Notes for Tuesday February 17, 1998

Some Notes and Intuitions on Split-Plot Designs and Analysis

Recall the basic model for the guayule data,

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} + \delta_l$$

where

$$\begin{aligned} \epsilon_{ijk} &\overset{iid}{\sim} N(0,\sigma^2) \text{ and} \\ \delta_l &\overset{iid}{\sim} N(0,\tau^2), \text{ independent of the } \epsilon_{ijk}\text{'s.} \end{aligned}$$

with the following meanings:

RESPONSE VARIABLE:

y_{ijkl}	is the number of germinations per batch of 100 seeds

FIXED EFFECTS:

μ	is the grand mean
α_i	represents deviations from μ due to each variety of guayule
β_j	represents deviations from μ due to each treatment
$(\alpha\beta)_{ij}$	represents deviations from μ due to interactions of variety and treatment

RANDOM EFFECTS (RANDOM ERRORS):

ϵ_{ijk}	represents error from individual batches of 100 seeds
δ_l	represents error from individual flats

A *fixed effect* is a parameter that is estimated directly in the model. A *random effect* is an error term. Models with a single error term are called *fixed effects models*. Models with a more than one error term (and some fixed effects) are called *mixed effects models*. Models with all error terms and no fixed effects (besides the grand mean) are called *random effects models*. We will learn more about these in the next couple of lectures.

For now, there are a couple of things to note about the design of this experiment, and the meanings of the parameters and variables above.

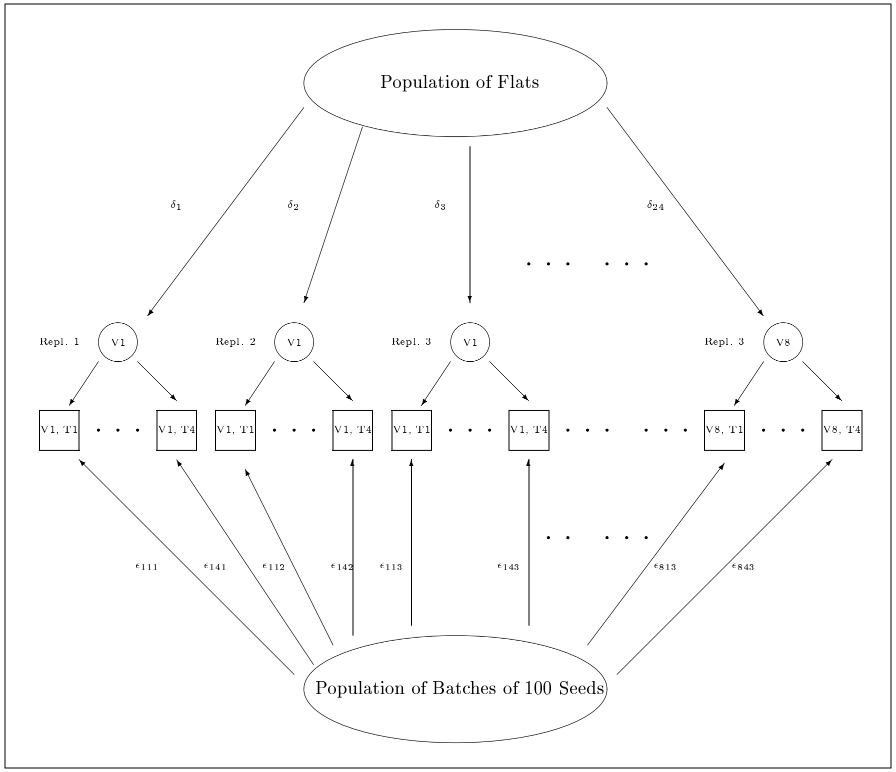
1. The index l for flats is a function of the indices i for variety and k for replications: $l = (i-1)*3+k$. We say that the replications k are *nested within* flats l, for example. Variety i is also nested within flats.

2. Since every combination of the indices j for treatment and l for flats occurs, we say treatment and flats are *crossed* or *completely crossed*. Similarly, treatment j and variety i are crossed. "Crossed" just means this part of the design is a "factorial" design. It is interesting to note that although i and j are crossed, and j and l are crossed, i and l are not crossed, they are nested.

3. In the above model there are two sources of error:
   - Error due to variation in batches of 100 seeds, modeled by ϵ_{ijk}, and
   - Error due to variation in the flats, modeled by δ_l.

   These sources "group" the data differently, that is, a different "draw" from the ϵ_{ijk} distribution is "attached" to every germination count y_{ijkl}, but the same "draw" from the δ_l distribution is "attached" to all four germination counts $y_{i1kl}, y_{i2kl}, y_{i3kl}, y_{i4kl}$ within the same flat. Because of the different groupings these sources are sometimes called *error strata* instead of error sources. Mixed and random effects models are also sometimes called *multistratum* or *multilevel* models.

Here is a picture of the data layout and sources of error:



$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} + \delta_l$$

There are really two "stacked" ANOVA problems here:

- The problem of assessing the effect of treatment, with variety as a blocking variable, contained in the bottom half of the graph (boxes and below); and
- The problem of assessing the effect of variety, with the variability in treatments removed (by averaging over treatments, say), contained in the top half of the graph (circles and above).

We will first do these two ANOVA's separately and then see how they are combined in the split-plot analysis.

Analysis of the bottom half ("Error: Within")

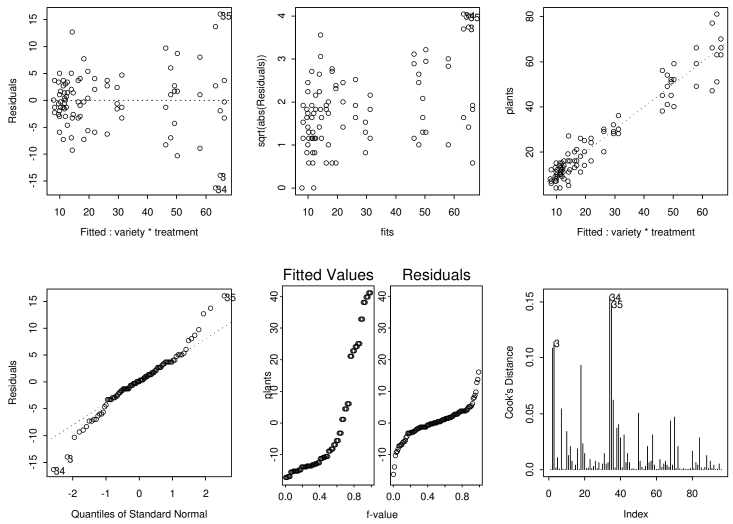
This is easy: we are just going to ignore the random effects of the flats in the analysis—in fact we are going to ignore the flats completely. The model and fit is simply:

```
402 > gua.below _ aov(plants ~ variety*treatment, data=guayule)
402 > anova(gua.below)
Analysis of Variance Table

Response: plants

Terms added sequentially (first to last)
                  Df Sum of Sq  Mean Sq  F Value      Pr(F)
          variety  7    763.16   109.02   2.7058 0.01604076
        treatment  3  30774.28 10258.09 254.5959 0.00000000
variety:treatment 21   2620.14   124.77   3.0966 0.00026666
        Residuals 64   2578.67    40.29
```

In this analysis there are 32 cells (8 varieties times 4 treatments) with three replications per cell. Essentially, "Variety" is being treated as a blocking variable, with the main interest being in "Treatment".

Note that "variety" seems to matter a lot, even though it is supposed to be just a blocking variable. The problem is the huge SS associated with "treatment", which cannot contribute to the residual SS used for testing "variety"; thus "treatment" is acting as a blocking variable for "variety" and (perhaps unnaturally) increasing the sensitivity of the model to "variety".

The result of `plot(gua.below)`, shown next,



suggests that (a) the residuals do not have constant variance across the fitted values, and (b) the residuals are a little fat-tailed, relative to the normal distribution.

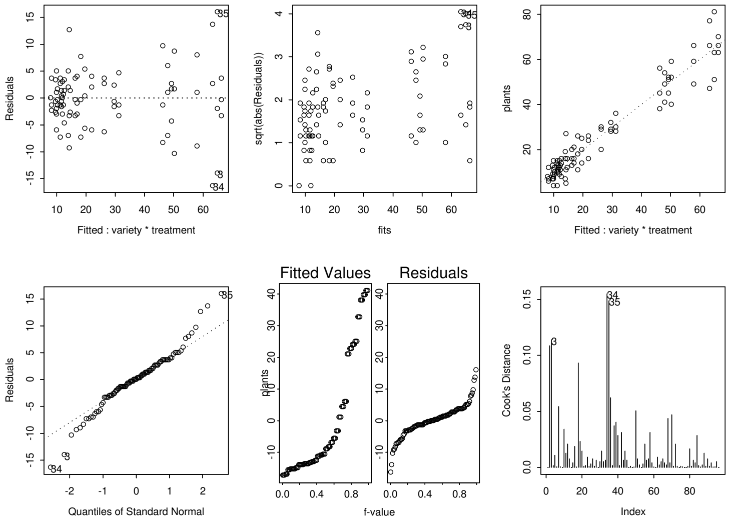
Analysis of the top half ("Error: Flats")

Now let's briefly consider a one-way ANOVA for the top half of the graph. For this analysis we reduce the raw germination counts in the boxes to averages over treatments, in the circles. That is, we want to consider how variable germination rates are across varieties, averaging out the variability across treatments.

```
402 > variety.flat _ data.frame(variety=factor(rep(1:8,rep(3,8))),
+ meanresp=sapply(split(guayule[,c(1,4)],guayule$flats),apply,2,mean)[2,])
There were 24 warnings (use warnings() to see them)
402 > anova(gua.above _ aov(meanresp ~ variety,data=variety.flat))
Analysis of Variance Table

Response: meanresp

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
  variety  7  190.7891 27.25558 1.232036 0.3420697
Residuals 16  353.9583 22.12240
```

In this analysis there are eight cells—one for each of the eight varieties, and three replications per cell. Note that "variety" is far from significant in this analysis, since variability due to "treatment" that seemed to be helping "variery" along has been averaged out.

The residual plots again suggest non-constant variance and extra-normal variability (fat tails) in the residuals (though unsurprisingly [why?] the problem is not as bad here).



Review of the full model analysis

Now we repeat the basic analysis of this data we saw from last time::

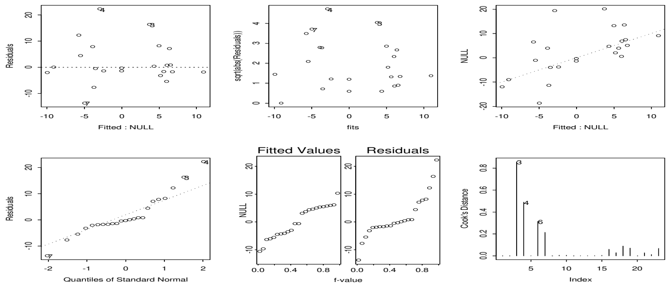
```
402 > gua.aov1 <- aov(plants ~ variety*treatment + Error(flats), guayule)
402 > summary(gua.aov1)
Error: flats
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
  variety  7   763.156 109.0223 1.232036 0.3420697
Residuals 16  1415.833  88.4896

Error: Within
                  Df Sum of Sq  Mean Sq  F Value        Pr(F)
        treatment  3  30774.28 10258.09 423.4386 0.00000e+000
variety:treatment 21   2620.14   124.77   5.1502 1.32674e-006
        Residuals 48   1162.83    24.23
```


As in the "top half" analysis above, "variety" is not a significant factor on averages across treatments within flats; except for a factor of four in the SS's (corresponding to averaging across the four treatments within each flat) this is identical to our "top half" analysis above.

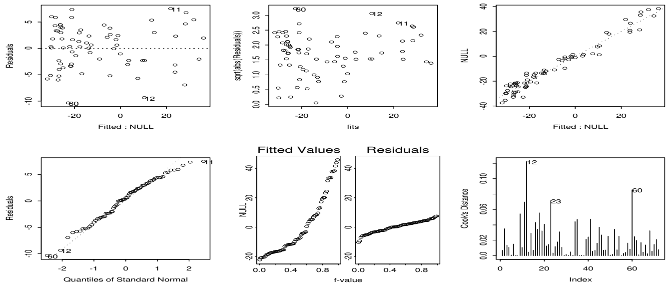
The remaining effects, "treatment" and "variety:treatment" are now tested using the same SS's as in our "bottom half" analysis above, but the residual SS's used is just the part attributable to individual batches of 100 seeds, after the residual SS's due to flats is taken out (note: $1415.833 + 1162.83 = 2578.67$, the residual SS from our "bottom half" model. This smaller residual SS makes the tests for "treatment" and "variety:treatment" more powerful (note the smaller p-values for example).

Finally, here are the residual diagnostic plots, for comparison with the plots based on the "above model" and the "below model" above.

```
402 > names(gua.fm1)
[1] "(Intercept)" "flats"       "Within"
402 > plot(gua.aov1$flats,title="Guayule Error(flats) flats")
```




```
402 > plot(gua.aov1$Within,title="Guayule Error(flats) within")
```



Split-plot vs. Blocking; Random Effects vs. Fixed Effects

We saw in the analyses above that an alternative to fitting the split-plot model

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} + \delta_l$$

would have been to fit the two-way ANOVA model

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where $i = 1, \ldots, 8$ indicates the variety, $j = 1, \ldots, 4$ indicates treatment, $k = 1, \ldots, 3$ indicates replications, and

$$\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

but that this model "ignored" the effects of the flats. An alternative not considered up to now would be to simply include the flats as another (say, additive) fixed effect:

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \tau_l + \epsilon_{ijl}$$

where

$$\epsilon_{ijl} \overset{iid}{\sim} N(0, \sigma^2)$$

When to do this, vs. the split-plot model? There are two questions to answer in deciding whether to treat a factor as a fixed effect or a random effect:

1. Suppose you were to obtain more replications of these observations. Could you do it without increasing the number of factor level parameters τ_l to estimate?
   - If "YES", then keep τ_l as a fixed effect;
   - If "NO", then replace τ_l with a random effect δ_l.
2. If you can get replications of the observations, would replications with the same value of l be independent of one another?
   - If "YES", then keep τ_l as a fixed effect;
   - If "NO", then replace τ_l with a random effect δ_l.

Split plot designs are problems where the answer to both questions is "no": we couldn't get more observations without introducing more flats, and thus increasing the number of levels of τ_l. Also, any replications we see are expeced to be dependent within flat (since they are influenced by common flat characteristics.

Other examples of mixed-effects linear models also arise because one or the other of the above questions is "no". We turn to one such example now.

An Example of a Mixed-Effects Problem

In repeated-measures data, multiple measurements are obtained from individuals on different occasions. For each individual, serial correlation usually exists, so a multivariate approach with structured covariance for individuals is very appealing. Although the distribution for the response is the same for each individual, some parameters might vary over individuals. The following examples show the features of such data and illustrate the new S-PLUS function lme.

Example: Orthodont Data. These repeated-measures data come from an orthodontic study presented in Potthoff and Roy (1964). They consist of four measurements of the distance in millimeters from the center of the pituitary to the pteryomaxillary fissure made at ages 8, 10, 12, and 14 years on 16 boys and 11 girls. The purpose of the study is to model the distance as a function of age, with consideration of gender difference. The data are available in a data frame that comes with SPLUS, called `Orthodont`, with columns `Subject`, `Sex`, `age`, and `distance` as indicated below.

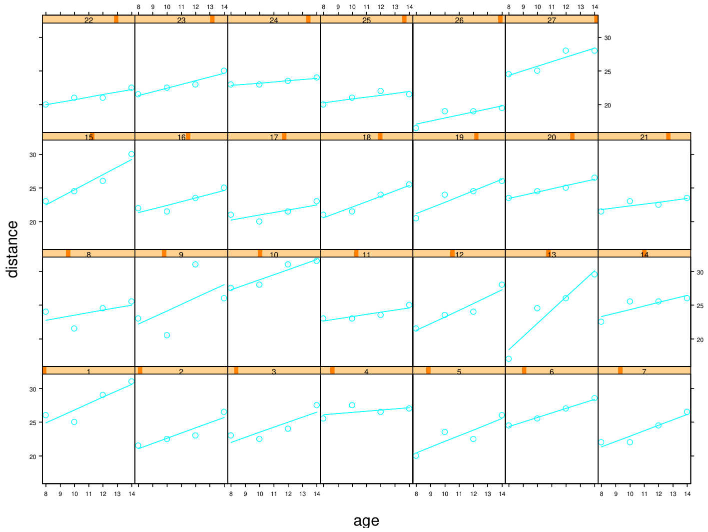
```
> Orthodont
Subject Sex age distance
1 1 0 8 26.0
2 1 0 10 25.0
3 1 0 12 29.0
4 1 0 14 31.0
.
.
.
105 27 1 8 24.5
106 27 1 10 25.0
107 27 1 12 28.0
108 27 1 14 28.0
```

In this data set, subjects are classified into two groups by Sex, an indicator variable assuming the value 0 for boys and 1 for girls. Each subject has four measures of distance, and the 108 total records are grouped into 27 clusters by Subject. They can be displayed graphically using the "Trellis" graphing functions in SPLUS.

```
402 > trellis.device(win.graph)
402 > xyplot(distance ~ age | Subject, data=Orthodont, type="b")
```

Here is a slightly fancier version

```
402 >  xyplot(distance ~ age | Subject, data=Orthodont,
+ panel=function(x,y) {
+ xy _ data.frame(x=x,y=y)
+ panel.xyplot(x,y)
+ panel.xyplot(x,fitted(lm(y ~ x,data=xy)),type="l")
+ }
+ )
```



Rather than fitting separate regression lines, as in the plot above, we would like to fit regression lines that are built up from the idea that each child has his/her own regression line that is randomly sampled from a population of regression lines (and then each child's growth data deviates randomly from his/her own line).

Two models are of special interest, because comparing them lets us explore the question of sex differences in these growth behaviors

- **Model 1.** For the distance measure y_{ij} of child i on occasion j

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) \cdot (age)_i + \epsilon_{ij}$$

  where

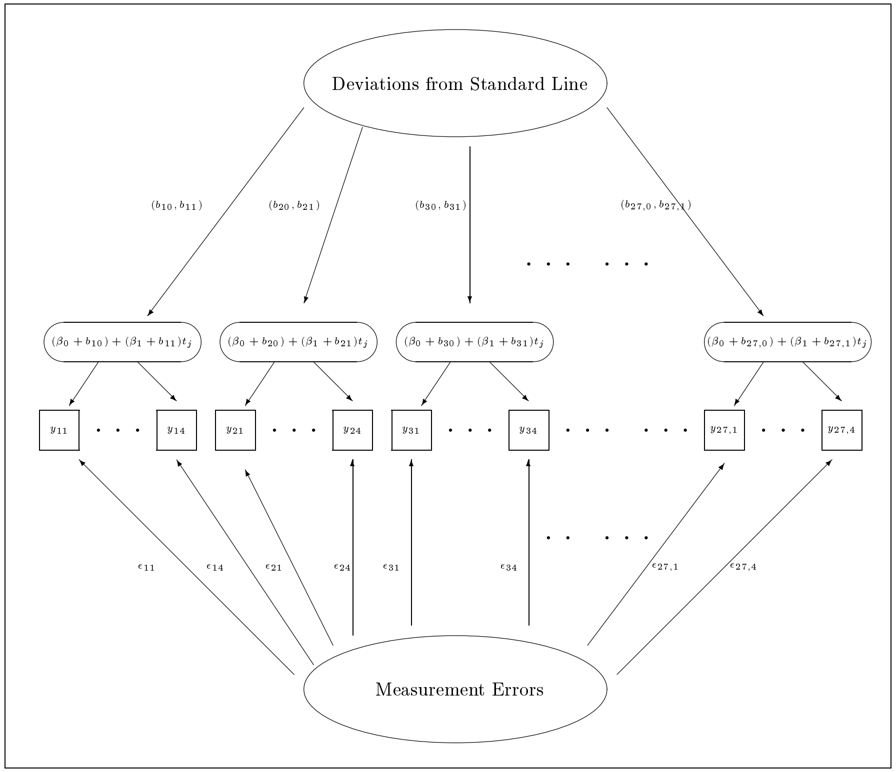
$$\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2)$$
$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \stackrel{iid}{\sim} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 D\right]$$

  and D has some simple structure (like, a constant times a 2×2 identity matrix).

- **Model 2.** Same model, but now with an additional *fixed effect* for Sex. Only the modeling equation changes, becoming:

$$y_{ij} = (\beta_{00} + \beta_{01} \cdot Sex_i + b_{i0}) + (\beta_{10} + \beta_{11} \cdot Sex_i + b_{i1}) \cdot (age)_i + \epsilon_{ij}$$

These are both *mixed effects linear models* since they have both parameters to be directly estimated (the fixed effects, represented by β's above) and random effects (represented by the b... parameters). As an illustration, we draw a diagram of the model for **Model 1.**



To fit these models we can use the specialized function `lme()` in SPLUS. `lme` stands for "linear mixed effects models. A template for an `lme()` call is

```
model.lme <- lme(fixed,random,cluster,data)
```

where

- `fixed` specifies the fixed effects part of the model
- `random` specifies the random effects part of the model
- `cluster` specifies over what units the random effects vary

- `data` specifies the data frame, as usual.

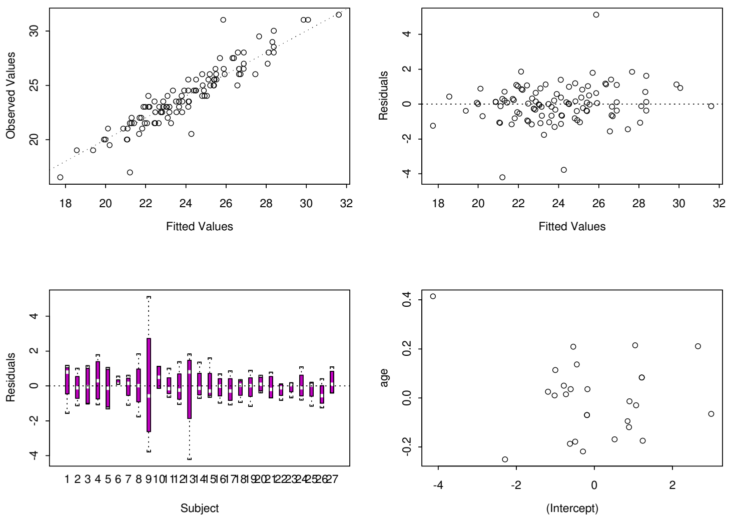
All except `data` are specified using the same modeling notation as in `aov()`. A response variable need only be indicated in the `fixed` portion of the command.

Fitting Model 1: $y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) \cdot (age)_i + \epsilon_{ij}$

```
402 > ortho.lme1 _ lme(fixed = distance ~ age, random= ~ age,
+ cluster = ~Subject, data = Orthodont)
402 > summary(ortho.lme1)
Call:
  Fixed: distance ~ age
 Random:  ~ age
Cluster:  ~ Subject
   Data: Orthodont

Estimation Method: RML
Convergence at iteration: 1
Restricted Loglikelihood: -221.3183
Restricted AIC: 454.6367
Restricted BIC: 470.7295

Variance/Covariance Components Estimate(s):
  Structure: unstructured
  Parametrization: matrixlog
  Standard Deviation(s) of Random Effect(s)
 (Intercept)       age
    2.327036 0.2264279
 Correlation of Random Effects
    (Intercept)
age -0.6093333

 Cluster Residual Variance: 1.716204

Fixed Effects Estimate(s):
                  Value Approx. Std.Error z ratio(C)
(Intercept) 16.7611111        0.77524619  21.620372
        age  0.6601852        0.07125327   9.265332

 Conditional Correlation(s) of Fixed Effects Estimates
    (Intercept)
age -0.8481509

Random Effects (Conditional Modes):
   (Intercept)          age
 1   1.0515836  0.215684513
 2  -0.7275021  0.014507873
```



```
 3  -0.1739026  0.035852373
 4   3.0004558 -0.065885037
 5  -1.1766686  0.025600404
 6   1.2176442  0.083191198
 7  -0.6079727  0.034900057
 8   0.8602982 -0.094736310
 9  -0.4443961  0.135908657
10   2.6532382  0.211146440
11   0.8904910 -0.118826004
12  -0.9979956  0.114564158
13  -4.1295488  0.413668943
14   0.9043454 -0.014119893
15  -0.5349747  0.208177738
16  -0.1877570 -0.068853738
17  -0.4859594 -0.178209676
18  -1.0118500  0.009858047
19  -0.7727913  0.050642414
20   1.0691640 -0.029862250
21   0.5168066 -0.167957707
22  -0.6205851 -0.186557012
23  -0.1877570 -0.068853738
24   1.2503210 -0.174400412
25  -0.2909479 -0.218041726
26  -2.2813834 -0.250590510
27   1.2176442  0.083191198


Standardized Population-Average Residuals:
       Min         Q1         Med       Q3      Max
 -3.223105 -0.4937612 0.007316628 0.472151 3.916034

Number of Observations: 108
Number of Clusters: 27
402 > help(plot.lme) # to learn about the various plot options
402 > par(mfrow=c(2,2))
402 > plot(ortho.lme1,option="r") # three residual plots
402 > plot(ortho.lme1,option="c") # plots of the random effects
```



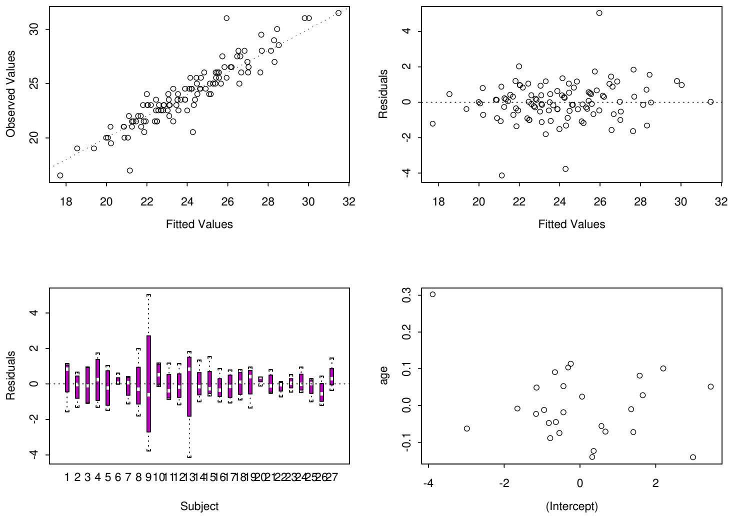
Fitting Model 2: $y_{ij} = (\beta_{00} + \beta_{01} \cdot Sex_i + b_{i0}) + (\beta_{10} + \beta_{11} \cdot Sex_i + b_{i1}) \cdot (age)_i + \epsilon_{ij}$

```
402 > ortho.lme2 _ lme(fixed = distance ~ age*Sex, random= ~ age,
+ cluster = ~ Subject, data = Orthodont)
402 > summary(ortho.lme2)
Call:
  Fixed: distance ~ age * Sex
 Random:  ~ age
Cluster:  ~ Subject
   Data: Orthodont

Estimation Method: RML
Convergence at iteration: 8
Restricted Loglikelihood: -216.2908
Restricted AIC: 448.5817
Restricted BIC: 470.0387

Variance/Covariance Components Estimate(s):
  Structure: unstructured
  Parametrization: matrixlog
  Standard Deviation(s) of Random Effect(s)
```



```
 (Intercept)     age
    2.405606 0.18035
 Correlation of Random Effects
    (Intercept)
age -0.6676482

 Cluster Residual Variance: 1.716195

Fixed Effects Estimate(s):
                   Value Approx. Std.Error  z ratio(C)
(Intercept)  16.3406250        1.01854580  16.0430930
        age   0.7843750        0.08599996   9.1206441
        Sex   1.0321023        1.59575459   0.6467801
    age:Sex  -0.3048295        0.13473604  -2.2624203

 Conditional Correlation(s) of Fixed Effects Estimates
        (Intercept)        age        Sex
    age -0.8801649
    Sex -0.6382847   0.5617958
age:Sex  0.5617958  -0.6382847 -0.8801649

Random Effects (Conditional Modes):
   (Intercept)          age
 1  1.58208521  0.081003095
 2 -1.15239408 -0.023557945
 3 -0.43307081 -0.018681211
 4  2.97680017 -0.140982294
 5 -1.64541912 -0.008467109
 6  1.35491126 -0.010655683
 7 -0.94674804 -0.011922991
 8  0.36710969 -0.123856428
 9 -0.43219508  0.053009997
10  3.45180025  0.050667879
11  0.32580533 -0.140521661
12 -1.15151834  0.048133263
13 -3.88161894  0.302028264
14  0.67601661 -0.070558393
15 -0.30828201  0.103005695
16 -0.78328209 -0.088644478
17 -0.64134523 -0.044752499
18 -0.66024593  0.090297301
19 -0.24895386  0.113567209
20  1.66118980  0.028206017
21  0.57100308 -0.054966600
22 -0.82633909 -0.048054836
23  0.05820159  0.023482827
24  1.41336367 -0.071785377
```

```
25 -0.53896008 -0.074780626
26 -2.98430888 -0.062685039
27  2.19639494  0.101471622


Standardized Population-Average Residuals:
       Min         Q1         Med        Q3      Max
 -3.168028 -0.3859308 0.007122444 0.4451512 3.849473

Number of Observations: 108
Number of Clusters: 27
402 > par(mfrow=c(2,2))
402 > plot(ortho.lme2,option="r")
402 > plot(ortho.lme2,option="c")
```



Comparing the Models

```
402 > anova(ortho.lme1,ortho.lme2)
Response: distance
ortho.lme1
 fixed: (Intercept), age
 random: (Intercept), age
 block: list(1:2)
 covariance structure: unstructured
 serial correlation structure: identity
 variance function: identity
ortho.lme2
 fixed: (Intercept), age, Sex, age:Sex
 random: (Intercept), age
 block: list(1:2)
 covariance structure: unstructured
 serial correlation structure: identity
 variance function: identity
           Model Df    AIC    BIC  Loglik    Test Lik.Ratio   P value
ortho.lme1     1  6 454.64 470.73 -221.32
ortho.lme2     2  8 448.58 470.04 -216.29 1 vs. 2    10.055 0.0065551
```