5. Model Selection

Specific learning objectives:

1. Implement the selection of an LME model by employing the relevant estimation methods for testing for random effects and fixed effects (LRT via ML vs. REML).

Model Selection

- Can be tricky: estimation of the fixed effects depends on the covariance matrix. Therefore, the selection of a different covariance matrix may result in different p-values.
- Why is this? The formula of the weighted least squares involves the Vi matrix.

Basic strategy for model selection:

1. Include the random effects to be included in the model.
2. Choose an initial variance-covariance model.
3. Compare various variance-covariance structures (via tests and plots)
4. Perform variable selection (e.g., via LRT) on the fixed effects using the selected variance-covariance structure.

Model Selection

Some strategies for modeling the time variable

1. If all subjects are measured at the same times: times may be treated as fixed.
   - In addition, if and all variables are categorical: try a saturated model (i.e., involves all variables and all their interactions).

2. If subjects are measured at different times, these must be treated as random: use smoothed average trends or individual profiles.
   - E.g.,a curvilinear pattern may be detected and a polynomial term for time included in the model.

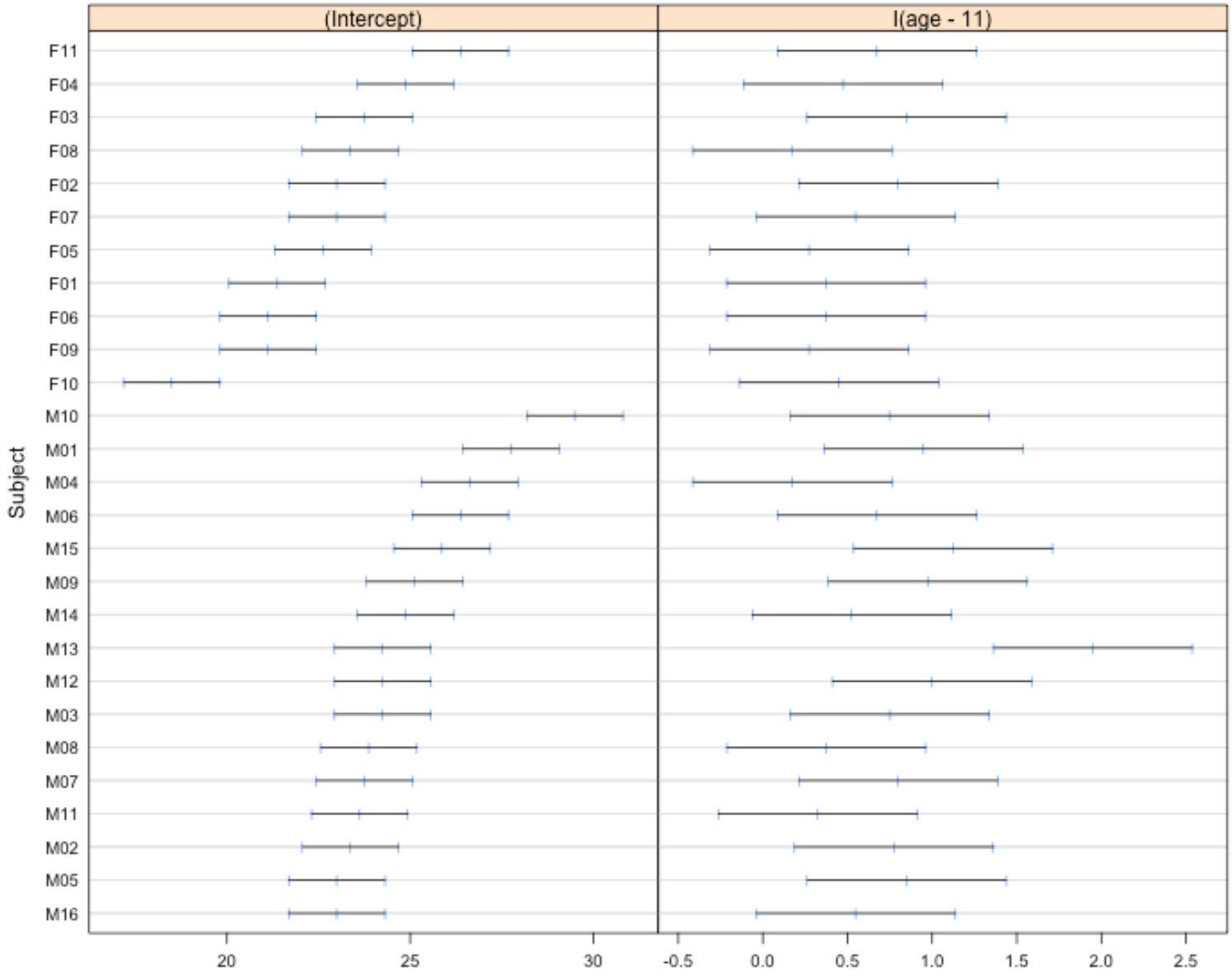
Some strategies for including random effects

1. Fit a model without the random effects and perform estimation via OLS. If the plot of the residuals vs. time show a trend, then random effects are needed.

2. Fit separate linear fits per subject and then examine the variation between the estimates of the intercept and slope for all subjects.

Maxillary Distance Data

95% CI's for Intercept and Slope, separate lm fits by subject



R Code for Plot

95% CI's for Intercept and Slope, separate lm fits
Maxillary Distance Example

The R function `lmList()` will fit separate `lm()` fits for each subject:

```
> sep.fit <- lmList(distance~I(age-11)|Subject,data=dat)
> sep.fit
Call:
  Model: distance ~ I(age-11) | Subject
   Data: dat

Coefficients:
    (Intercept) I(age - 11)
M16      23.000       0.550
M05      23.000       0.850. . .
```

R Code for Plot

95% CI's for Intercept and Slope, separate lm fits

Maxillary Distance Example

```
> summary(sep.fit)
Call:
  Model: distance ~ I(age-11) | Subject
   Data: dat

Coefficients:
   (Intercept)
    Estimate Std. Error  t value Pr(>|t|)
M16   23.000  0.6550198 35.11344        0
M05   23.000  0.6550198 35.11344        0
. . .


. . .
I(age - 11)
    Estimate Std. Error   t value
Pr(>|t|)
M16    0.550  0.2929338 1.8775576
6.584707e-02
M05    0.850  0.2929338 2.9016799
5.361639e-03. . .
```

R Code for Plot
95% CI's for Intercept and Slope, separate lm fits
Maxillary Distance Example

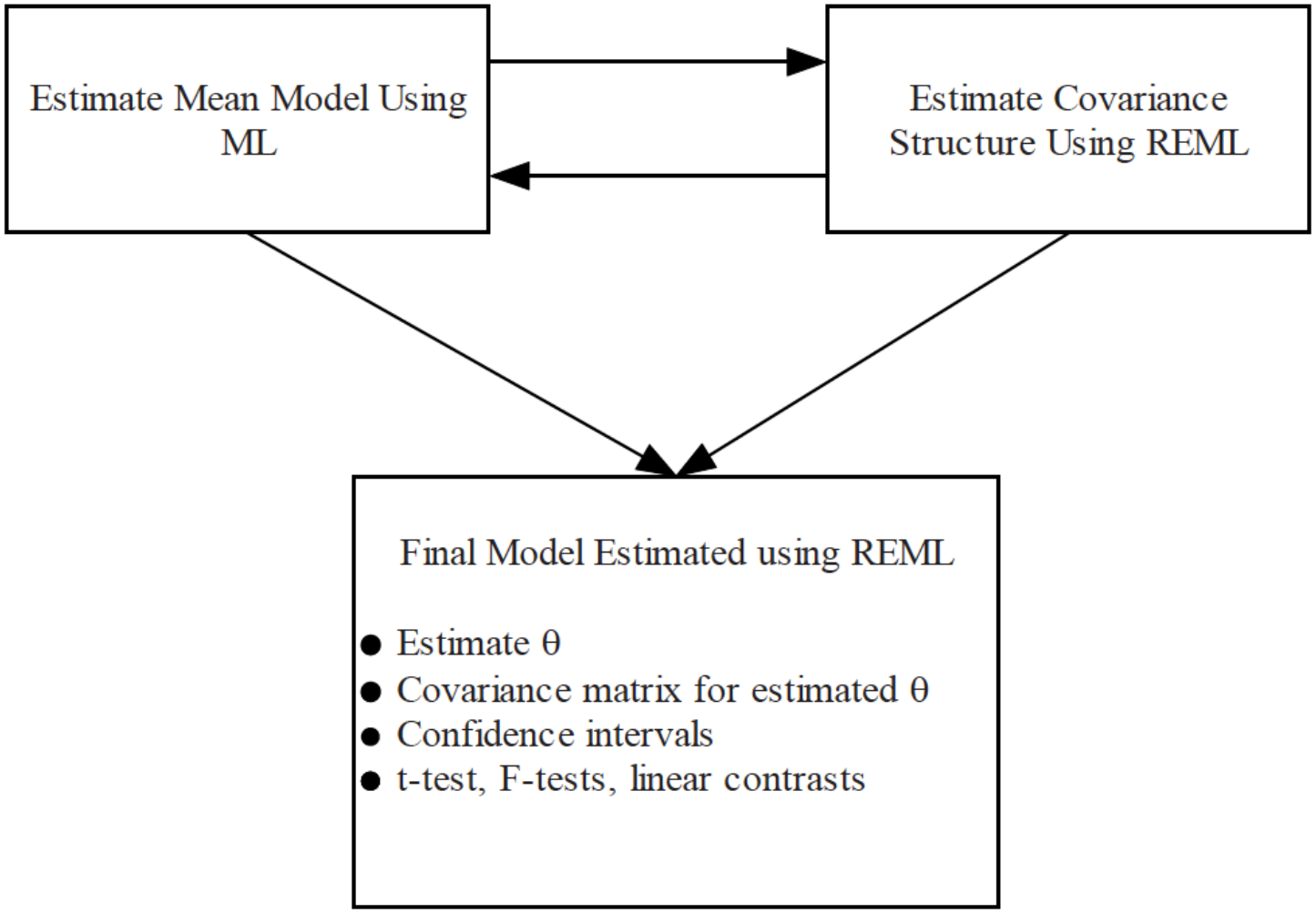
The R function `intervals()` will fit give 95% CI's for the intercepts and slopes:

```
> intervals(sep.fit)
, , (Intercept)

       lower   est.    upper
M16 21.68676 23.000 24.31324
M05 21.68676 23.000 24.31324
. . .

, , I(age - 11)

           lower  est.     upper
M16 -0.03729682 0.550 1.1372968
M05  0.26270318 0.850 1.4372968
. . .
```

Simply plotting the `intervals()` output will give the desired plot:

```
plot(intervals(sep.fit.cent))
```

Summary of Model Selection Process



Bonate, Ch. 6

6. Model checks
Goodness of Fit

Specific learning objectives:

1. State the model assumptions for the random effects and the residuals.
2. Assess the distribution of the residuals.
3. Assess the distribution of the random effects.
4. Identify the assumptions that are to be assessed in the various residuals and predicted random effects plots.

Residual Analysis and Goodness of Fit

- As before, residual analysis can be used to
  - Assess the adequacy of the fitted model
  - Identify outliers.

- Assumptions:
  1. The within-subject errors are independent and identically normally distributed, with mean zero and variance σ^2, and they are independent of the random effects.
  2. The random effects are normally distributed, with mean zero and covariance matrix G (not depending on the group) and are independent for different groups (e.g., Females vs. Males).

Section 4.3, Pinheiro & Bates, 2011.

Accessing Residuals in R

The `resid()` function gives the deviations of the observations Y_{ij} vs. the subject-specific mean (deviations of each subject's data points around that subject-specific line).

(Called "raw" residuals in Pinheiro & Bates, 2011)

```
> fitt.sexage <- update(fitt.sex,.~.+Sex:I(age-11))
> resid(fitt.sexage,level=1)
         M01          M01          M01          M01
1.154282542 -1.576485812  0.692745834  0.961977480 . . .
```

The option `type="pearson"` or `type="p"` in `resid()` gives the "raw" residuals standardized by the within-subject standard deviation (Pearson Residual):

```
> resid(fitt.sexage,level=1,type="p")
         M01          M01          M01          M01
0.881105017 -1.203387824  0.528797593  0.734311706. . .
```

Residual Analysis and Goodness of Fit

Assumption 1: The within-subject errors are independent and identically normally distributed, with mean zero and variance σ^2, and they are independent of the random effects.

- Used to assess that errors are centered at zero and have constant variance across subjects:
  - Box plots of Within-subject residuals:
  - Scatterplots of Fitted values vs standardized residuals
- To assess normality:
  - Normal Q-Q plots of residuals

Residual Analysis and Goodness of Fit

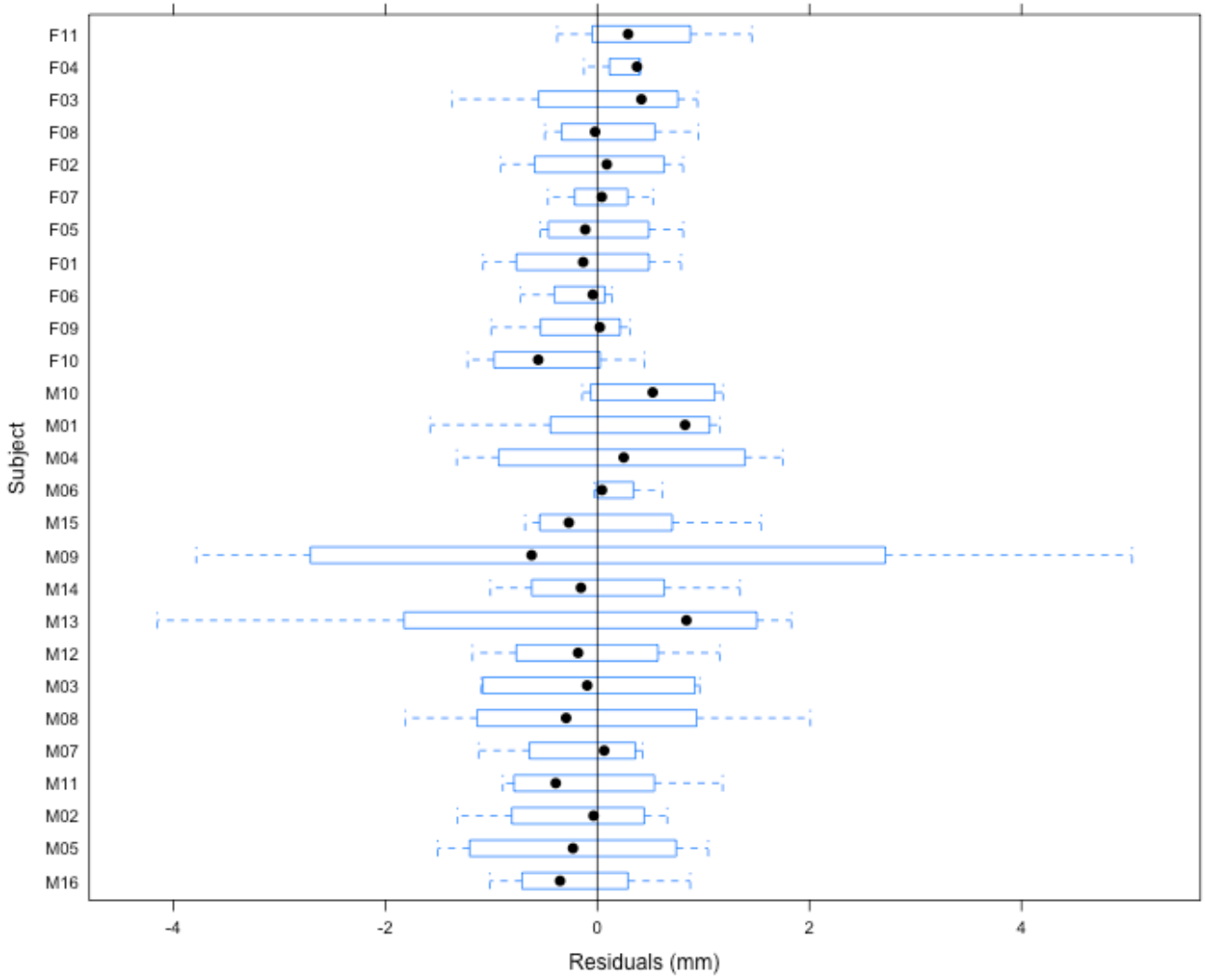
Box plots of Residuals by Subject
Maxillary Distances Data
Mixed model with Sex and Age interaction

Used to examine within-subject residuals:

1. Mostly centered at zero;
however, since there are only 4 measurements per subject we cannot rely too much to infer about within-subject variances.

2. Observation M09 has very large residuals.

3. Males residuals have larger variability than females which could violate the variance homogeneity assumption.

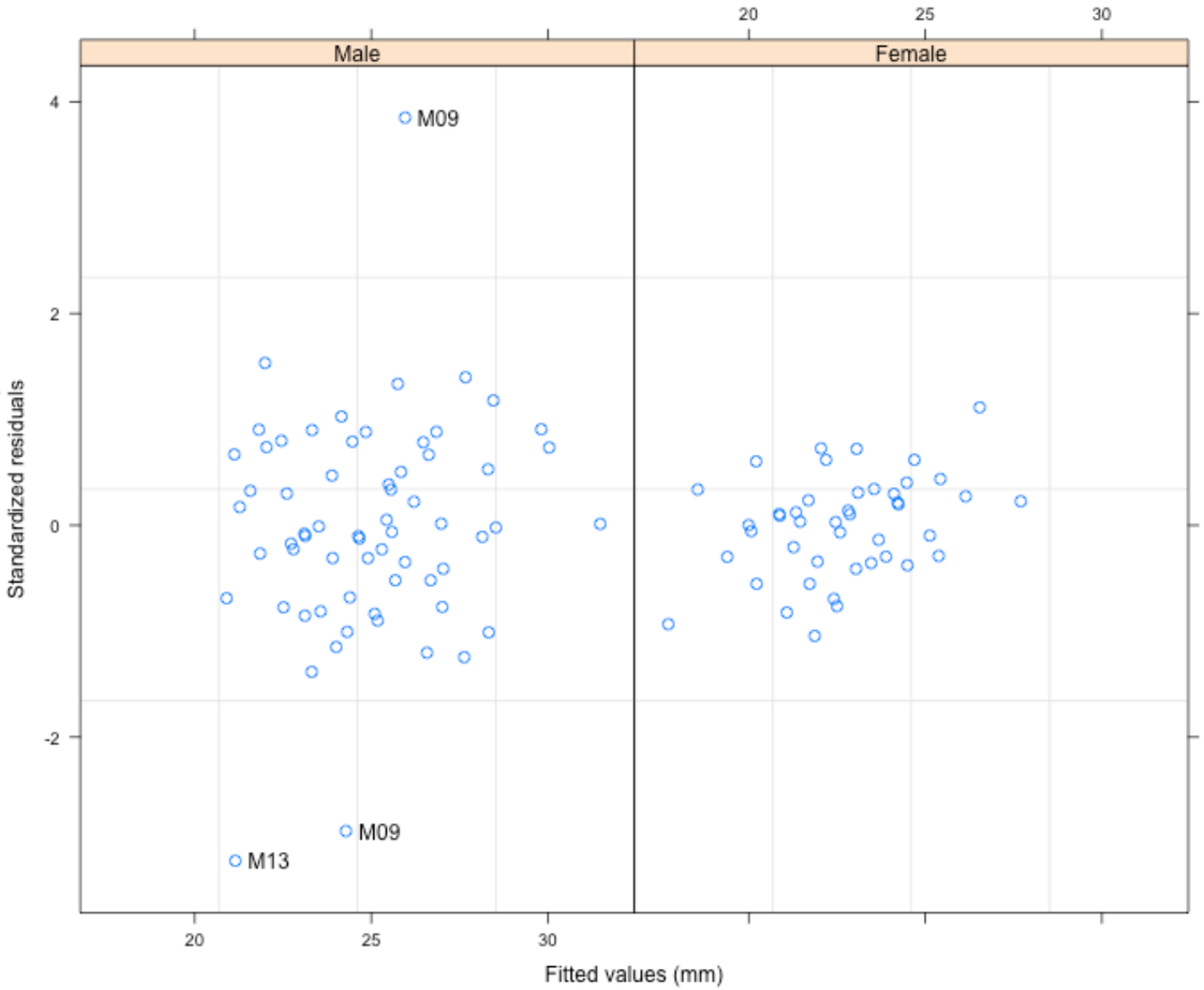



```
plot(fitt.sexage,Subject~resid(.),abline=0)
```

Residual Analysis and Goodness of Fit

Greater variability among males is also apparent here.

Two observations from subject M09 and one from M13 appear as outliers.




```
plot(fitt.sexage,resid(.,type="p")~fitted(.)|Sex,id=0.05,adj=-0.3)
```

Residual Analysis and Goodness of Fit

… more on the plot of standardized residuals:

. ref object: 'fitt.sexage'

```
plot(fitt.sexage,
             resid(.,type="p")~fitted(.)|Sex,
             id=0.05,
             adj=-0.3)
```

The `id=0.05` option is used to identify those residuals above or below the value of a Standard Normal quantile given by `1-id/2`. In this case, 1-0.5/2=0.975.

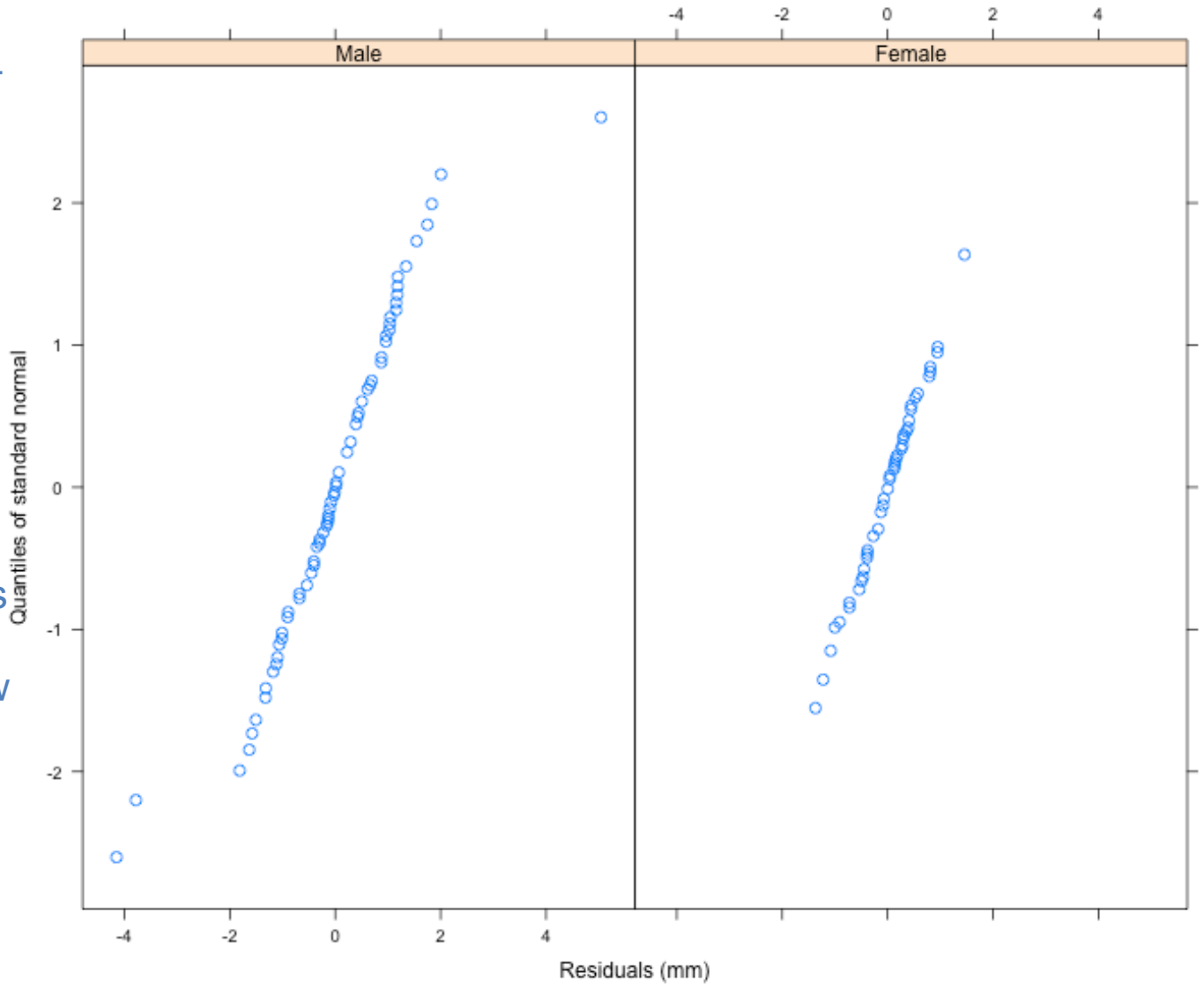
The `adj=-0.3` controls the position of the identifying labels.

Residual Analysis and Goodness of Fit

Used to examine within-subject residuals:

Again, we see three outlying points that we correspond to the male subjects previously identified as M09 and M13.

Aside from these three outliers, the distributions for both males and females appear to follow a straight line, hence the normality assumption is met.




```
qqnorm(fitt.sexage, ~resid(.)|Sex)
```

Residual Analysis and Goodness of Fit

The typical call for the qqnorm() function in LME modeling in R is:

```
qqnom( object ,  formula)
```

Where:

`object` is an `lme` fit.
`formula` is a one-sided formula of the form `~x|g`.

- The `x` term can be either the residuals or the predicted random effects associated with the `lme` fit specified through `object`.
- The `g` defines an optional grouping factor determining the panels of the display.

```
qqnorm(fitt.sexage, ~resid(.)|Sex)
```

Pinheiro & Bates, 2011.

Residual Analysis and Goodness of Fit

random effects are the uij's, 个体的值跟总体平均的差值

Assumption 2: The random effects are normally distributed, with mean zero and covariance matrix G (not depending on the group) and are independent for different groups (e.g., Females vs. Males)

Var(ui)+Var(eij) = Matrix [sigma^2] = Matrix [g] =G

- To assess normality of the distribution of the predicted random effects and identifying outliers:
  - Normal Q-Q plots.

eij 是 每个个体内部的数据跟个体平均的差值
uij 是个体平均的数据跟总体平均的差值

- To assess homogeneity of the predicted random effects covariance matrix and identifying outliers:
  - Scatterplot matrix of the predicted random effects

In R, predicted random effects are accessed through:

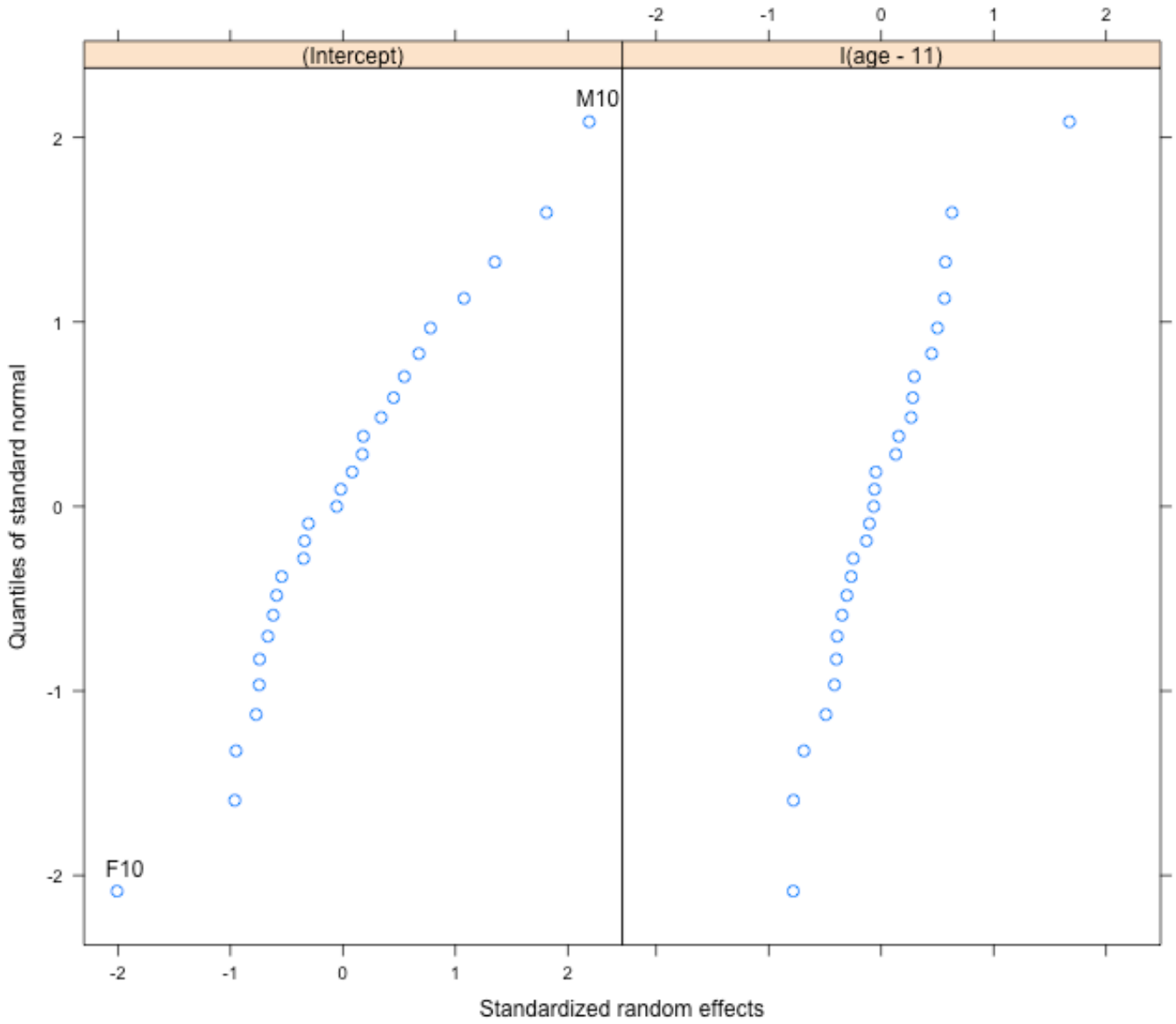
```
fitt.sexage$coef$random
ranef(fitt.sexage)
Ranef(fitt.sexage,standard=T)
```

Standardized to the estimated standard deviation, allows for direct comparison with a N(0,1).

Residual Analysis and Goodness of Fit

The normality assumptions seems reasonable for the predicted random effects, though there is some asymmetry in the distribution corresponding to the Intercept.

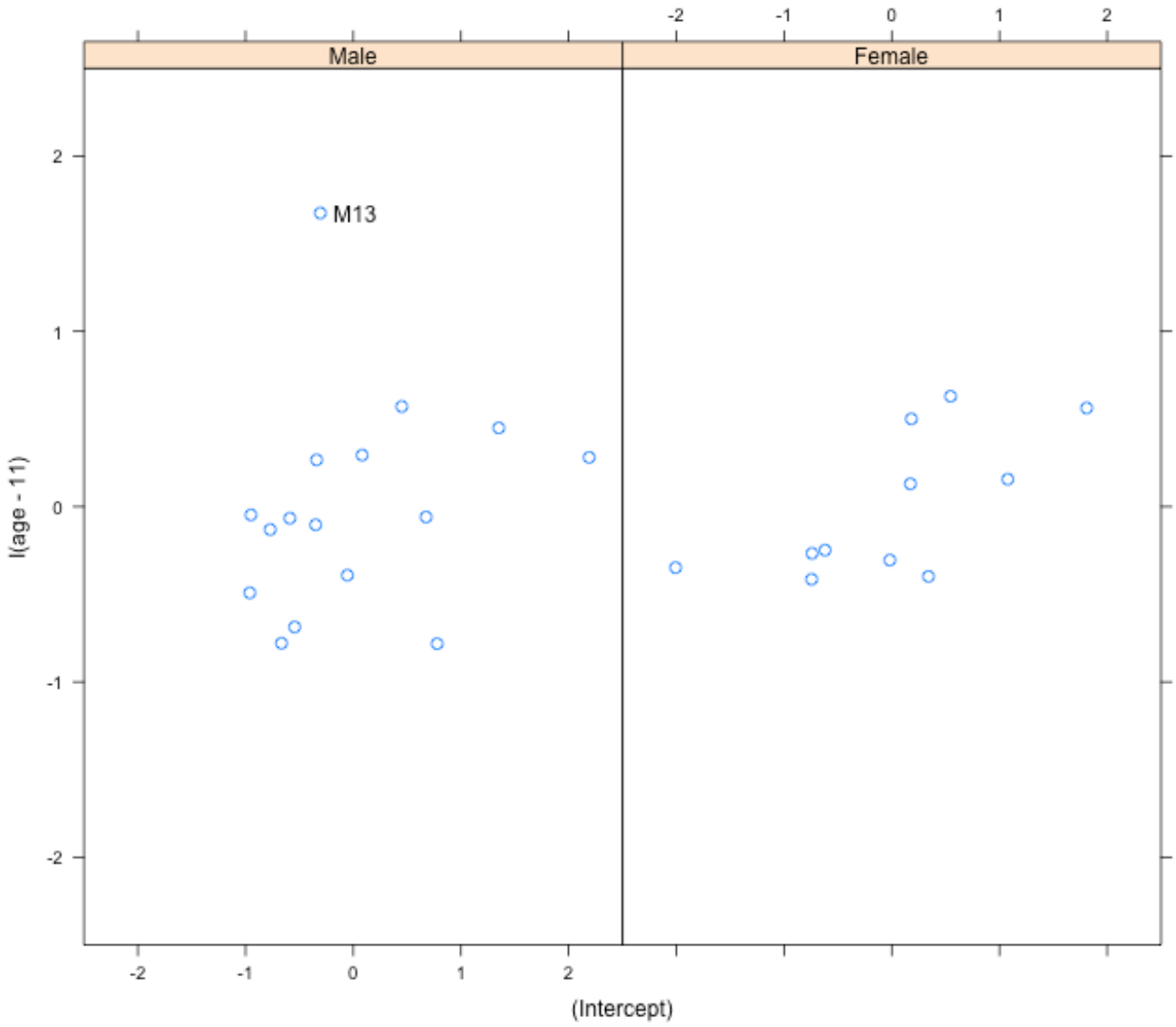
A few outliers appear to be present in both panels, F10, F11, M10 for the intercept and M13 for the slope.




```
qqnorm(fitt.sexage, ~ranef(.,standard=T), id=0.05, adj=c(.3,-1))
```

Residual Analysis and Goodness of Fit

Except for the value for M13, the predicted random effects seem to have similar distributions across Females and Males groups.




```
pairs(fitt.sexage, ~ranef(.,standard=T)|Sex, id=~Subject=="M13",
                          adj=-0.3, ylim=c(-2.5,2.5),xlim=c(-2.5,2.5))
```

Residual Analysis and Goodness of Fit

Influence diagnostics for LME in R

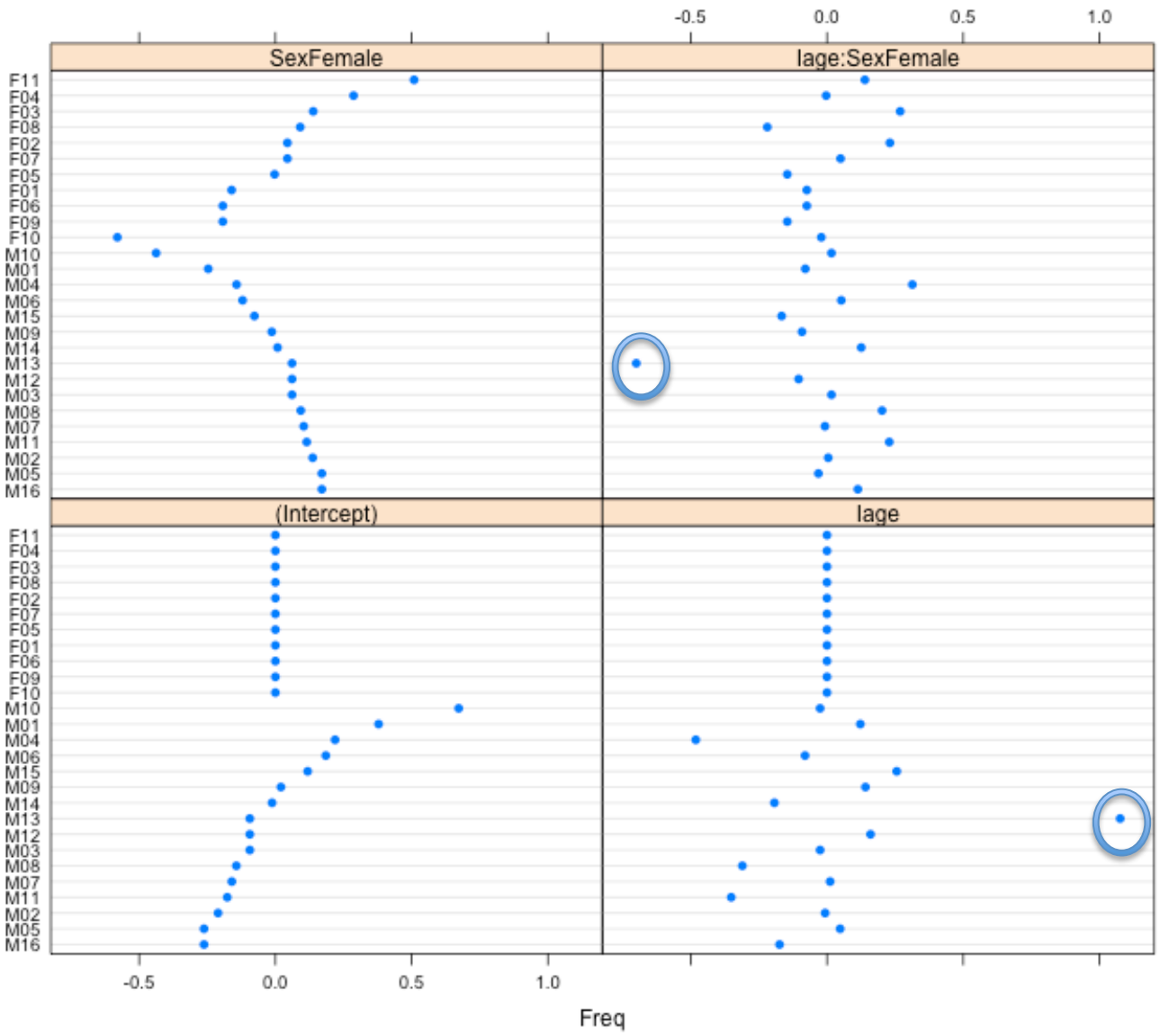
新的package

```
library(lme4)
library(influence.ME)

dat$Iage <- dat$age-11
fitt.sexage2 <- lmer(distance~Iage+Sex+Iage:Sex + (Iage|Subject),data=dat)

summary(fitt.sexage2)
fitted(fitt.sexage2)
resid(fitt.sexage2,standard=T)
ranef(fitt.sexage2)


# obtain plots of dfbetas and cook's distances
alt.est <- influence(fitt.sexage2, "Subject")
plot(alt.est, which="dfbetas")
plot(alt.est, which="cook",sort=T)
```

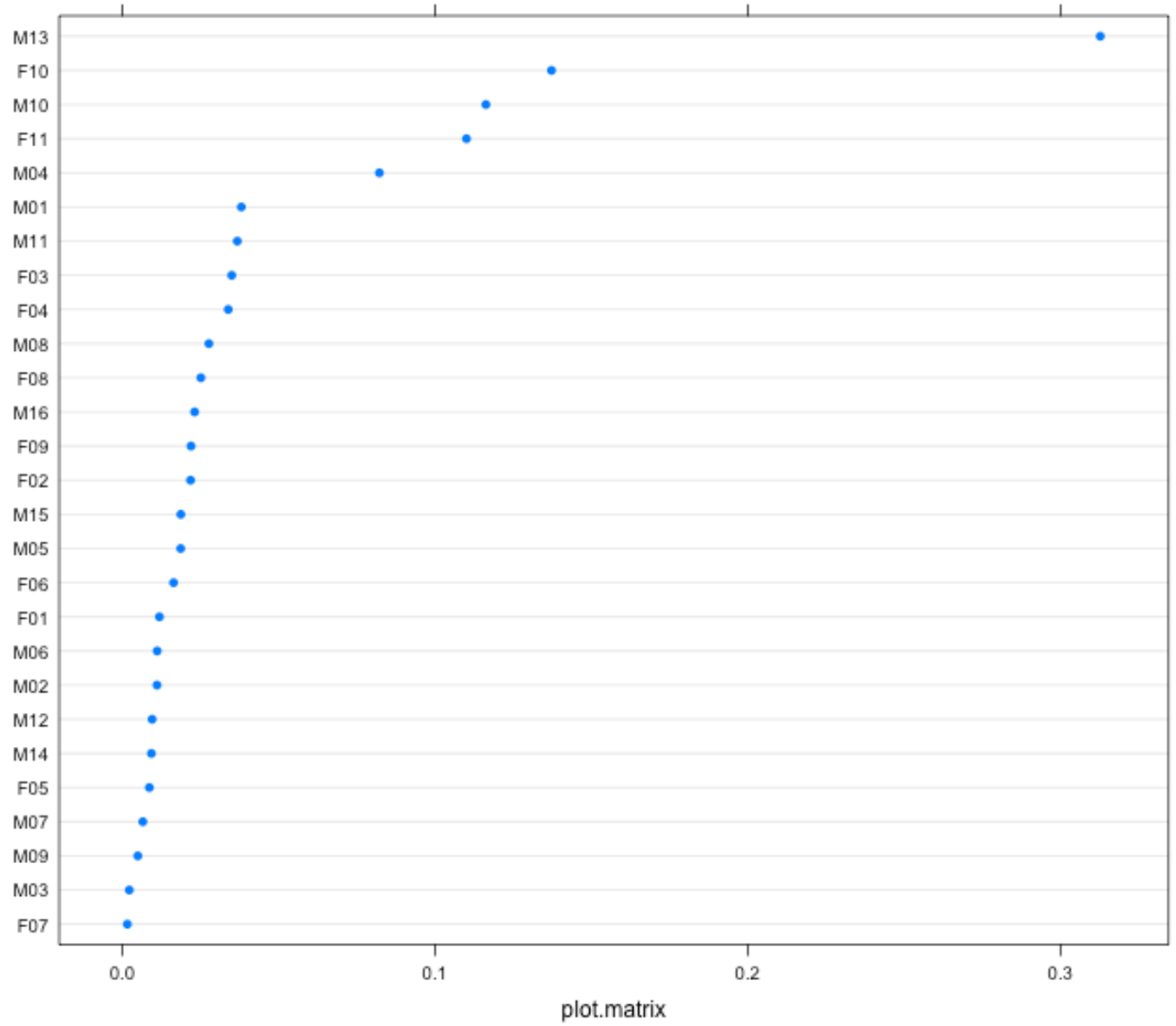
lmer() is analogue to lm() function

Access to various lmer objects

Residual Analysis and Goodness of Fit



Residual Analysis and Goodness of Fit



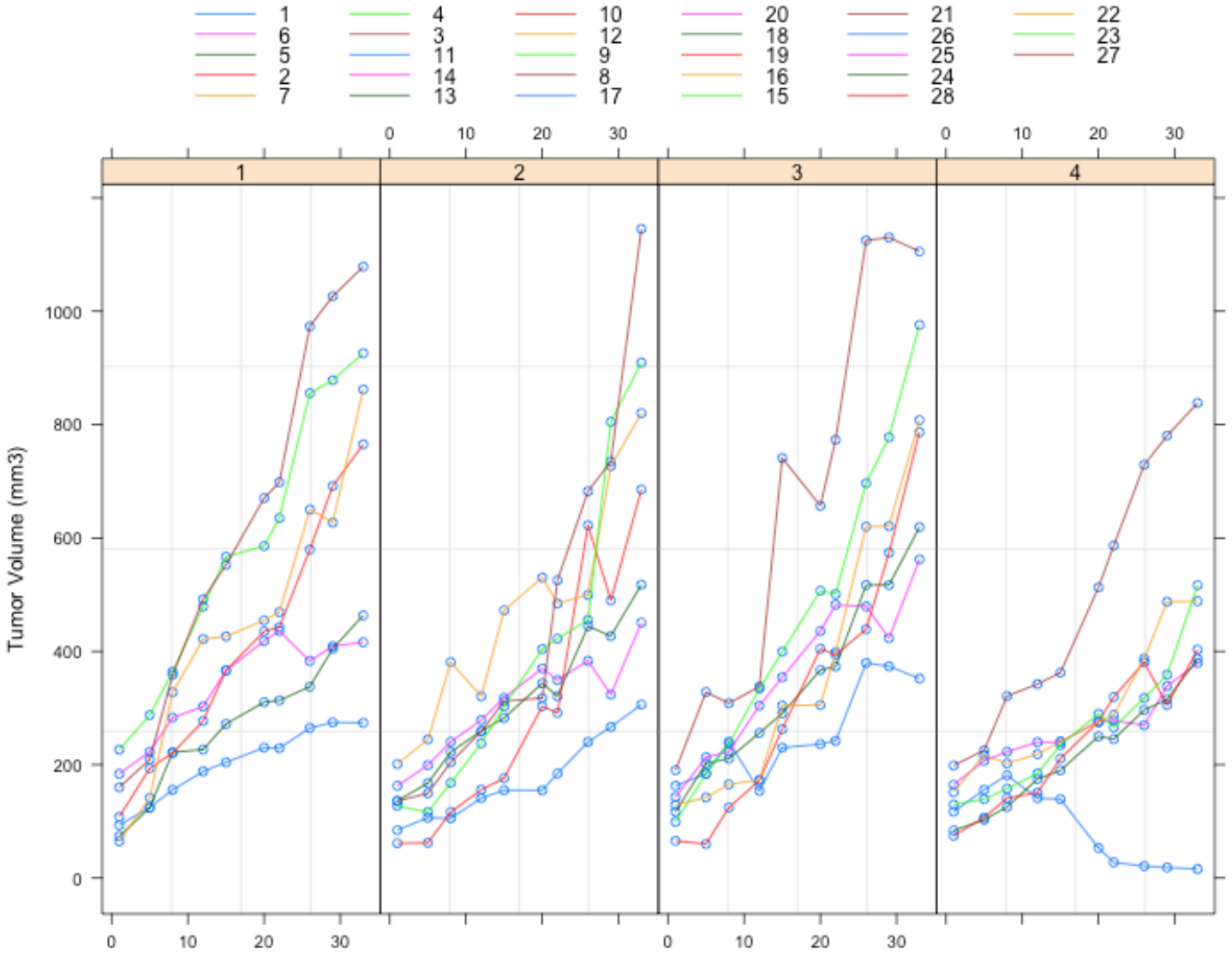
7. Implementation in R

Exercise: Tumor Growth Data
(Bonate, 2011)

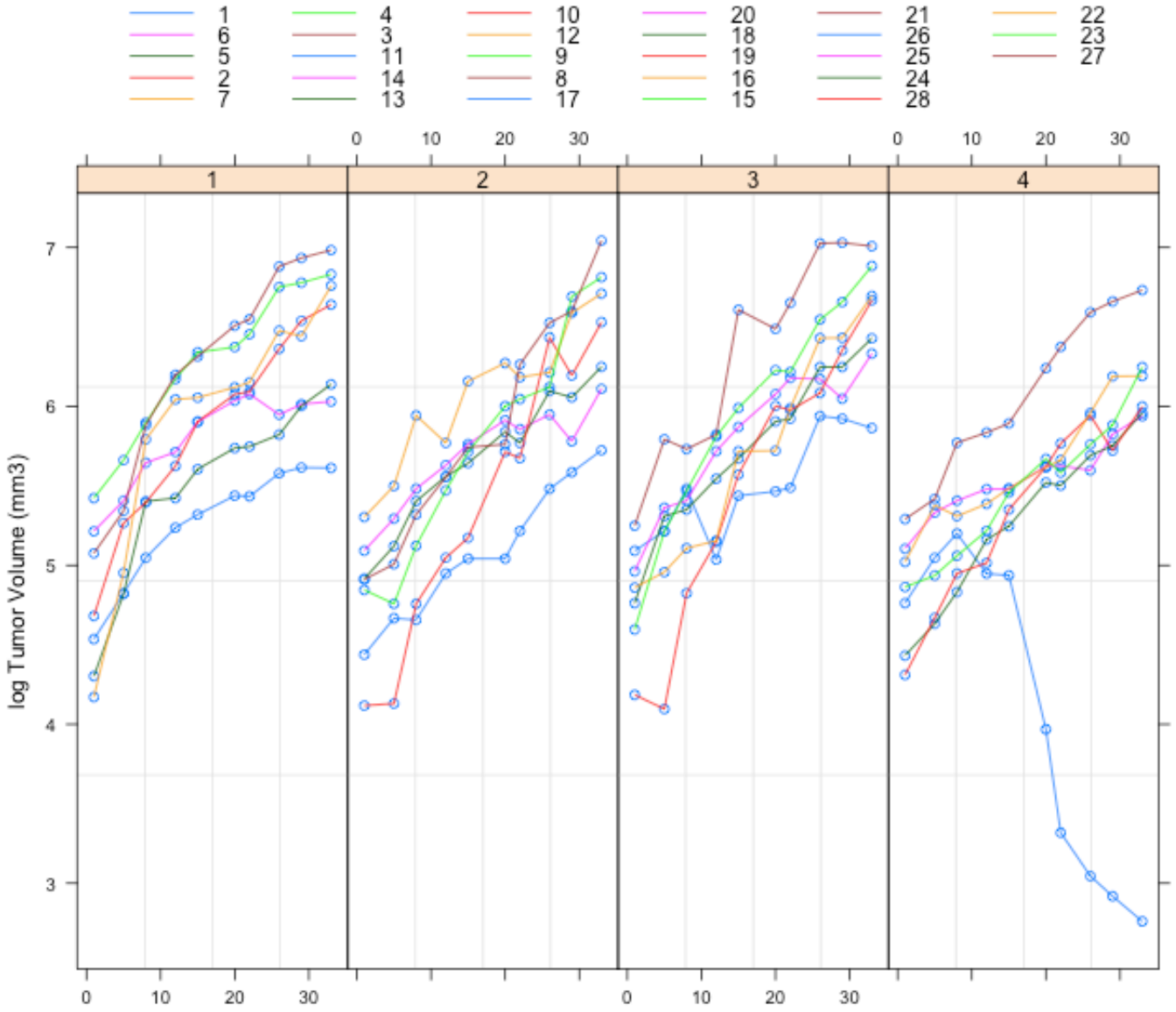
In the development of a new oncolytic, an experiment was conducted:

- Mice subcutaneously implanted with A549 human lung tumors
- Were randomized to four treatment groups:
    1. Saline once daily by intraperitoneal (IP) administration (control group)
    2. Drug 10 mg/kg once daily by oral (PO) administration for 28 days
    3. Drug 100 mg/kg once daily by PO administration for 28 days
    4. Drug 10 mg/kg once daily by IP administration for 28 days
- Response measure: Tumor Volume (mm3) based on length and width.
- Secondary variable: Weight
- Times of measurements: Days 1, 5, 8, 12, 15, 20, 22, 26, 29, and 33
- Seven (7) mice were randomized into each treatment group.

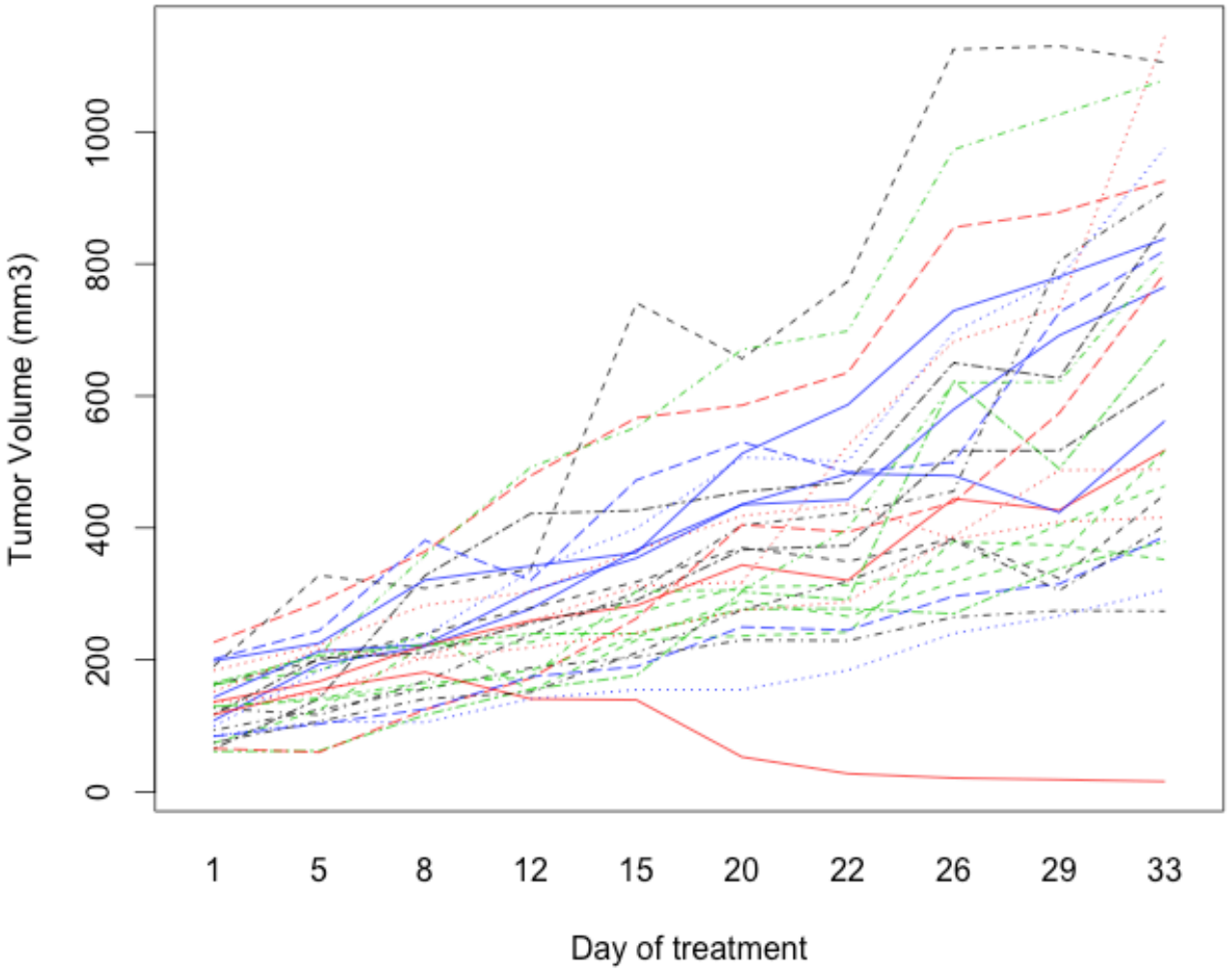
Exercise: Tumor Growth Data
(Bonate, 2011)



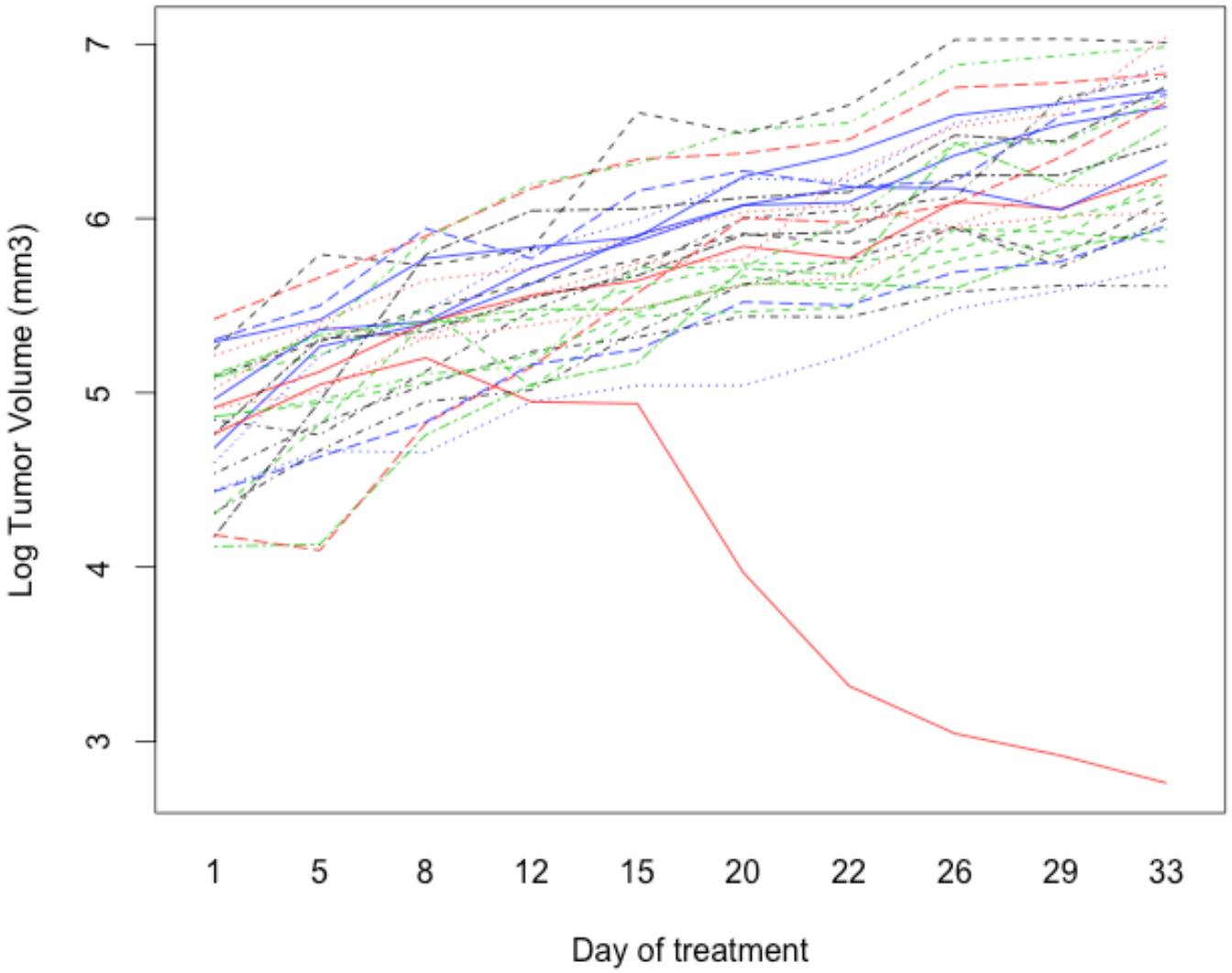
Exercise: Tumor Growth Data
(Bonate, 2011)



Exercise: Tumor Growth Data
(Bonate, 2011)



Exercise: Tumor Growth Data
(Bonate, 2011)



Exercise: Tumor Growth Data
(Bonate, 2011)

Based on the model below,

1. Assess the significance of the random effects ("REML" based LRT)
2. Perform the backwards elimination method for the fixed effects ("ML" based LRT) using the random effects structure determined in (1).
3. Fit the final model via REML to obtain unbiased estimates.
4. Plot the individual fitted trajectories.
5. Analyze the residuals.

Group 1 is control group

$$\begin{aligned}\log Vol_{ij} &= \beta_0 + \beta_1 G2_i + \beta_2 G3_i + \beta_3 G4_i + \beta_4 Wt_{ij} \\ &+ \beta_5 Day_{ij} + \beta_6 Day^2{}_{ij} \\ &+ \beta_7 \left(G2 \times Day\right)_{ij} + \beta_8 \left(G3 \times Day\right)_{ij} + \beta_9 \left(G4 \times Day\right)_{ij} \\ &+ \beta_{10} \left(G2 \times Day^2\right)_{ij} + \beta_{11} \left(G3 \times Day^2\right)_{ij} + \beta_{12} \left(G4 \times Day^2\right)_{ij} \\ &+ u_{i1} + u_{i2} Day_{ij} + \varepsilon_{ij}.\end{aligned}$$

Exercise: Tumor Growth Data
(Bonate, 2011)

Final Model:

$$\log Vol_{ij} = \beta_0 + \beta_1 G2_i + \beta_2 G3_i + \beta_3 G4_i$$
$$+ \beta_4 Day_{ij} + \beta_5 Day^2{}_{ij}$$
$$+ \beta_6 \left(G2 \times Day\right)_{ij} + \beta_7 \left(G3 \times Day\right)_{ij} + \beta_8 \left(G4 \times Day\right)_{ij}$$
$$+ \beta_9 \left(G2 \times Day^2\right)_{ij} + \beta_{10} \left(G3 \times Day^2\right)_{ij} + \beta_{11} \left(G4 \times Day^2\right)_{ij}$$
$$+ u_{i1} + u_{i2} Day_{ij} + \varepsilon_{ij}.$$

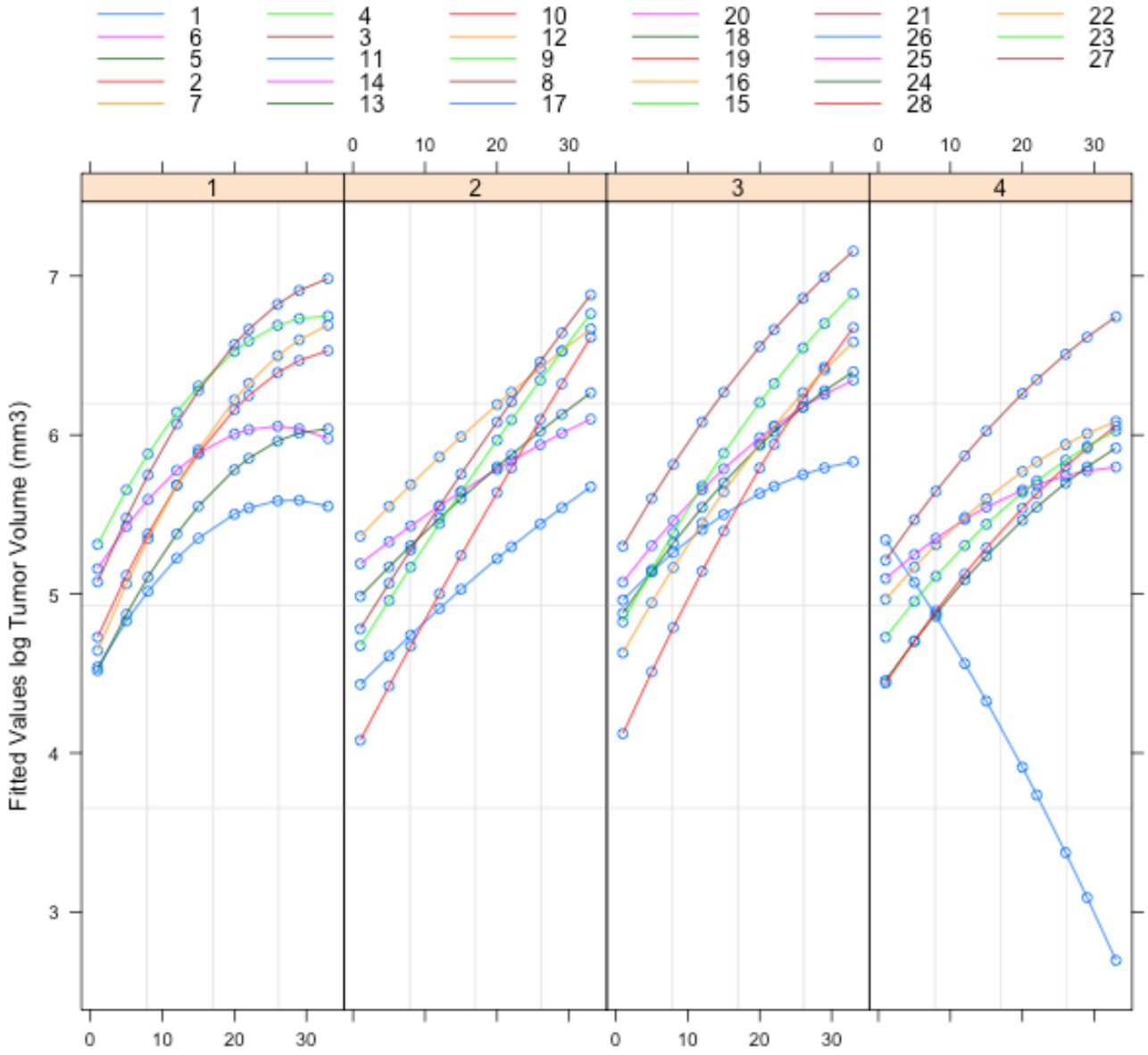
Ui1 is the intercept
Ui2 is the slope

(DarlSubject) 是 Day 的 Ui，slop random effect 自动出现

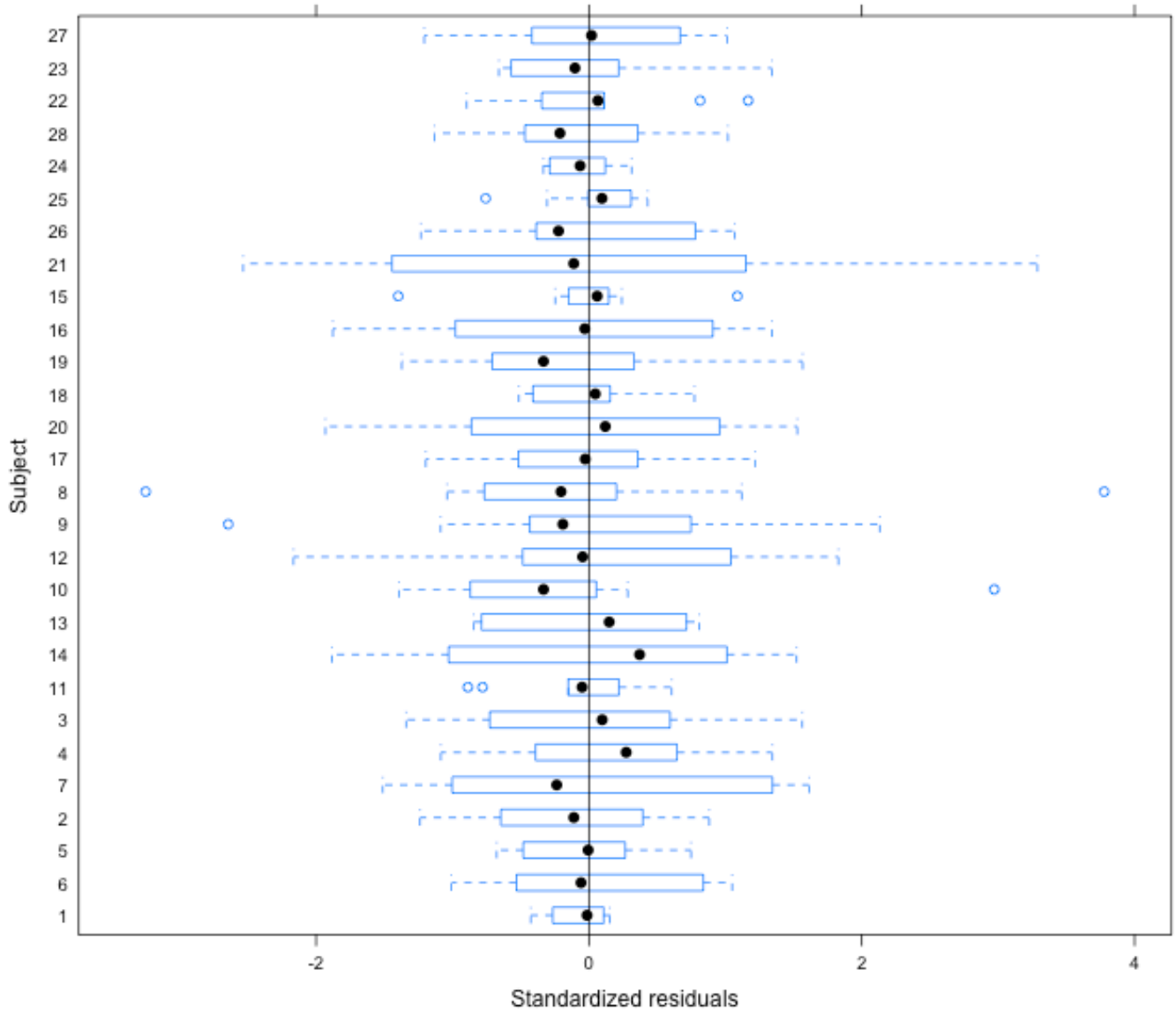
Exercise: Tumor Growth Data
(Bonate, 2011)

```
Linear mixed-effects model fit by REML
 Data: tg.dat
       AIC      BIC    logLik
  75.38893 132.8447 -21.69447

Random effects:
 Formula: ~Day | Subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 0.40285609 (Intr)
Day         0.02826608 -0.571
Residual    0.15452737

Fixed effects: logVol ~ Group + Day + Day2 + Group:Day + Group:Day2
                Value  Std.Error  DF   t-value p-value
(Intercept)  4.758996 0.16101395 244 29.556419  0.0000
Group2      -0.030807 0.22770812  24 -0.135290  0.8935
Group3      -0.006177 0.22770812  24 -0.027125  0.9786
Group4       0.091124 0.22770812  24  0.400179  0.6926
Day          0.096695 0.01285924 244  7.519514  0.0000
Day2        -0.001460 0.00020403 244 -7.154388  0.0000
Group2:Day  -0.038142 0.01818571 244 -2.097343  0.0370
Group3:Day  -0.021523 0.01818571 244 -1.183510  0.2378
Group4:Day  -0.054748 0.01818571 244 -3.010516  0.0029
Group2:Day2  0.001242 0.00028855 244  4.304967  0.0000
Group3:Day2  0.000836 0.00028855 244  2.898289  0.0041
Group4:Day2  0.000894 0.00028855 244  3.097391  0.0022
```

Exercise: Tumor Growth Data
(Bonate, 2011)

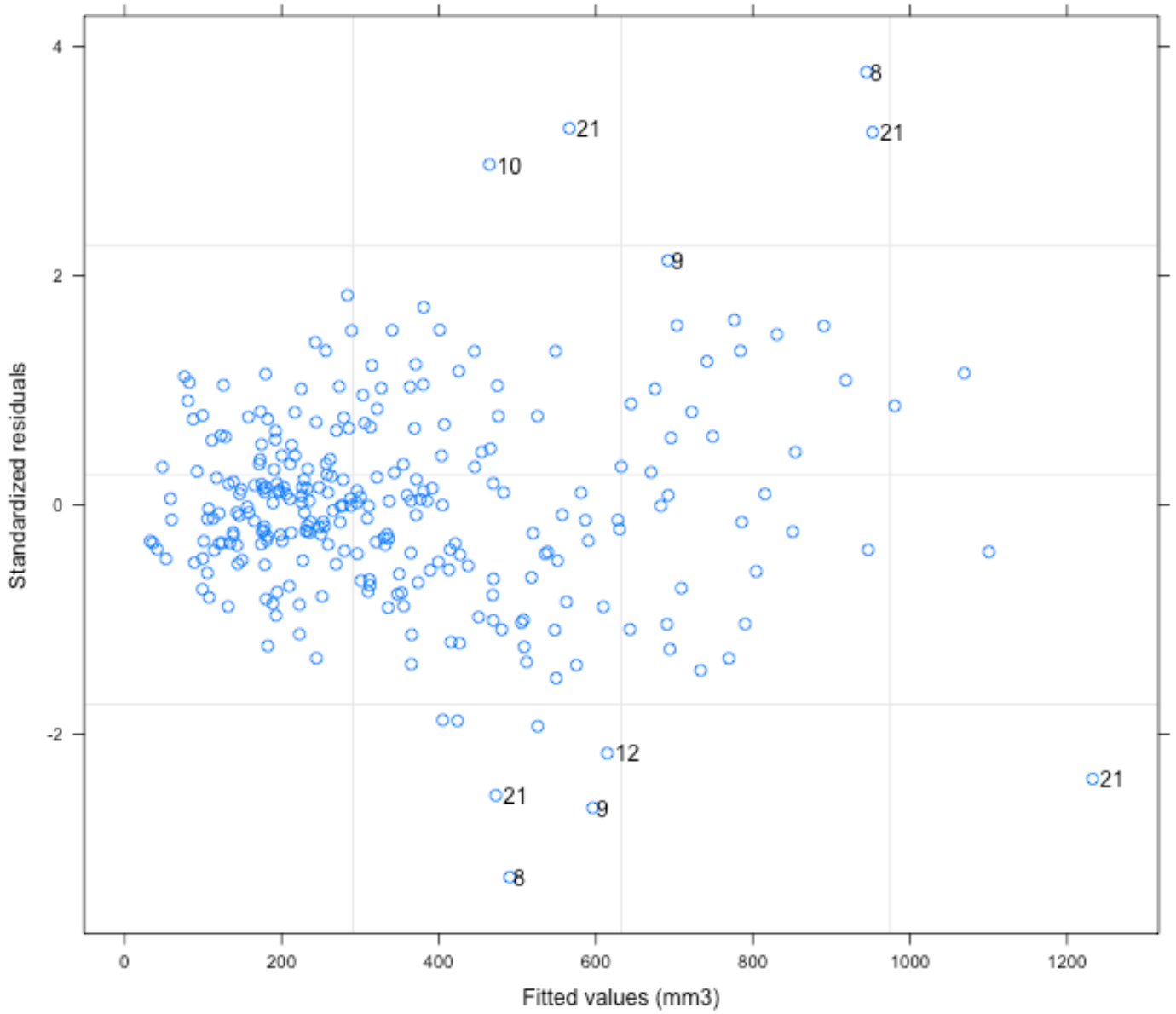


Exercise: Tumor Growth Data
(Bonate, 2011)



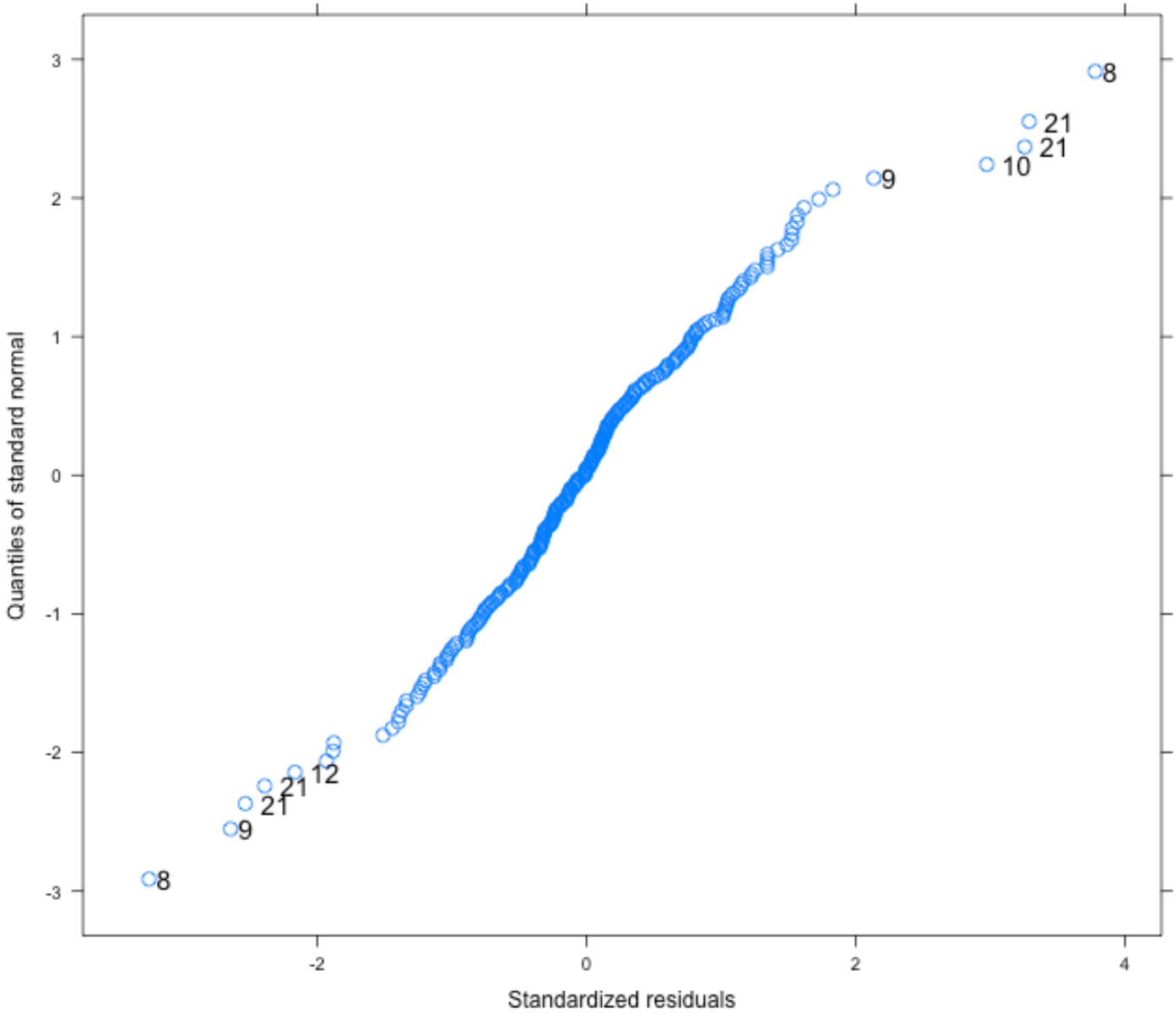

```
plot(final.fit,Subject~resid(.,type="p"),abline=0)
```

Exercise: Tumor Growth Data
(Bonate, 2011)



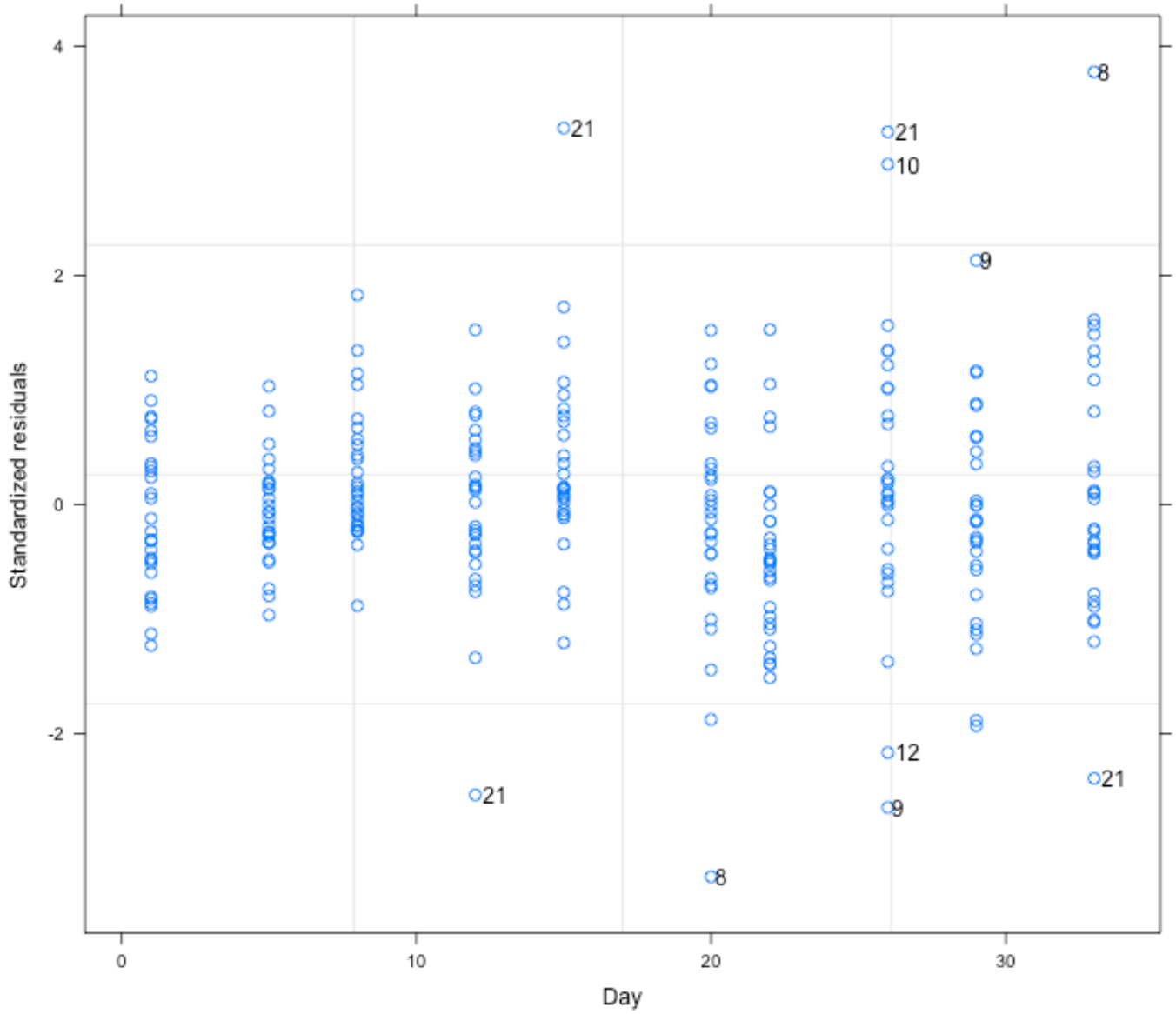

```
plot(final.fit,resid(.,type="p")~fitted(.),id=0.05,adj=-0.3)
```

Exercise: Tumor Growth Data
(Bonate, 2011)



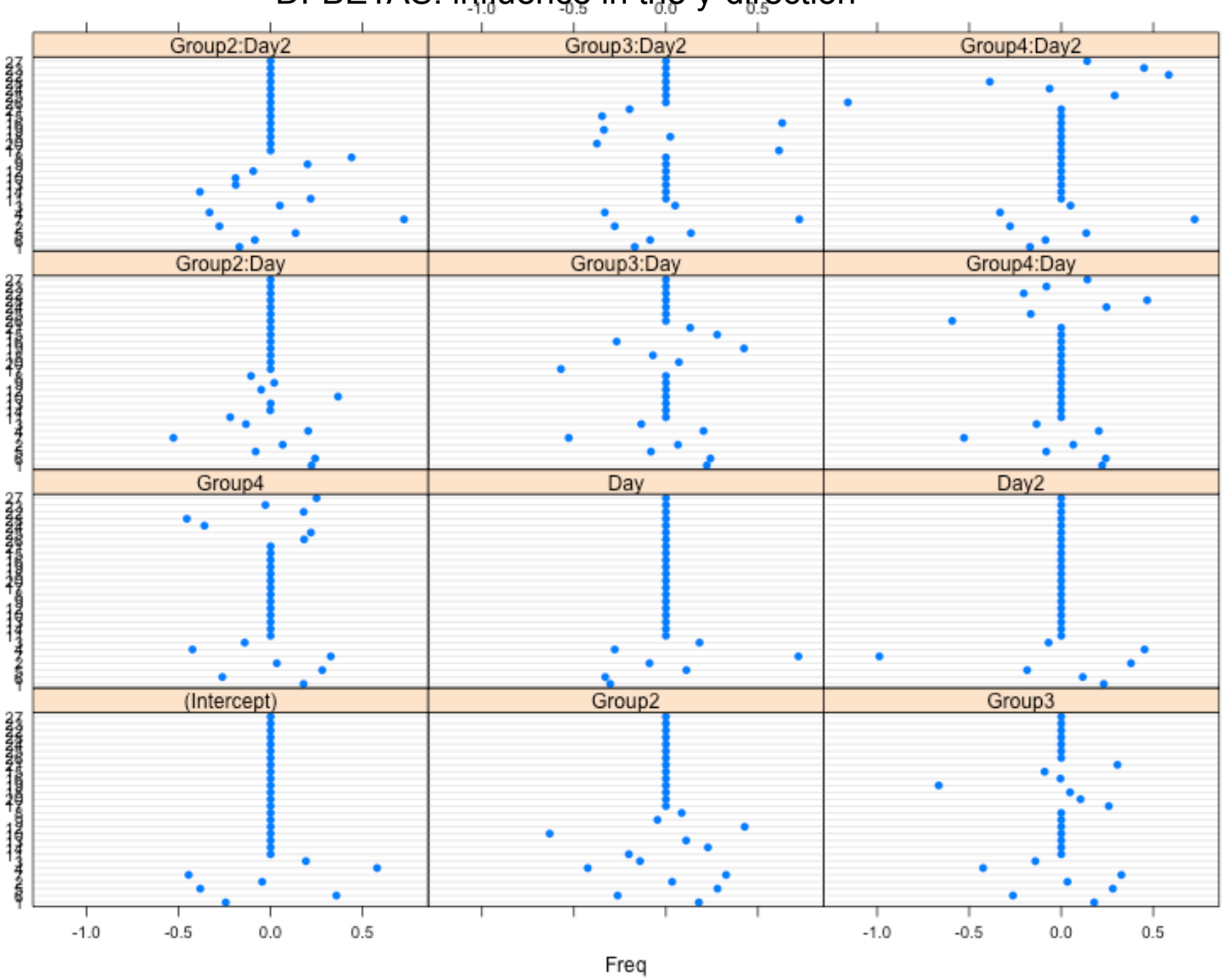

```
qqnorm(final.fit, ~resid(.,type="p"), id=0.05, cex=1.2, adj=-.5)
```

Exercise: Tumor Growth Data
(Bonate, 2011)

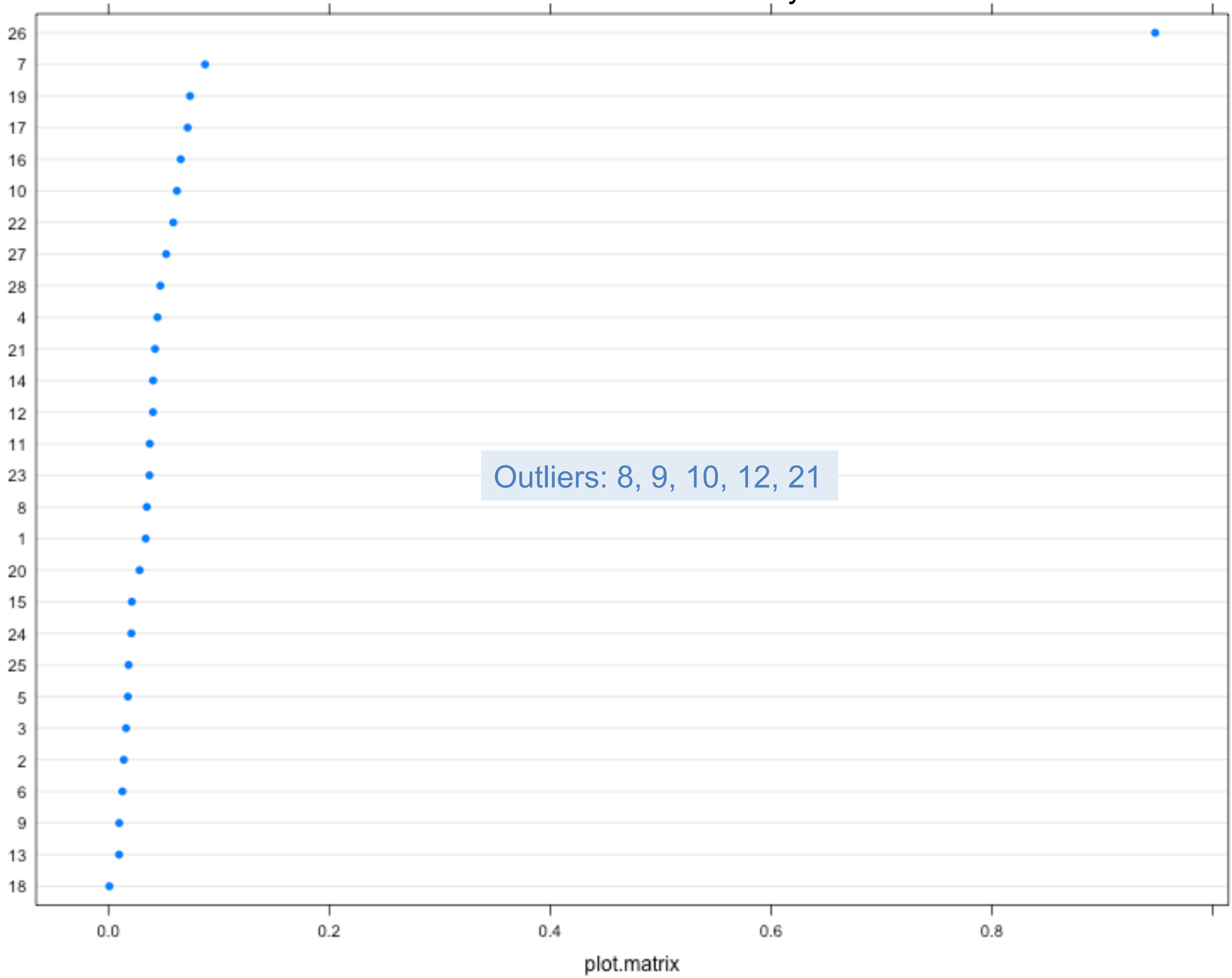



```
plot(final.fit,resid(.,type="p")~Day,id=0.05,adj=-0.3,xlab="Day")
```

DFBETAS: influence in the y-direction



Cook’s Distance: influence in the y-direction



```
final.fit2 <- lmer(logVol ~ Group + Day + Day2 +
          Group:Day + Group:Day2 + (Day|Subject),
          data=tg.dat)

alt.est <- influence(final.fit2, "Subject")

plot(alt.est, which="dfbetas")

plot(alt.est, which="cook",sort=T)
```