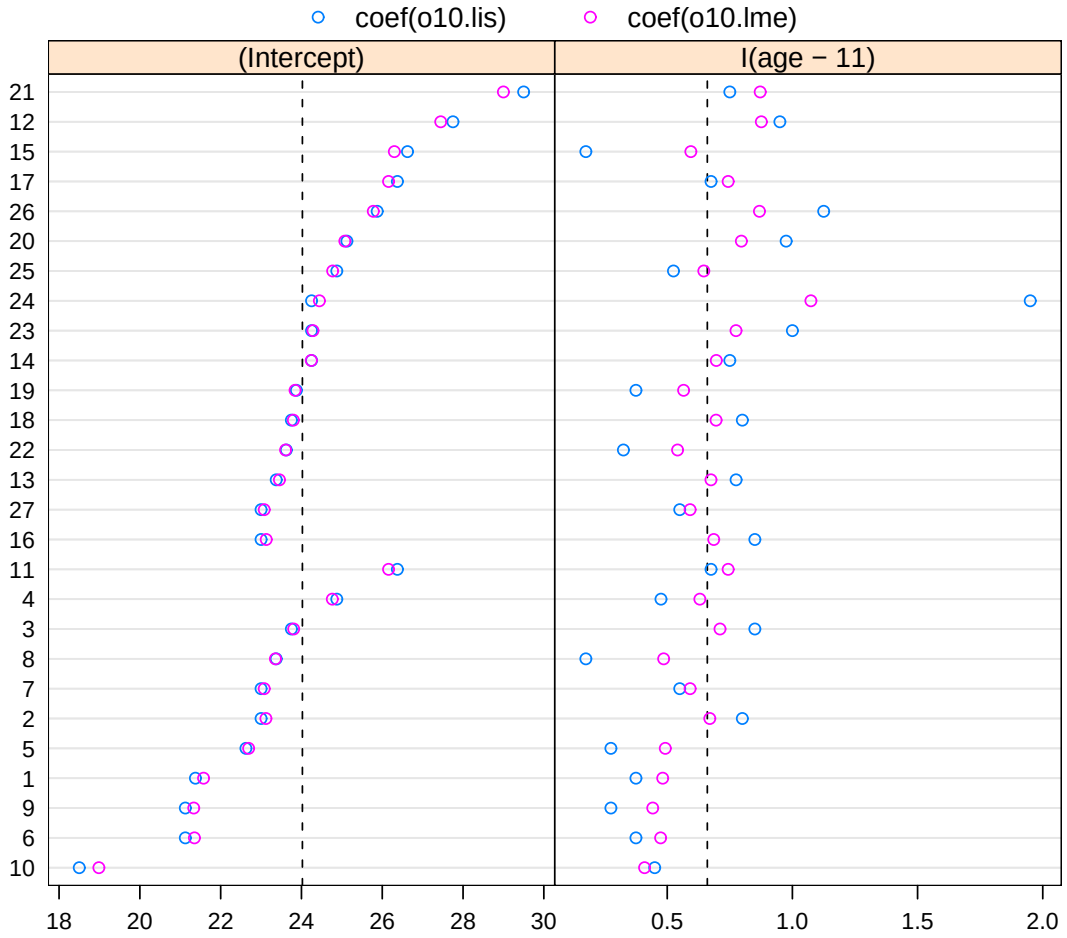
```
> fitted (o20.lme, level = 0:1)
       fixed    child
1   21.20909 20.20973
2   22.16818 21.07931
3   23.12727 21.94889
4   24.08636 22.81848
5   21.20909 21.27124
6   22.16818 22.41092
......

> resid (o20.lme, level = 1, type = "pearson")
           1            1            1            1            2            2
 0.603239771 -0.823878245 -0.342657072  0.138564100 -0.207046322 -0.695335585
           2            2            3            3            3            3
 0.343046503  0.618092916 -1.044766724  0.721423153  0.197606004  0.437124530
           4            4            4            4            5            5
 0.308458803  0.296614352 -0.096897937  0.272925449  0.121867070  0.618673749
           5            5            6            6            6            6
......

attr(,"label")
[1] "Standardized residuals"
```

Prediction

```
> Orthodont[Orthodont[,2]==11,]
   obs child age distance male
41  41    11   8     24.5    0
42  42    11  10     25.0    0
43  43    11  12     28.0    0
44  44    11  14     28.0    0

> newO <- data.frame (child = rep (c ("11", "19"), each = 3),
+                     male = rep (c (0, 1), each = 3),
+                     age = rep (16:18, 2))

> predict (o20.lme, newdata = newO)
      11       11       11       19       19       19
28.86544 29.44646 30.02749 27.27604 27.93656 28.59708
attr(,"label")
[1] "Predicted values"

> predict (o20.lme, newdata = newO, level = 0:1)
  child predict.fixed predict.child
1    11      25.04545      28.86544
2    11      25.52500      29.44646
3    11      26.00455      30.02749
4    19      28.89062      27.27604
5    19      29.67500      27.93656
6    19      30.45938      28.59708
```

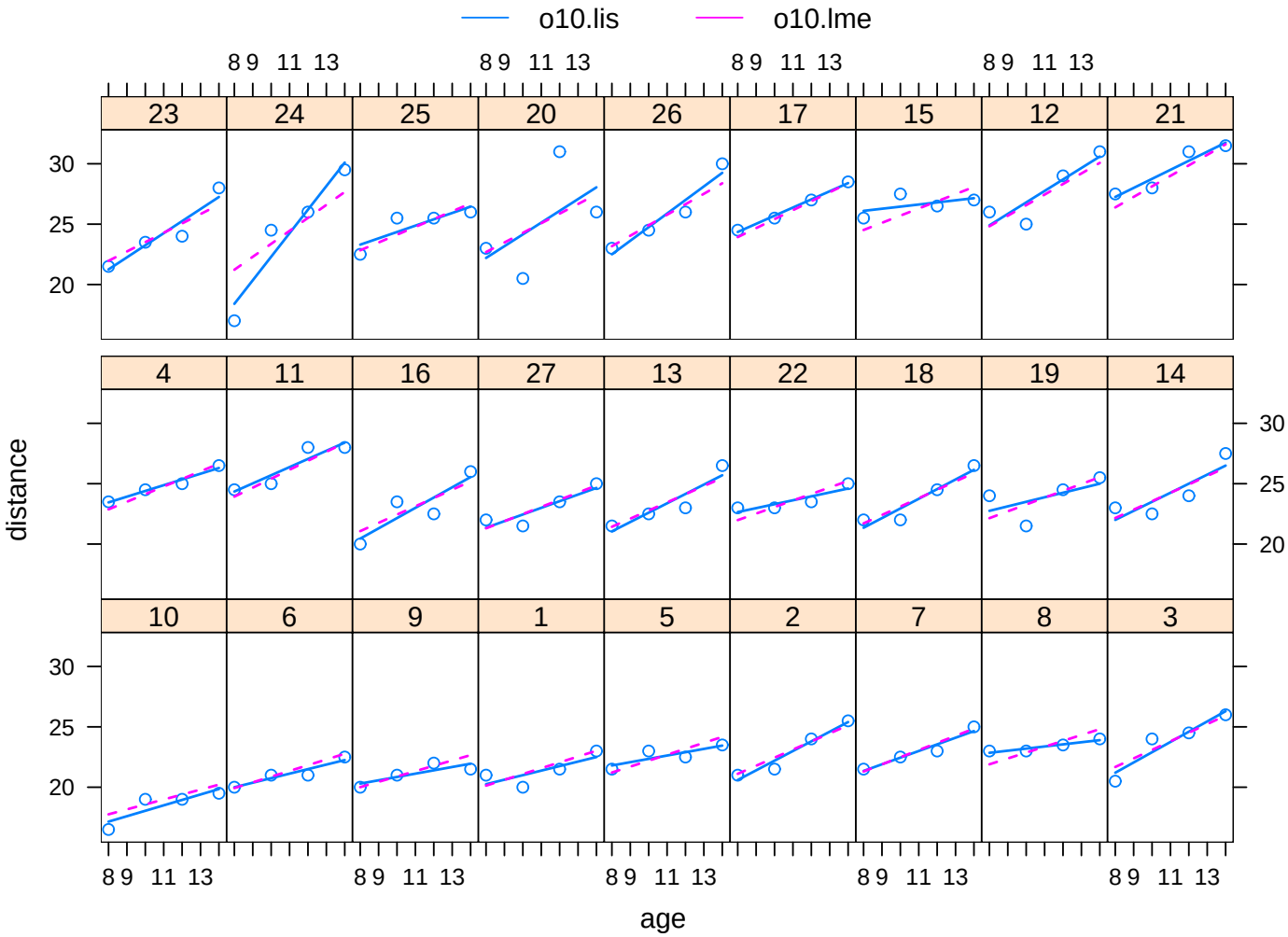
We can see the shrinkage when comparing the predicted random effects from lme with the individual regression coefficients.

```
> comp0 <- compareFits (coef (o10.lis), coef (o10.lme))
> plot (comp0, mark = fixef (o10.lme))
```



Comparing predicted values

```
> plot (comparePred (o10.lis, o10.lme), length.out = 2,
+       lty = 1:2, lwd = 1.5,layout = c(9, 3), between = list (y = c(0, 0.5)))
```



`lme` and `lm` models can be compared using:

```
> o20.lmeM <- update (o20.lme, method = "ML")
> o10.lm2 <- lm (distance ~ I(age - 11) * Sex, data = Orth.new)
> anova (o20.lmeM, o10.lm2)

          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
o20.lmeM      1  8 443.8060 465.2630 -213.9030
o10.lm2       2  5 488.2418 501.6524 -239.1209 1 vs 2 50.43581  <.0001
```

(It is important to have the `lme` object to be the first argument to `anova`.)

Patterned Variance-Covariance Matrices for the Random Effects

The default variance-covariance matrix for the random effects is `pdSymm`, a symmetric positive-definite matrix. We can force the random effects to be uncorrelated by using `pdDiag`.

```
> o30.lme <- update (o20.lme, random = pdDiag (~ I(age - 11)))
> summary(o30.lme)
Linear mixed-effects model fit by REML
 Data: Orth.new
       AIC      BIC    logLik
  446.8426 465.3533 -216.4213

Random effects:
 Formula: ~I(age - 11) | child
 Structure: Diagonal
        (Intercept) I(age - 11) Residual
StdDev:    1.830327   0.1803455  1.31004

Fixed effects: distance ~ I(age - 11) + male + I(age - 11):male
                    Value Std.Error DF  t-value p-value
(Intercept)      22.647727 0.5861390 79 38.63884  0.0000
I(age - 11)       0.479545 0.1037193 79  4.62349  0.0000
male              2.321023 0.7614169 25  3.04829  0.0054
I(age - 11):male  0.304830 0.1347353 79  2.26243  0.0264
 Correlation:
                 (Intr) I(g-11) male
```

```
I(age - 11)       0.00
male             -0.77   0.00
I(age - 11):male  0.00  -0.77    0.00

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.06658937 -0.39982537  0.02559617  0.43693649  3.85940305

Number of Observations: 108
Number of Groups: 27
> anova(o20.lme, o30.lme)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
o20.lme     1  8 448.5817 469.7368 -216.2908
o30.lme     2  7 446.8426 465.3533 -216.4213 1 vs 2 0.260933  0.6095
```

Other possible choice for the covariance structure of the random effects are: `pdBlocked`, `pdCompSymm`, and `pdIdent`. Remember that the default is `pdSymm` (a general symmetric positive-definite matrix).

Fitting Multilevel Model (for illustration purpose only)

```
> o40.lme <- update (o10.lme, random = ~ 1 | male / child)
> summary (o40.lme)
Linear mixed-effects model fit by REML
 Data: Orth.new
       AIC      BIC    logLik
  452.0344 465.3516 -221.0172

Random effects:
 Formula: ~1 | male
        (Intercept)
StdDev:    1.550378

 Formula: ~1 | child %in% male
        (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:
           male child %in% male
              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

> ranef (o40.lme, levels = 1:2)
Level: male
  (Intercept)
0   -1.035618
1    1.035618

Level: child %in% male
      (Intercept)
0/10 -3.713345877
0/9  -1.444234541
0/6  -1.444234541
0/1  -1.228128700
0/5  -0.147599492
0/8   0.500718032
0/7   0.176559270
0/2   0.176559270
0/3   0.824876794
0/4   1.797353081
0/11  3.093988130
1/22 -1.073600862
1/27 -1.613865466
1/19 -0.857495021
1/16 -1.613865466
...
```

Model Diagnosis

Two important assumptions to check:

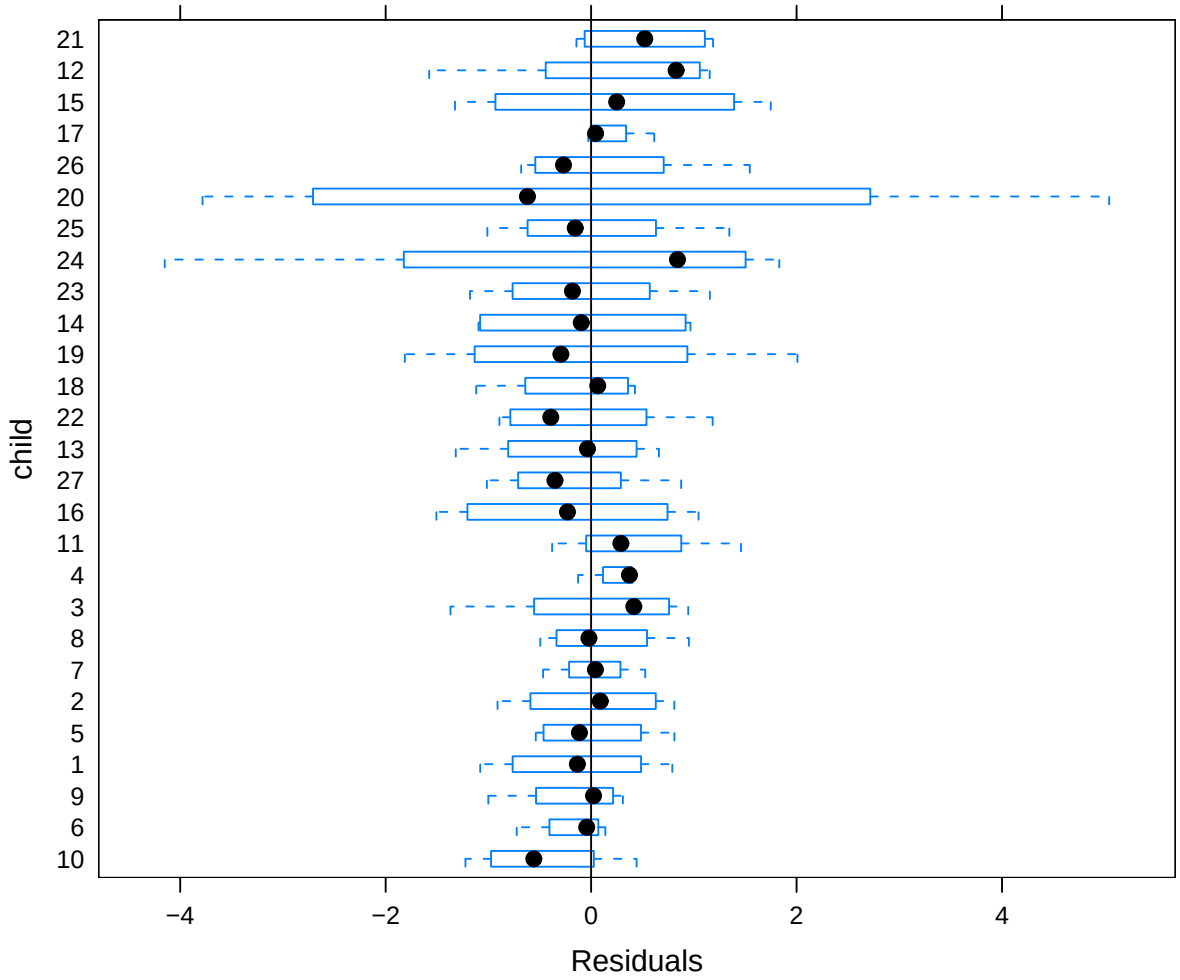
1. The within-group errors are iid $\mathcal{N}\left(0, \sigma^2\right)$ and independent of the random effects.

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 for different subjects.

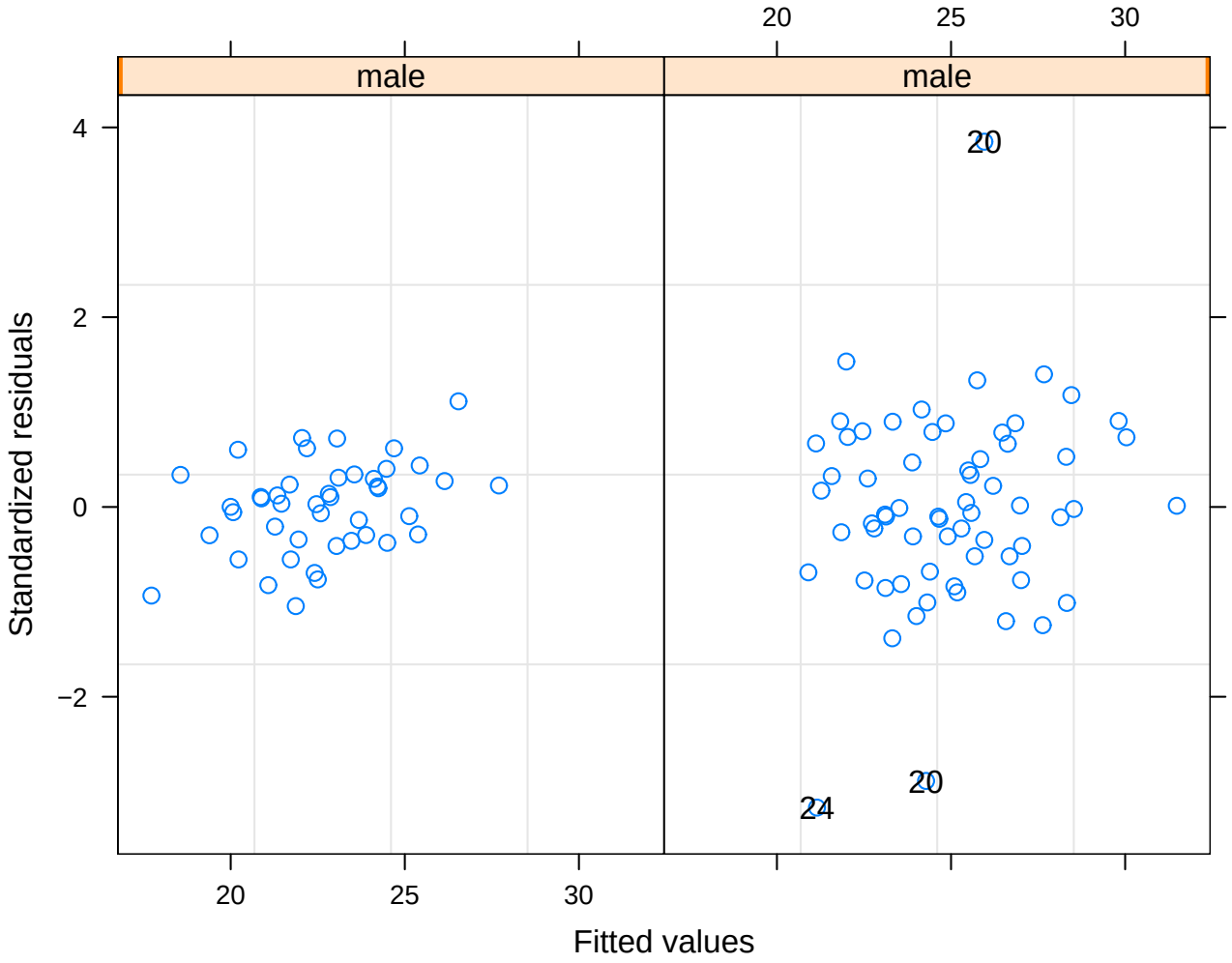
The most useful methods are based on plots of the residuals, the fitted values, and the estimated random effects.

Checking Within-Group Errors

```
> plot (o20.lme, Subject ~ resid (.), abline = 0)
```



```
> plot (o20.lme, resid (., type = "p") ~ fitted (.) | male, id = 0.05)
```



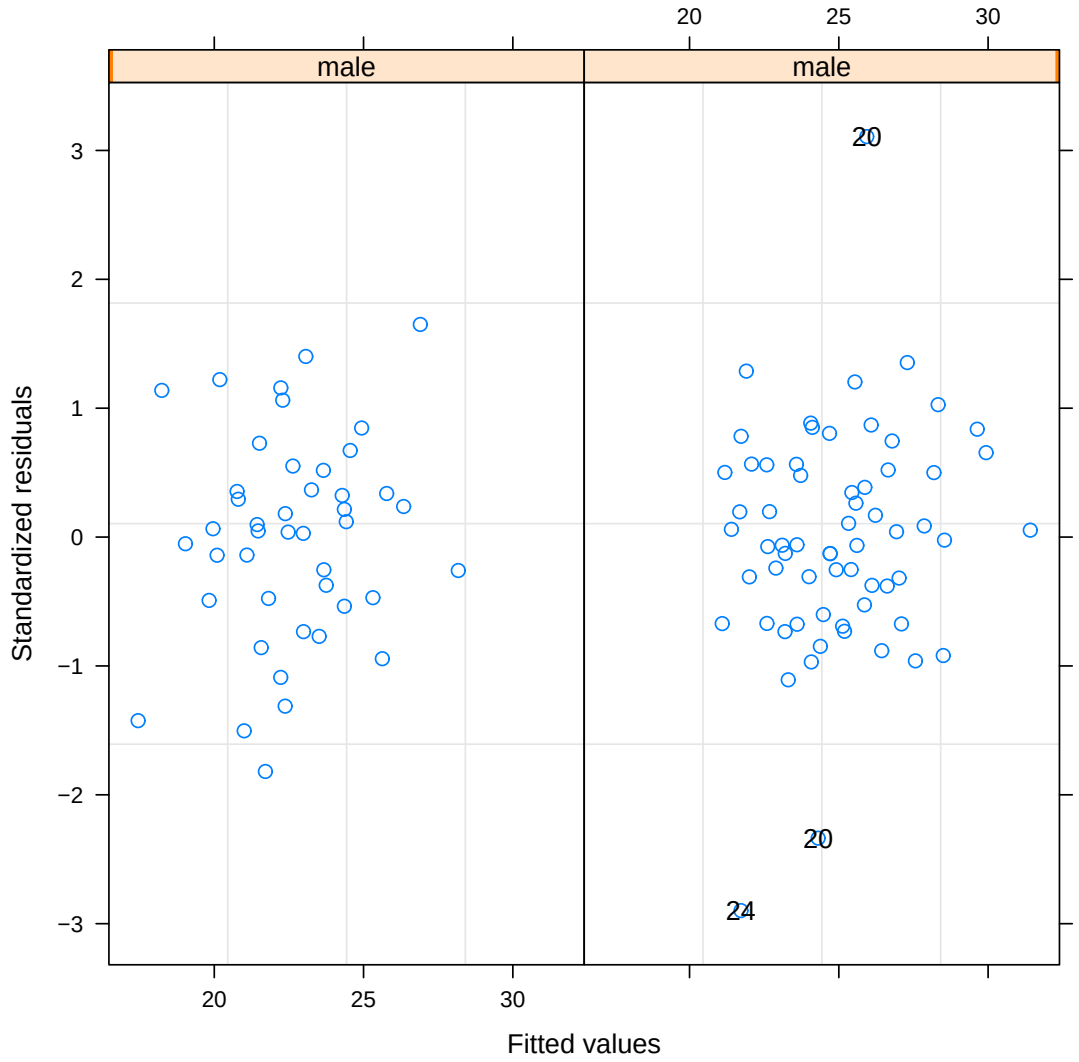
- It seems that the boys are more variable than the girls.

Now we allow the within-group variance to be different between girls and boys.

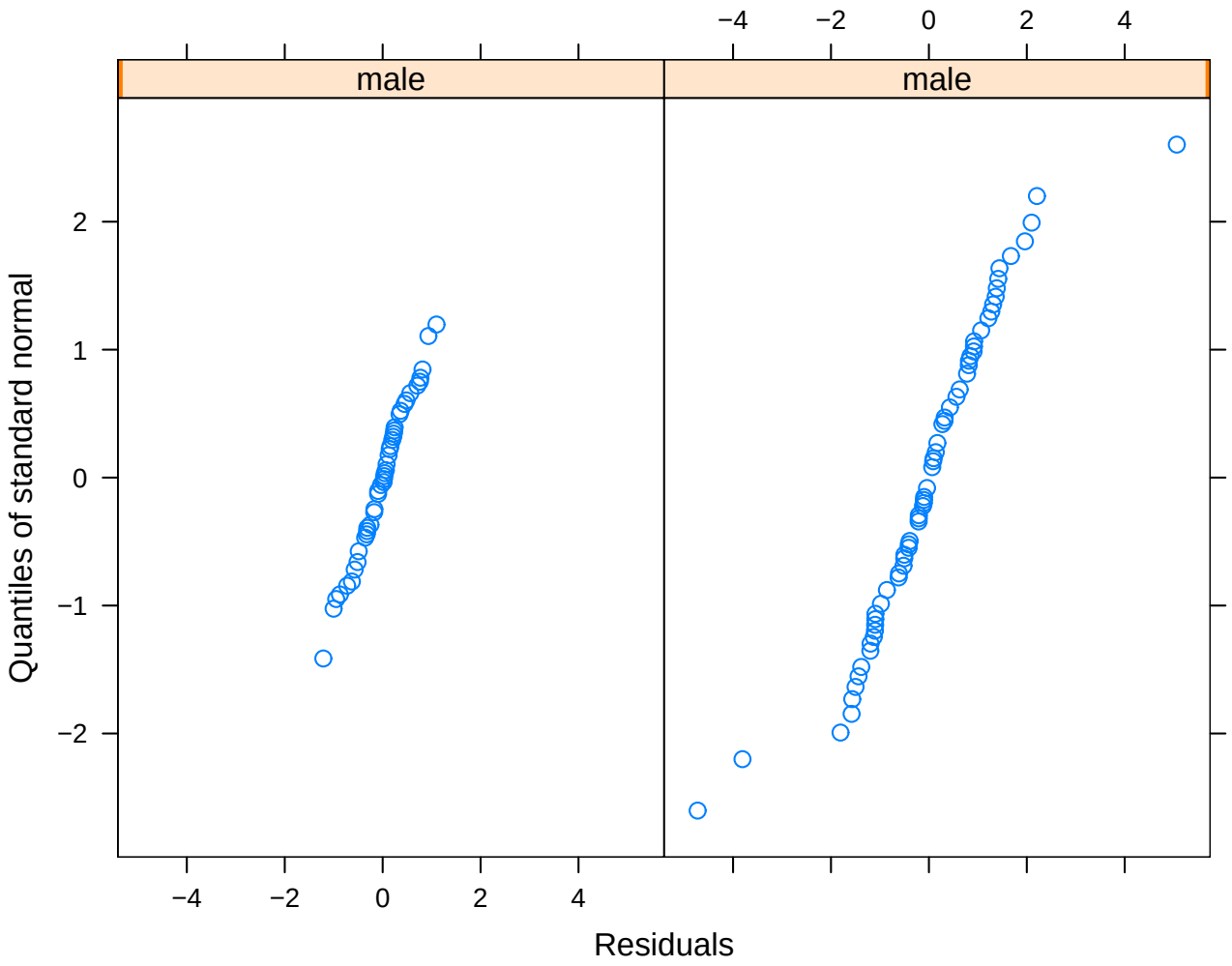
```
> o25.lme <- update (o20.lme, weights = varIdent (form = ~ 1 | male))
> summary (o25.lme)
Linear mixed-effects model fit by REML
 Data: Orthodont
       AIC     BIC    logLik
  429.5225 453.322 -205.7612

Random effects:
 Formula: ~I(age - 11) | child
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 1.8549769 (Intr)
I(age - 11) 0.1565178 0.394
Residual    0.6662499

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | male
 Parameter estimates:
       0        1
1.000000 2.445906
Fixed effects: distance ~ I(age - 11) + male + I(age - 11):male
                    Value Std.Error DF  t-value p-value
(Intercept)      22.647727 0.5682438 79 39.85565  0.0000
I(age - 11)       0.479545 0.0651518 79  7.36044  0.0000
```

```
male              2.321023 0.7612179 25  3.04909  0.0054
I(age - 11):male  0.304830 0.1186358 79  2.56946  0.0121
 Correlation:
                 (Intr) I(g-11) male
I(age - 11)       0.281
male             -0.746 -0.209
I(age - 11):male -0.154 -0.549   0.194

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.8984545 -0.5001207  0.0398503  0.5183388  3.1071955

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
```

```
> plot (o25.lme, resid (., type = "p") ~ fitted (.) | male,
+       id = 0.05)
```



```
> qqnorm (o25.lme, ~ resid (.) | male)
```



Variance functions for errors in `nlme`

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

$$\text{Var}(\epsilon_{ij} \mid \boldsymbol{b}_i) = \sigma^2 g^2(\mu_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{\delta}),$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_j$, where $\mu_{ij} = \text{E}(Y_{ij} \mid \boldsymbol{b}_i)$, and $\boldsymbol{v}_{ij}$ are covariates and $\boldsymbol{\delta}$ are parameters.
In `nlme`, different variance functions (from the `varFunc` classes) can be provided to the `weights=` argument (the default is the homoscedasticity variance structure, i.e. $g() =?$).

Some heteroscedasticity variance examples are:

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

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

  It is represented as `varFixed (~ Age)`.

- Different variances per stratum (`varIdent`): the within-group variances are different for each level of a class variable s:

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

  where by default $\delta_1 = 1$, so that $\delta_l > 0, l = 2, \ldots, S$ represent the ratio between the standard deviations of the lth stratum and the first stratum. For example, `weights=varIdent(form= ~ 1|male)`.

- 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 for errors in `nlme`

The general within-group correlation structure is expressed

$$\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = h[d(p_{ij}, p_{ij'}), \rho],$$

where h is a correlation function taking values between -1 and 1, and ρ is a correlation parameter (or a vector of parameters). In the context of time series data, $h()$ is referred as the **autocorrelation function**.

To investigate serial correlation structure,

- we first calculate the standardized residuals from a fitted mixed-effects model, $r_{ij} = \epsilon_{ij}/\hat{\sigma}_{ij} = (y_{ij} - \hat{y}_{ij})/\hat{\sigma}_{ij}$, where $\sigma^2_{ij} = \text{Var}(\epsilon_{ij})$.
- Then calculate the **empirical autocorrelation** at time lag u,

$$\hat{h}(u) = \hat{\text{Corr}}(r_t, r_{t-u}) = \frac{\sum_{i=1}^{m} \sum_{|t_{ij} - t_{ij'}|=u} r_{ij} r_{ij'}/N(u)}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} r^2_{ij}/N(0)},$$

  where $N(u)$ is the number of residual pairs of lag u. The autocorrelation function is useful for equally spaced data.

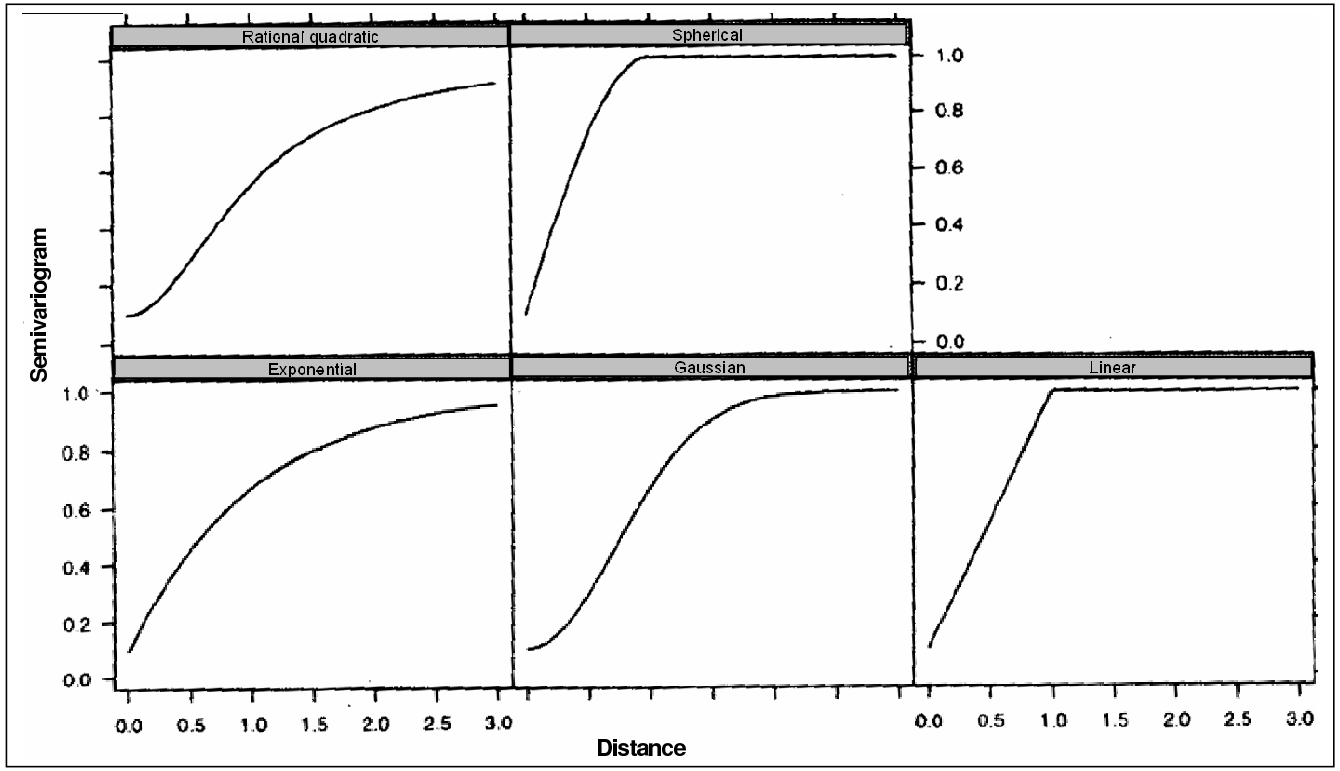
For unequally spaced data, variogram is found more useful. Recall that for a stationary process, the **variogram** is $\gamma(u) = \sigma^2(1 - h(u))$.

- The classical estimator of variogram is

$$\hat{\gamma}(u) = \frac{1}{2N(u)} \sum_{i=1}^{m} \sum_{|t_{ij} - t_{ij'}| = u} (r_{ij} - r_{ij'})^2.$$

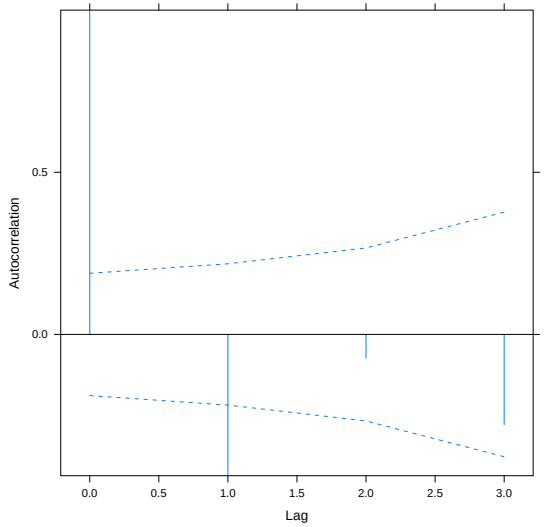
- With highly irregular sample times, the variogram can be estimated from the data pairs $\{\frac{1}{2}(r_{ij} - r_{ij'})^2, t_{ij} - t_{ij'}\}$ by fitting a non-parametric (i.e. smooth) curve.
- Pinheiro and Bates (2000, Figure 5.9 and Table 5.2) showed the plots of variogram versus time lag for different correlation models (rational quadratic, spherical, exponential, gaussian, and linear).

Exponential	$\gamma(u;\rho) = 1 - \exp(-u/\rho)$
Gaussian	$\gamma(u;\rho) = 1 - \exp[-(u/\rho)^2]$
Linear	$\gamma(u;\rho) = 1 - (1 - u/\rho)I(u < \rho)$
Rational quadratic	$\gamma(u;\rho) = (u/\rho)^2/[1 + (u/\rho)^2]$
Spherical	$\gamma(u;\rho) = 1 - [1 - 1.5(u/\rho) + 0.5(u/\rho)^3]I(u < \rho)$

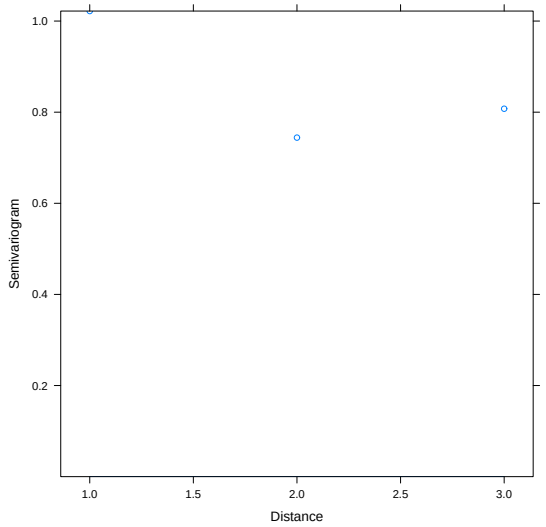


`nlme` provides functions to calculate auto-correlation function (`ACF`) and the variogram (`Variogram`) to help investigate appropriate correlation structure.
(We use the orthodontic data for illustration purpose only)

```
> ACF(o25.lme)
  lag         ACF
1   0  1.00000000
2   1 -0.43568486
3   2 -0.07302671
4   3 -0.27813481
> plot(ACF(o25.lme),alpha=0.05)
```



```
> Variogram(o25.lme)
     variog dist n.pairs
1 1.0219517    1       81
2 0.7441602    2       54
3 0.8075576    3       27
> plot(Variogram(o25.lme))
```



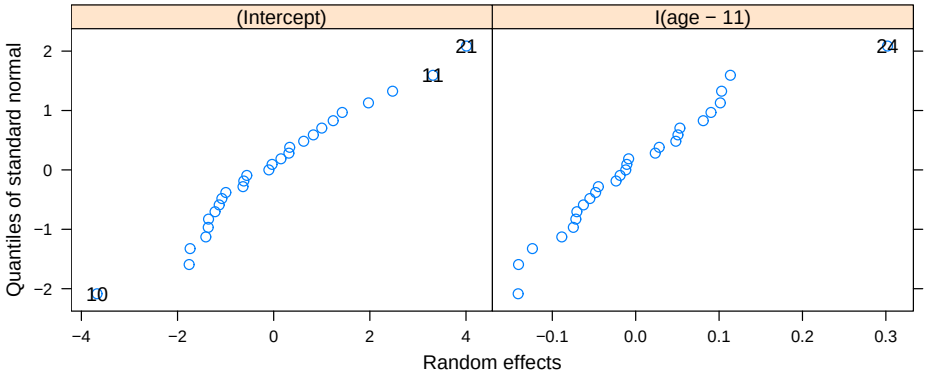
In `nlme`, correlation structures (from the `corStruct` class) are specified using the `corr=` argument (the default is the independent correlation structure, i.e. $h() =?$). Some non-independent correlation examples are:

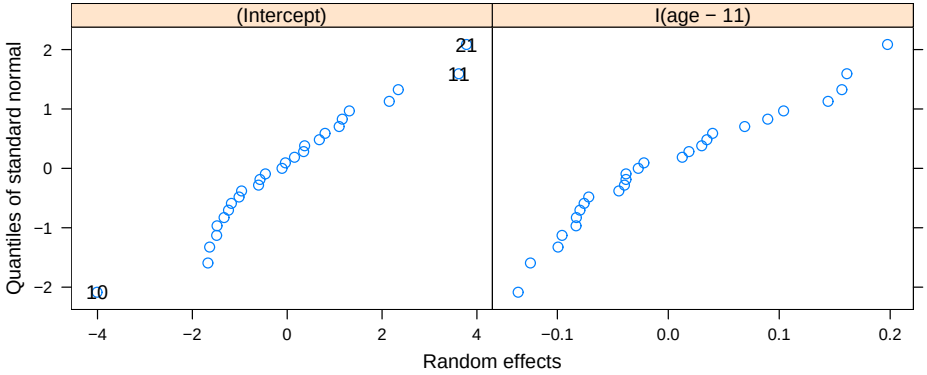
- Compound symmetry : `corCompSymm (~ 1 | Subject)` corresponds to $\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = \rho$.
- Autocorrelation of order 1 (AR1) for integer position vectors: `corAR1(~ 1 | Subject)` corresponds to $\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = \rho^{|t_{ij} - t_{ij'}|}$.
- Other plausible correlation structures are `corSymm`, `corCAR1`, `corARMA`, `corExp`,`corGaus`, `corLin`, `corRatio`, and `corSpher`.

Checking the Random Effects

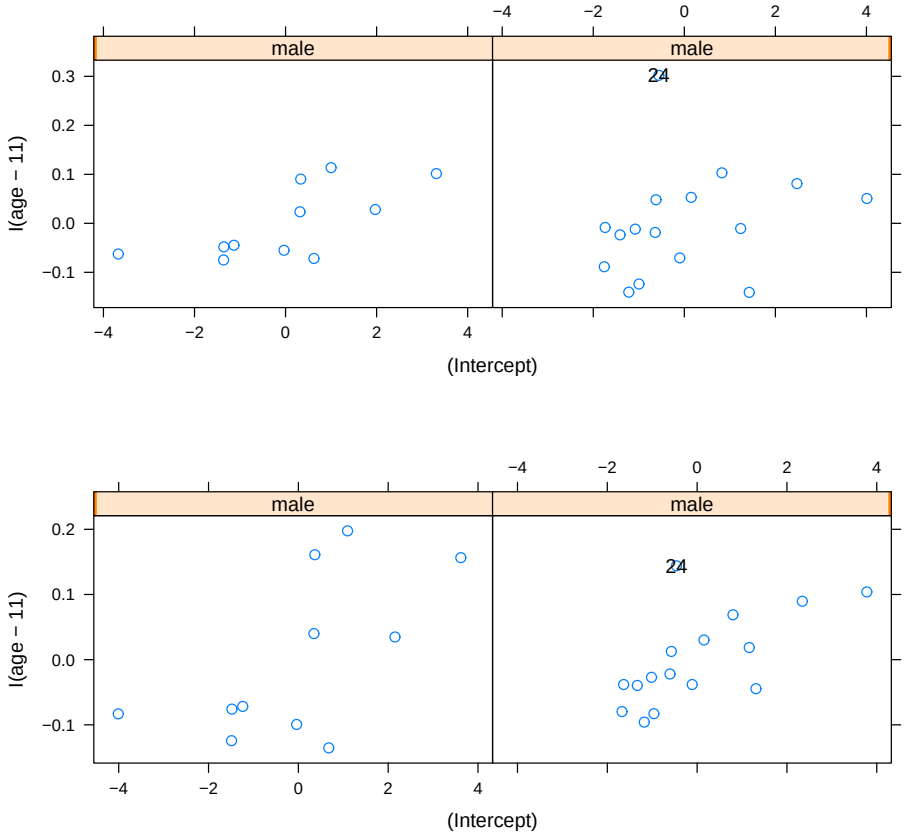
We can use Q-Q plots and conditional plots to check the normality and homogeneity of the random effects. However, as we cautioned earlier, these assumptions are harder to check.
Comparing the models assuming equal or unequal variances for boys and girls:

```
> qqnorm (o20.lme, ~ ranef (.), id = 0.10)
> qqnorm (o25.lme, ~ ranef (.), id = 0.10)
```





```
> pairs (o20.lme, ~ ranef (.) | male, id = ~ Subject == "24")
> pairs (o25.lme, ~ ranef (.) | male, id = ~ Subject == "24")
```



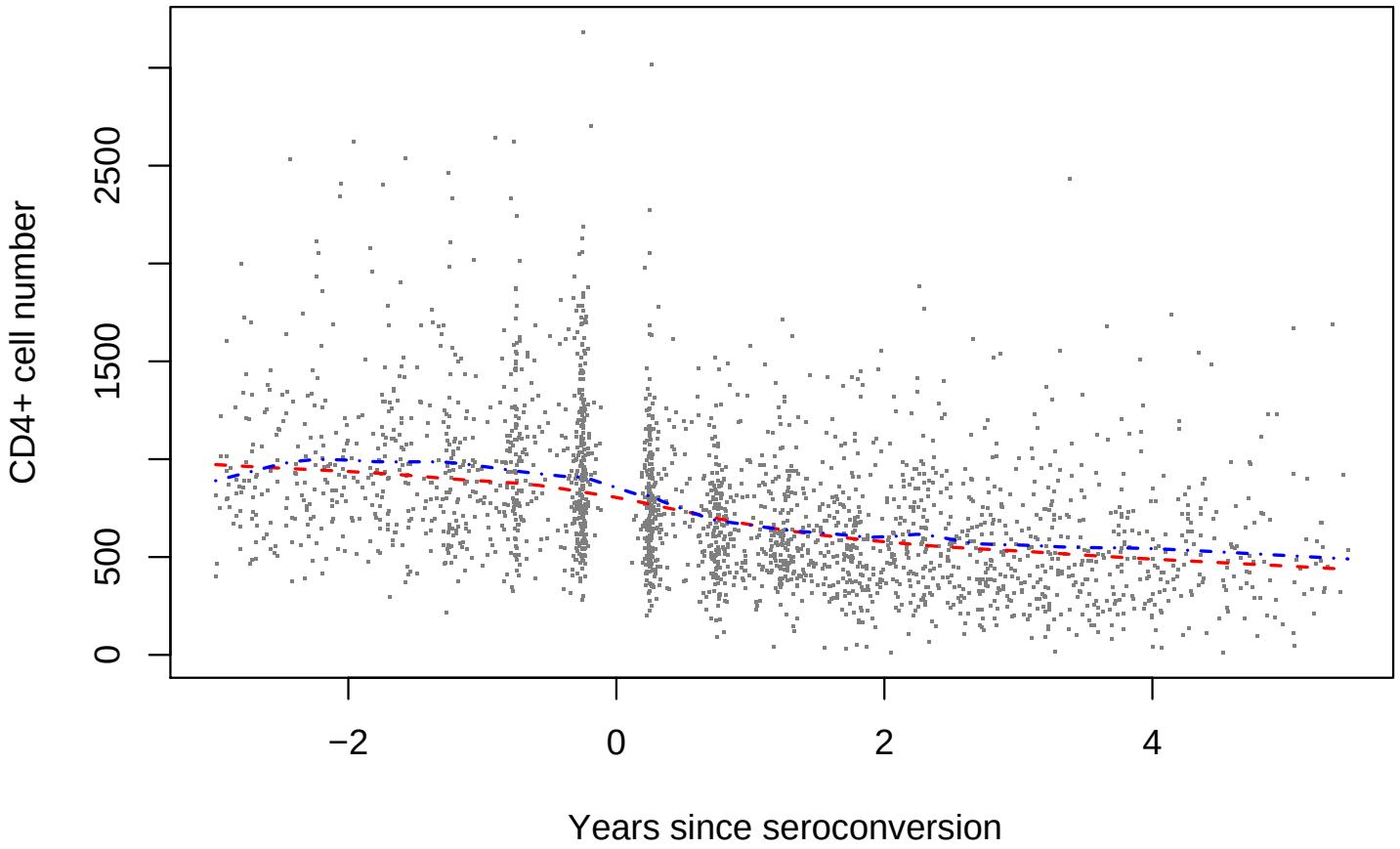
- The heteroscesdasticity model (`o25.lme`) accommodates the boys' outlying observations with increasing within-group error variance, thus reducing the between-group variance, thus more shrinkage.
- Note here everyone has the same set of covariates so the random effects should be iid. In general it might be necessary to standardize the random effects.

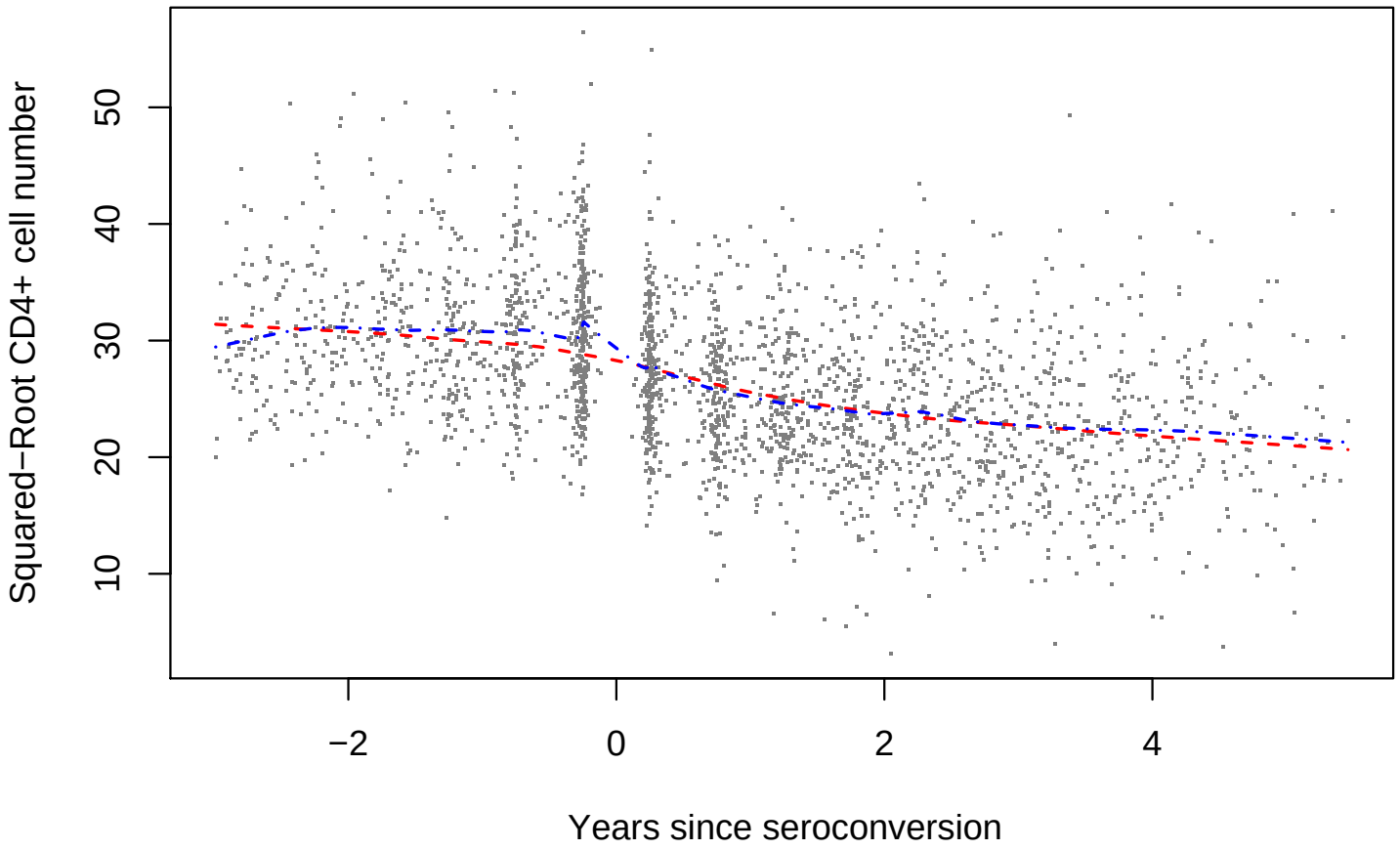
Multicenter AIDS Cohort Study: CD4+ Data

```
> library (nlme)
> CD4 <- read.table (file.path ("..", "data", "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, FUN = function (x) max (x, na.rm = TRUE))[1:5,]
          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

> gsummary (CD4g, FUN = function (x) min (x, na.rm = TRUE))[1:5,]
           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
```

The Mean Structure

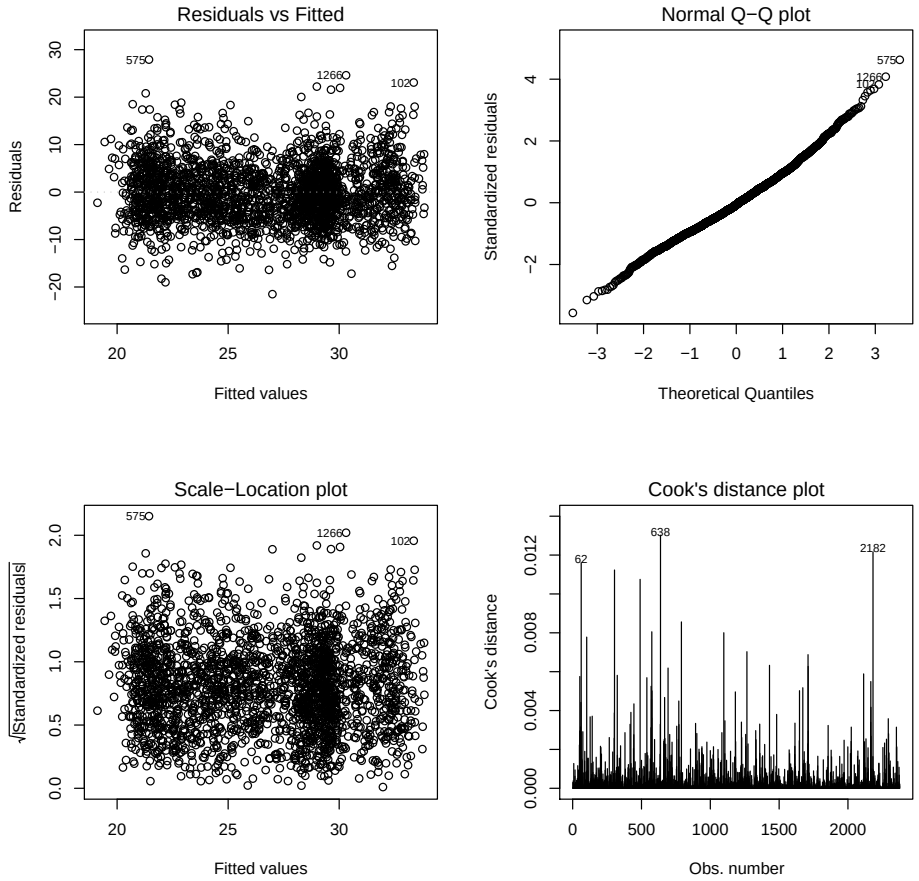
- Relatively flat before seroconversion, then a quick drop, and slower but steady decline.
- The trend is perhaps more apparent using squared-root transformed response.



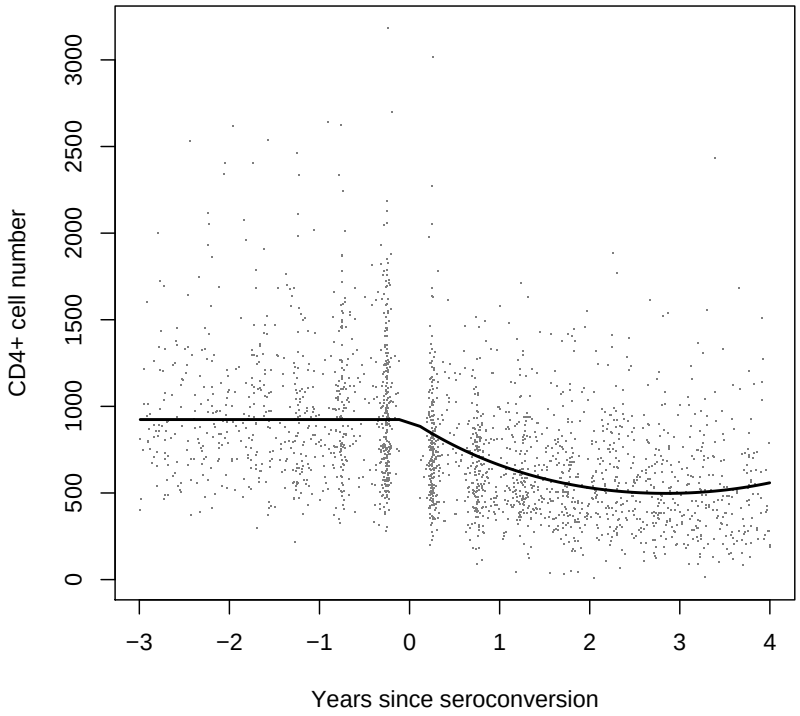

Squared−Root CD4+ cell number
10
20
30
40
50
−2
0
2
4
Years since seroconversion

```
> 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(Time2^2), 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(Time2^2)   0.75434    0.06654  11.337  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 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)
```

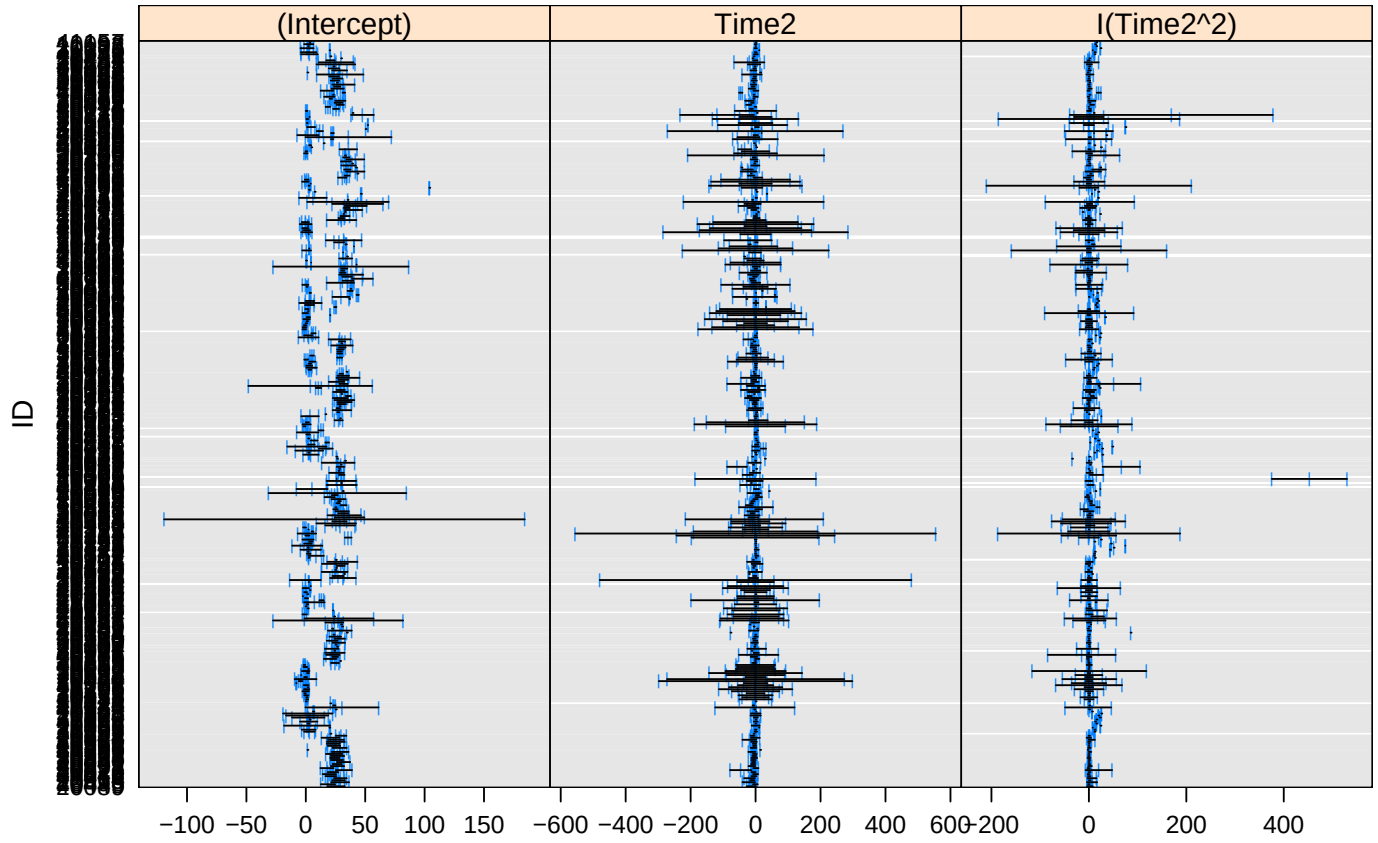


```
> temp <- subset (CD4, Time < 4)
> cd4.lmt <- lm (I(sqrt (CD4)) ~ Time2 + I(Time2^2), data = temp)
> temp$fitted <- fitted (cd4.lmt)^2
> temp <- temp[order (temp$Time),]
> plot (CD4 ~ Time, data = temp, col = "gray50", pch = ".",
+       xlab = "Years since seroconversion",
+       ylab = "CD4+ cell number")
> lines (temp$fitted ~ temp$Time)
```



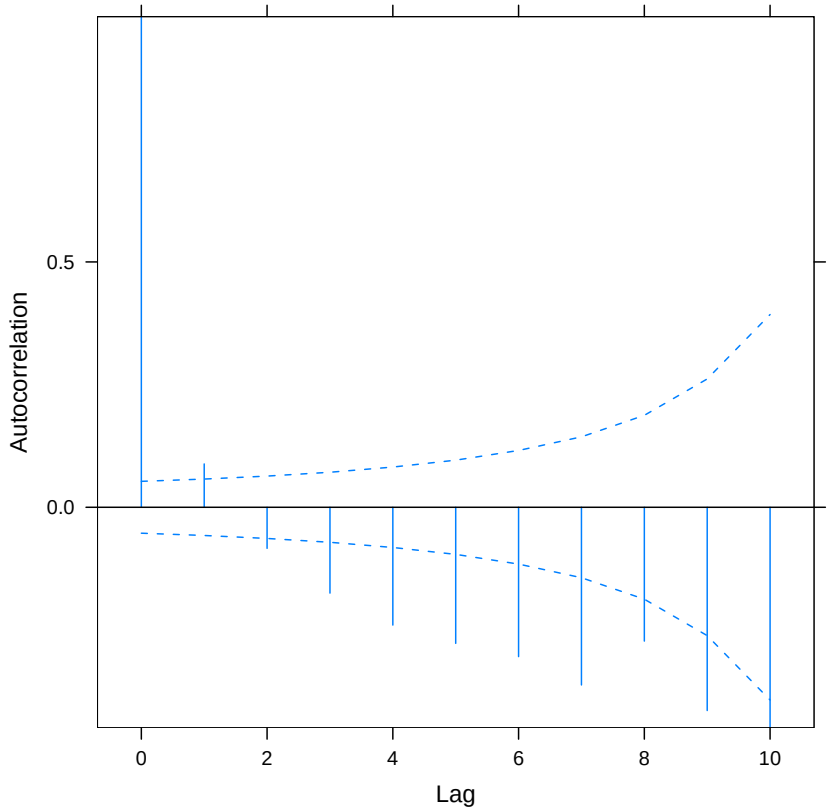
Random Effects

```
> CD4g$Time2 <- ifelse (CD4g$Time < 0, 0, CD4g$Time)
> CD4.lst <- lmList (I(sqrt (CD4)) ~ Time2 + I(Time2^2), data = CD4g)
> plot (intervals (CD4.lst), layout = c(3, 1))
```



Serial Correlation

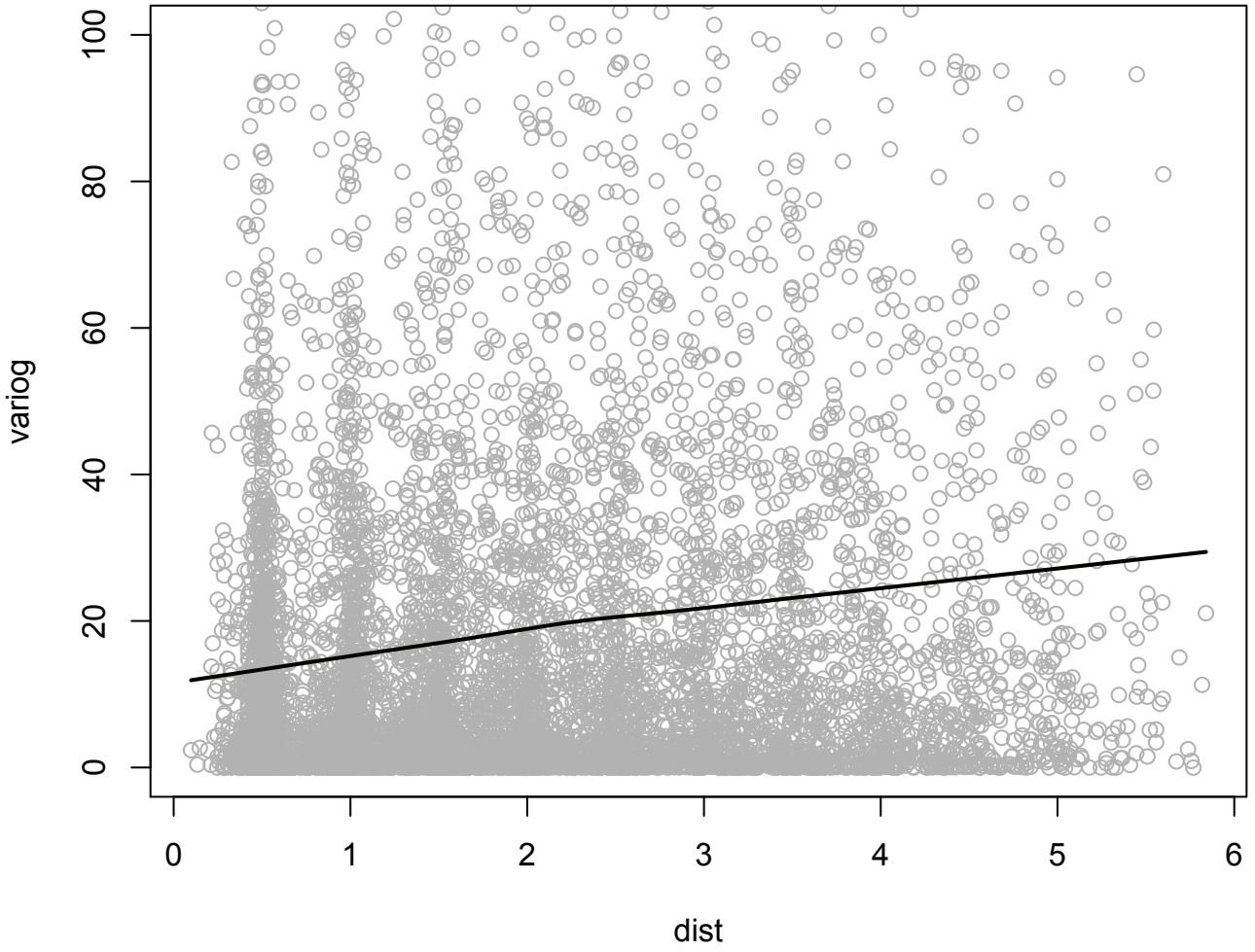
```
> CD4.lme <- lme (I(sqrt (CD4)) ~ Cesd + Drugs + Sex + Packs +
+                 Time2 + I(Time2^2), data = CD4g,
+                 random = ~ 1 | ID)
> plot (ACF (CD4.lme), alpha = 0.01)
```



```
> Variogram (CD4.lme)
      variog dist n.pairs
1  1.1926927    0     346
2  0.8642451    1    2298
3  1.0158060    2    1659
4  1.0052477    3    1177
5  1.0066143    4     817
6  0.9856498    5     577
7  0.9705400    6     387
8  1.0874431    7     260
9  1.0043419    8     161
10 0.7068785    9      86
11 0.6551067   10      41
12 0.6369616   11       8
```

Ignore the previous output from `Variogram`, let the data speak for themselves.

```
> r  <- tapply (resid (CD4.lme), CD4$ID, function (x) x)
> dt <- tapply (CD4$Time, CD4$ID, function (x) {
+     tmp <- outer (x, x, "-")
+     abs (tmp[lower.tri(tmp)])
+     })
> non.singles <- which (sapply (r, length) != 1)
> r <- r[non.singles]
> dt <- dt[non.singles]
> CD4.v <- mapply (function (x, y) Variogram (x, y), r, dt,
>                  SIMPLIFY = FALSE)
> CD4.v <- do.call ("rbind", CD4.v)
> temp <- loess.smooth (x = CD4.v$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)
```


variog
0
20
40
60
80
100
0
1
2
3
4
5
6
dist

Exponential Correlation

```
> CD4.lme2 <- lme (I(sqrt (CD4)) ~ Cesd + Drugs + Sex + Packs +
+                  Time2 + I(Time2^2), data = CD4g,
+                  random = ~ 1 | ID,
+                  correlation = corExp (form = ~ Time, value = 0.1))
> summary (CD4.lme2)
Linear mixed-effects model fit by REML
 Data: CD4g
       AIC      BIC    logLik
  14316.81 14374.52 -7148.407

Random effects:
 Formula: ~1 | ID
        (Intercept) Residual
StdDev:    3.911397 4.674152

Correlation Structure: Exponential spatial correlation
 Formula: ~Time | ID
 Parameter estimate(s):
    range
0.5057501
Fixed effects: I(sqrt(CD4)) ~ Cesd + Drugs + Sex + Packs
+ Time2 + I(Time2^2)
                Value Std.Error   DF   t-value p-value
(Intercept) 29.243383 0.3957119 2001  73.90069  0.0000
Cesd        -0.044400 0.0137543 2001  -3.22808  0.0013
```

```
Drugs        0.404469 0.3158645 2001   1.28051  0.2005
Sex          0.050516 0.0380137 2001   1.32888  0.1840
Packs        0.539584 0.1246205 2001   4.32982  0.0000
Time2       -4.686639 0.2698162 2001 -17.36975  0.0000
I(Time2^2)   0.626026 0.0625698 2001  10.00524  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
-3.616632119 -0.546344171  0.005234834  0.563662168  4.387449674

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) 28.46733273 29.24338326 30.01943379
Cesd        -0.07137434 -0.04440004 -0.01742573
Drugs       -0.21498910  0.40446864  1.02392637
Sex         -0.02403475  0.05051578  0.12506631
Packs        0.29518460  0.53958417  0.78398373
Time2       -5.21578924 -4.68663912 -4.15748900
I(Time2^2)   0.50331717  0.62602597  0.74873477
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: ID
                   lower     est.    upper
sd((Intercept)) 3.519412 3.911397 4.347041

 Correlation structure:
```

```
          lower      est.     upper
range 0.4303168 0.5057501 0.5944065
attr(,"label")
[1] "Correlation structure:"

 Within-group standard error:
   lower     est.    upper
4.480838 4.674152 4.875806
```

Further Reading: optional

- Chapter 3-5 of Pinheiro and Bates (2000) Mixed-effects models in S and S-PLUS. Springer.
- Chapters 9 and 10 of Verbeke and Molenberghs (2000).