Mixed Models with R:
Longitudinal Data Analysis

The Special Roles of Time

ICPSR 2017 at York University

Georges Monette

random@yorku.ca

Contents

Summary:

At first sight a mixed model for longitudinal data analysis does not look very different from a mixed model for hierarchical data. In matrices:

Linear Model	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$	
Mixed Model for Hierarchical Data:	$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\gamma} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\varepsilon}_j$ $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \sigma^2\mathbf{I}_j)$ $\perp \quad \mathbf{u}_j \sim N(\mathbf{0}, \mathbf{G})$	$\mathbf{y}_j = \begin{bmatrix} y_{j1} \\ y_{j2} \\ \vdots \\ y_{jn_j} \end{bmatrix}$ Observations in **j**th cluster (students in **j**th school)
Mixed Model for Longitudinal Data:	$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\gamma} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\varepsilon}_j$ $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \mathbf{R}_j)$ $\perp \quad \mathbf{u}_j \sim N(\mathbf{0}, \mathbf{G})$	$\mathbf{y}_j = \begin{bmatrix} y_{j1} \\ y_{j2} \\ \vdots \\ y_{jn_j} \end{bmatrix}$ Observations over time on **j**th subject

Formally, mixed models for hierarchical data and for longitudinal data look almost the same. In practice, longitudinal data introduces some fascinating challenges:

1) The observations within a cluster are **not necessarily independent**. This is the reason for the broader conditions that $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \mathbf{R}_j)$ (where $\mathbf{R}_j$ is a variance matrix) instead of merely the special case: $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \sigma^2\mathbf{I}_j)$. Observations close in time might depend on each other in ways that are different from those that are far in time. Note that if all observations have equal variance and are equally positively correlated – what is called a Compound Symmetry variance structure – this is entirely accounted for by the random intercept model on the G side. The purpose of the **R** matrix is to potentially capture interdependence that is more complex than compound symmetry.

2) The **mean response** may depend on time in ways that are far more complex than is typical for other types of predictors. Depending on the time

scale of the observations, it may be necessary to use **polynomial models**, **asymptotic models**, **Fourier analysis** (orthogonal trigonometric functions) or **splines** that adapt to different features of the relationship in different periods of times.

3) There can be **many** partially confounded **'clocks'** in the same analysis: period-age-cohort effects, age and time relative to a focal event such as giving birth, injury, arousal from coma, etc.

4) Some **periodic patterns** can be modeled either with **fixed effects** or with **random effects** through the R matrix. Periodic patterns with a **fixed period** (e.g. seasonal periodic variation) are naturally modeled with **fixed effects** (FE) using, for example, trigonometric functions. Periodic patterns with a **randomly varying period** (e.g. sunspot cycles, are more appropriately modeled with the **R matrix** which can be used to model the kind of pattern encountered in time series analysis, **autoregressive** and **moving average** models for residuals
These slides focus on the simple functions of time and the **R** side. Lab 3 introduces more complex forms for functions of time.

Take 1: The basic ideas

A traditional example

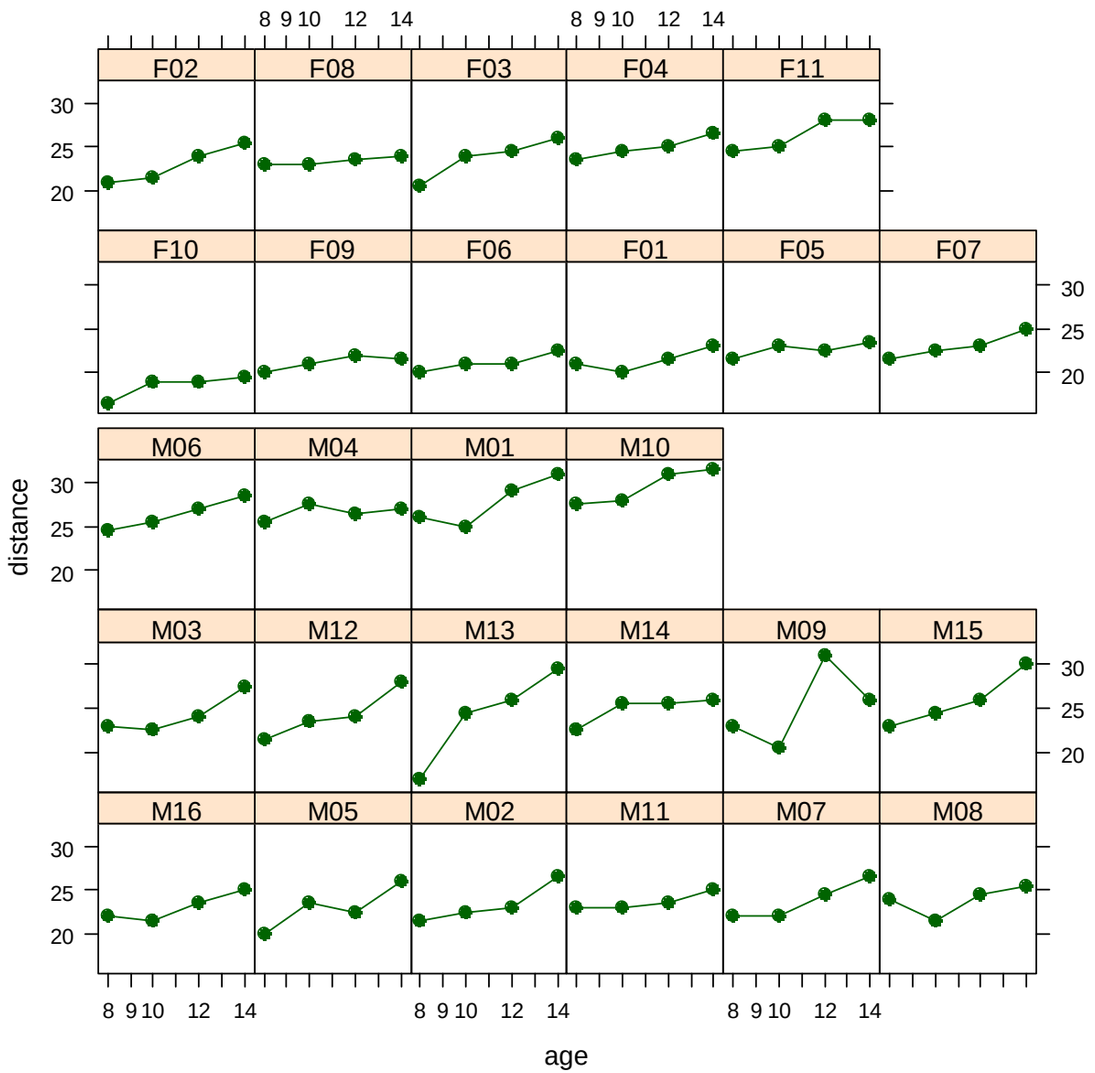


Figure 1: Pothoff and Roy dental measurements in boys and girls.

Balanced data:

- everyone measured at the same set of ages
- could use a classical repeated measures analysis

Some terminology:

Cluster: the set of observations on one subject

Occasion: observations at a given time for each subject

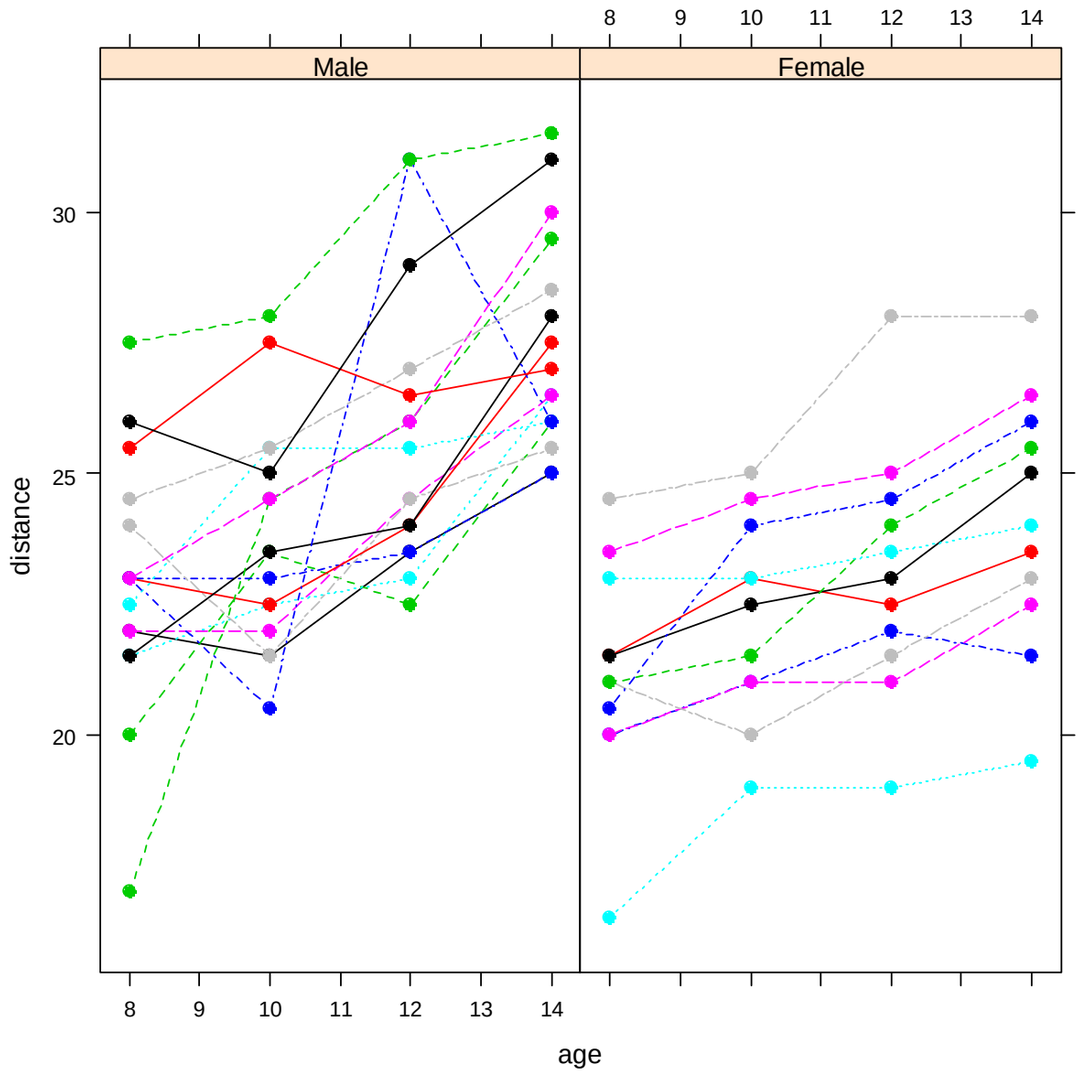


Figure 2: A different view by sex

Viewing by sex helps to see pattern between sexes:

Note:

Slopes are relatively consistent within each sex – except for a few anomalous male curves.

BUT

Intercept is highly variable.

An analysis that pools the data ignores this feature. Slope estimates will have excessively large SEs and 'level' estimates too low SEs.

Pooling the data ('wrong' analysis)

Ordinary least-squares on pooled data:

$$y_{it} = \gamma_0 + \gamma_{age} age_{it} + \gamma_{sex} sex_i + \gamma_{age \times sex} age_{it} sex_i + \varepsilon_{it}$$

$i = 1, \ldots, N$ (number of subjects [clusters])

$t = 1, \ldots, T_i$ (number of occasions for *i*th subject)

Note that it is customary to use *i* for subjects (clusters) and *t* for occasions (time). This means that 'i' assumes the role previously played by 'j'.

R:

```
> library( spida )   # see notes on installation
> library( nlme )    # loaded automatically
> library( lattice ) # ditto
> data ( Orthodont ) # without spida
```

```
> head(Orthodont)
  distance age Subject  Sex
1     26.0   8     M01 Male
2     25.0  10     M01 Male
3     29.0  12     M01 Male
4     31.0  14     M01 Male
5     21.5   8     M02 Male
6     22.5  10     M02 Male
> dd <- Orthodont

> tab(dd, ~Sex)
Sex
  Male Female  Total
    64     44    108
```

```
> dd$Sub <- reorder( dd$Subject, dd$distance)
# for plotting

## OLS Pooled Model

>      fit <- lm ( distance ~ age * Sex , dd)
>      summary(fit)
   .   .   .
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

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
```

Note that both SexFemale and age:SexFemale have large p-values. Are you tempted to just drop both of them?

Check the joint hypothesis that they are BOTH 0.

```
> wald( fit, "Sex")

     numDF denDF  F.value p.value
Sex      2   104 14.97688 <.00001

Coefficients      Estimate Std.Error  DF   t-value p-value
  SexFemale       1.032102  2.218797 104  0.465163 0.64279
  age:SexFemale -0.304830  0.197666 104 -1.542143 0.12608
```

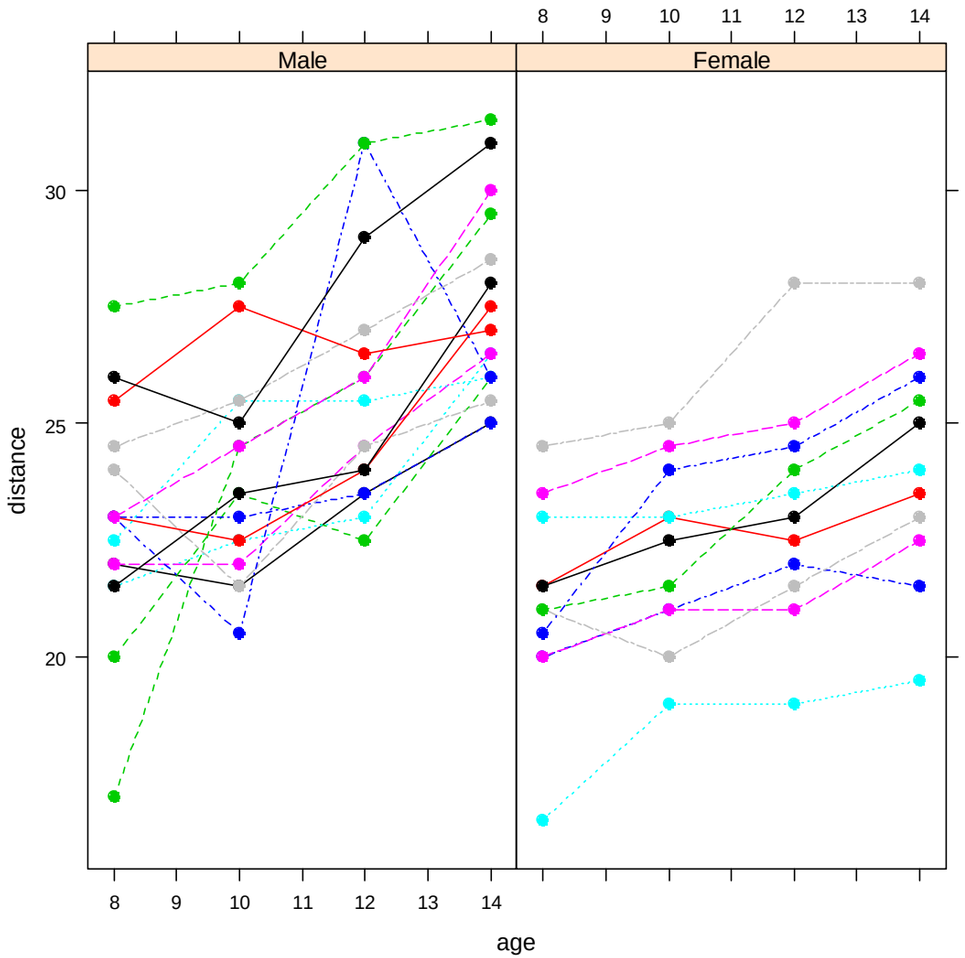
This analysis suggests that we could drop one or the other but not both! Which one should we choose? To respect the principle of marginality

we should drop the interaction, not the main effect of Sex. This leads us to:

```
> fit2 <- lm( distance ~ age + Sex ,dd )
> summary( fit2 )

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 ***
```

and we conclude there is an effect of Sex of jaw size but did not find evidence that the rate of growth is different.

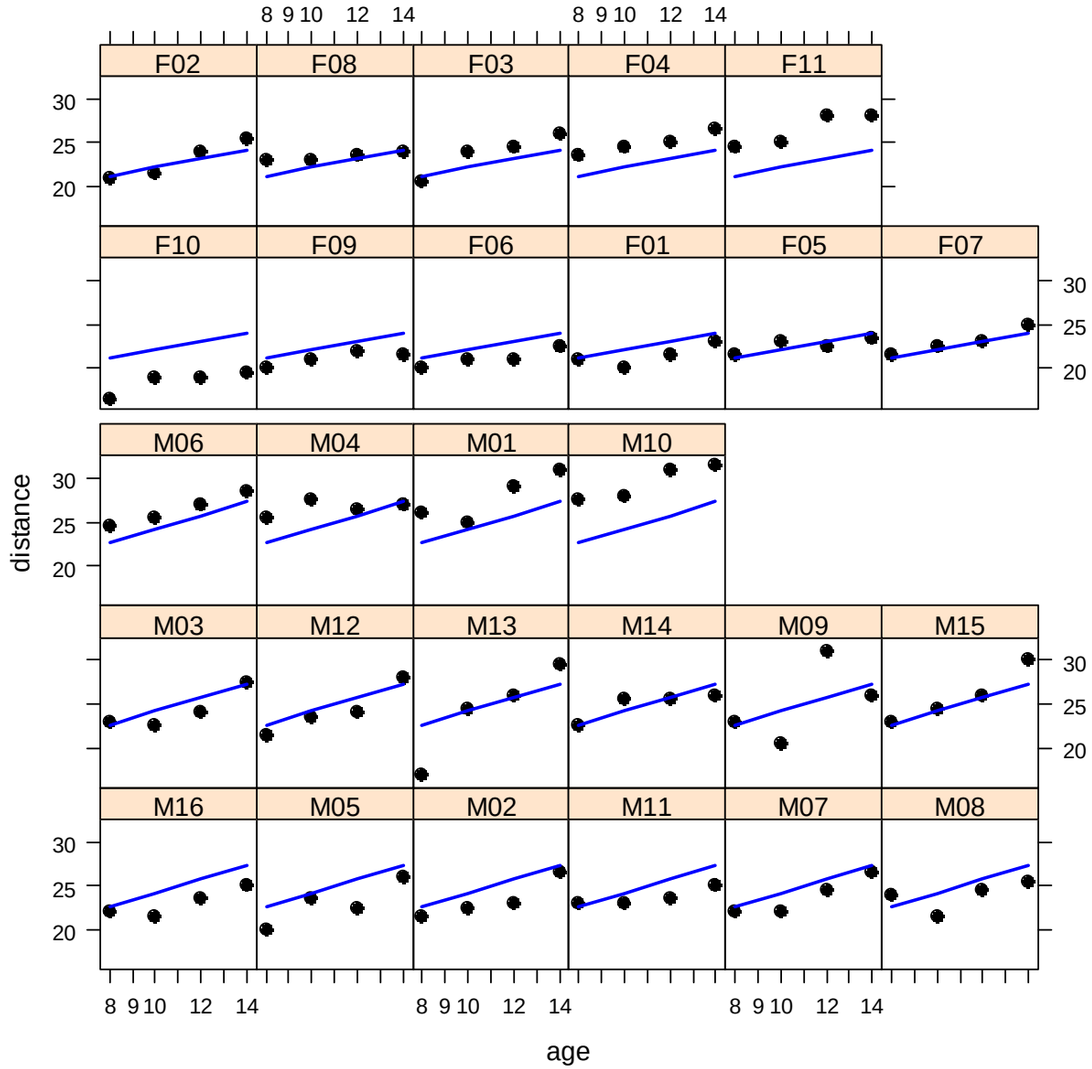
Revisiting the graph we saw earlier:



The analysis is inconsistent with what we see. It fails Tukey's interocular test. Except for a few irregular male trajectories, the male trajectories appear steeper than the female ones. On the other hand, if we extrapolate the curves back to age 0, there would not much difference in levels. estimates too low SEs.

Figure 3: A different view by sex

OLS cannot exploit the consistency in slopes to recognize that hypotheses about slopes should have a relatively smaller SE than hypotheses about the levels of the curves. OLS is blind to the lines.



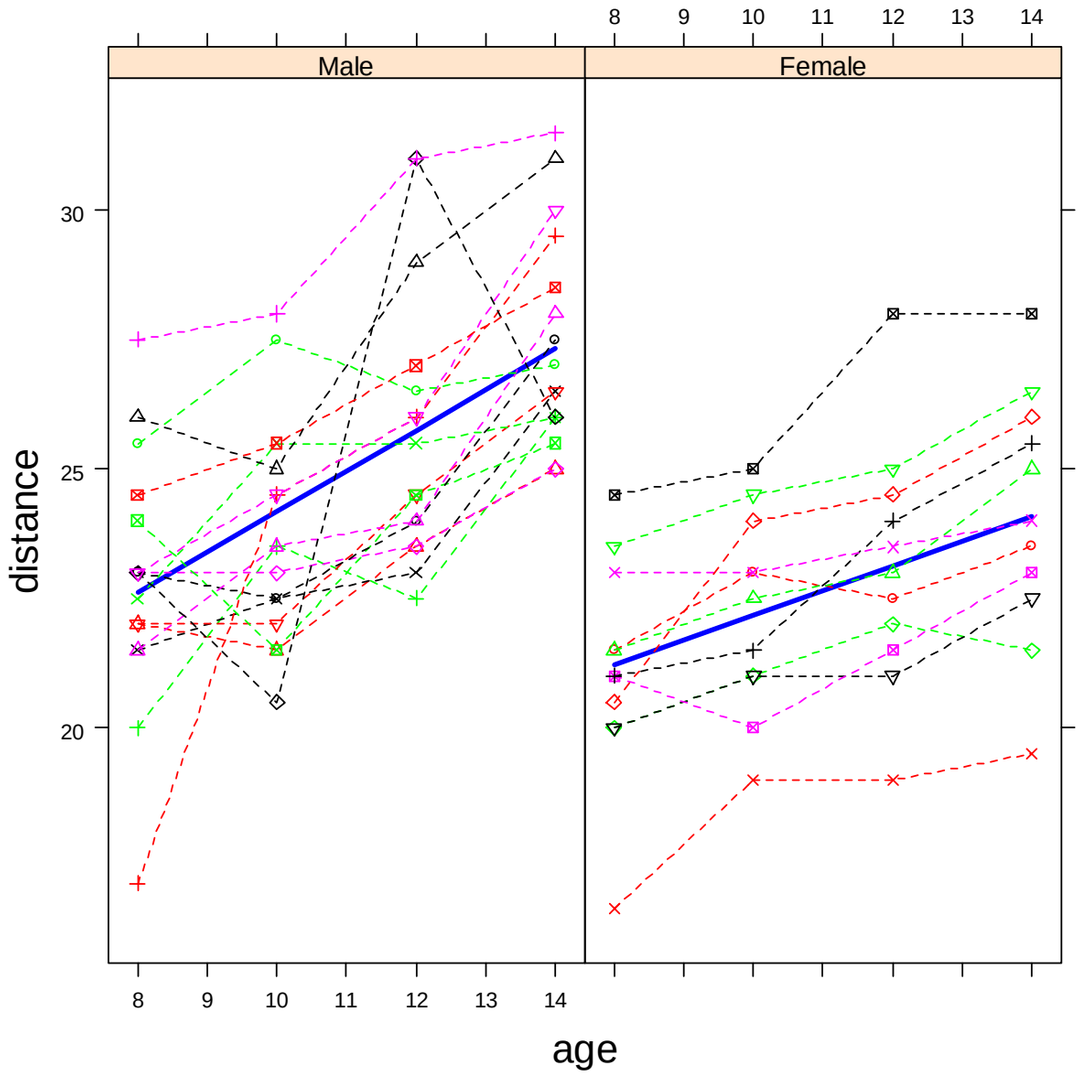
From the first OLS fit:

Estimated variance within each subject:

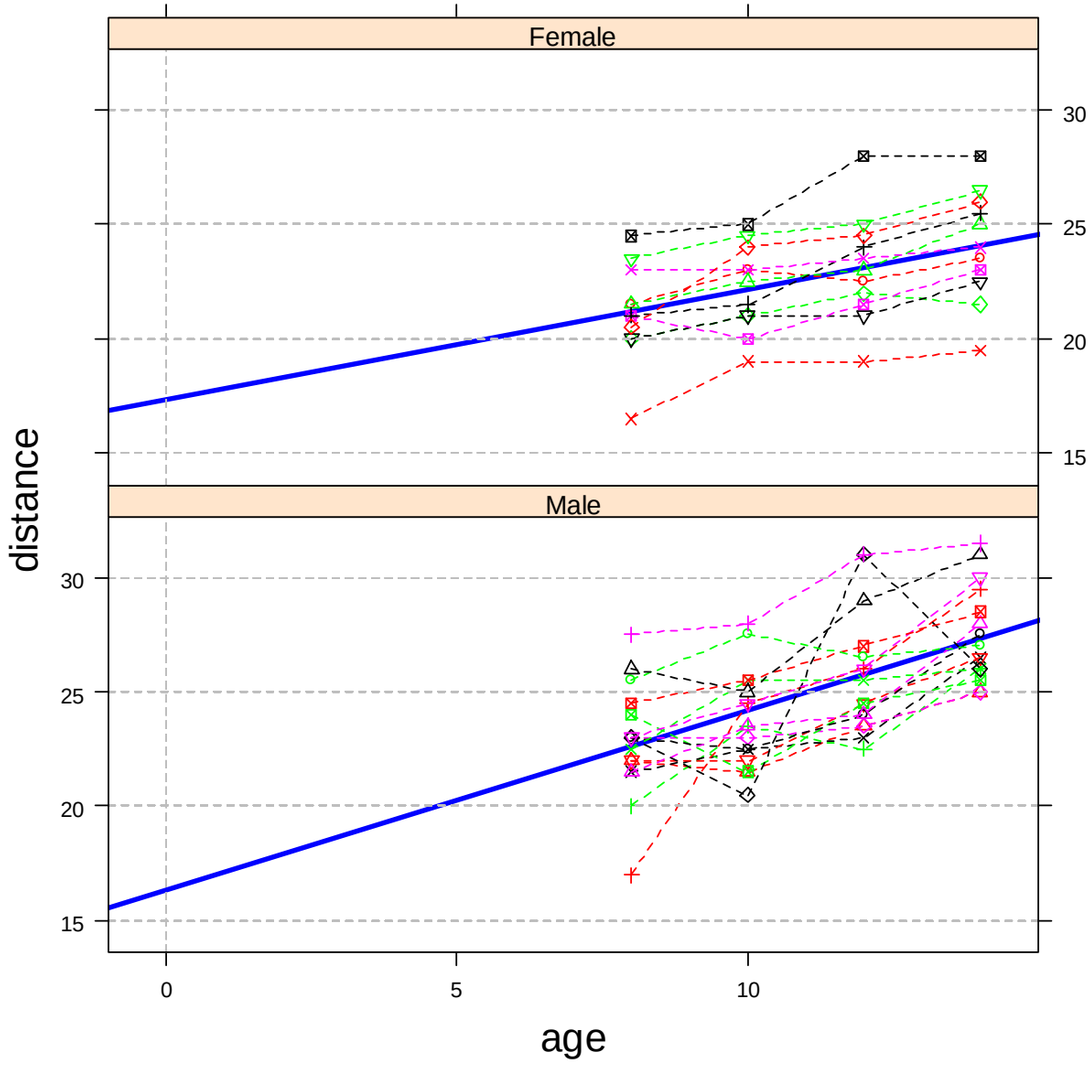
$$\begin{bmatrix} 2.26^2 & . & . & . \\ . & 2.26^2 & . & . \\ . & . & 2.26^2 & . \\ . & . & . & 2.26^2 \end{bmatrix}$$

Why is this wrong?

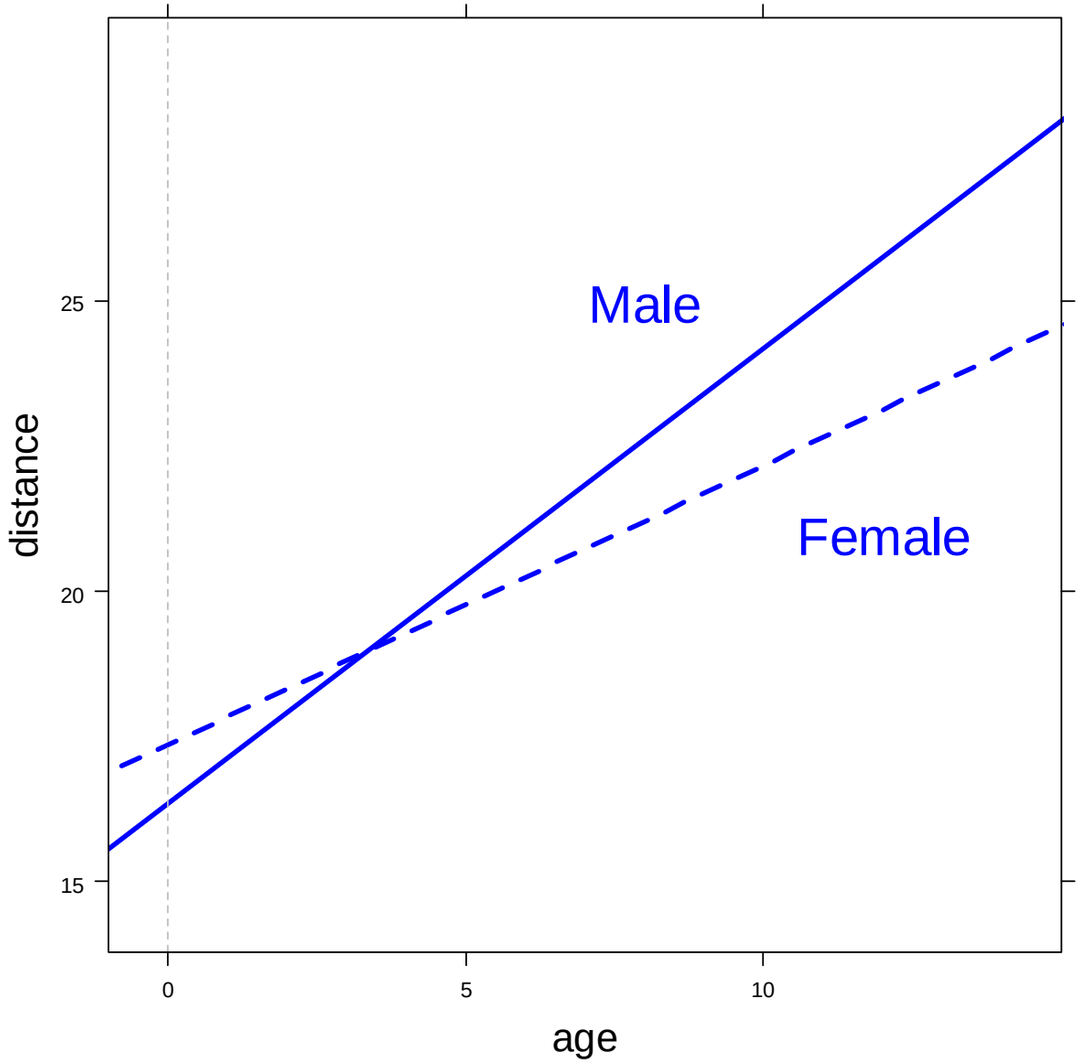
- Residuals within clusters are not independent; they tend to be highly correlated with each other



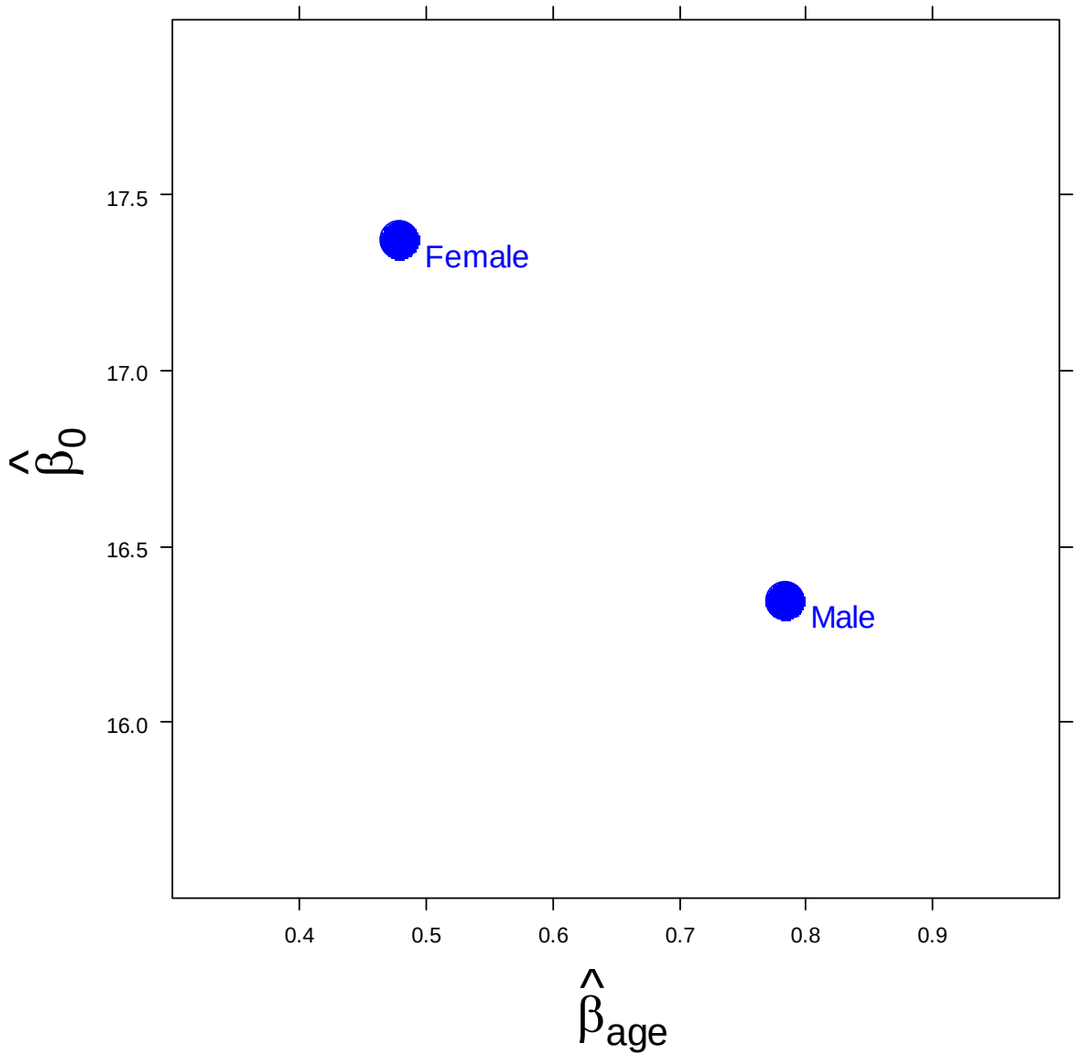
Fitted lines in ‘data space’.



Determining the intercept and slope of each line



Fitted lines in
‘data’ space



Fitted ‘lines’ in ‘beta’ space

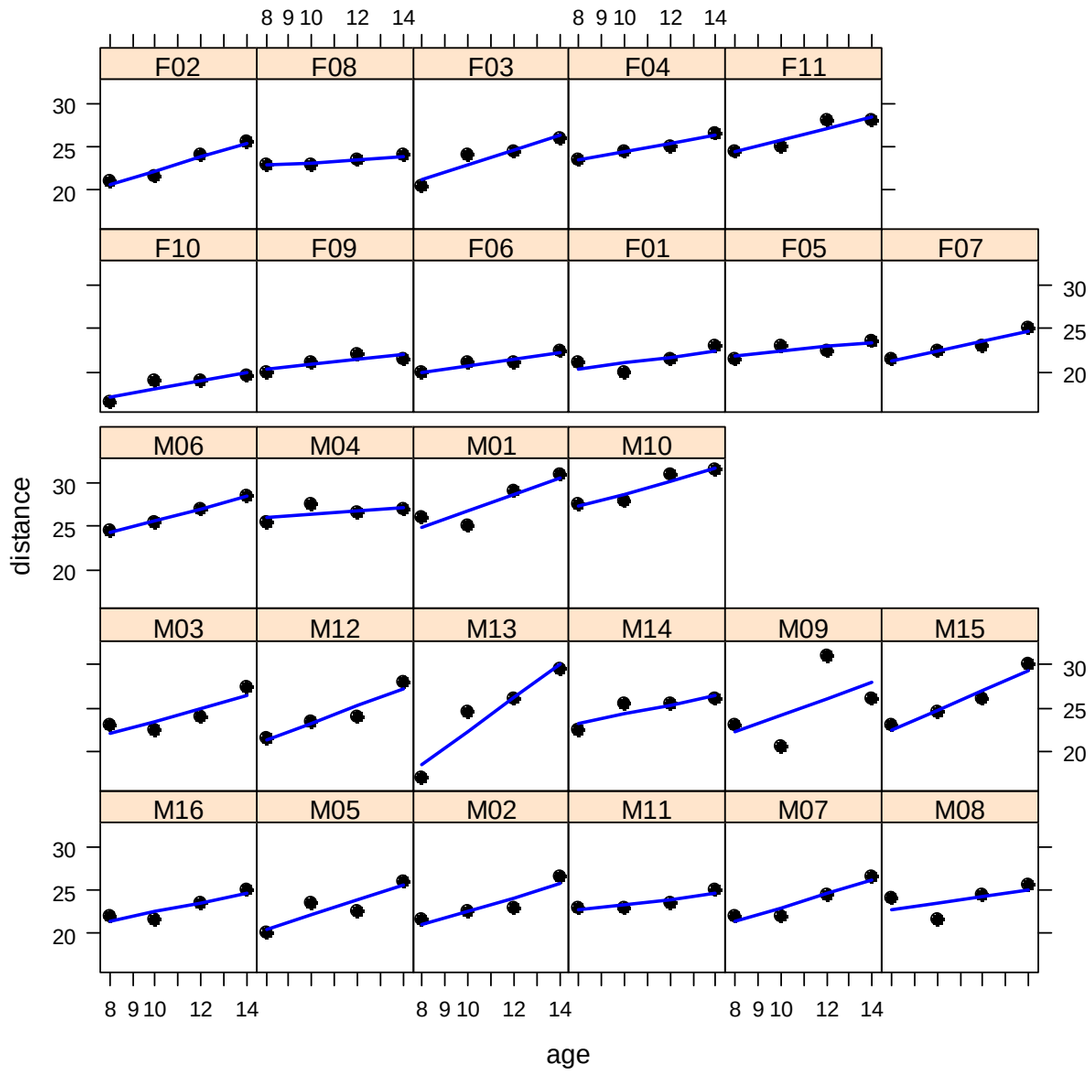
Fixed effects regression model

See
Paul D. Allison (2005) *Fixed Effects Regression Methods for Longitudinal Data Using SAS.* SAS Institute – a great book on basics of **mixed models**!

- Treat Subject as a factor
- Lose Sex unless it is constructed as a Subject contrast
- Fits a separate OLS model to each subject:

$$y_{it} = \beta_{i\,0} + \beta_{i\,age} age_{it} + \varepsilon_{it}$$

See the wiki for details: SPIDA 2010: Longitudinal Models: Additional material

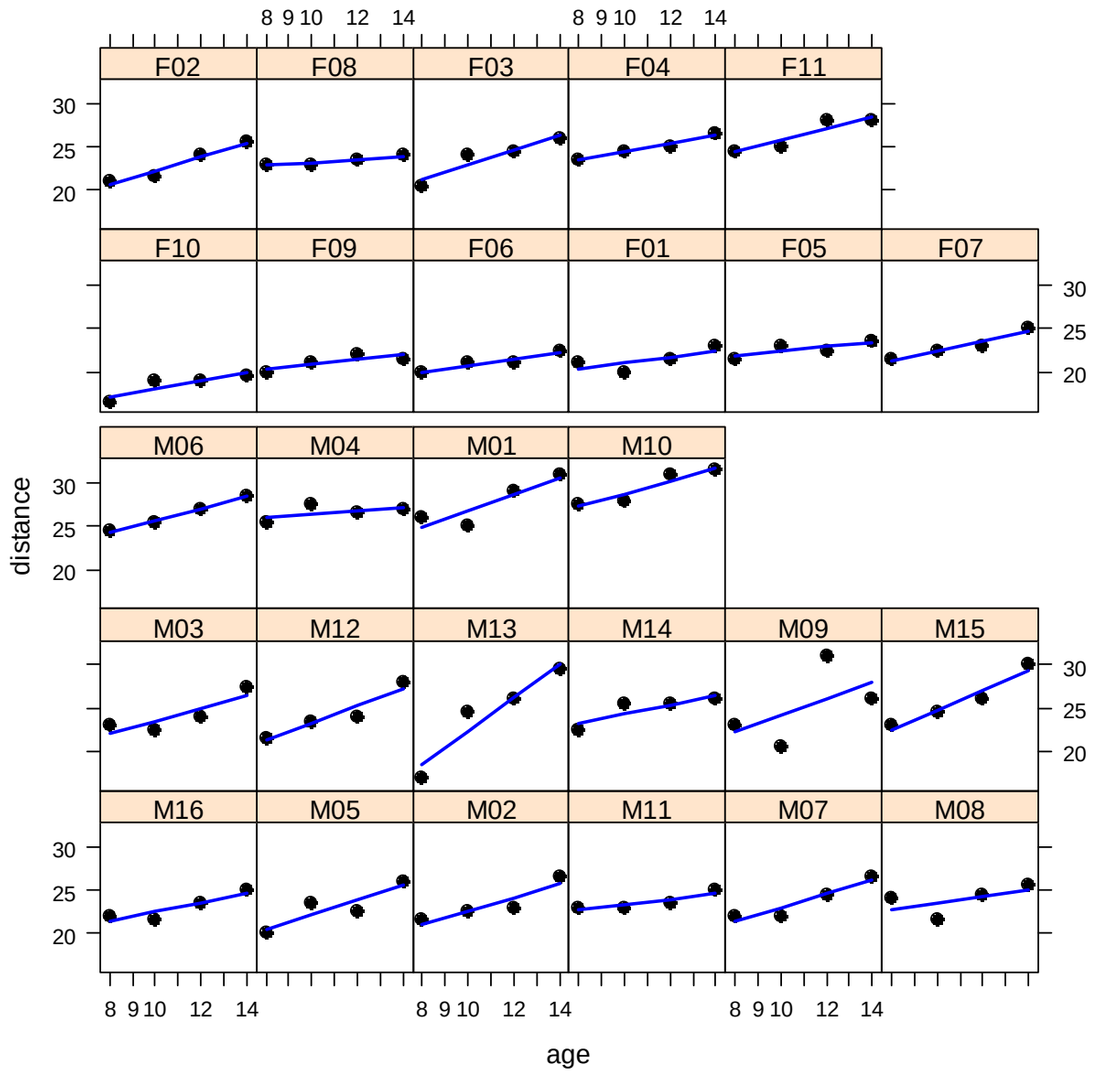


Estimated variance for each subject:

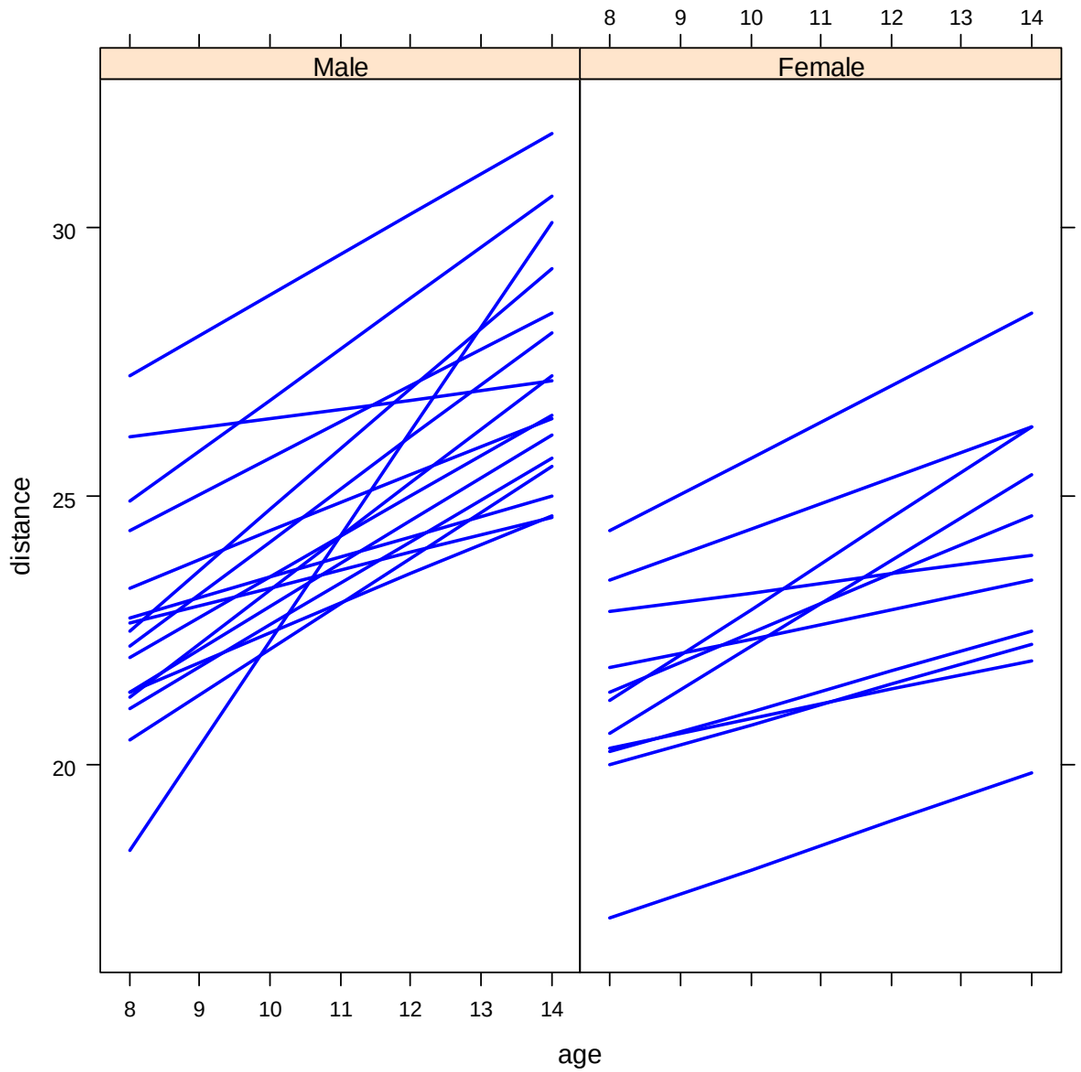
$$\begin{bmatrix} 1.31^2 & . & . & . \\ . & 1.31^2 & . & . \\ . & . & 1.31^2 & . \\ . & . & . & 1.31^2 \end{bmatrix}$$

Problems:

- No estimate of sex effect
- Can't generalize to population, only to 'new' observations from same subjects



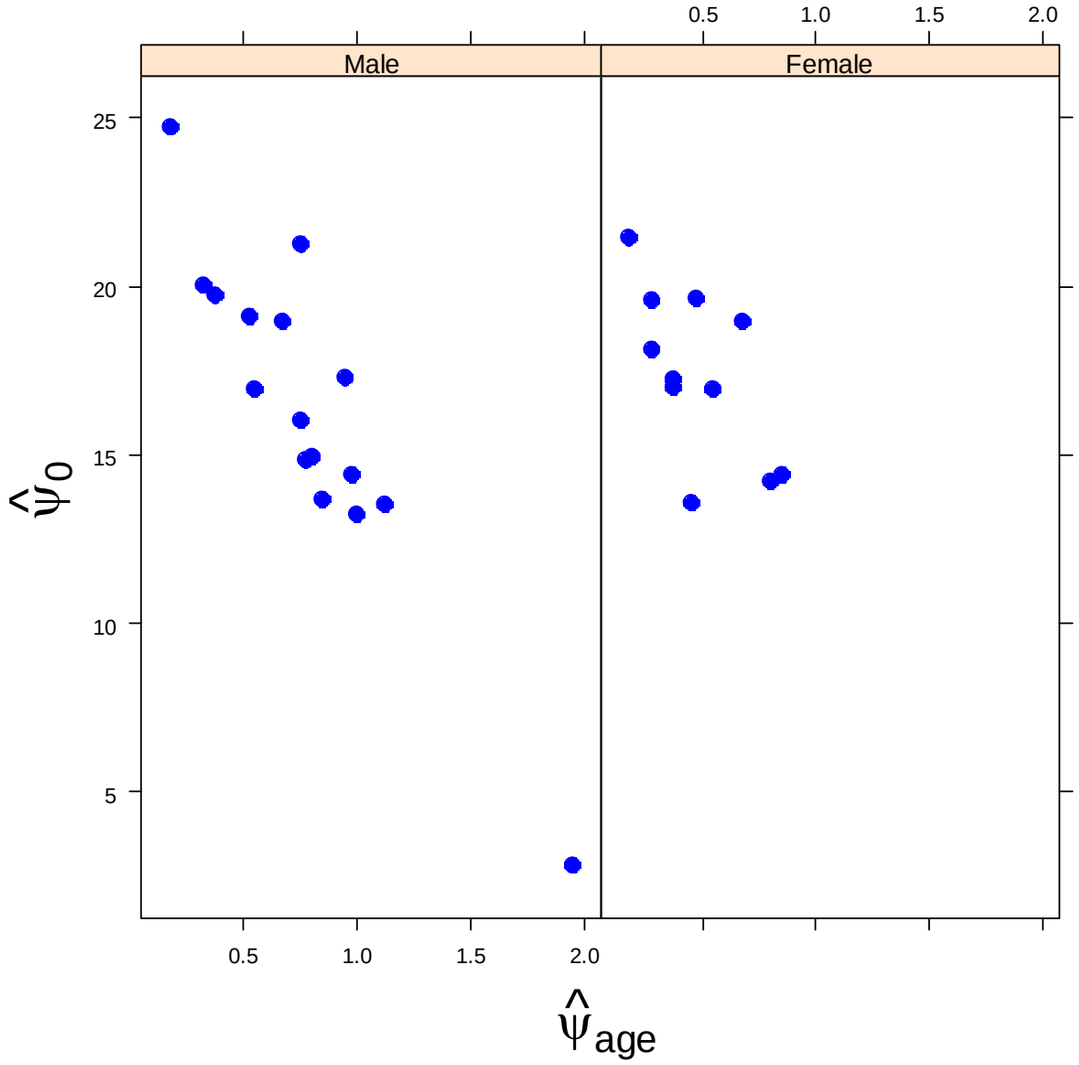
- Can't predict for new subject.
- Can construct sex effect but CI is for difference between sexes in **this** sample
- No autocorrelation in time



Fitted lines in data space

- Female lines lower and less steep

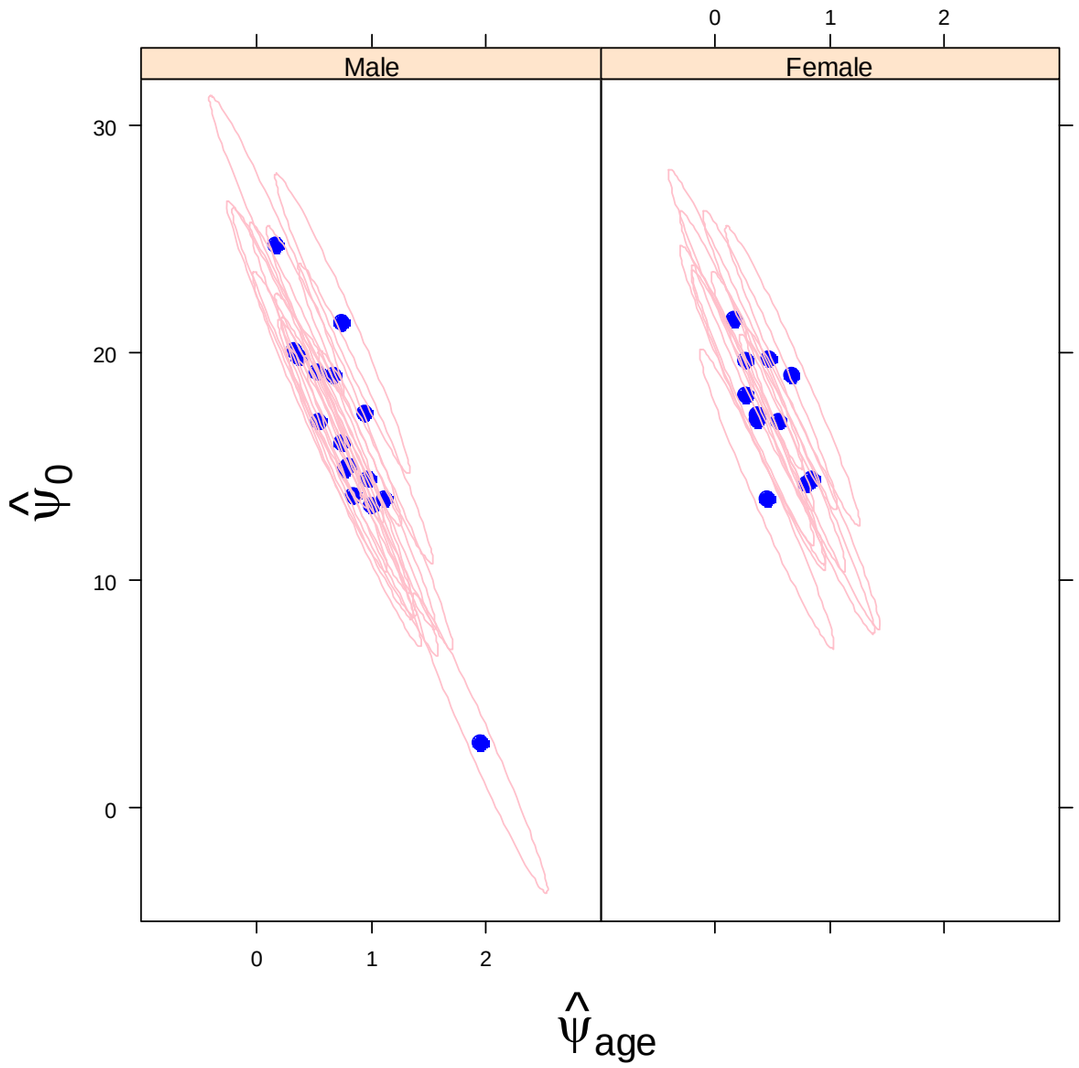
- Patterns within Sexes not so obvious.



Fitted lines in beta space

- Patterns within sexes more obvious: steeper slope associated with smaller intercept.

- Single male outlier stands out



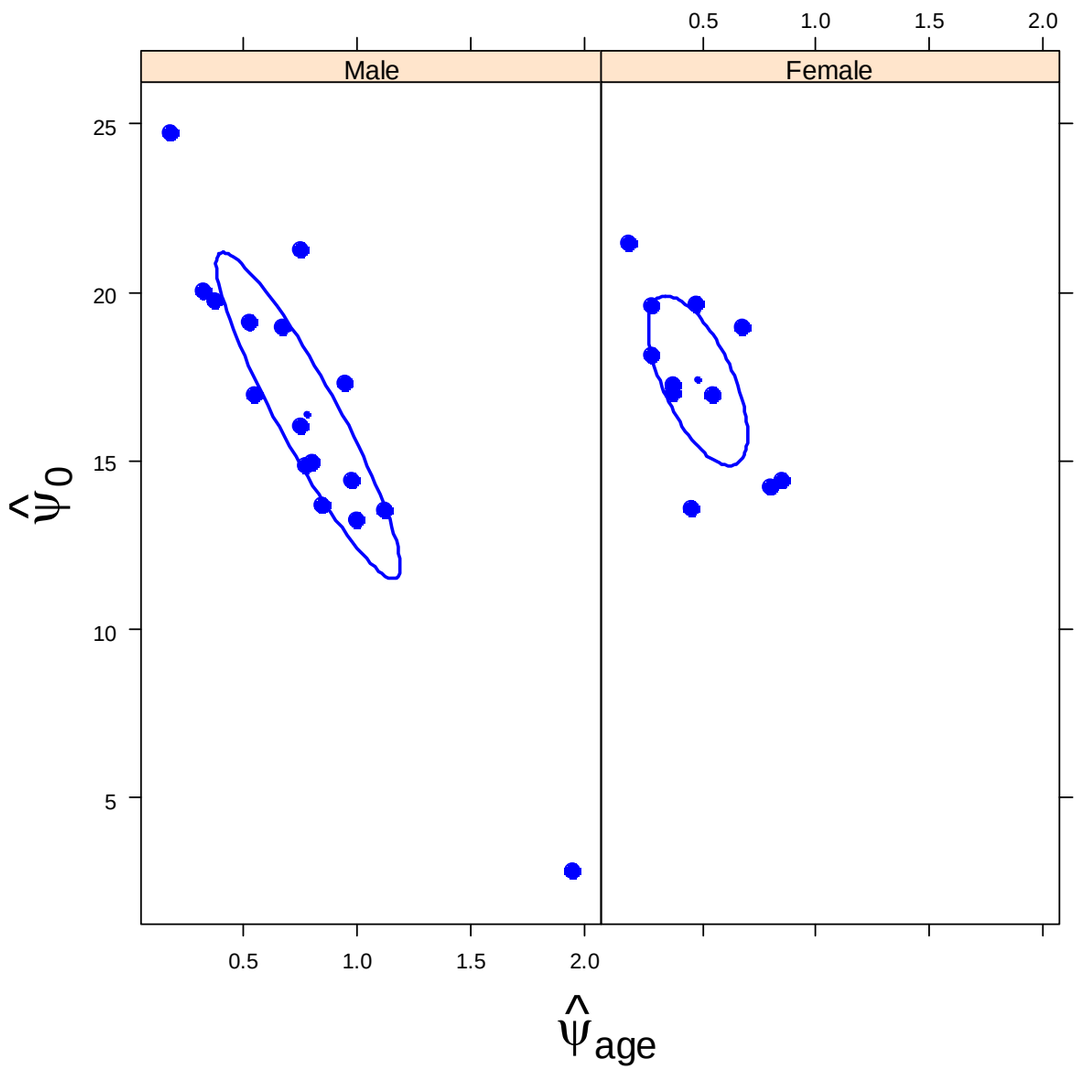
Each within-subject least squares estimate

$$\hat{\beta}_i = \begin{bmatrix} \hat{\beta}_{i0} \\ \hat{\beta}_{i\,age} \end{bmatrix}$$

has variance $\sigma^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$ which is used to construct a confidence ellipse for the 'fixed effect'

$$\beta_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i\,age} \end{bmatrix}$$

for the ith subject.
Each CI uses only the information from that subject (except for the estimate of σ^2)



Differences between subjects such as the dispersion of $\hat{\beta}_i$s and the information they provide on the dispersion of the true β_is is ignored in this model.

The standard error of the estimate of each average Sex line uses the sample distribution of ε_{it}s within subjects but not the variability in β_is between subjects.

Other approaches

- Repeated measures (univariate and multivariate)
  - Need same times for each subject, no other time-varying (Level 1) variables
- Two-stage approach: use $\hat{\beta}_i$s in second level analysis:
  - If design not balanced, then $\hat{\beta}_i$s have different variances, and would need different weights, Using $\sigma^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$ does not work because the relevant weight is based on the marginal variance $\mathbf{G}+\sigma^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$, not the conditional variance given the *i*th subject, $\sigma^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$.

Multilevel Models

Start with the fixed effects model:

Within-subject model (same as fixed effects model above):

$$y_{it} = \beta_{i0} + \beta_{i1} X_{it} + \varepsilon_{it} \qquad \varepsilon_{i\bullet} \sim N(0, \sigma^2 I)$$
$$i = 1, \ldots, N \quad t = 1, \ldots, T_i$$

β_{i0} is the 'true' intercept and β_{i1} is the 'true' slope with respect to X.

σ^2 is the within-subject residual variance.

X (*age* in our example) is a time-varying variable. We could have more than one.

Then add:

Between-subject model (new part):

We suppose that β_{i0} and β_{i1} vary randomly from subject to subject.

But the distribution might be different for different Sexes (a 'between-subject' or 'time-invariant' variable). So we assume a multivariate distribution:

$$\beta_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i1} \end{bmatrix} = \begin{bmatrix} \gamma_{00} + \gamma_{01} W_i \\ \gamma_{10} + \gamma_{11} W_i \end{bmatrix} + \begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix} \quad i = 1, \ldots, N$$

$$= \begin{bmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{bmatrix} \begin{bmatrix} 1 \\ W_i \end{bmatrix} + \begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix}$$

$$\begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{bmatrix} \right) \sim N(\mathbf{0}, \mathbf{G}) \quad i = 1, \ldots, N$$

where W_i is a coding variable for Sex, e.g. 0 for Males and 1 for Females.

$$E\left(\begin{bmatrix}\beta_{i0}\\ \beta_{i\,age}\end{bmatrix}\right)=\begin{bmatrix}\gamma_{00}\\ \gamma_{10}\end{bmatrix} \quad \text{among Males}$$

$$=\begin{bmatrix}\gamma_{00}+\gamma_{01}\\ \gamma_{10}+\gamma_{11}\end{bmatrix} \quad \text{among Females}$$

Some software packages use the formulation of the multilevel model, e.g. MLWin.
SAS and R use the 'mixed model' formulation. It is very useful to know how to go from one formulation to the other.

From Multilevel Model to Mixed Model

Combining the two levels of the multilevel model by substituting the between subject model into the within-subject model. Then gather together the fixed terms and the random terms:

$$
\begin{aligned}
y_{it} &= \beta_{i0} + \beta_{i1} X_{it} + \varepsilon_{it} \\
&= \left(\gamma_{00} + \gamma_{01} W_i + u_{i0}\right) + \left(\gamma_{10} + \gamma_{11} W_i + u_{i1}\right) X_{it} + \varepsilon_{it} \\
&= \left(\gamma_{00} + \gamma_{01} W_{ii}\right) + \left(\gamma_{10} + \gamma_{11} W_{ii}\right) X_{it} + u_{i0} + u_{i1} X_{it} + \varepsilon_{it} \\
&= \gamma_{00} + \gamma_{01} W_i + \gamma_{10} X_{it} + \gamma_{11} W_i X_{it} \text{ (fixed part of the model)} \\
&\quad + u_{i0} + u_{i1} X_{it} + \varepsilon_{it} \text{ (random part of the model)}
\end{aligned}
$$

Anatomy of the fixed part:

γ_{00} (Intercept)

$+\gamma_{01}W_i$ (between-subject, time-invariant variable)

$+\gamma_{10}X_{it}$ (within-subject, time-varying variable)

$+\gamma_{11}W_iX_{it}$ (cross-level interaction)

Interpretation of the fixed part: the parameters reflect population average values.

Anatomy of the random part:

For one occasion:

$$\delta_{it} = u_{i0} + u_{i1}X_{it} + \varepsilon_{it}$$

Putting the observations of one subject together:

$$\begin{bmatrix} \delta_{i1} \\ \delta_{i2} \\ \delta_{i3} \\ \delta_{i4} \end{bmatrix} = \begin{bmatrix} 1 & X_{i1} \\ 1 & X_{i2} \\ 1 & X_{i3} \\ 1 & X_{i4} \end{bmatrix} \begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}$$

$$\delta_{i\bullet} = \mathbf{Z}_i u_{i\bullet} + \varepsilon_{i\bullet}$$

Note: the random-effects design uses only time-varying variables

Distribution assumption:

$$u_{i\bullet} \sim N(0, \mathbf{G}) \quad \text{independent of} \quad \varepsilon_{i\bullet} \sim N(0, \mathbf{R}_i)$$

where, so far, $\mathbf{R}_i = \sigma^2 I$

Notes:

- **G** (usually) does not vary with i. It is usually a free positive definite matrix or it may be a structured pos-def matrix. More on **G** later.

- $\mathbf{R}_i$ (usually) does change with i – as it must if T_i is not constant. $\mathbf{R}_i$ is expressed as a function of parameters. The simplest example is $\mathbf{R}_i = \sigma^2 I_{n_i \times n_i}$. Later we will use $\mathbf{R}_i$ to include auto-regressive parameters for longitudinal modeling.

- We can't estimate **G** and **R** directly. We estimate them through:

$$\mathbf{V}_i = \mathrm{Var}(\delta_{i\bullet}) = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i$$

- Some things can be parametrized either on the G-side or on the R-side. If they're done in both, you lose identifiability. Ill-conditioning due "collinearity" between the G- and R-side models is a common problem.

Mixed Model for Longitudinal Data in R:

```
>  fit <- lme( distance ~ age * Sex, dd,
+          random = ~ 1 + age | Subject,
+          correlation
+           = corAR1 ( form = ~ 1 | Subject))
```

- Model formula: `distance ~ age * Sex`

  - specifies the fixed model
  - includes the intercept and marginal main effects by default
  - contains time-varying, time-invariant and cross-level variables together

- Random argument: `~ 1 + age | Subject`

  - Specifies the variables in the random model and the variable defining clusters.

  - The G matrix is the variance covariance matrix for the random effect. Here $\mathbf{G} = \mathrm{Var}\begin{pmatrix} \beta_{0i} \\ \beta_{age,i} \end{pmatrix} = \mathrm{Var}\begin{pmatrix} u_{0i} \\ u_{age,i} \end{pmatrix} = \begin{bmatrix} g_{00} & g_{0,age} \\ g_{age,0} & g_{age,age} \end{bmatrix}$

  - Normally, the random model only contains an intercept and, possibly, time-varying variables

- Correlation argument: Specifies the model for the $\mathbf{R}_i$ matrices

  - Omit to get the default: $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i \times n_i}$

  - Here we illustrate the use of an AR(1) structure producing for example

$$R_i = \sigma^2 \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$$ in a cluster with 4 occasions.

```
#   Mixed Model in R:

>  fit <- lme( distance ~ age * Sex, dd,
+       random = ~ 1 + age | Subject,
+       correlation
+             = corAR1 ( form = ~ 1 | Subject))

>  summary(fit)
Linear mixed-effects model fit by REML
 Data: dd
       AIC      BIC    logLik
  446.8076 470.6072 -214.4038

Random effects:
 Formula: ~1 + age | Subject
 Structure: General positive-definite, Log-
Cholesky parametrization
```

```
                StdDev    Corr
(Intercept) 3.3730482 (Intr)
age         0.2907673 -0.831
```

$\sqrt{g_{00}}$

$\sqrt{g_{11}}$ $\quad r_{01} = g_{01} / \sqrt{g_{00} g_{11}}$

```
Residual    1.0919754

Correlation Structure: AR(1)
 Formula: ~1 | Subject
 Parameter estimate(s):
     Phi
-0.47328
```

correlation between adjoining obervations

```
Fixed effects: distance ~ age * Sex
                  Value Std.Error DF   t-value p-value
(Intercept)   16.152435 0.9984616 79 16.177323  0.0000
age            0.797950 0.0870677 79  9.164702  0.0000
SexFemale      1.264698 1.5642886 25  0.808481  0.4264
age:SexFemale -0.322243 0.1364089 79 -2.362334  0.0206
 Correlation: among gammas
              (Intr) age    SexFml
age           -0.877
SexFemale     -0.638  0.559
age:SexFemale  0.559 -0.638 -0.877

Standardized Within-Group Residuals:
         Min            Q1           Med
Q3           Max
-3.288886631 -0.419431536 -0.001271185
0.456257976  4.203271248

Number of Observations: 108
```

```
Number of Groups: 27
```

Confidence intervals for all parameters:

```
>     intervals( fit )
Approximate 95% confidence intervals

 Fixed effects:
                      lower       est.       upper
(Intercept)   14.1650475 16.1524355 18.13982351
age            0.6246456  0.7979496  0.97125348
SexFemale     -1.9570145  1.2646982  4.48641100
age:SexFemale -0.5937584 -0.3222434 -0.05072829
attr(,"label")
[1] "Fixed effects:"
```

Do NOT use CIs for SDs to test whether they are 0. Use anova + simulate. Cf Lab 1.

```
Random Effects:
  Level: Subject
                          lower       est.      upper
sd((Intercept))       2.2066308  3.3730482  5.1560298
sd(age)               0.1848904  0.2907673  0.4572741
cor((Intercept),age) -0.9377008 -0.8309622 -0.5808998


 Correlation structure:
         lower     est.        upper
Phi -0.7559617 -0.4728 -0.04182947
attr(,"label")
[1] "Correlation structure:"
```

Negative!

```
 Within-group standard error:
    lower      est.     upper
0.9000055 1.0919754 1.3248923

>     VarCorr( fit )        from the G matrix
Subject = pdLogChol(1 + age)
            Variance   StdDev    Corr
(Intercept) 11.3774543 3.3730482 (Intr)
age          0.0845456 0.2907673 -0.831
Residual     1.1924103 1.0919754

To get the G matrix itself
in a form that can be used
in matrix expressions

>      getVarCov( fit )
```

```
Random effects variance covariance matrix
            (Intercept)       age
(Intercept)    11.37700 -0.814980
age            -0.81498  0.084546
  Standard Deviations: 3.373 0.29077
```

Notes on interpreting autocorrelation

The estimated autocorrelation is negative. Although most natural processes would be expected to produce positive autocorrelations, occasional large measurement errors can create the appearance of a negative autocorrelation. Some processes correct so that a unusually large value tends to be compensated by an unusually small one, e.g. residue in chemical process.

Some issues concerning autocorrelation

1. Lack of fit will generally contribute positively to autocorrelation. For example, if trajectories are quadratic but you are fitting a linear trajectory, the residuals will be positively autocorrelated. **Strong positive autocorrelation can be a symptom of lack of fit.** This is an example of poor identification between the FE model and the R model, that is, between the deterministic and the stochastic aspects of the model. See Lab 3 for a similar discussion of seasonal (FE) versus cyclical variation (R-side) periodic patterns.

2. As mentioned above, occasional large measurement errors will contribute negatively to the estimate of autocorrelation.

3. In a well fitted OLS model, the residuals are expected to be negatively correlated, more so if there are few observations per subject.

4. With few observations per subject, the estimate of autocorrelation (R side) can be poorly identified and highly correlated with G-side parameters. [See 'Additional Notes']

Looking at the data we suspect that M09 might be highly influential for autocorrelation. We can refit without M09 to see how the estimate changes.

What happens when we drop M09?

```
> fit.dropM09 <- update( fit,
+          subset = Subject != "M09")
```

```
> summary( fit.dropM09 )
Linear mixed-effects model fit by REML
 Data: dd
  Subset: Subject != "M09"
       AIC      BIC    logLik
  406.3080 429.7545 -194.1540
.  .  .  .  .
Correlation Structure: AR(1)
 Formula: ~1 | Subject
 Parameter estimate(s):
       Phi
-0.1246035    still negative
. . . . . .
> intervals( fit.dropM09 )
Approximate 95% confidence intervals
. . . .
 Correlation structure:
         lower       est.     upper
Phi -0.5885311 -0.1246035 0.4010562
attr(,"label")     but not significantly
[1] "Correlation structure:"
```

Mixed Model in Matrices

In the ith cluster:

$$y_{it} = \gamma_{00} + \gamma_{01} W_i + \gamma_{10} X_{it} + \gamma_{11} W_i X_{it} + u_{i0} + u_{i1} X_{it} + \varepsilon_{it}$$

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \begin{bmatrix} 1 & W_i & X_{i1} & W_i X_{i1} \\ 1 & W_i & X_{i2} & W_i X_{i2} \\ 1 & W_i & X_{i3} & W_i X_{i3} \\ 1 & W_i & X_{i4} & W_i X_{i4} \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{bmatrix} + \begin{bmatrix} 1 & X_{i1} \\ 1 & X_{i2} \\ 1 & X_{i3} \\ 1 & X_{i4} \end{bmatrix} \begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}$$

$$\mathbf{y}_{i\bullet} = \mathbf{X}_i \boldsymbol{\gamma} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

[*Could we fit this model in cluster i?*]

where

$$\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{G}) \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$$

$$\delta_i = \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i \quad \delta_i \sim N(\mathbf{0}, \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i' + \mathbf{R}_i)$$

For the whole sample

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} \mathbf{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_N \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_N \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{bmatrix}$$

Finally making the complex look deceptively simple:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\delta} \end{aligned}$$

with

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{R}_N \end{bmatrix}$$

$$\mathrm{Var}(\mathbf{u}) = \mathbf{G}$$

$$\boldsymbol{\delta} = \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

$$\mathrm{Var}(\boldsymbol{\delta}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

Fitting the mixed model

Use Generalized Least Squares on

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\gamma}, \mathbf{ZGZ}' + \mathbf{R})$$

$$\hat{\boldsymbol{\gamma}}^{GLS} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

We need $\hat{\boldsymbol{\gamma}}^{GLS}$ to get $\hat{\mathbf{V}}$ and vice versa so one algorithm iterates from one to the other until convergence.

There are two main ways of fitting mixed models with normal responses:

1. Maximum Likelihood (**ML**)

   a. Fits all of $\boldsymbol{\gamma}, \mathbf{G}, \mathbf{R}$ at the same time.

   b. Two ML fits can be used in the 'anova' function to test models that differ in their FE models or in their RE models: G and/or R.

   c. ML fits tend to underestimate **V** and Wald tests will tend to err on the liberal side.

2. Restricted (or Residual) Maximum Likelihood

   a. Maximum likelihood is applied to the '*residual space*' with respect to the **X** matrix and only estimates **G** and **R**, hence **V**. The estimated **V** is then used to obtain the GLS estimate for $\boldsymbol{\gamma}$.

b. 'anova' can only be used to compare two models with identical FE models. Thus 'anova' (Likelihood Ratio Tests) can only be applied to hypotheses about REs.

c. The estimate of **V** tends to be better than with ML and Wald tests are expected to be more accurate than with ML.

d. Thus with REML, you sacrifice the ability to perform LRTs for FEs but improve Wald tests for FEs. Also, LRTs for REs are expected to be more accurate.

e. REML is the default for 'lme' and PROC MIXED in SAS.

Comparing GLS and OLS

We used OLS above:

$$\hat{\boldsymbol{\gamma}}^{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

$$\text{instead of } \hat{\boldsymbol{\gamma}}^{GLS} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

How does OLS differ from GLS?

Do they differ only in that GLS produces more accurate standard errors? Or can $\hat{\boldsymbol{\beta}}^{OLS}$ be very different from $\hat{\boldsymbol{\beta}}^{GLS}$?

With balanced data they will be the same. With unbalanced data they can be dramatically different. OLS is an estimate based on the pooled data. GLS provides an estimate that is closer to that of the unpooled

data. Estimation of the FE model and of RE model are highly related in contrast with OLS and GLMs with canonical links where they are orthogonal.

Testing linear hypotheses in R

Hypotheses involving linear combination of the fixed effects coefficients can be tested with a Wald test. The Wald test is based on the normal approximation for maximum likelihood estimators using the estimated variance-covariance matrix.

Using the 'wald' function alone displays the estimated fixed effects coefficients and Wald-type confidence intervals as well as a test that all true coefficients are equal 0 (this is rarely of any interest).

```
> wald (fit)
 numDF denDF F.value p.value
     4    24 952.019 <.00001

Coefficients     Estimate Std.Error DF   t-value p-value
  (Intercept)   16.479081  1.050894 76 15.681019 <.00001
  age            0.769464  0.089141 76  8.631956 <.00001
  SexFemale      0.905327  1.615657 24  0.560346 0.58044
  age:SexFemale -0.290920  0.137047 76 -2.122774 0.03703
```

Continuation:

```
Coefficients    Lower 0.95 Upper 0.95
  (Intercept)    14.386046  18.572117
  age             0.591923   0.947004
  SexFemale      -2.429224   4.239878
  age:SexFemale  -0.563872  -0.017967
```

We can estimate the response level at age 14 for Males and Females by specifying the appropriate linear transformation of the coefficients.

```
>  L <- rbind( "Male at 14" = c( 1, 14, 0, 0),
+                "Female at 14" = c( 1, 14, 1, 14))
> L
              [,1] [,2] [,3] [,4]
Male at 14       1   14    0    0
Female at 14     1   14    1   14

>     wald ( fit, L )
  numDF denDF  F.value p.value
1     2    24 1591.651 <.00001

              Estimate Std.Error DF  t-value p-value Lower 0.95
  Male at 14   27.25157  0.605738 76 44.98908 <.00001   26.04514
  Female at 14 24.08403  0.707349 24 34.04829 <.00001   22.62413
                Upper 0.95
  Male at 14     28.45800
  Female at 14   25.54392
```

To estimate the gap at 14:

```
>   L.gap <- rbind( "Gap at 14" = c( 0, 0, 1, 14))
>     L.gap
           [,1] [,2] [,3] [,4]
Gap at 14    0    0    1   14

>     wald ( fit, L.gap)
  numDF denDF  F.value p.value
1     1    24 11.56902 0.00235

            Estimate Std.Error DF   t-value p-value Lower 0.95
 Gap at 14 -3.167548  0.931268 24 -3.401327 0.00235  -5.089591

            Upper 0.95
 Gap at 14  -1.245504
```

To simultaneously estimate the gap at 14 and at 8 we can do the following. Note that the overall (simultaneous) null hypothesis here is equivalent to the hypothesis that there is no difference between the sexes.

```
>     L.gaps <- rbind( "Gap at 14" = c( 0, 0, 1, 14),
+                      "Gap at 8" = c( 0,0,1, 8))
>     L.gaps
           [,1] [,2] [,3] [,4]
Gap at 14    0    0    1   14
Gap at 8     0    0    1    8

>      wald ( fit, L.gaps)
  numDF denDF F.value p.value
1     2    24 5.83927 0.00858

             Estimate Std.Error DF   t-value p-value Lower 0.95
  Gap at 14 -3.167548  0.931268 24 -3.401327 0.00235  -5.089591
  Gap at 8  -1.422030  0.844256 24 -1.684359 0.10508  -3.164489
```

An equivalent hypothesis that there is no difference between the sexes is the hypothesis that the two coefficients for sex are simultaneously equal to 0. The 'wald' function simplifies this by allowing a string as a second argument that is used to match coefficient names. The test conducted is that all coefficients whose name has been matched are simultaneously 0.

```
>      wald ( fit, "Sex" )
    numDF denDF F.value p.value
Sex     2    24 5.83927 0.00858

Coefficients     Estimate Std.Error DF   t-value p-value
  SexFemale      0.905327  1.615657 24  0.560346 0.58044
  age:SexFemale -0.290920  0.137047 76 -2.122774 0.03703

Coefficients    Lower 0.95 Upper 0.95
  SexFemale      -2.429224   4.239878
  age:SexFemale  -0.563872  -0.017967
```

Note the equivalence of the two F-tests above.

Modeling dependencies in time

The main difference between using mixed models for multilevel modeling as opposed to longitudinal modeling are the assumptions about ε_{it}, plus the more complex functional forms for time effects. For observations observed in time, part of the correlation between εs could be related to their distance in time.

R-side model allows the modeling of temporal and spatial dependence.

Correlation argument	R

Autoregressive of order 1:
`corAR1( form =`
`~ 1 | Subject)`
	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$

Correlation argument	R
Autoregressive Moving Average of order (1,1) `corARMA( form = ~ 1 | Subject, p = 1, q =1)`	$\sigma^2 \begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix}$
AR(1) in continuous time e.g. supposing a subject with times 1,2, 5.5 and 10[2] `corCAR1( form = ~ time | Subject)`	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^{4.5} & \rho^9 \\ \rho & 1 & \rho^{3.5} & \rho^8 \\ \rho^{4.5} & \rho^{3.5} & 1 & \rho^{4.5} \\ \rho^9 & \rho^8 & \rho^{4.5} & 1 \end{bmatrix}$

[2] Note that the times and the number of times – hence the indices – can change from subject to subject but σ^2 and ρ have the same value.

G-side vs. R-side

- A few things can be done with either side. But don't do it with both in the same model. The redundant parameters will not be identifiable. For example, the G-side random intercept model is 'almost' equivalent to the R-side compound symmetry model.

- With OLS the linear parameters are orthogonal to the variance parameter. Collinearity among the linear parameters is determined by the design, $\mathbf{X}$, and does not depend on values of parameters. Computational problems due to collinearity can be addressed by orthogonalizing the $\mathbf{X}$ matrix.

- With mixed models the variance parameters are generally not orthogonal to each other and, with unbalanced data, the linear parameters are not orthogonal to the variance parameters.

- G-side parameters can be highly collinear even if the **X** matrix is orthogonal. Centering the variables of the RE model around the "point of minimal variance" will help but the resulting design matrix may be highly collinear.

- G-side and R-side parameters can be highly collinear. The degree of collinearity may depend on the value of the parameters.

- For example, our model identifies ρ through:

$$\hat{\mathbf{V}} = \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -3 & -1 & 1 & 3 \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

  For values of ρ above 0.5, the Hesssian is very ill-conditioned. The lesson may be that to use AR, ARMA models effectively, you need at least some subjects observed on many occasions.

- R-side only: population average models

- G-side only: hierarchical models with conditionally independent observations in each cluster

- Population average longitudinal models can be done on the R-side with AR, ARMA structures, etc.

- The absence of the G-side may be less crucial with balanced data.

- The G-side is not enough to provide control for unmeasured between subject confounders if the time-varying predictors are unbalanced (more on this soon).

- A G-side random effects model DOES NOT provide the equivalent of temporal correlation.

Simpler Models

The model we've looked at is deliberately complex including examples of the main typical components of a mixed model. We can use mixed models for simpler problems.
Using X as a generic time-varying (within-subject) predictor and W as a generic time-invariant (between-subject) predictor we have the following:

	MODEL	***RANDOM***	***Formula***
One-way ANOVA with random effects	~ 1	~ 1 | Sub	$y_{it} = \gamma_{00} + u_{i0} + \varepsilon_{it}$
Means as outcomes	~ 1 + W (~W)[3]	~ 1 | Sub	$y_{it} = \gamma_{00} + \gamma_{01} W_i + u_{i0} + \varepsilon_{it}$

[3] The model in () is equivalent in R

One-way ANCOVA with random effects	~ 1 + X (~X)	~ 1 | Sub	$y_{it} = \gamma_{00} + \gamma_{10} X_{it} + u_{i0} + \varepsilon_{it}$
Random coefficients model	~ 1 + X (~X)	~ 1+ X|Sub	$y_{it} = \gamma_{00} + \gamma_{10} X_{it} + u_{i0} + u_{i1} X_{it} + \varepsilon_{it}$
Intercepts and slopes as outcomes	~ 1 + X + W + X:W (~ X*W)	~ 1+ X|Sub	$y_{it} = \gamma_{00} + \gamma_{01} W_i + \gamma_{10} X_{it} + \gamma_{11} W_i X_{it} + u_{i0} + u_{i1} X_{it} + \varepsilon_{it}$
Non- random slopes	~ 1 + X + W + X:W (~ X*W)	~ 1 | Sub	$y_{it} = \gamma_{00} + \gamma_{01} W_i + \gamma_{10} X_{it} + \gamma_{11} W_i X_{it} + \gamma_{i0} + \varepsilon_{it}$

BLUPS: Estimating Within-Subject Effects

We've seen how to estimate β, G and R. Now we consider $\beta_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i1} \end{bmatrix}$.

We've already estimated β_i using the fixed-effects model with a OLS regression within each subject. Call this estimator: $\hat{\beta}_i$. How good is it?

$$\text{Var}(\hat{\beta}_i - \beta_i) = \sigma^2 \left(\mathbf{X}_i' \mathbf{X}_i \right)^{-1}$$

Can we do better? We have another 'estimator' of β_i.

Suppose we know γs for the population. We could also **predict**[4] β_i by using the within Sex mean intercepts and slopes, e.g. for Males we could use: $\begin{bmatrix}\gamma_{00}\\ \gamma_{10}\end{bmatrix}$ with error variance:

$$\text{Var}\left(\begin{bmatrix}\beta_{0i}\\ \beta_{1i}\end{bmatrix}-\begin{bmatrix}\gamma_{00}\\ \gamma_{10}\end{bmatrix}\right)=\mathbf{G}$$

We could then combine $\hat{\beta}_i$ and $\begin{bmatrix}\gamma_{00}\\ \gamma_{10}\end{bmatrix}$ by weighting then by inverse variance (= precision). This yields the BLUP (Best Linear Unbiased Predictor):

$$\left\{\mathbf{G}^{-1}+\left[\sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right]^{-1}\right\}^{-1}\left\{\mathbf{G}^{-1}\begin{bmatrix}\gamma_{00}\\ \gamma_{10}\end{bmatrix}+\left[\sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right]^{-1}\hat{\beta}_i\right\}$$

---

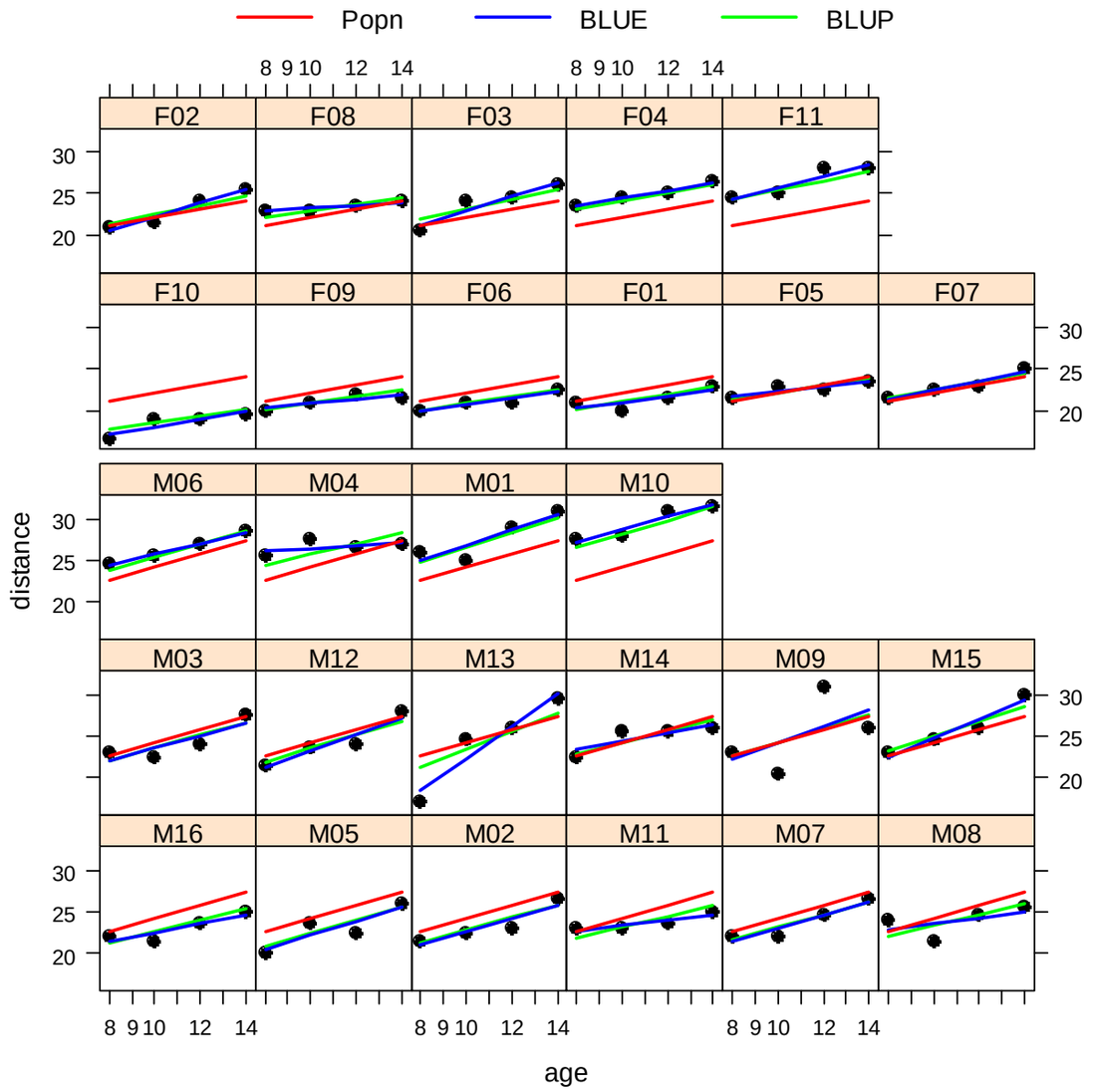
[4] Non-statisticians are always thrown for a loop when we 'predict' something that happened in the past. We use 'predict' for things that are random, 'estimate' for things that are 'fixed'. Orthodox Bayesians always predict.

If we replace the unknown parameters with their estimates, we get the EBLUP (Empirical BLUP):

$$\tilde{\beta}_i = \left\{ \hat{\mathbf{G}}^{-1} + \left[\hat{\sigma}^2 \left(\mathbf{X}_i' \mathbf{X}_i \right)^{-1} \right]^{-1} \right\}^{-1} \left\{ \hat{\mathbf{G}}^{-1} \begin{bmatrix} \hat{\gamma}_{00} \\ \hat{\gamma}_{10} \end{bmatrix} + \left[\hat{\sigma}^2 \left(\mathbf{X}_i' \mathbf{X}_i \right)^{-1} \right]^{-1} \hat{\beta}_i \right\}$$

The EBLUP 'optimally' combines the information from the ith cluster with the information from the other clusters. We **borrow strength** from the other clusters.

The process 'shrinks' $\hat{\beta}_i$ towards $\begin{bmatrix} \hat{\gamma}_{00} \\ \hat{\gamma}_{10} \end{bmatrix}$ along a path determined by the locus of osculation of the families of ellipses with shape $\hat{\mathbf{G}}$ around $\begin{bmatrix} \hat{\gamma}_{00} \\ \hat{\gamma}_{10} \end{bmatrix}$ and shape $\left[\hat{\sigma}^2 \left(\mathbf{X}_i' \mathbf{X}_i \right)^{-1} \right]$ around $\hat{\beta}_i$.

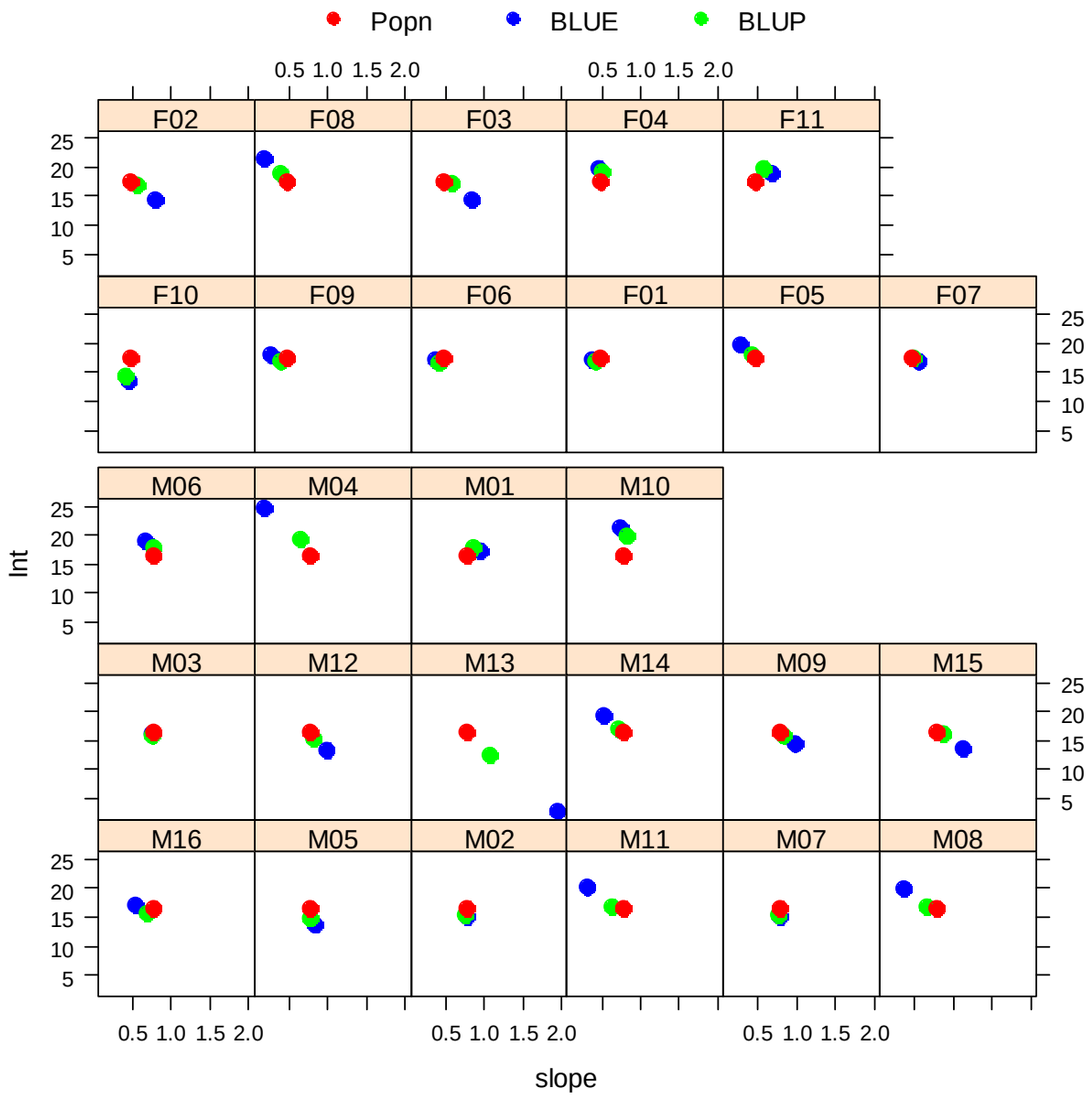


The slope of the BLUP is close to the population slope

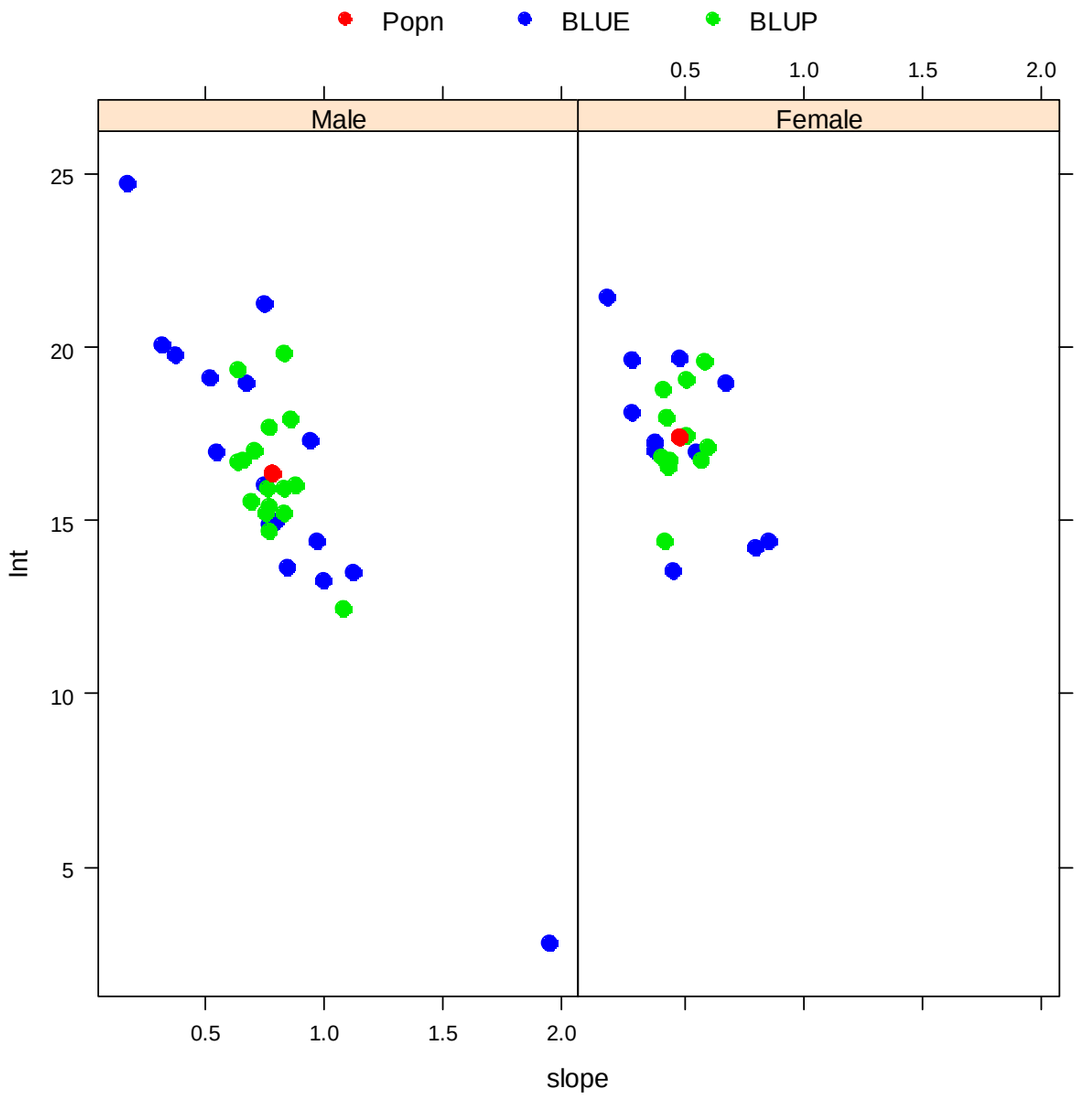
but

the level of the BLUP is close to the level of the BLUE

This suggests that **G** has a large variance for intercepts and a small variance for slopes



Population estimate BLUE and BLUP in beta space

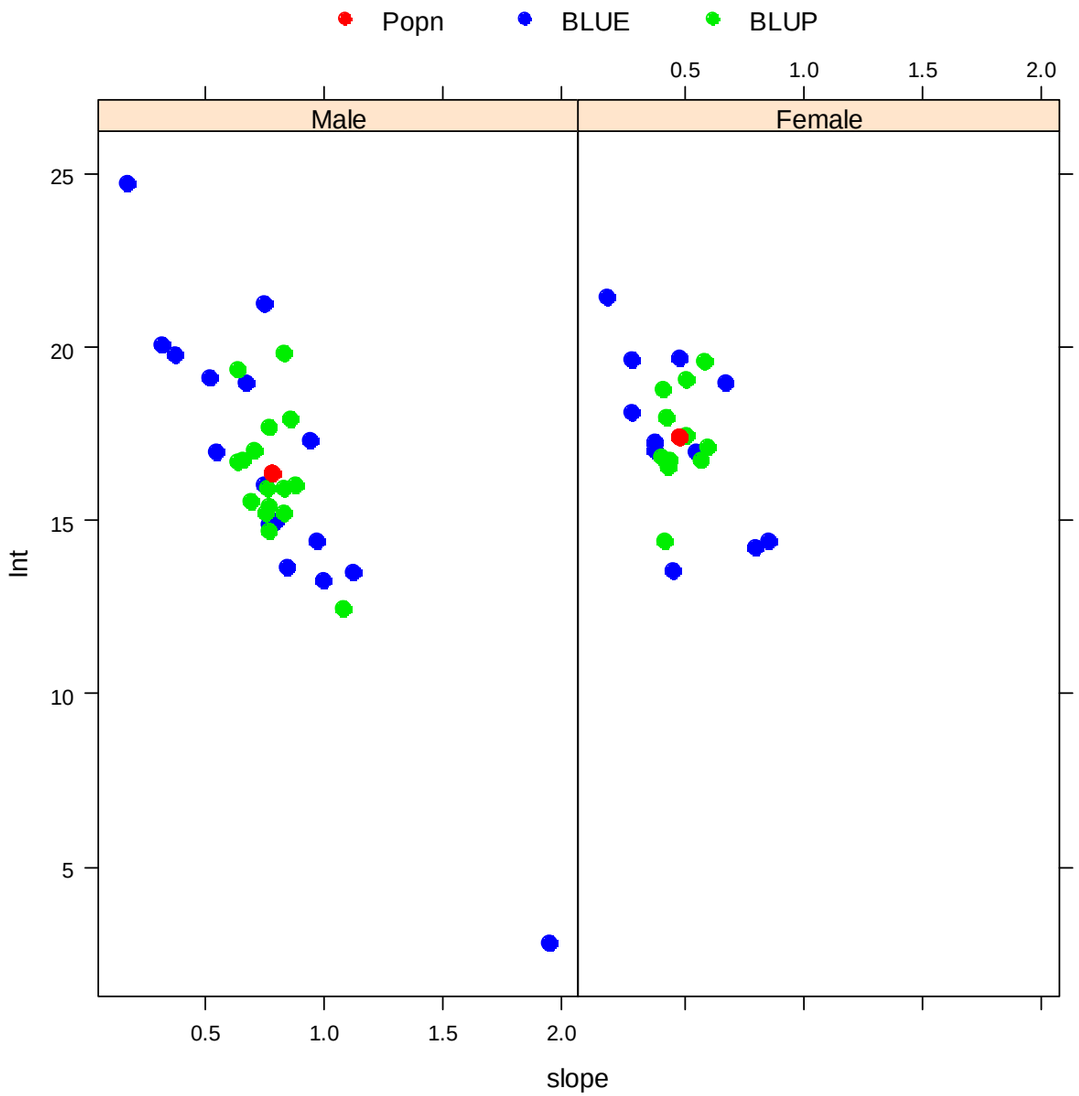


The marginal dispersion of BLUEs comes from:

$$\mathrm{Var}(\hat{\beta}_i) = \mathbf{G} + \sigma^2 (\mathbf{X}_i' \mathbf{X}_i)^{-1}$$

$$\mathrm{Var}(\hat{\beta}_{i1}) \approx g_{11} + \frac{\sigma^2}{T_i \times S_{X_i}^2}$$

- $\mathrm{Var}(\beta_i) = \mathbf{G}$ [population var.]
- $\mathrm{Var}(\hat{\beta}_i \mid \beta_i) = \sigma^2 (\mathbf{X}_i' \mathbf{X}_i)^{-1}$ [cond'l var. resampling from i^{th} subject]
- $\mathrm{E}(\hat{\beta}_i \mid \beta_i) = \beta_i$ [BL**U**E]

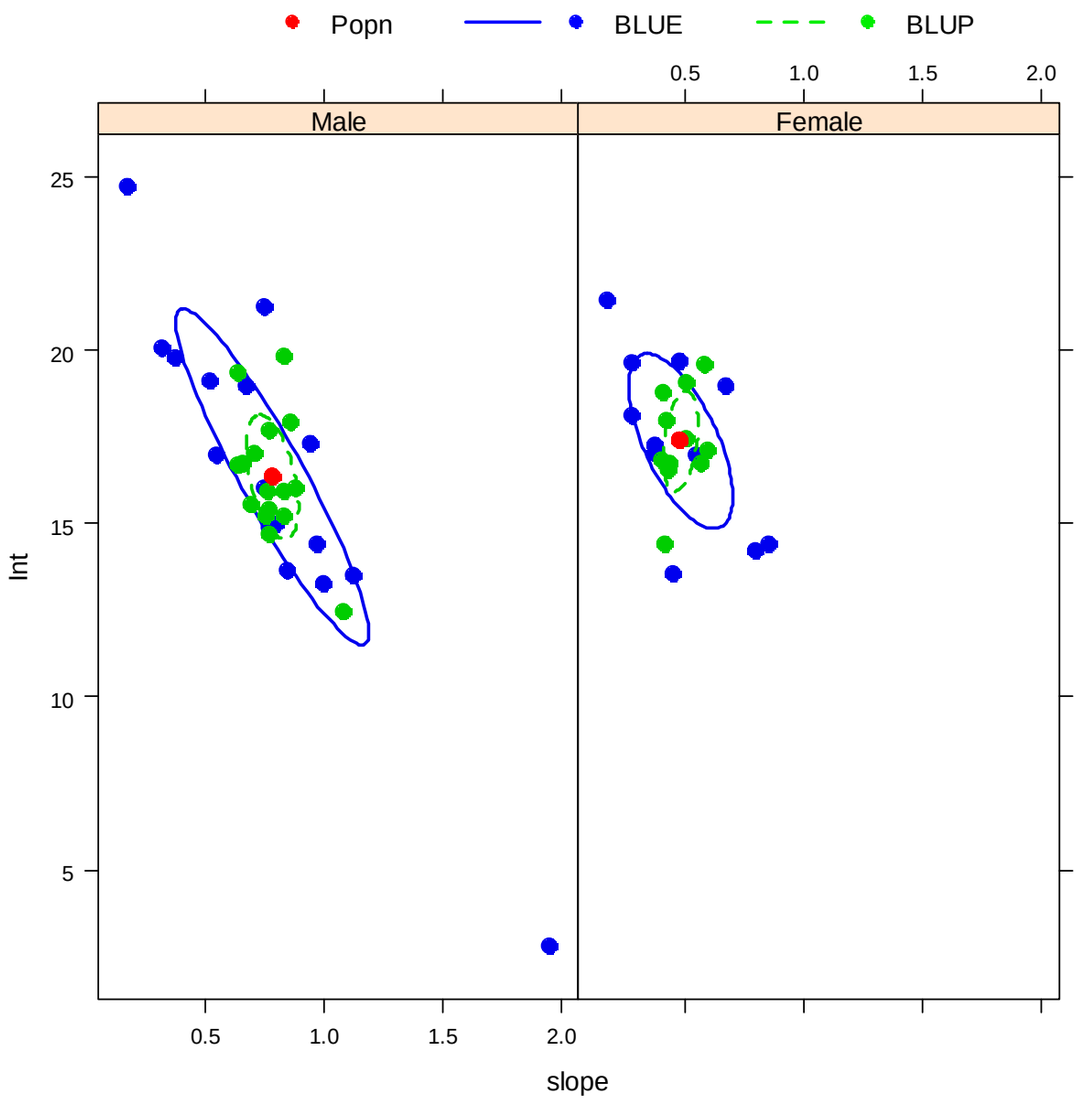


So:

$$\mathrm{Var}(\hat{\beta}_i) = \mathrm{Var}(\mathrm{E}(\hat{\beta}_i \mid \beta_i)) + \mathrm{E}\left\{\mathrm{Var}(\hat{\beta}_i \mid \beta_i)\right\}$$

$$= \mathrm{Var}(\beta_i) + \mathrm{E}\left\{\mathrm{Var}(\hat{\beta}_i \mid \beta_i)\right\}$$

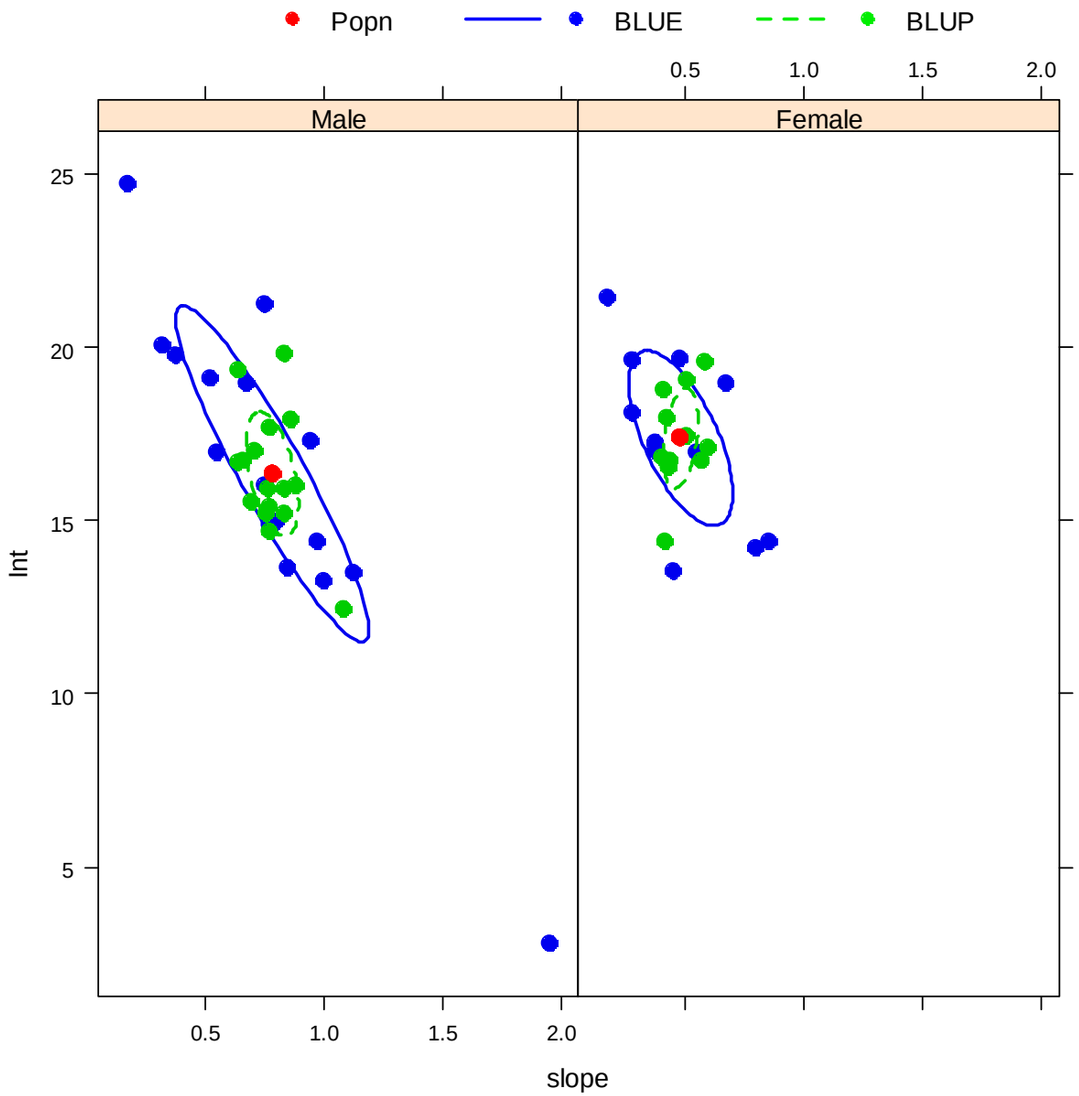
$$= \mathbf{G} + \sigma^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$$



While the expected variance of the BLUEs is larger than **G**

the expected variance of the BLUPs is smaller than **G.**

Beware of drawing conclusions about **G** from the dispersion of the BLUPs.

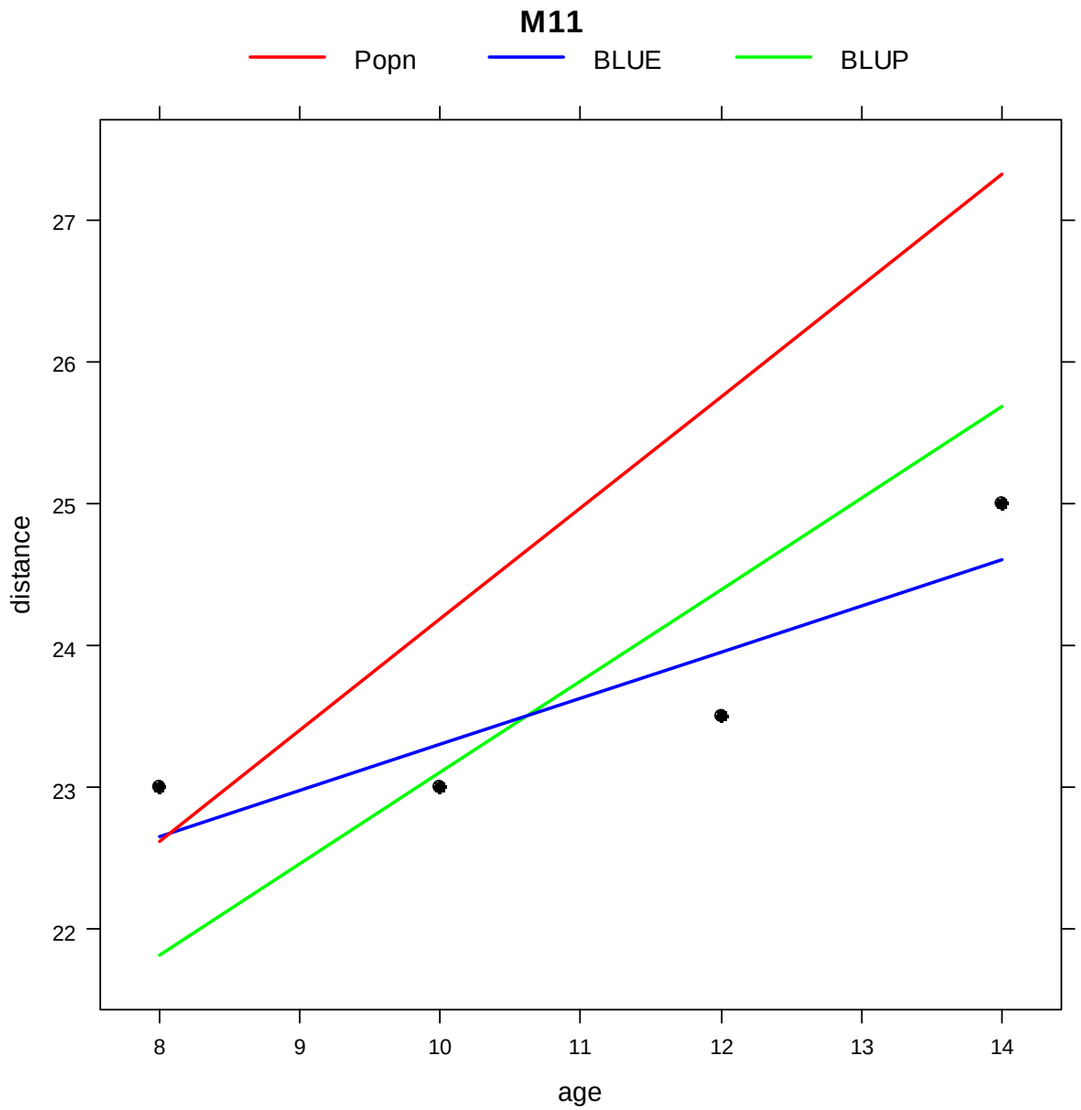


The estimate of **G** can be unstable and often collapses to singularity leading to non-convergence for many methods.

Possible remedies:

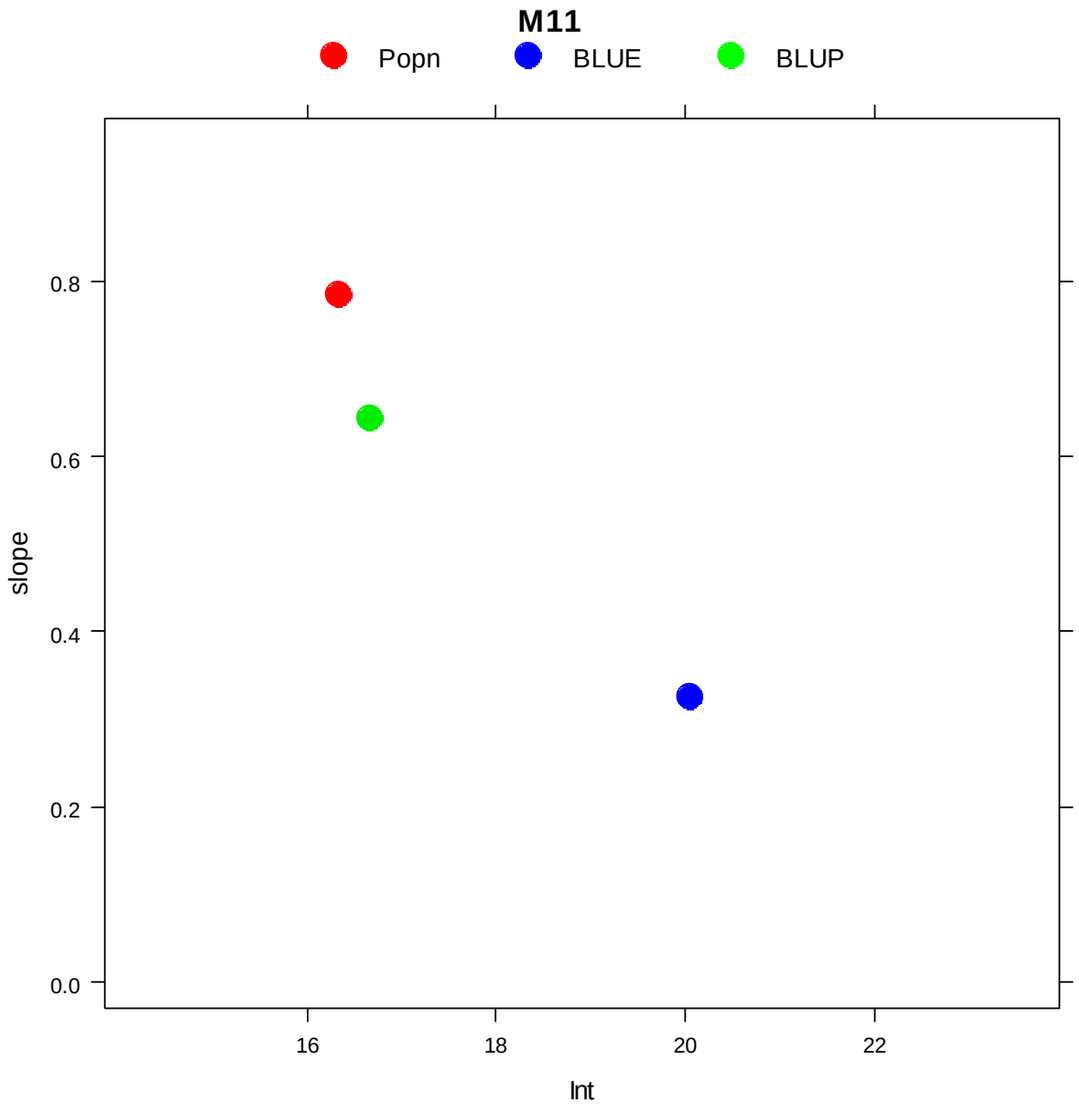
- Recentre X near point of minimal variance,
- Use a smaller **G**
- Change the model

Where the EBLUP comes from : looking at a single subject

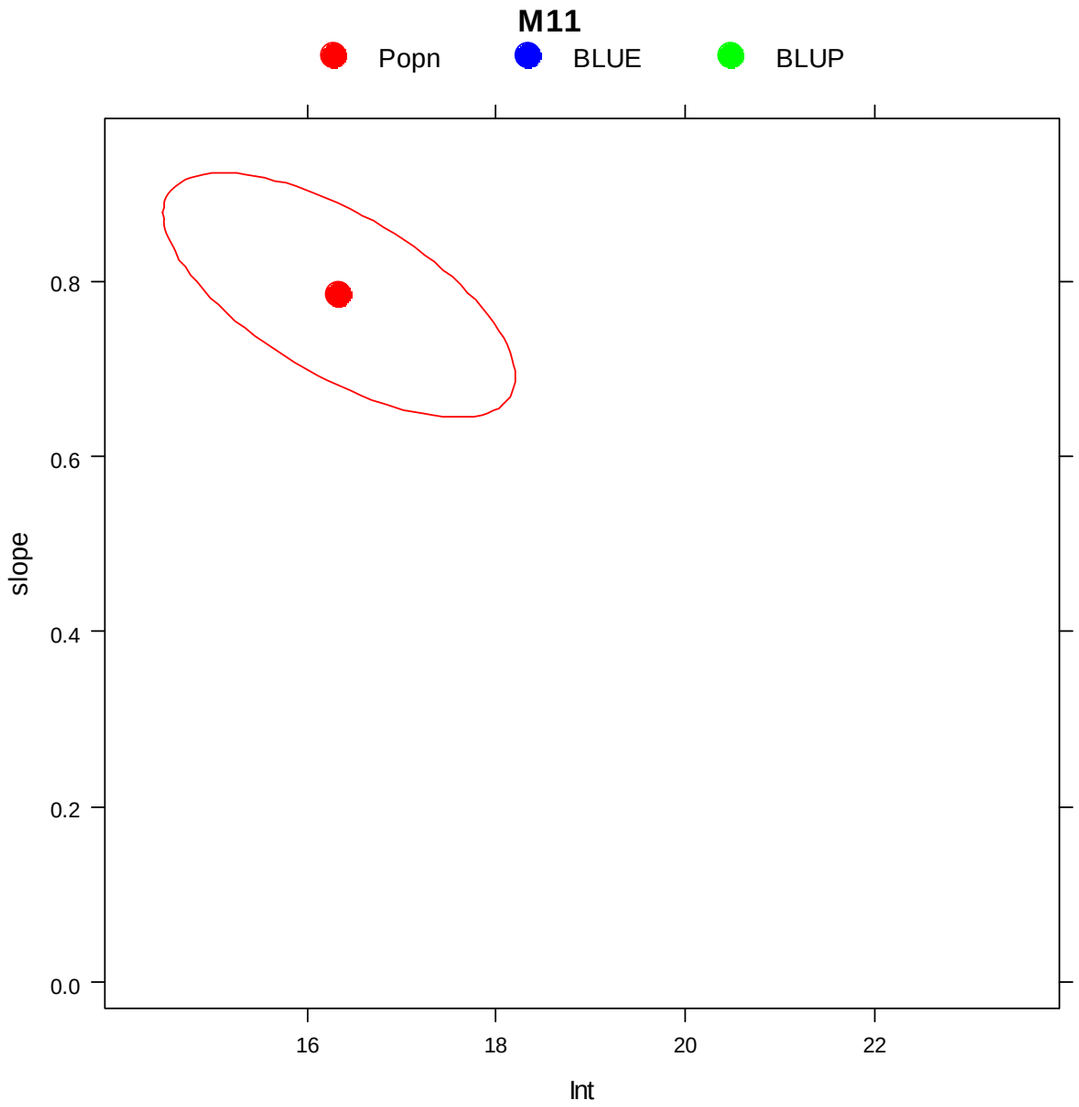


Note that the EBLUP's slope is close to the slope of the population estimate (i.e. the male population conditioning on between-subject predictors) while the level of the line is close to level of the BLUE.

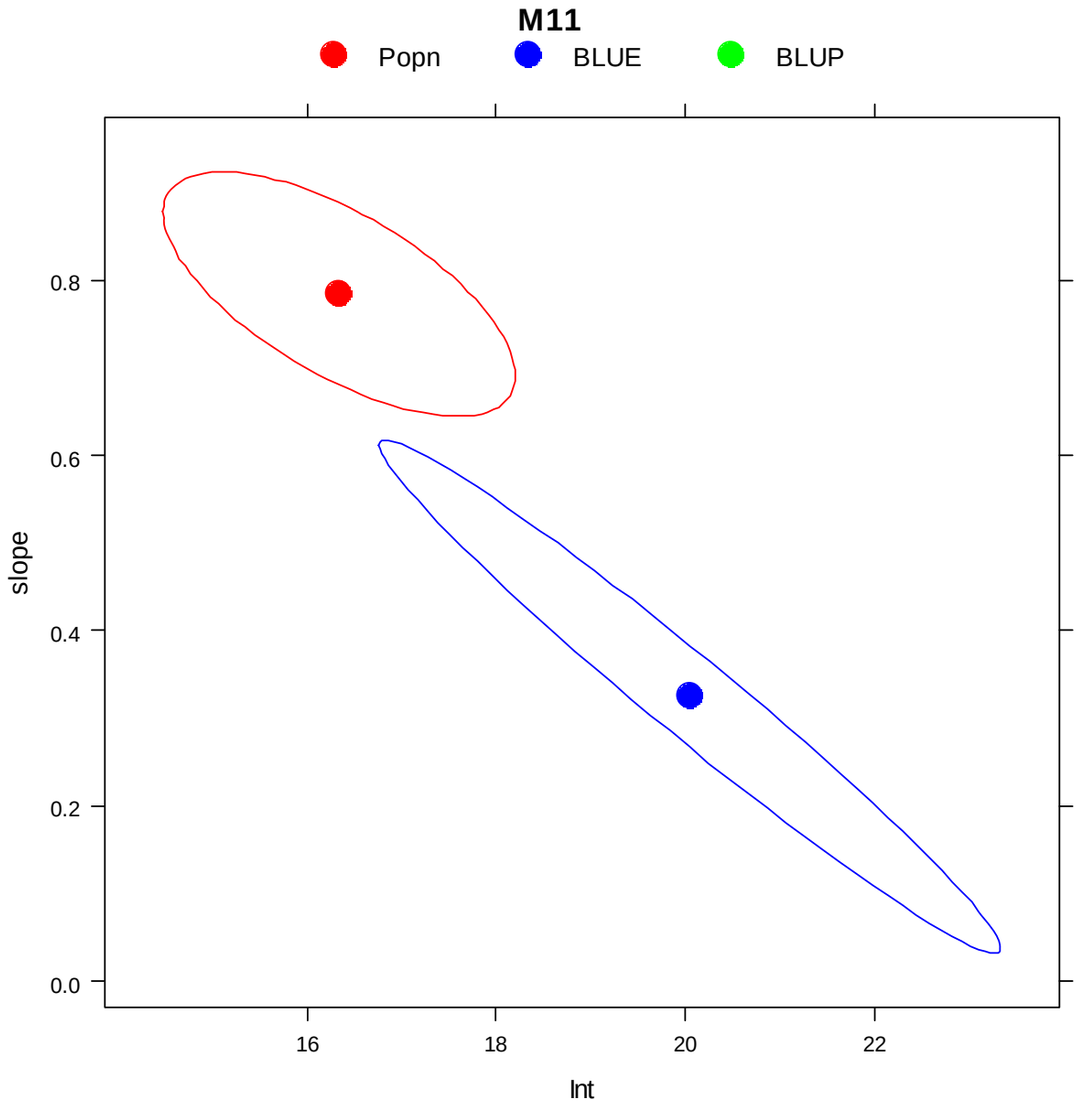
The relative precisions of the BLUE and of the population estimate on slope and level are reflected through the shapes of $\mathbf{G}$ and $\sigma^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$



The same picture in "beta-space"



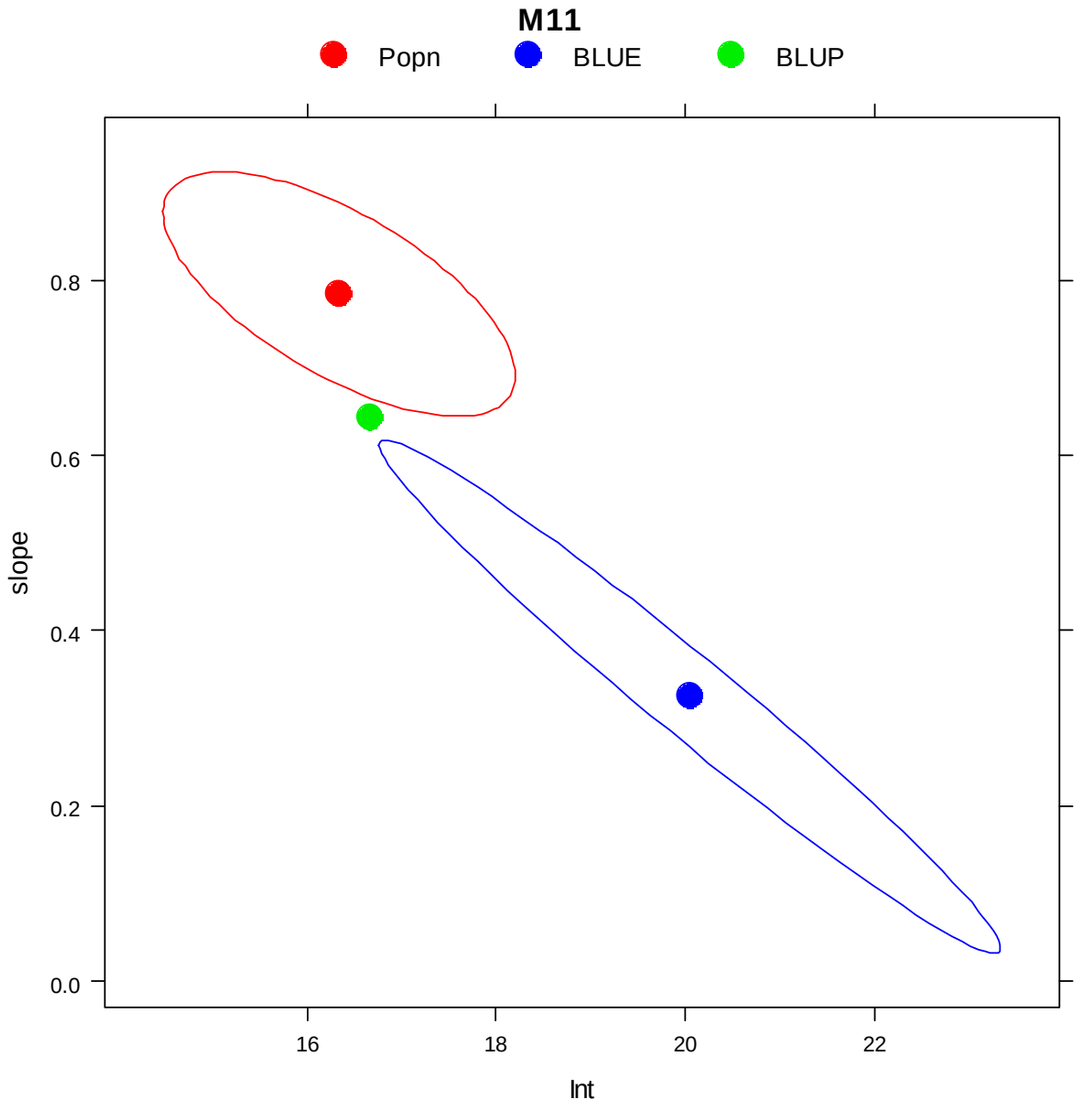
The population estimate with a SD ellipse.



The population estimate with a SD ellipse

and

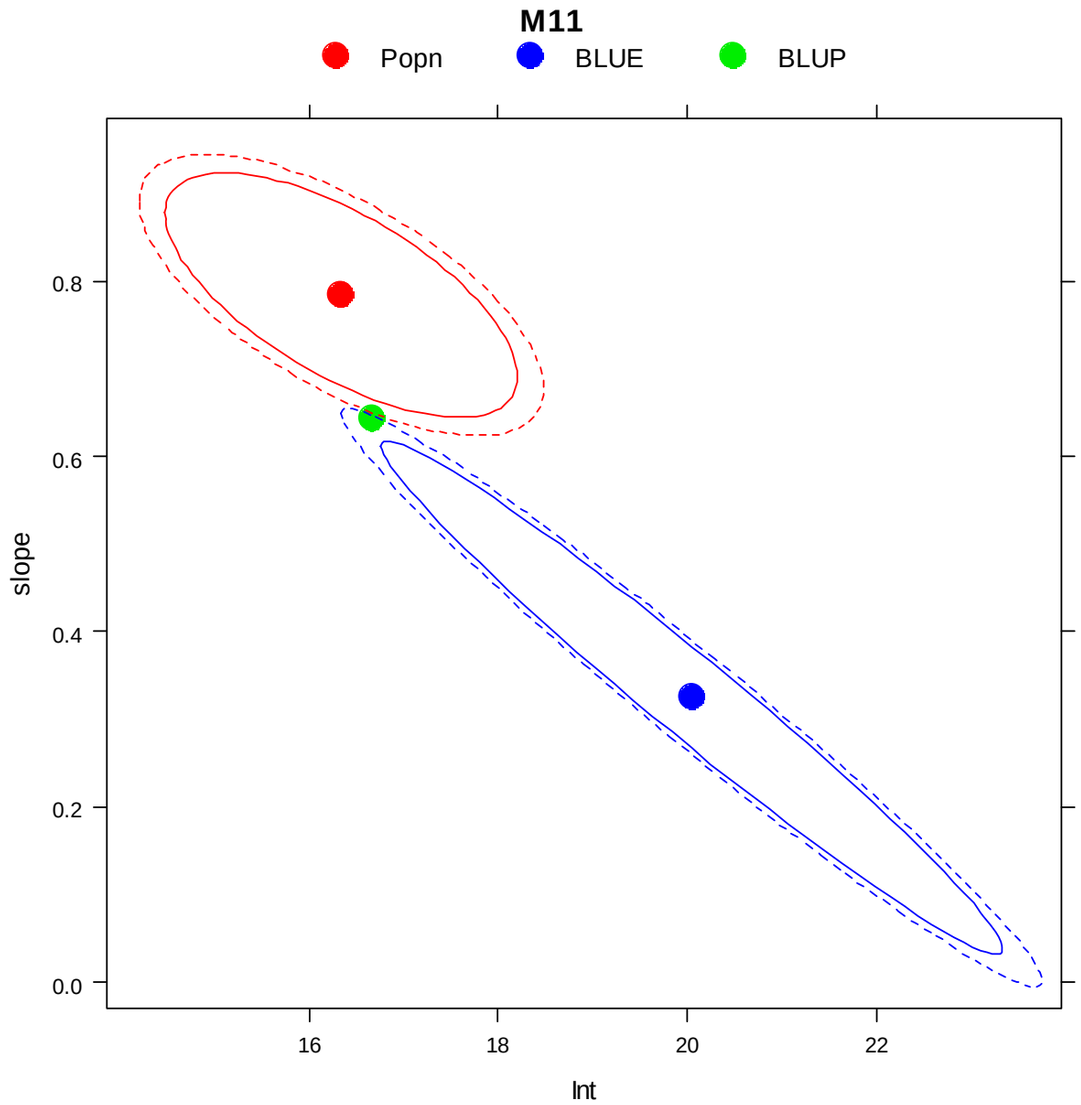
the BLUE with its SE ellipse



The EBLUP is an Inverse Variance Weighted mean of the BLUE and of the population estimate.

We can think of taking the BLUE and ‘shrinking’ it towards the population estimate along a path that optimally combines the two components.

The path is formed by the osculation points of the families of ellipses around the BLUE and the population estimate.



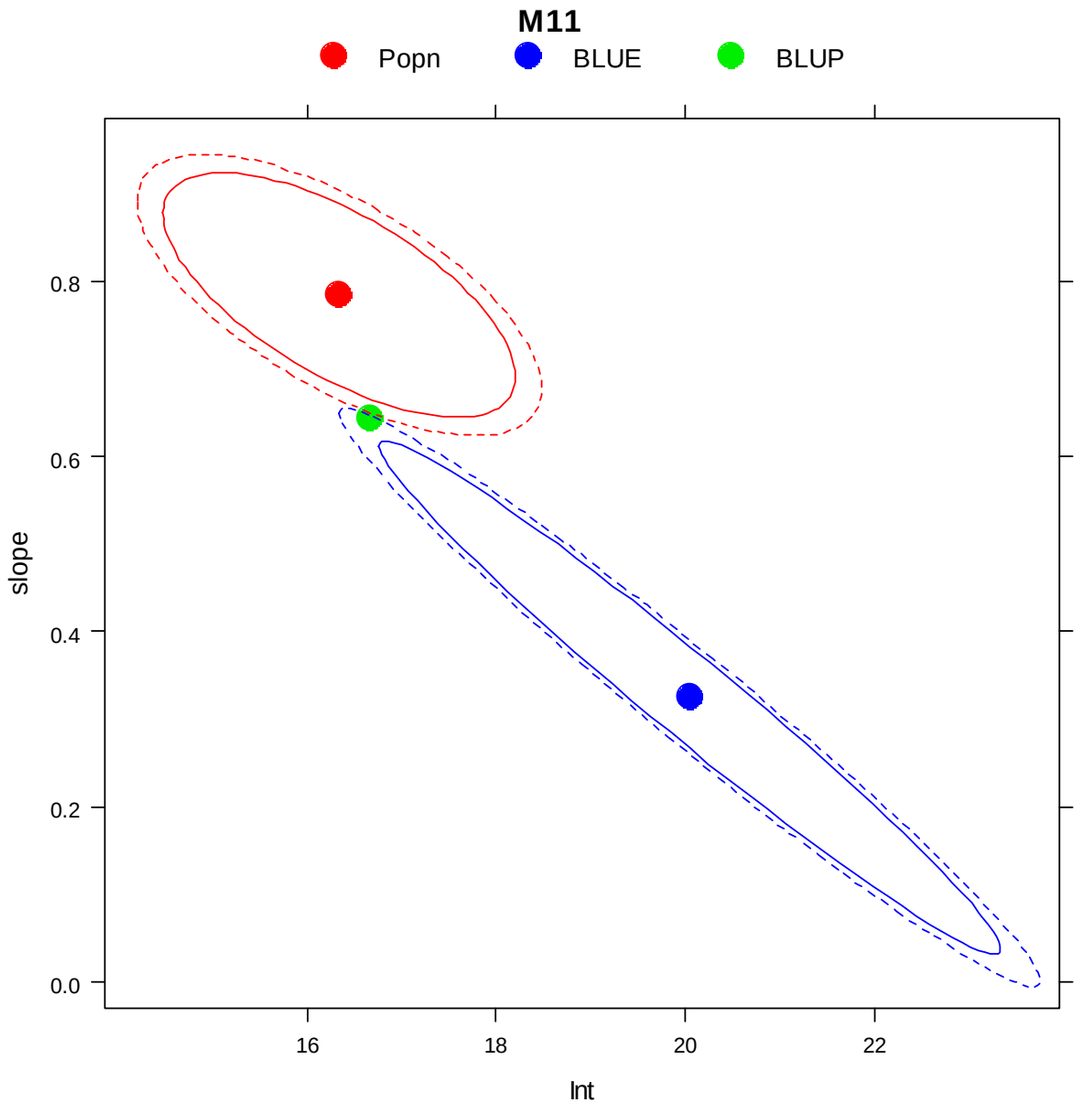
The amount and direction of shrinkage depends on the relative shapes and sizes of

G

and

$$\mathrm{Var}(\hat{\beta}_i \mid i) = \sigma^2 (\mathbf{X}_i' \mathbf{X}_i)^{-1} \approx \frac{\sigma^2}{T_i} \mathbf{S}_{\mathbf{X}_i}^{-1}$$

The BLUP is at an osculation point of the families of ellipses generated around the BLUE and population estimate.



Imagine what could happen if **G** were oriented differently:

Paradoxically, both the slope and the intercept could be far outside the population estimate and the BLUE.

When is a BLUP a BLUPPER?

The rationale behind BLUPs is based on ***exchangeability***. No outside information should make this cluster stand out from the others and the mean of the population deserves the same weight in prediction for this cluster as it deserves for any other cluster that doesn't stand out.

If a cluster stands out somehow, then the BLUP might be a BLUPPER.

Interpreting G

The parameters of **G** give the variance of the intercepts, the variance of the slopes and the covariance between intercepts and the slopes.

Would it make sense to assume that the covariance is 0 to reduce the number of parameters in the model? To address this, consider that the variance of the heights of individual regression lines a fixed value of X is:

$$\begin{aligned}\mathrm{Var}(\beta_{i0}+\beta_{i1}X) &= \mathrm{Var}\left(\begin{bmatrix}1 & X\end{bmatrix}\begin{bmatrix}\beta_{i0}\\ \beta_{i1}\end{bmatrix}\right)\\ &= \begin{bmatrix}1 & X\end{bmatrix}\begin{bmatrix}g_{00} & g_{01}\\ g_{10} & g_{11}\end{bmatrix}\begin{bmatrix}1\\ X\end{bmatrix}\\ &= g_{00}+2g_{01}X+g_{11}X^2\end{aligned}$$

Summarizing:

$$\text{Var}(\beta_{\iota 0} + \beta_{\iota 1} X) = g_{00} + 2g_{01}X + g_{11}X^2$$

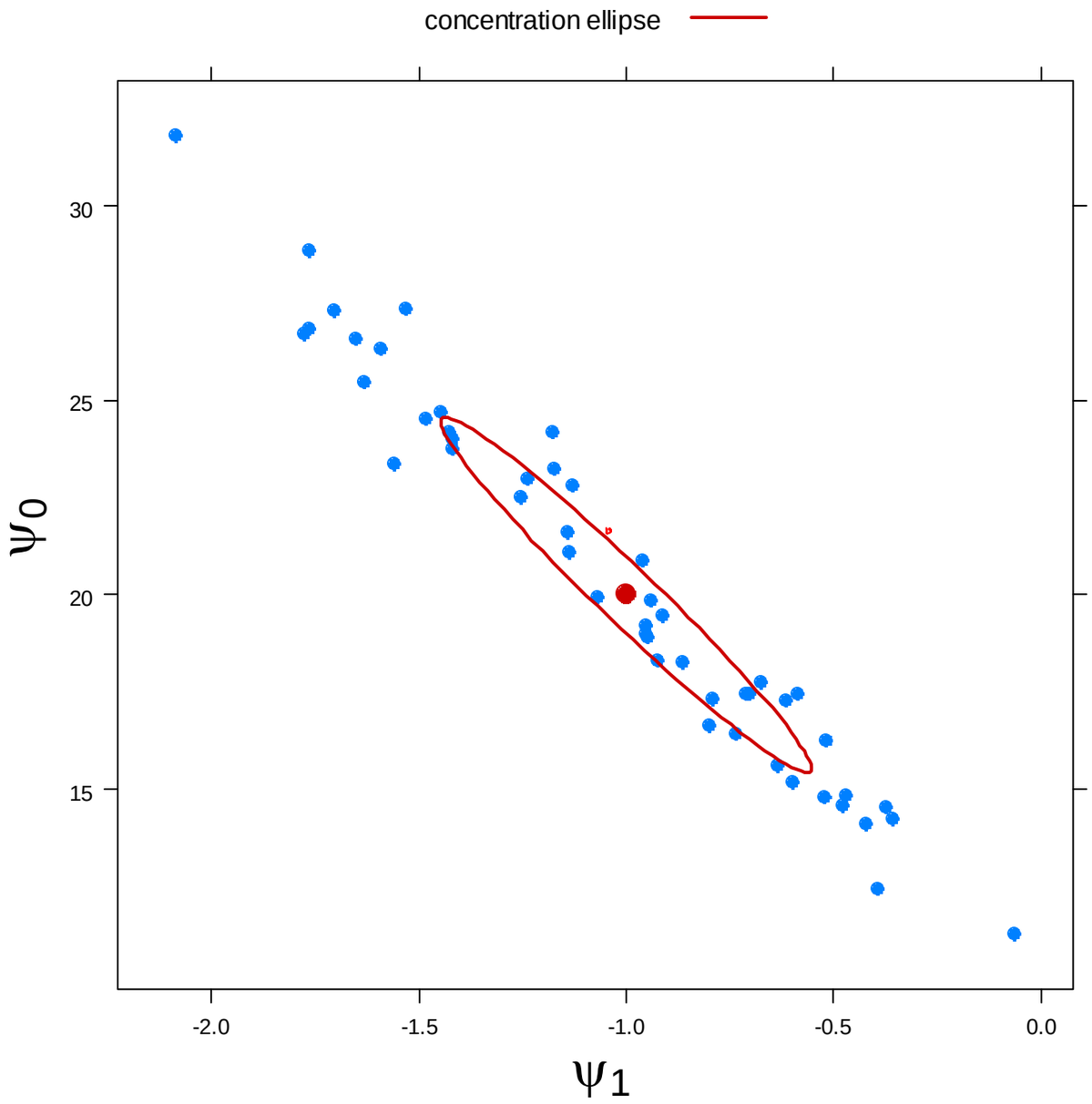
is quadratic function of X.

So $\text{Var}(\beta_{\iota 0} + \beta_{\iota 1} X)$ has a minimum at $-\frac{g_{01}}{g_{11}}$

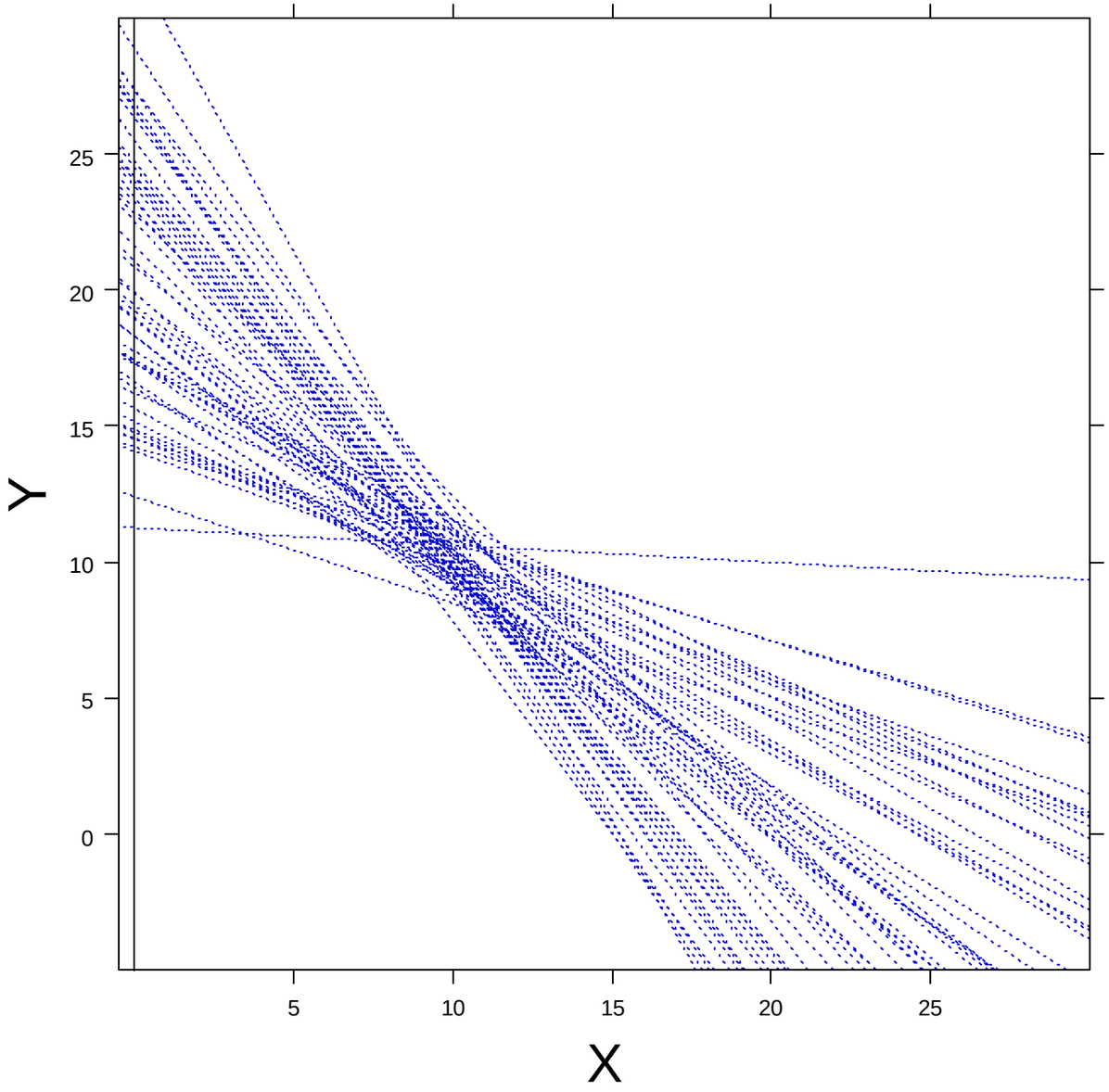
and the minimum variance is $g_{00} - \frac{g_{01}^2}{g_{11}}$

Thus, assuming that the covariance is 0 is equivalent to assuming that the minimum variance occurs when $X = 0$. This is an assumption that is not invariant with location transformations of X. It is similar to removing a main effect that is marginal to an interaction in a model, something that should not be done without a thorough understanding of its consequences.

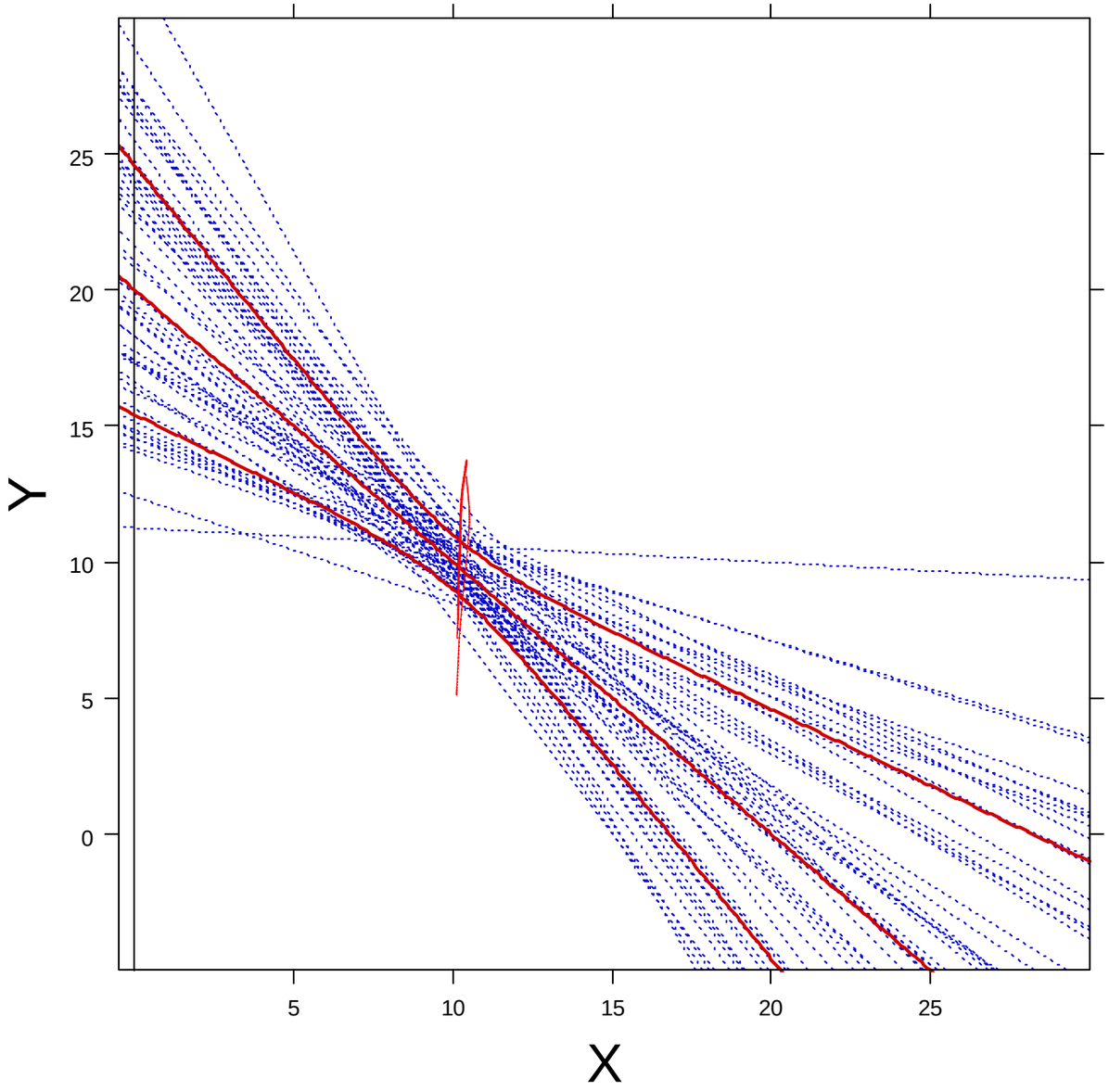
Example: Let $\begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} 20 \\ -1 \end{bmatrix}$ and $\mathbf{G} = \begin{bmatrix} 10.5 & -1 \\ -1 & 0.1 \end{bmatrix}$



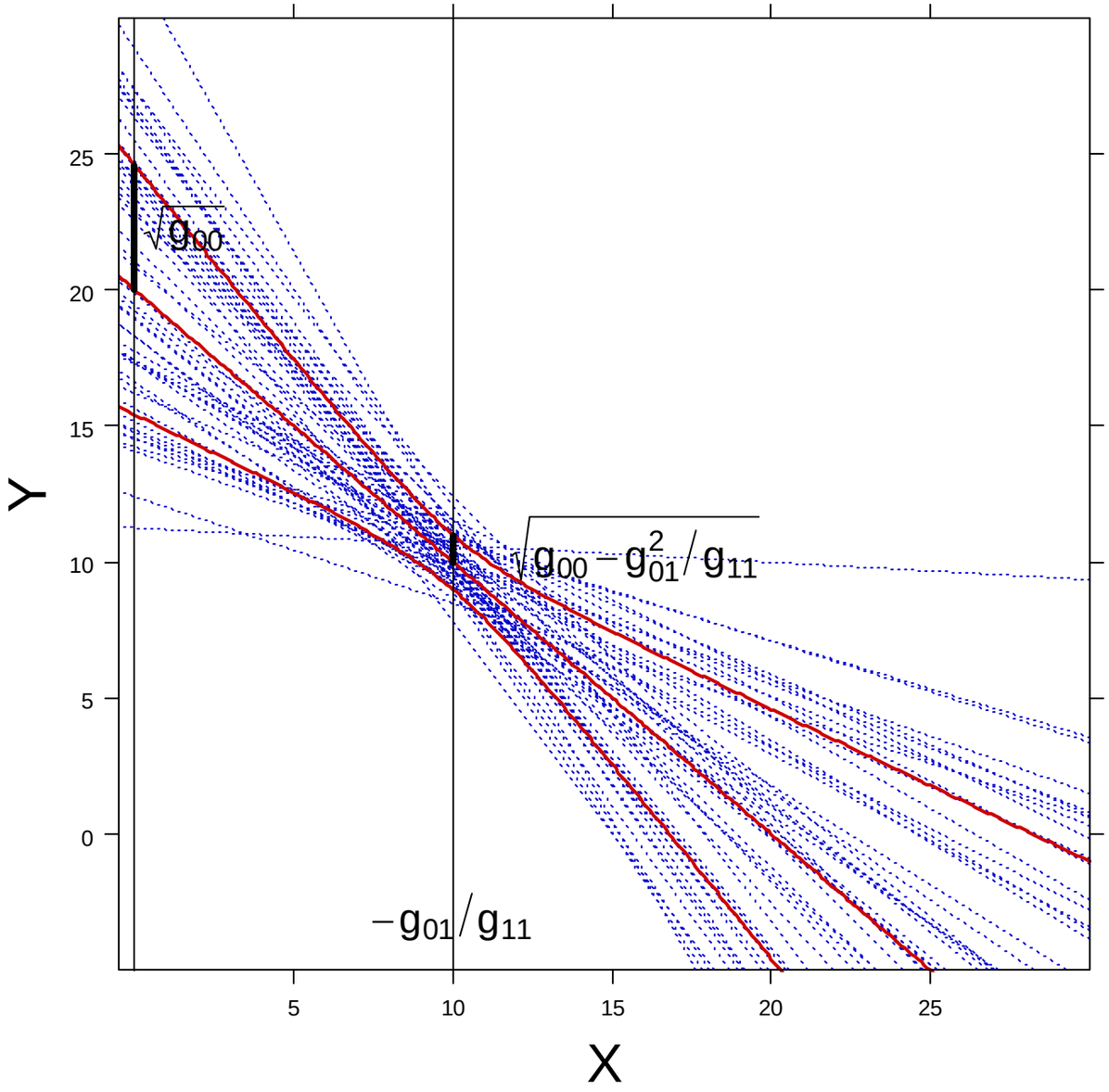
A sample of lines in beta space



The same lines in data space.



The same lines in data space with the population mean line and lines at one SD above and below the population mean line



The parameters of **G** determine the location and value of the minimum standard deviation of lines

With two time-varying variables with random effects, the **G** matrix would look like:

$$\text{Var}\left(\begin{bmatrix}\beta_{i0}\\ \beta_{i1}\\ \beta_{i2}\end{bmatrix}\right)=\begin{bmatrix}g_{00} & g_{01} & g_{02}\\ g_{10} & g_{11} & g_{12}\\ g_{20} & g_{21} & g_{22}\end{bmatrix}$$

The point of minimum variance is located at:

$$-\begin{bmatrix}g_{11} & g_{12}\\ g_{21} & g_{22}\end{bmatrix}^{-1}\begin{bmatrix}g_{10}\\ g_{20}\end{bmatrix}$$

Differences between lm (OLS) and lme (mixed model) with balanced data

Just looking at regression coefficients:

```
> fit.ols <- lm( distance ~ age * Sex, dd)
> fit.mm <- lme( distance ~ age * Sex, dd,
+         random = ~ 1 + age | Subject)
> summary(fit.ols)

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

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
. . .
> summary(fit.mm)
. . .
Fixed effects: distance ~ age * Sex
                    Value Std.Error DF   t-value p-value
(Intercept)    16.340625 1.0185320 79 16.043311  0.0000
age             0.784375 0.0859995 79  9.120691  0.0000
SexFemale       1.032102 1.5957329 25  0.646789  0.5237
age:SexFemale  -0.304830 0.1347353 79 -2.262432  0.0264
```

Note that going from OLS to MM, precision shifts from between-subject comparisons to within-subject comparisons. When data are balanced, lm (OLS) and lme (mixed models) produce the same $\hat{\gamma}$s with different SEs.

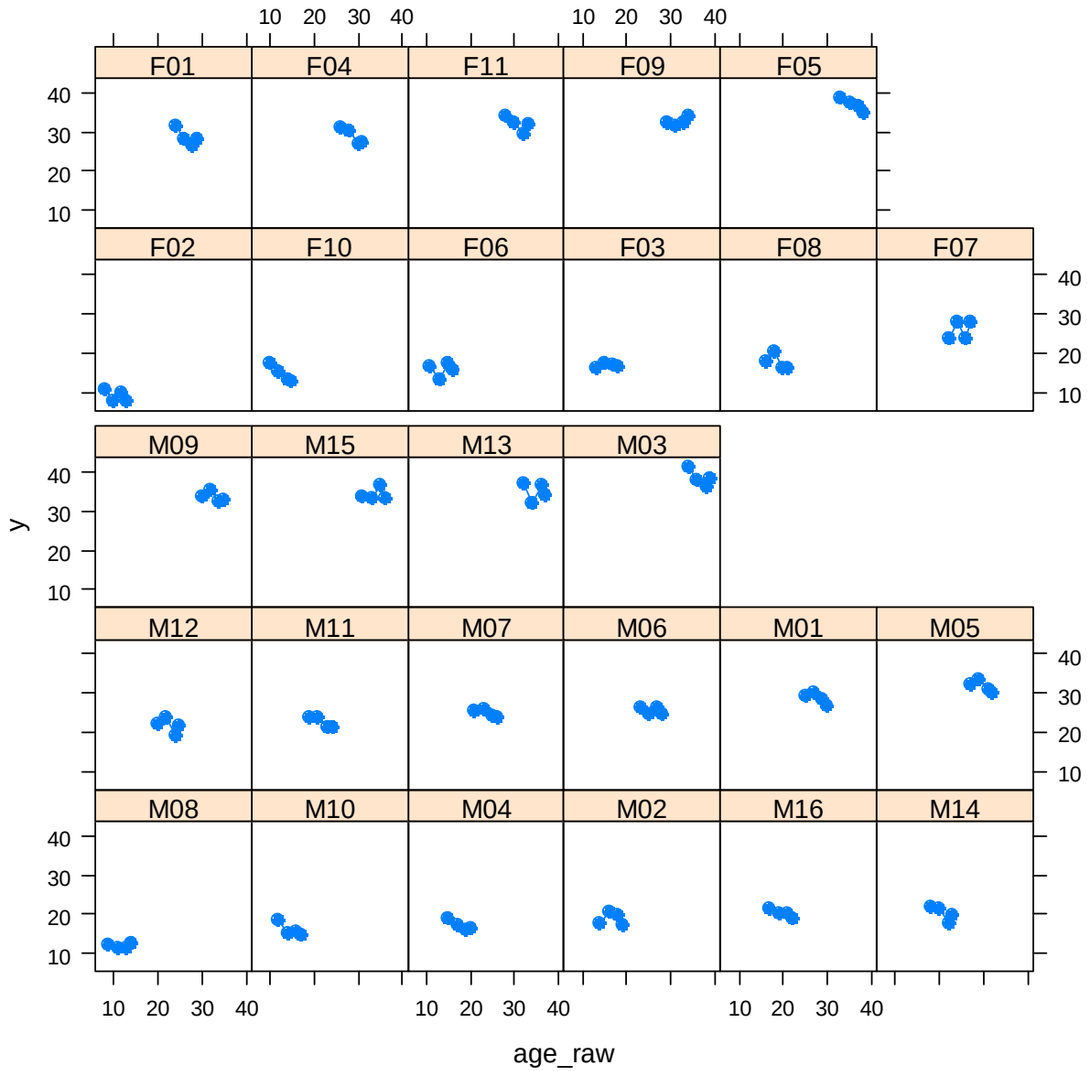
Take 2: Learning lessons from unbalanced data

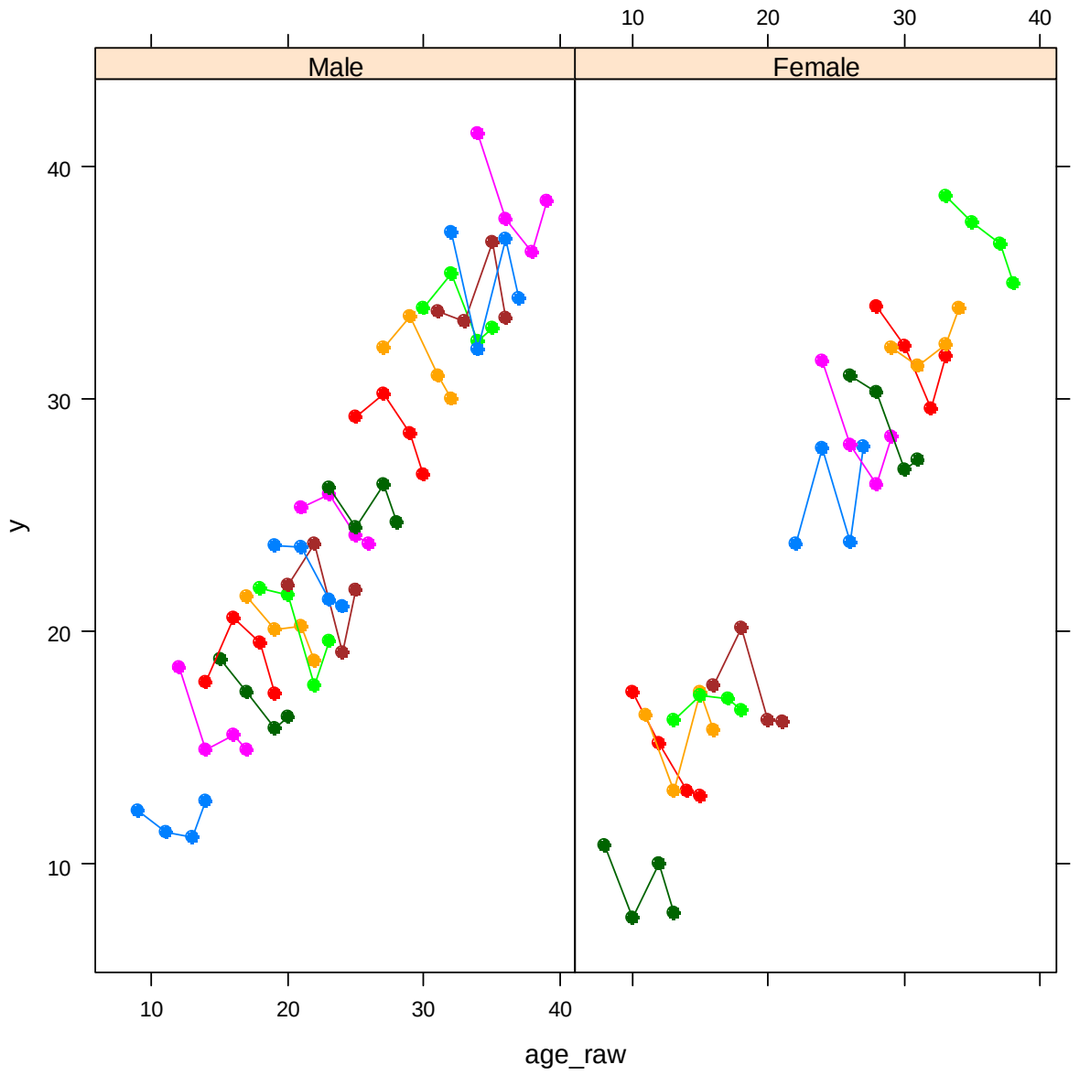
What can happen with unbalanced data?

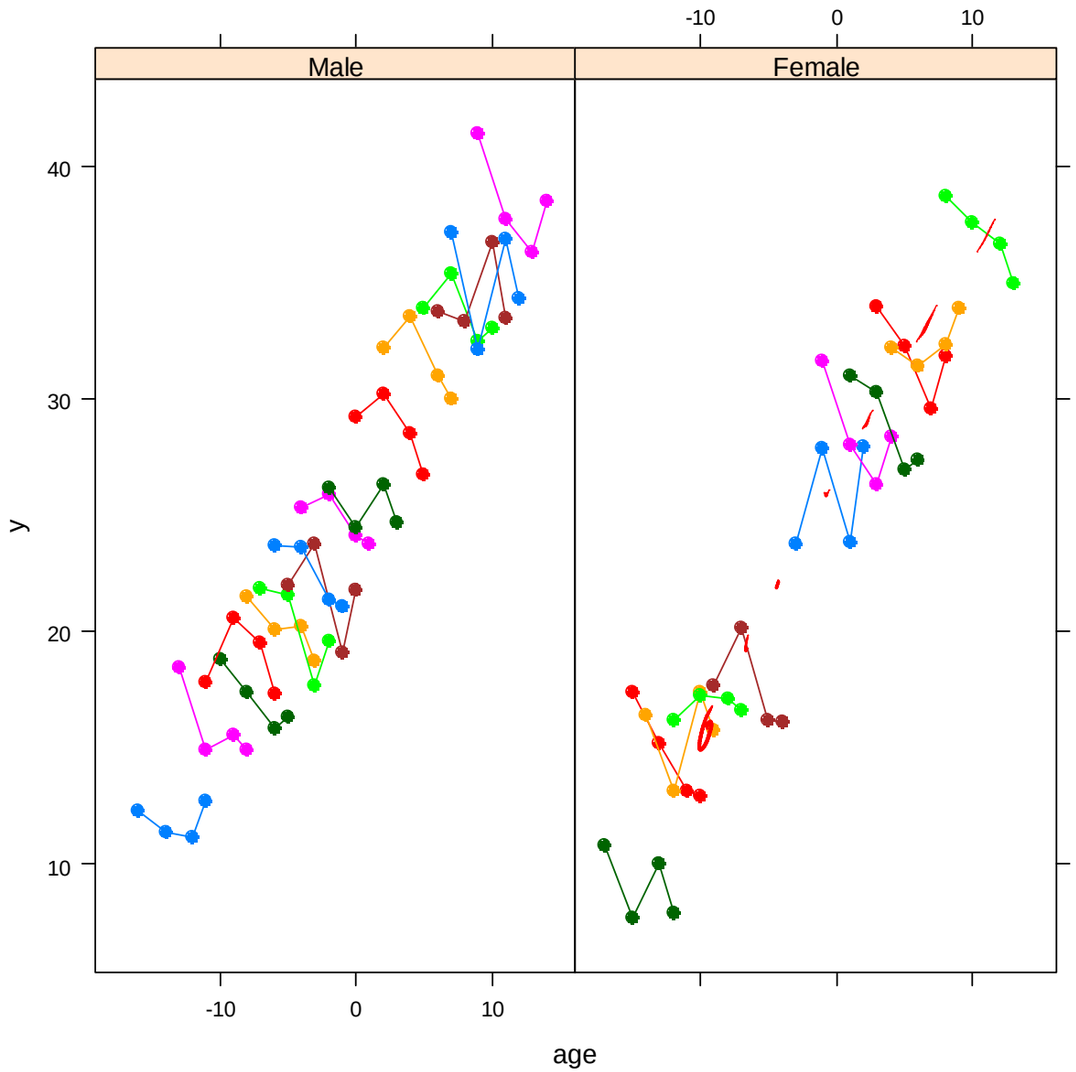
Here is some data that is similar to the Pothoff and Roy data but with:

- different age ranges for different subjects
- a between-subject effect of age that is different from the within-subject effect of age

```
> head(du)
          y  x id xb xw Subject    Sex age
1 12.37216  8  1 11 -3     F09 Female   8
2 11.20801 10  1 11 -1     F09 Female  10
3 10.44755 12  1 11  1     F09 Female  12
4 10.43831 13  1 11  2     F09 Female  13
5 14.13549  9  2 12 -3     F11 Female   9
6 13.47965 11  2 12 -1     F11 Female  11


> tail(du)
            y  x id xb xw Subject  Sex age
103 35.67045 37 26 36  1     M08 Male  37
104 35.70928 38 26 36  2     M08 Male  38
105 38.81624 34 27 37 -3     M10 Male  34
106 37.87866 36 27 37 -1     M10 Male  36
107 36.22499 38 27 37  1     M10 Male  38
108 35.62520 39 27 37  2     M10 Male  39
```


F01
F04
F11
F09
F05
F02
F10
F06
F03
F08
F07
M09
M15
M13
M03
M12
M11
M07
M06
M01
M05
M08
M10
M04
M02
M16
M14
y
age_raw


Male
Female
y
age_raw
10
20
30
40



Using age centered at 25.

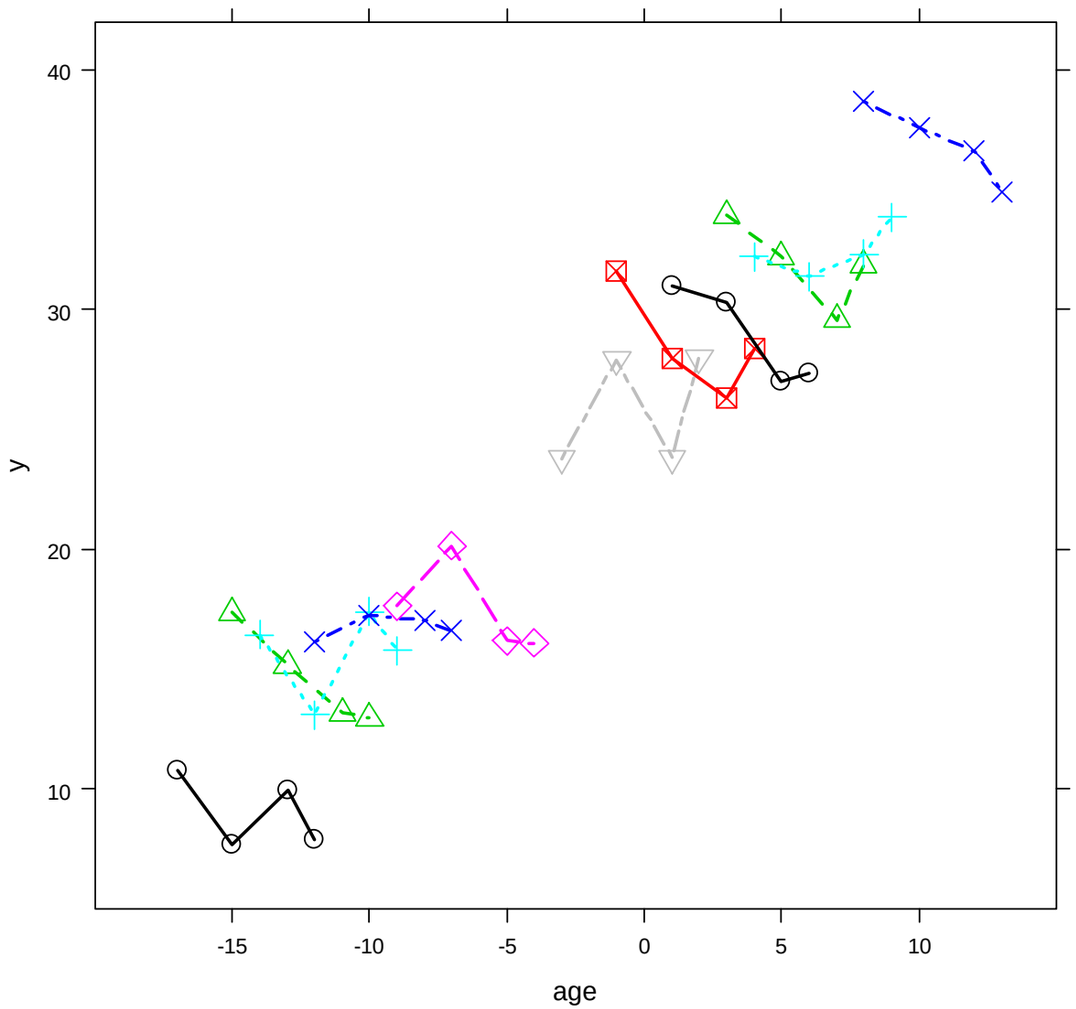
Why? Like the ordinary regression model, the mixed model is **equivariant** under **global centering** but convergence may be improved because the **G** matrix is less eccentric.

R code and output

```
> fit <- lme( y ~ age * Sex, du,
+             random = ~ 1 + age| Subject)
> summary( fit )
Linear mixed-effects model fit by REML
 Data: du
       AIC      BIC    logLik
  374.6932 395.8484 -179.3466

Random effects:
 Formula: ~1 + age | Subject
 Structure: General positive-definite, Log-
Cholesky parametrization
             StdDev     Corr
(Intercept) 9.32672995 (Intr)
age         0.05221248 0.941
Residual    0.50627022
```

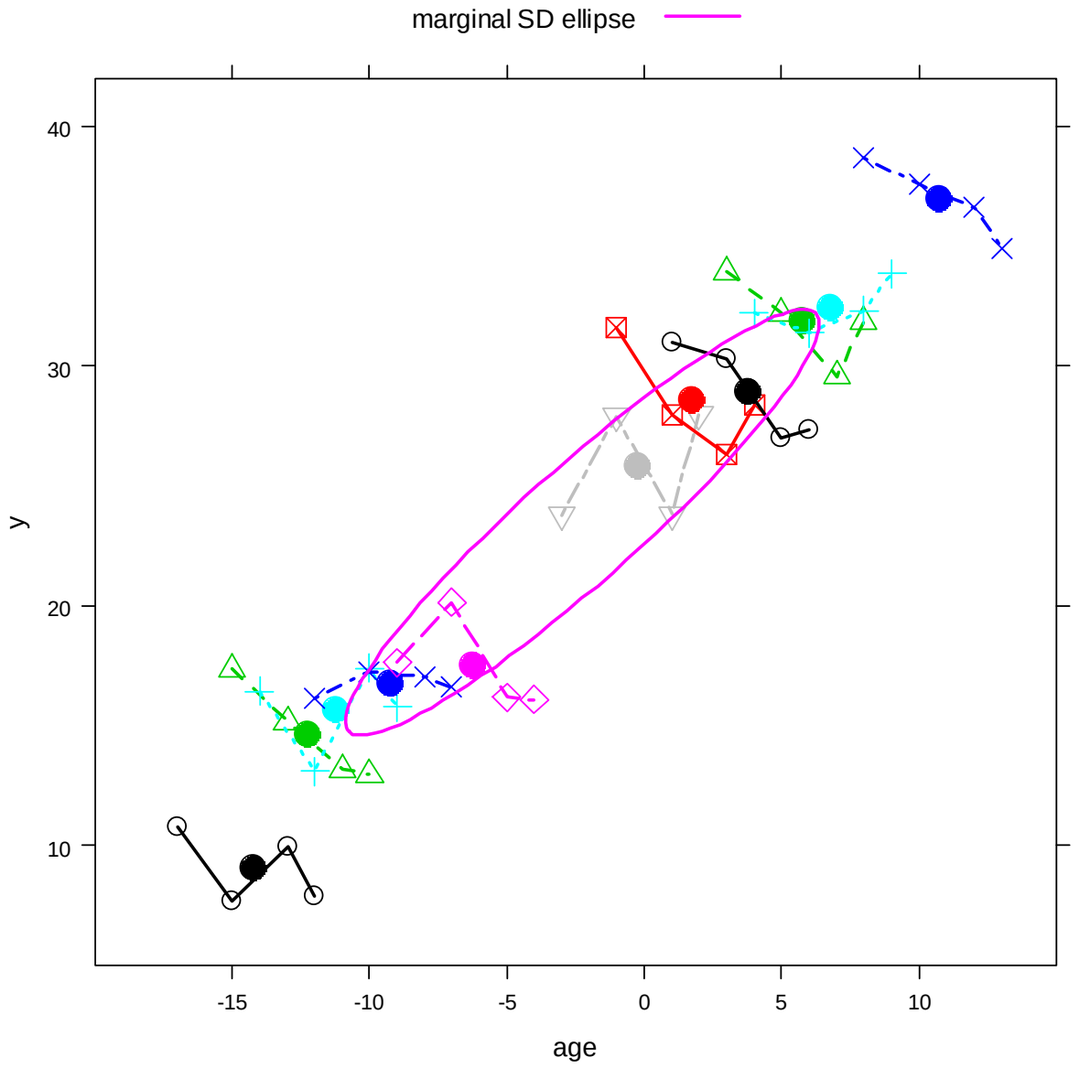
```
Fixed effects: y ~ age * Sex
                  Value Std.Error DF    t-value p-value
(Intercept)    40.26568  2.497546 79  16.122095  0.0000
age            -0.48066  0.035307 79 -13.613685  0.0000
SexFemale     -14.01875  3.830956 25  -3.659333  0.0012
age:SexFemale   0.05239  0.055373 79   0.946092  0.3470

 Correlation:
              (Intr) age    SexFml
age           -0.007
SexFemale     -0.652  0.005
age:SexFemale  0.005 -0.638  0.058

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.10716969 -0.54148659 -0.02688422  0.59030024  2.14279806

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

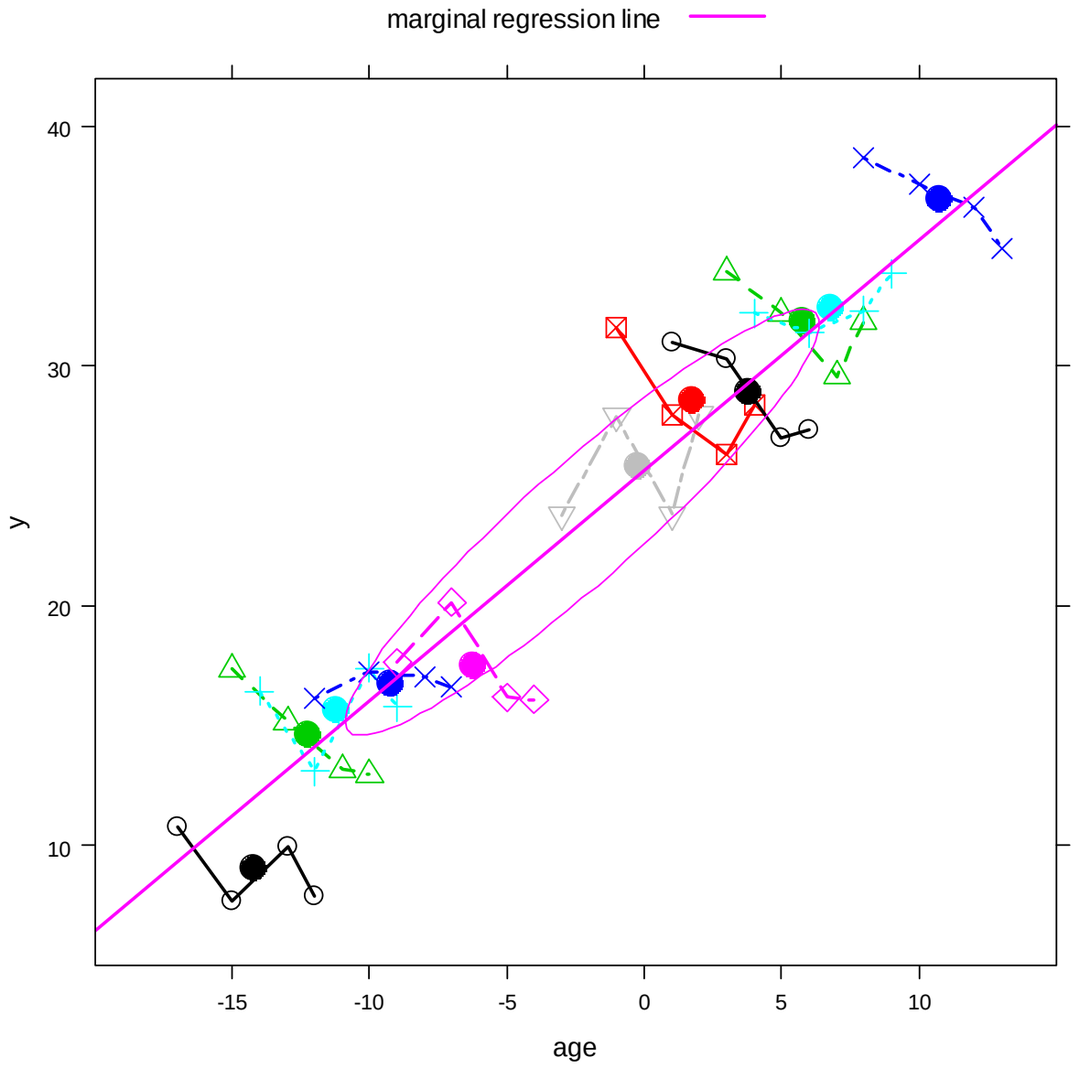
Between, Within and Pooled Models



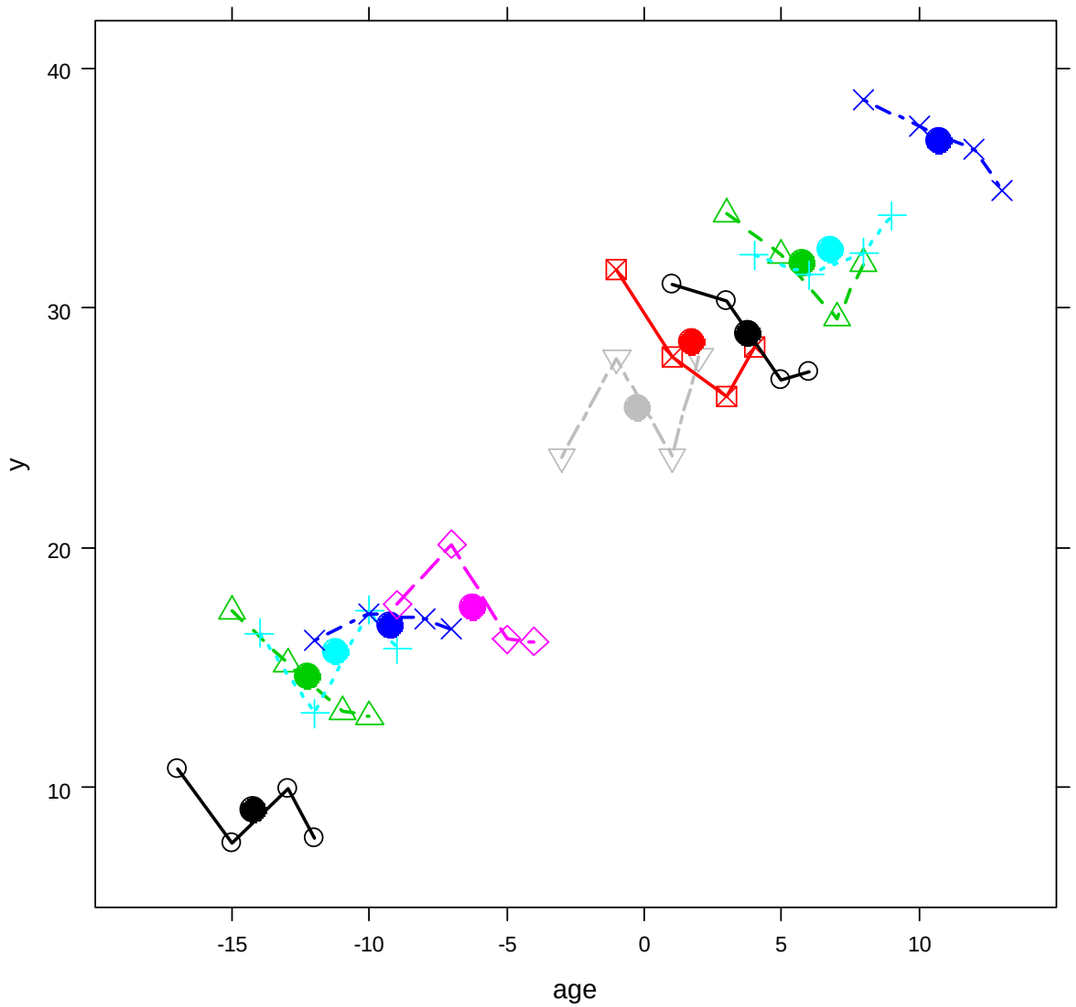
We first focus on one group, the female data:

What models could we fit to this data?

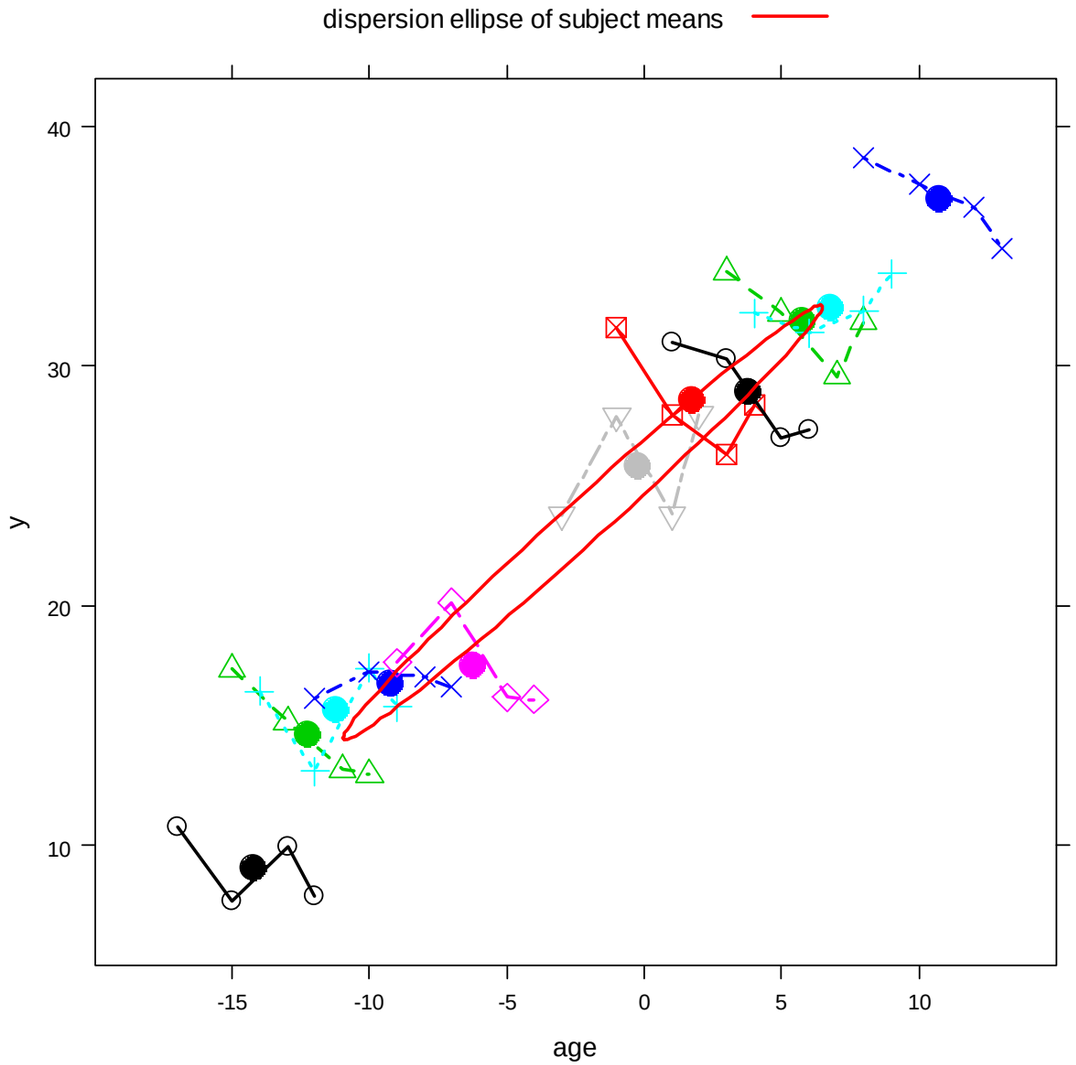

marginal SD ellipse
40
30
20
10
y
-15
-10
-5
0
5
10
age

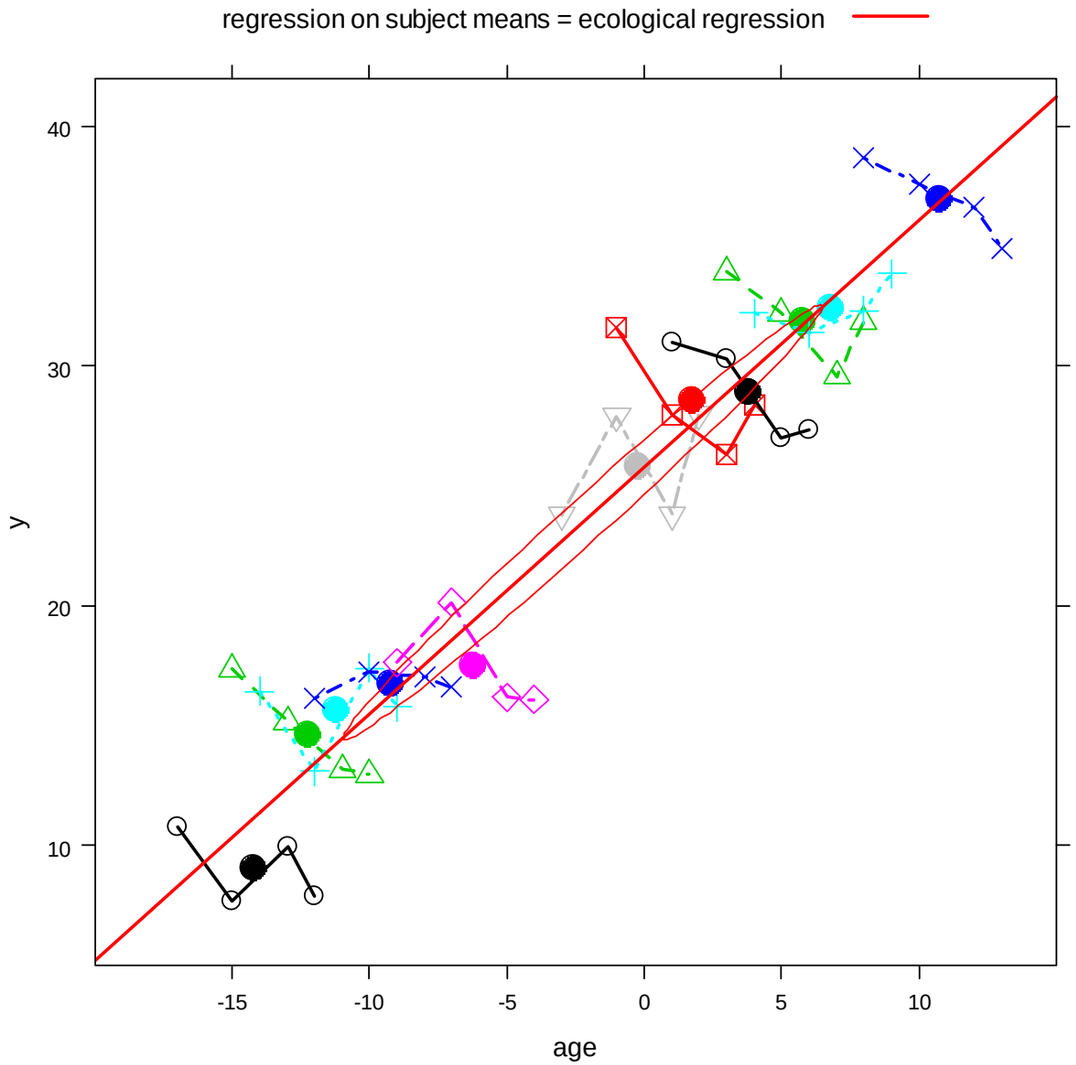


Regressing with the pooled data – ignoring Subject – yields the marginal (unconditional) estimate of the slope: $\hat{\beta}_P$

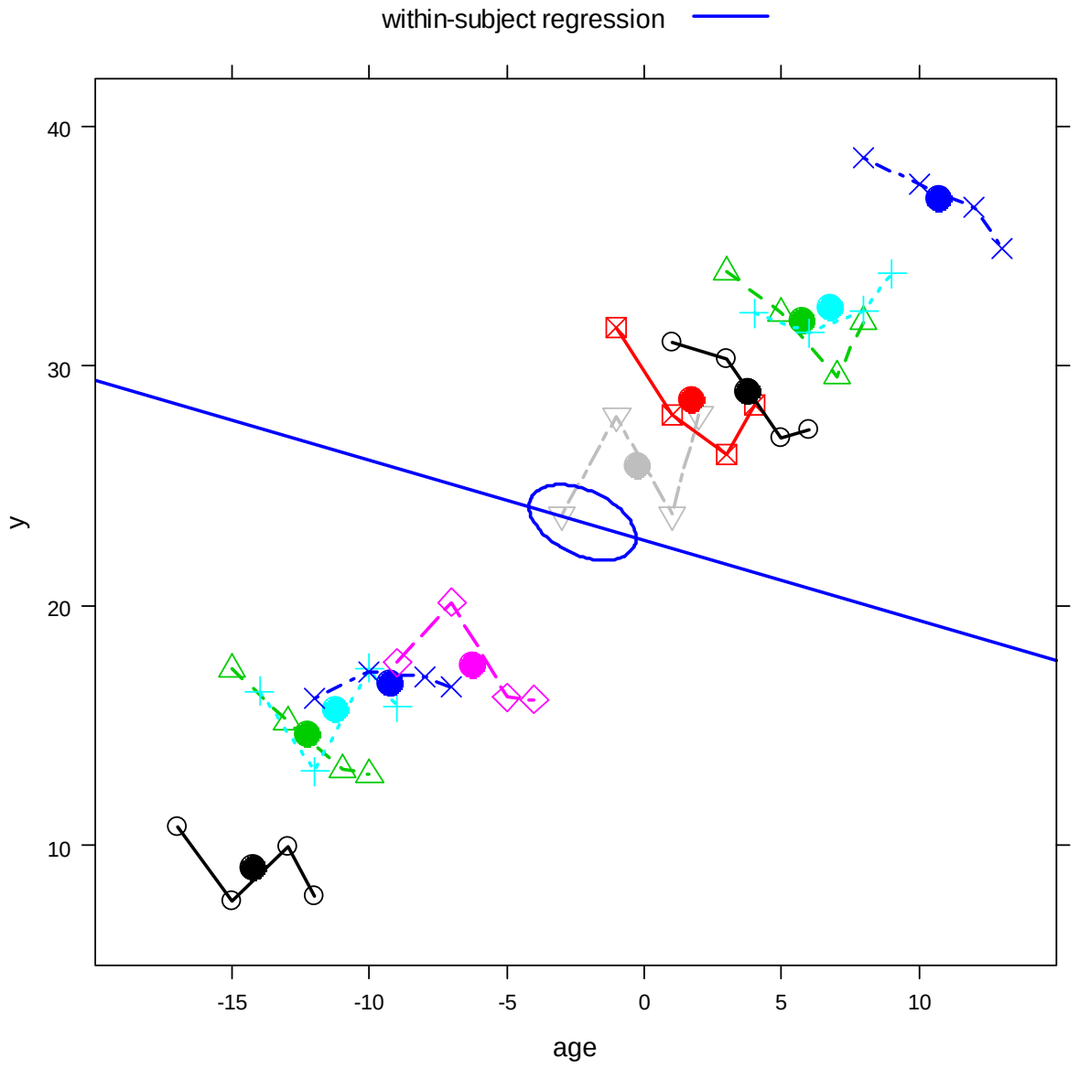


We could replace each Subject by its means for x and y and use the resulting aggregated data with one point for each Subject.


dispersion ellipse of subject means
40
30
20
10
y
-15
-10
-5
0
5
10
age



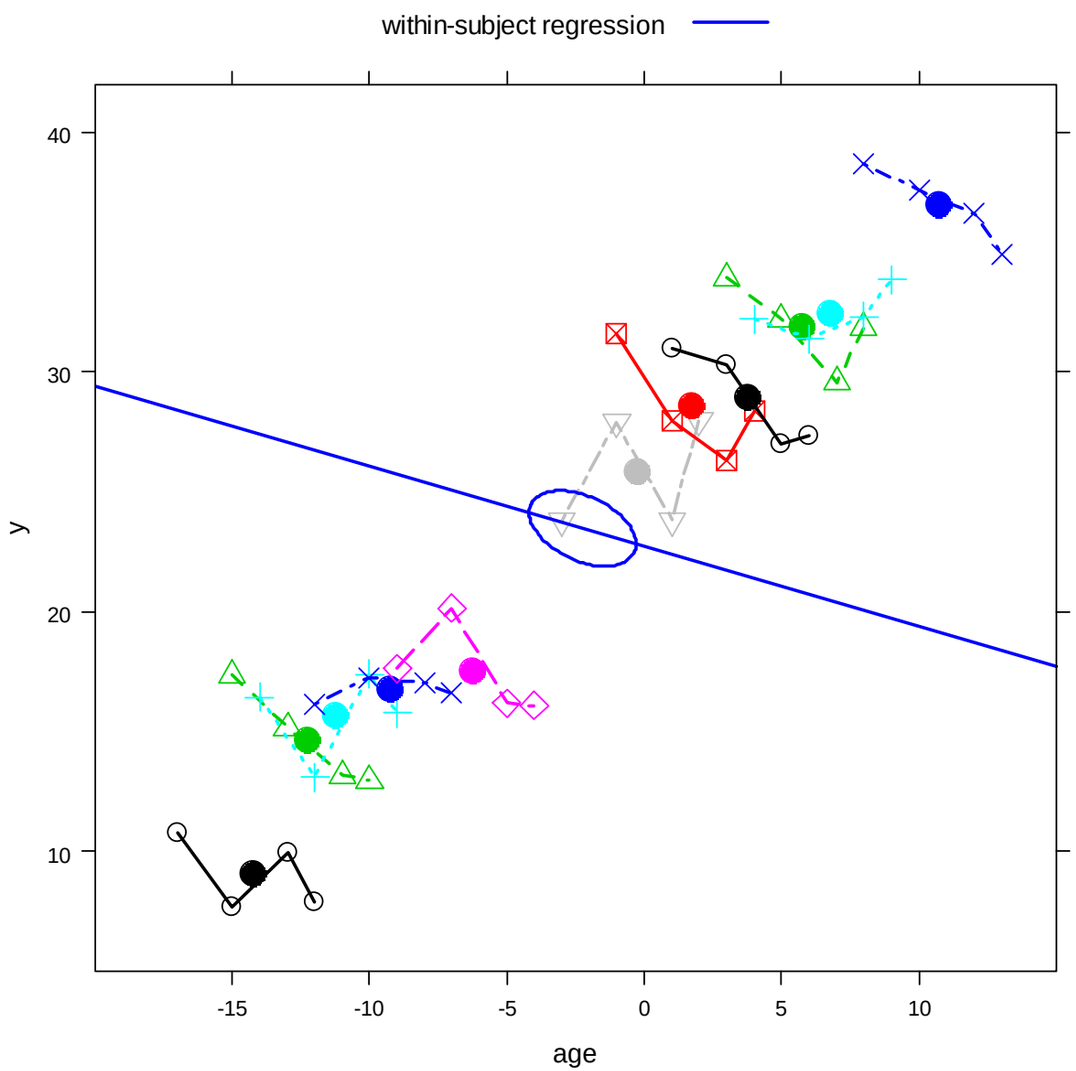
Performing a regression on the aggregated data yields the ‘between-subject’ regression, in some contexts called an ‘ecological regression’ estimating, in some contexts, the 'compositional effect' of age.



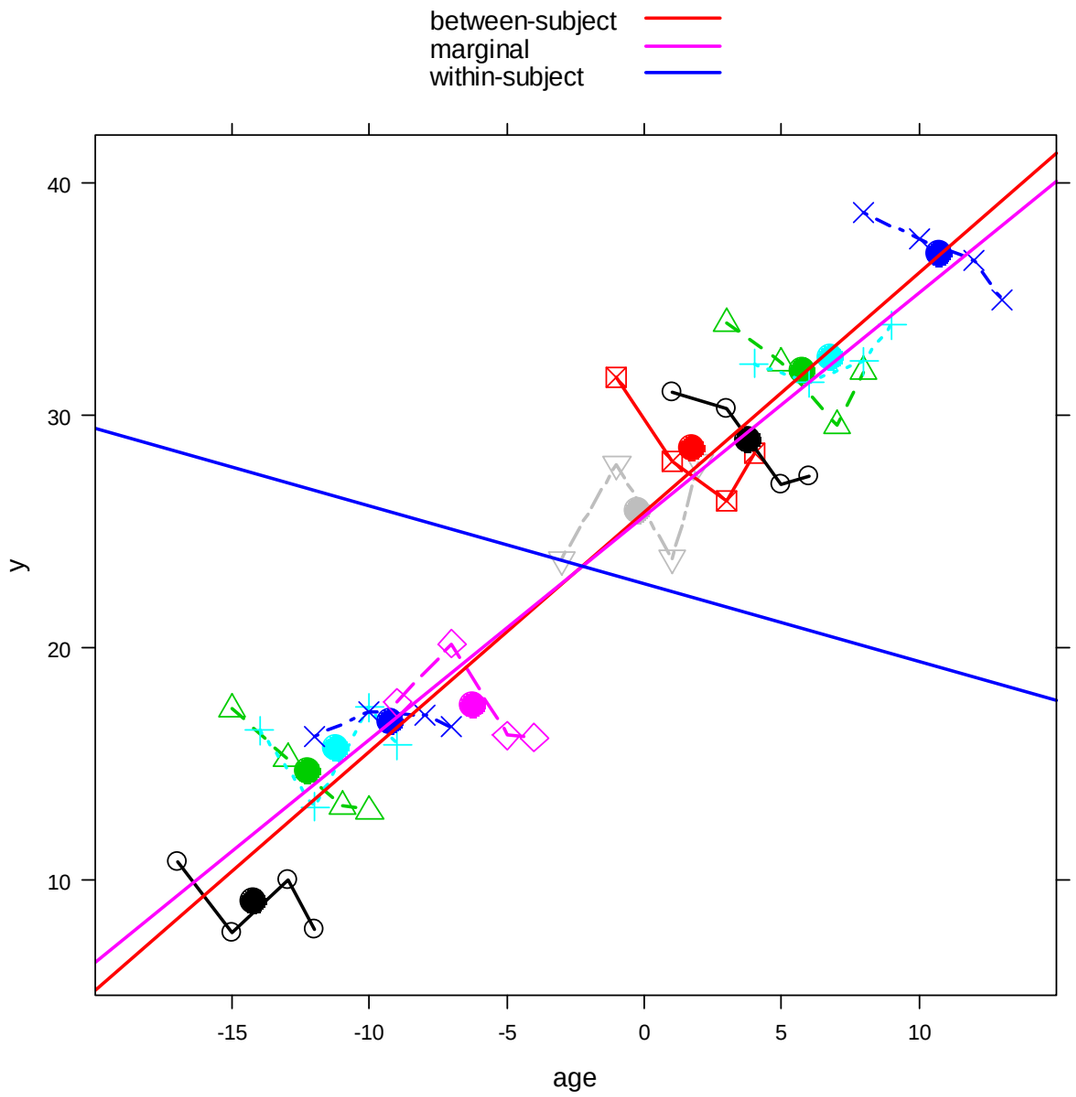
We can combine all within-subject regressions to get a combined estimate of the within-subject slope. This is the estimate obtained with a fixed-effects model using age and Subject additively. Equivalently, we can perform a regression using (the within-subject residuals of y minus mean y) on (age minus mean age).

Q: Which is better: $\hat{\beta}_{\mathrm{B}}, \hat{\beta}_{\mathrm{W}}$ or $\hat{\beta}_{\mathrm{P}}$?

.



A: None. They answer different questions. Typically, $\hat{\beta}_P$ would be used for prediction across the population; $\hat{\beta}_W$ for 'causal' inference controlling for between-subject confounders, assuming that all confounders affect all observations similarly.



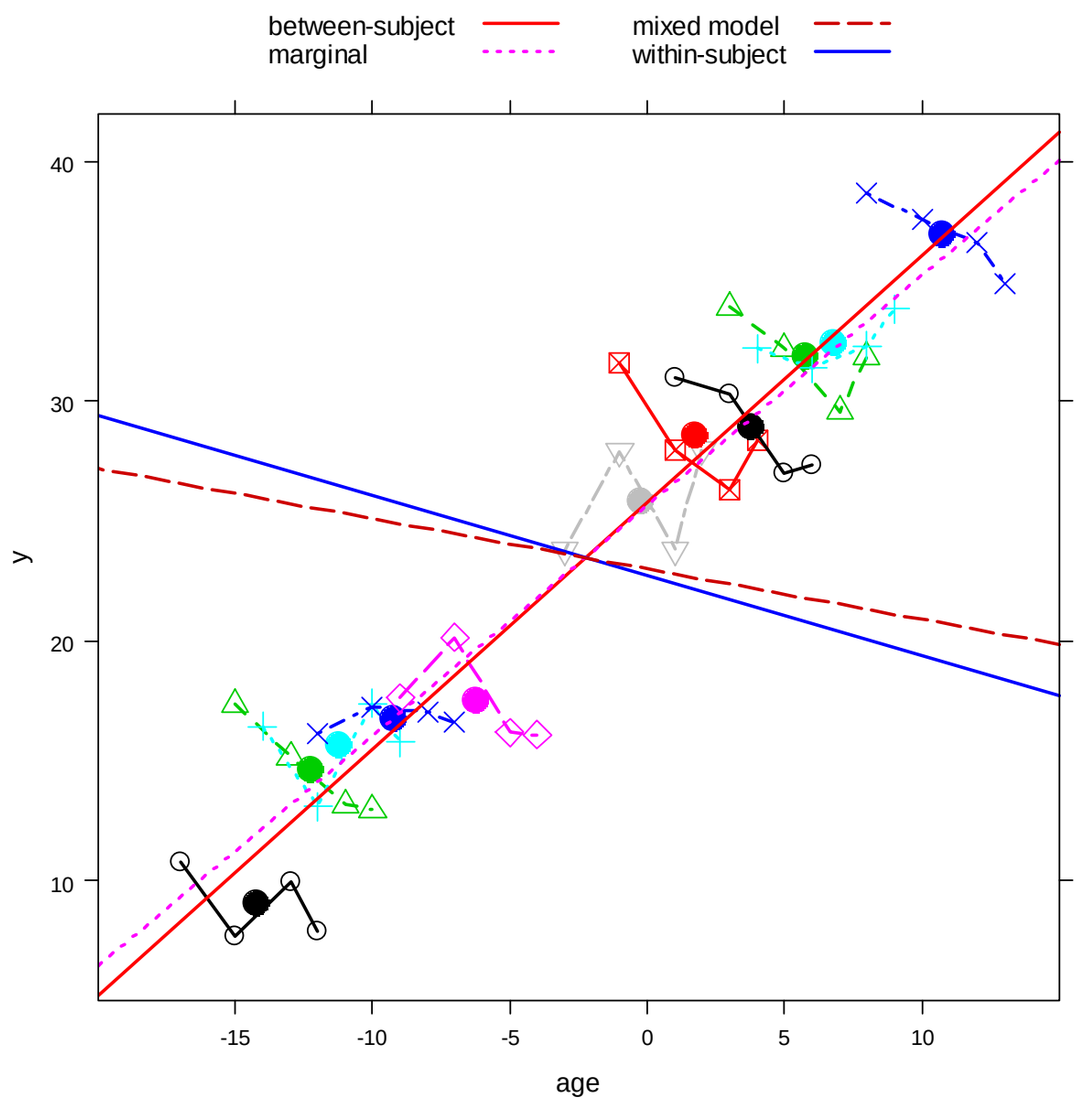
The relationship among estimators:

$\hat{\beta}_{\mathrm{P}}$ combines $\hat{\beta}_{\mathrm{B}}$ and $\hat{\beta}_{\mathrm{W}}$:

$$\hat{\beta}_{\mathrm{P}} = \left(W_{\mathrm{B}} + W_{\mathrm{W}}\right)^{-1}\left(W_{\mathrm{B}}\hat{\beta}_{\mathrm{B}} + W_{\mathrm{W}}\hat{\beta}_{\mathrm{W}}\right)$$

The weights depend only on the design (**X** matrix), not of estimated variances of the response.

The Mixed Model



The mixed model estimate[5] also combines $\hat{\beta}_B$ and $\hat{\beta}_W$:

$$\hat{\beta}_{MM} = \left(W_B^{MM} + W_W\right)^{-1} \left(W_B^{MM}\hat{\beta}_B + W_W\hat{\beta}_W\right)$$

but with a lower weight on $\hat{\beta}_B$:

$$W_B^{MM} = \frac{\sigma^2/T}{\sigma^2/T + g_{00}} W_B = \frac{\frac{\sigma^2/T}{g_{00}}}{1+\frac{\sigma^2/T}{g_{00}}} W_B$$

Note that

$$W_B^{MM} \leq W_B$$

[5] Using a random intercept model

- The mixed model estimator is a variance optimal combination of $\hat{\beta}_B$ and $\hat{\beta}_W$.
- It makes perfect sense if $\hat{\beta}_B$ and $\hat{\beta}_W$ estimate the same thing, i.e. if $\beta_B = \beta_W$!
- Otherwise, it's an arbitrary combination of estimates that estimate different things. The weights in the combination have no substantive interpretation.
- i.e. it's an optimal answer to a meaningless question.

Summary of the relationships among 4 models:

Model	Estimate of slope	Precision
Between Subjects	$\hat{\beta}_B$	W_B
Marginal (pooled data)	$\hat{\beta}_P$	
Mixed Model	$\hat{\beta}_{MM}$	
Within Subjects	$\hat{\beta}_W$	W_W

The pooled estimate combines $\hat{\beta}_{\mathrm{B}}$ and $\hat{\beta}_{\mathrm{W}}$:

$$\hat{\beta}_{\mathrm{P}} = \left(W_{\mathrm{B}} + W_{\mathrm{W}}\right)^{-1}\left(W_{\mathrm{B}}\hat{\beta}_{\mathrm{B}} + W_{\mathrm{W}}\hat{\beta}_{\mathrm{W}}\right)$$

Mixed model

With a random intercept model:

$$y_{it} = \beta_0 + \beta_1 X_{it} + u_{i0} + \varepsilon_{it}, \quad u_{i0} \sim N(0, g_{00}), \quad \varepsilon_{it} \sim N(0, \sigma^2)$$

with g_{00}, σ^2 known $\hat{\beta}_{\mathrm{MM}}$ is **also a weighted combination** of $\hat{\beta}_{\mathrm{B}}$ and $\hat{\beta}_{\mathrm{W}}$ but with **less weight** on $\hat{\beta}_{\mathrm{B}}$:

$$W_{\mathrm{B}}^{\mathrm{MM}} = \frac{\sigma^2/T}{\sigma^2/T + g_{00}} W_{\mathrm{B}}$$

$$= f_{monotone}\left(\frac{\text{Between-Subject Information}}{\text{Within-Subject Information}}\right) \times W_{\mathrm{B}}$$

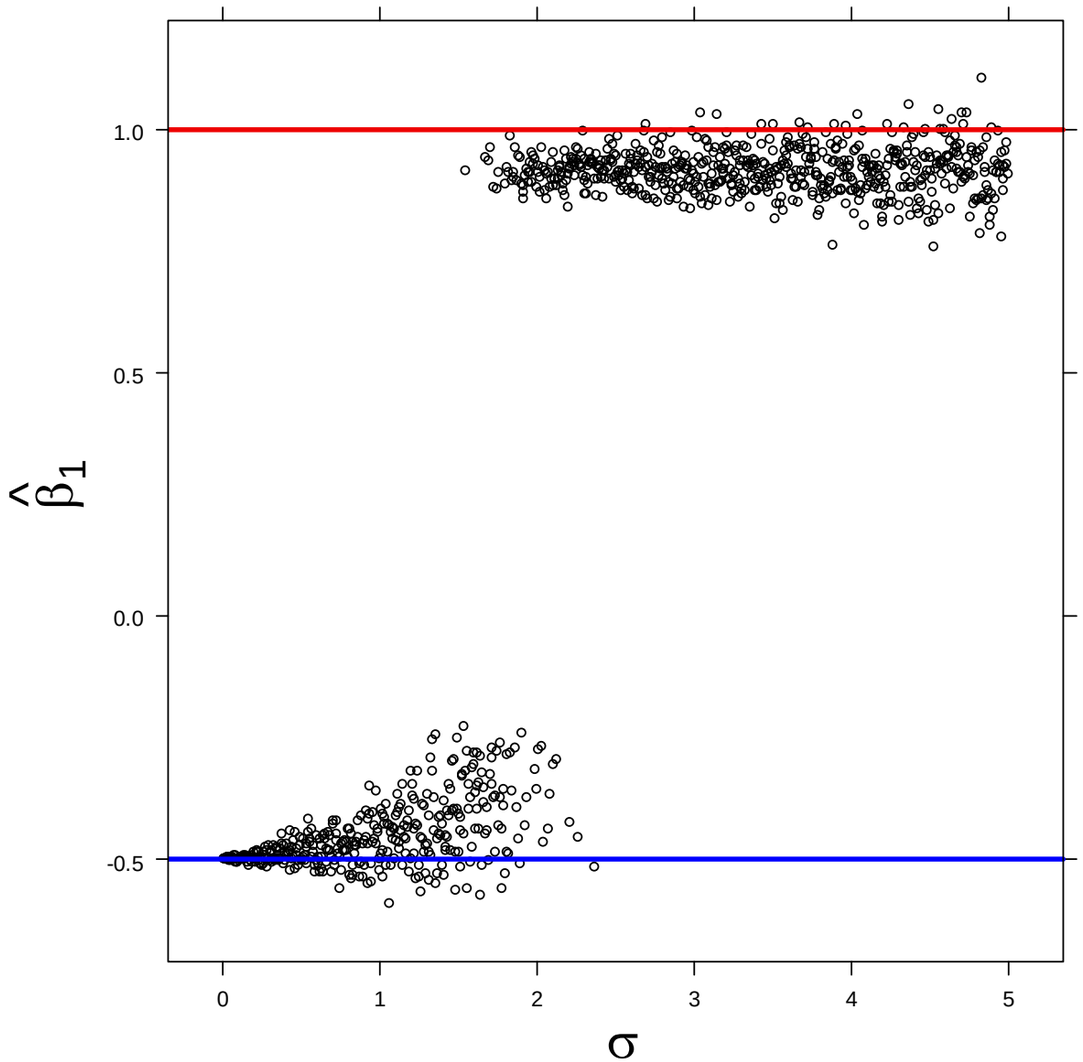
- $\hat{\beta}_{\mathrm{MM}}$ is between $\hat{\beta}_{\mathrm{W}}$ and $\hat{\beta}_{\mathrm{P}}$, i.e. it does better than $\hat{\beta}_{\mathrm{P}}$ in the sense of being closer to $\hat{\beta}_{\mathrm{W}}$ but is not equivalent to $\hat{\beta}_{\mathrm{W}}$.

- With balanced data $\hat{\beta}_{\mathrm{W}} = \hat{\beta}_{\mathrm{MM}} = \hat{\beta}_{\mathrm{P}} = \hat{\beta}_{\mathrm{P}}$

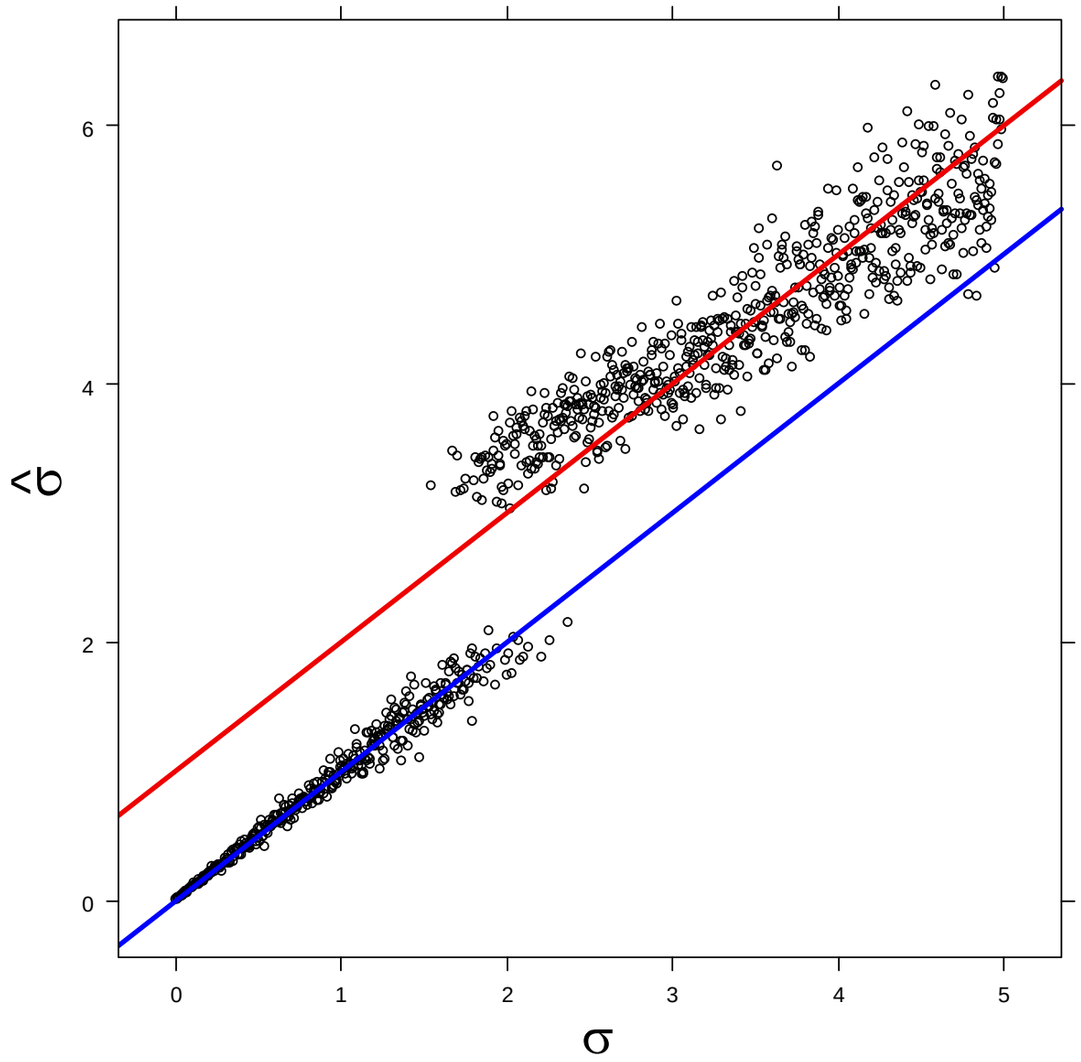
- As $\frac{\sigma^2}{T} \times \frac{1}{g_{00}} \to 0$, $\hat{\beta}_{\mathrm{MM}} \to \hat{\beta}_{\mathrm{W}}$, so a mixed model estimates the within effect asymptotically in T – which is the cluster size NOT the number of clusters.

- As $\frac{\sigma^2}{T} \times \frac{1}{g_{00}} \to \infty, \quad \hat{\beta}_{\mathrm{MM}} \to \hat{\beta}_{\mathrm{B}}$. Thus the mixed model estimate fails to control for between-subject confounding factors. Note that this does not capture the whole story because $\hat{\beta}_{\mathrm{W}}$ and $\hat{\beta}_{\mathrm{B}}$ are not independent of g_{00}. If $g_{00} = 0$ then $\beta_{\mathrm{B}} = \beta_{\mathrm{W}}$ so that $\beta_{\mathrm{MM}} = \beta_{\mathrm{B}} = \beta_{\mathrm{W}}$

A serious a problem? a simulation

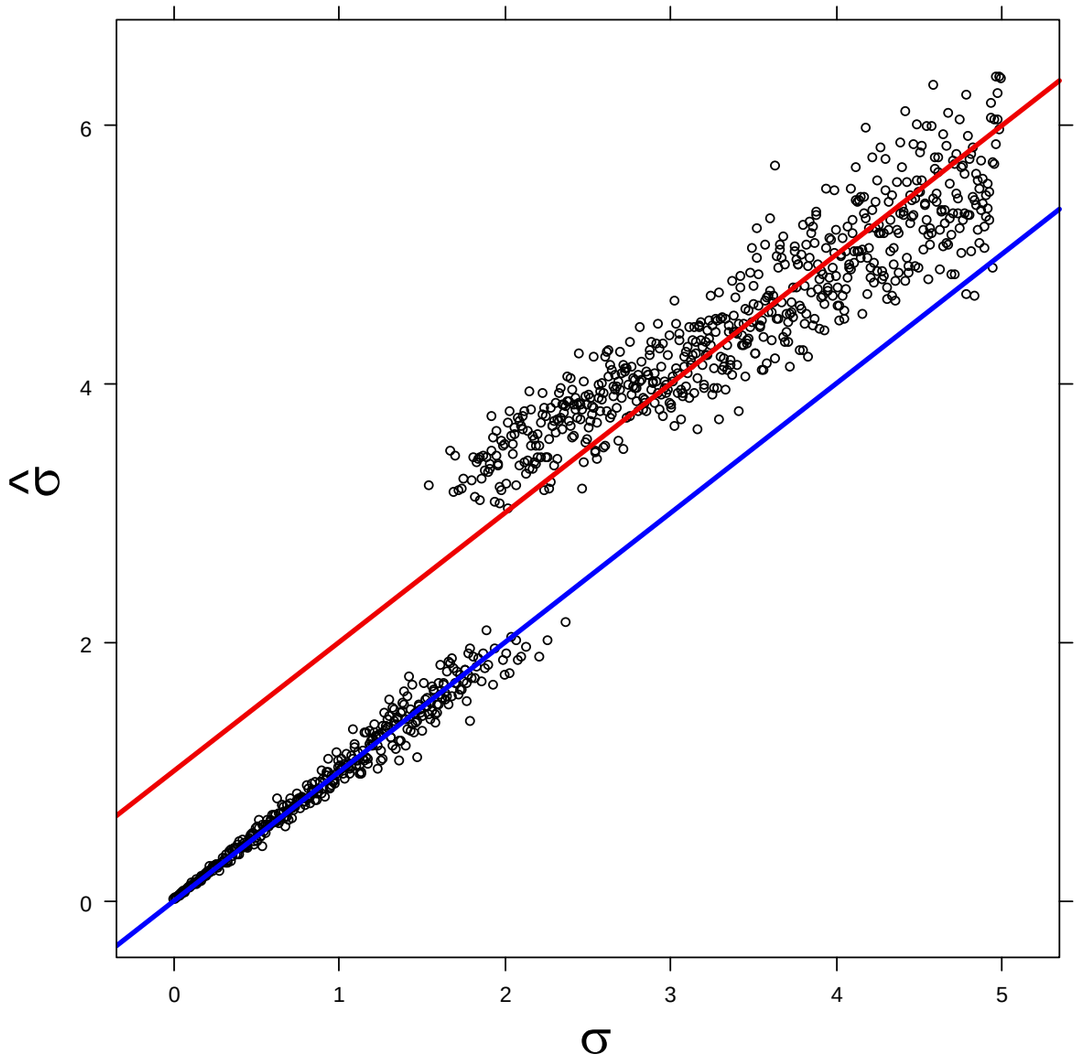


1,000 simulations showing mixed model estimates of slope using the same configuration of Xs with $\beta_W = -1/2$ and $\beta_B = 1$, keeping $g_{00} = 1/2$ and allowing σ^2 to vary from 0.005 to 5



What happened?

As σ gets larger,
the relatively small
value of g_{00} is harder to
identify
and
both sources of
variability
(within-subject and
between-subject)
are attributed to σ.



The blue line is the diagonal $\hat{\sigma} = \sigma$ and the equation of the red line is $\hat{\sigma} = \sigma + 1$.

When $\hat{g}_{00} \approx 0$, the between- subject relationship is treated as if it has very high precision and it dominates in forming the mixed model estimate.

Splitting age into two variables

Since age has a within-subject effect that is inconsistent with its between-subject effect we can split it into two variables:

1. Between-subject 'contextual predictor': e.g. *age.mean* of each subject (or the starting age), and

2. within-subject predictor:

    a. *age* itself or
    b. within-subject residual: $age.resid = age - age.mean$

So we model:

$$\mathrm{E}(y_{it}) = \beta_0 + \beta_{age.mean} age.mean_i + \beta_{age} age_{it} + \delta_{it}$$

or

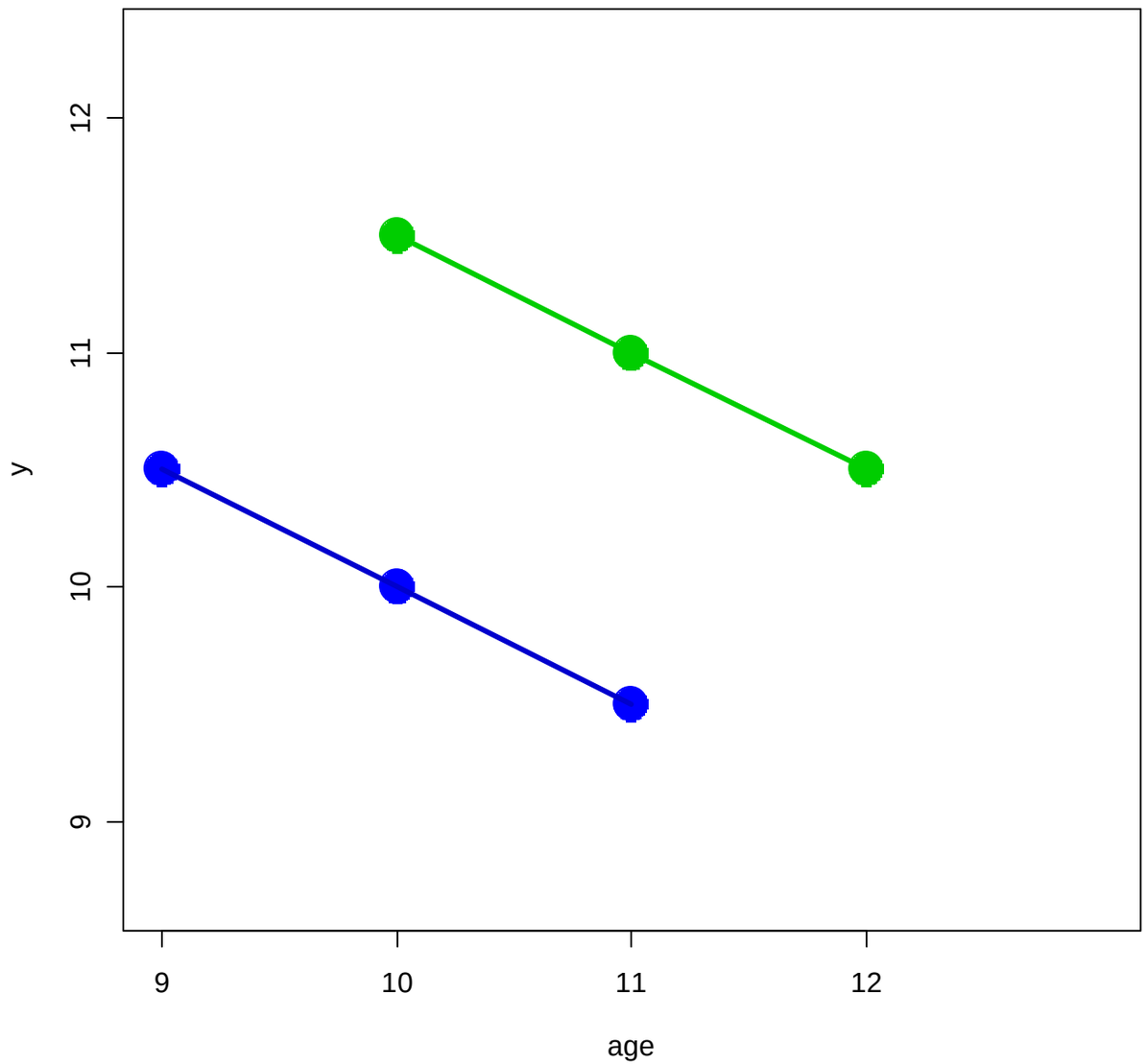
$$\mathrm{E}(y_{it}) = \beta_0^* + \beta_{age.mean}^* age.mean_i + \beta_{age.diff}^* age.diff_{it} + \delta_{it}$$

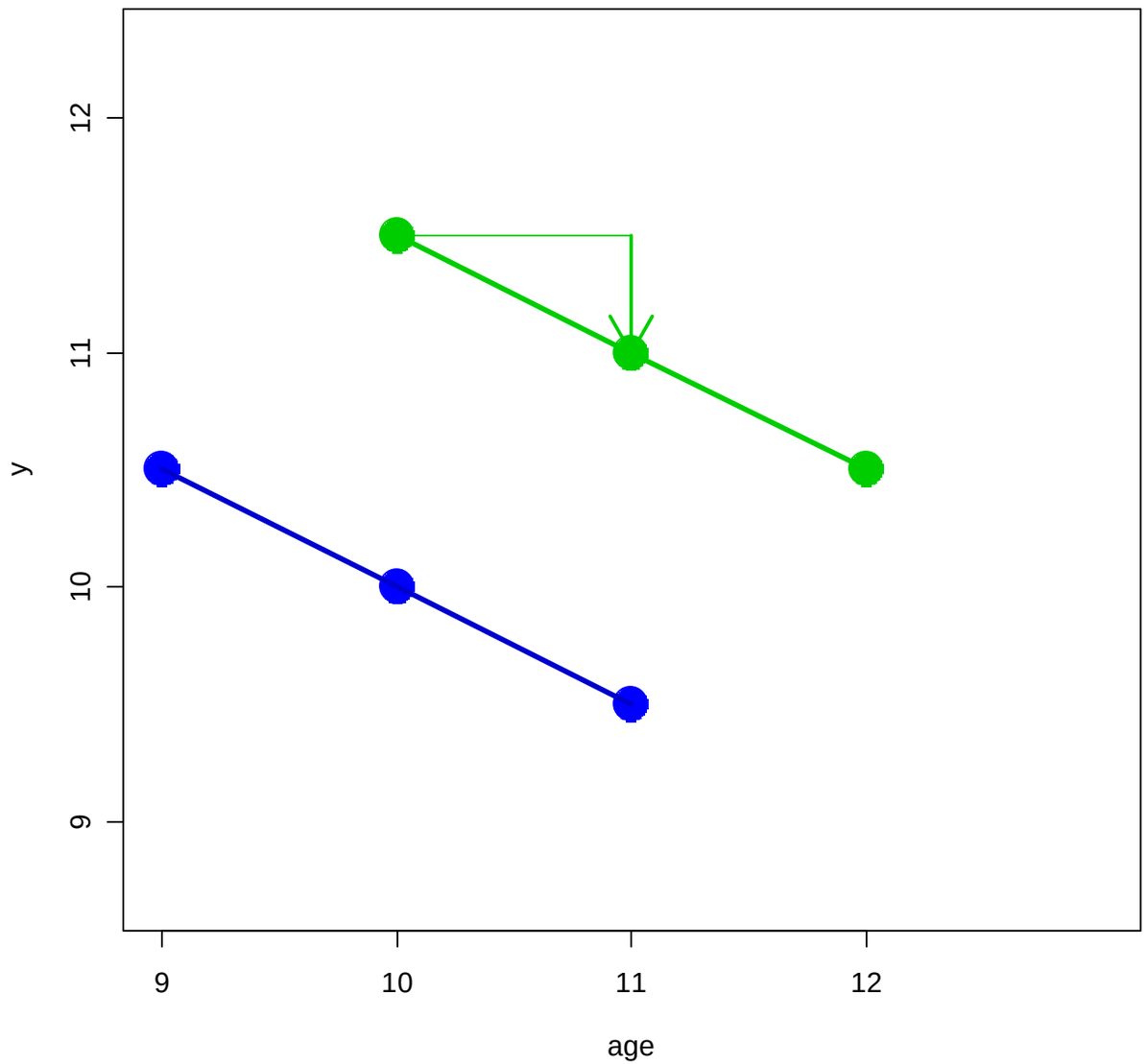
Surprisingly

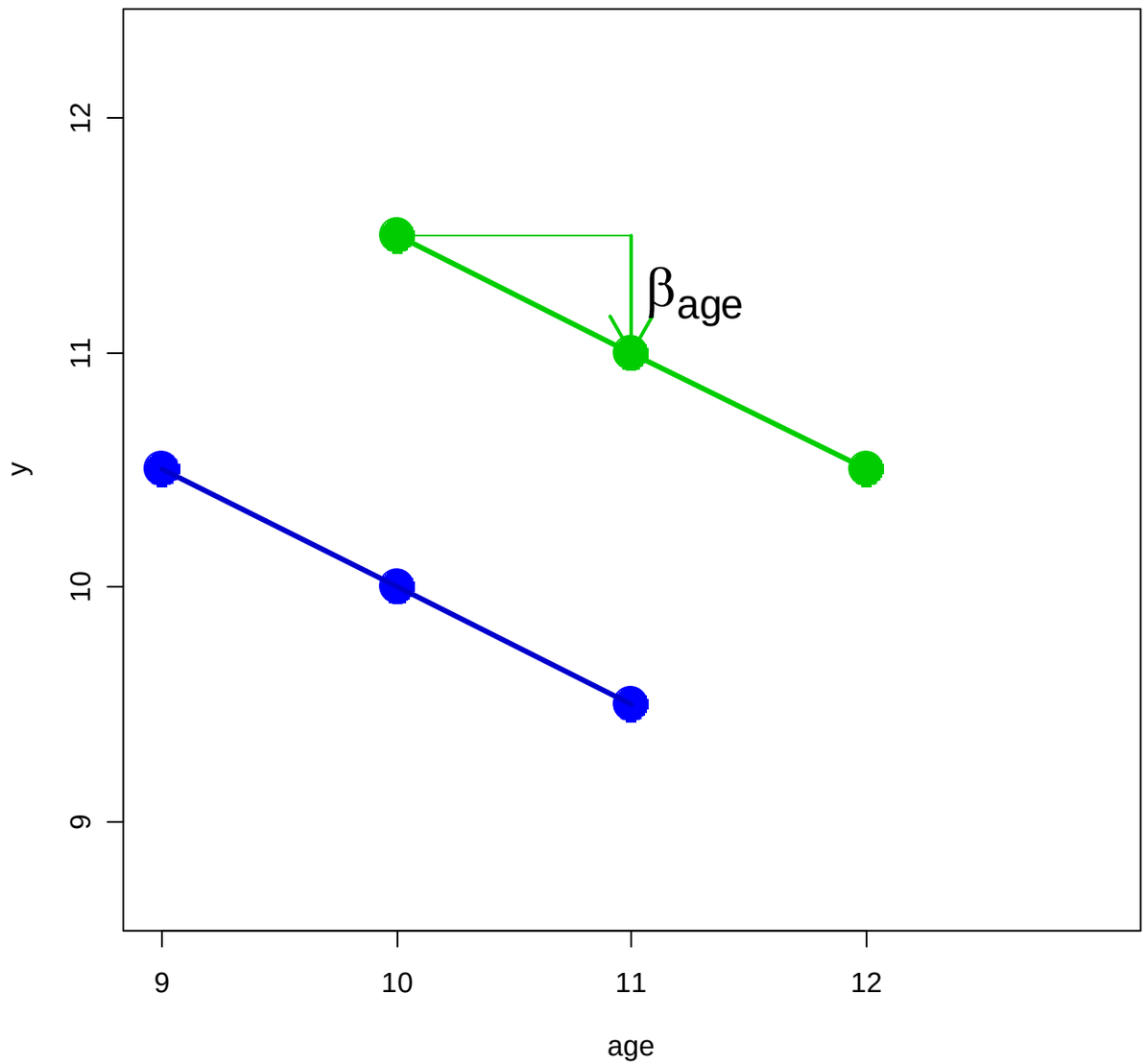
$$\beta_{age.diff}^* = \beta_{age}$$

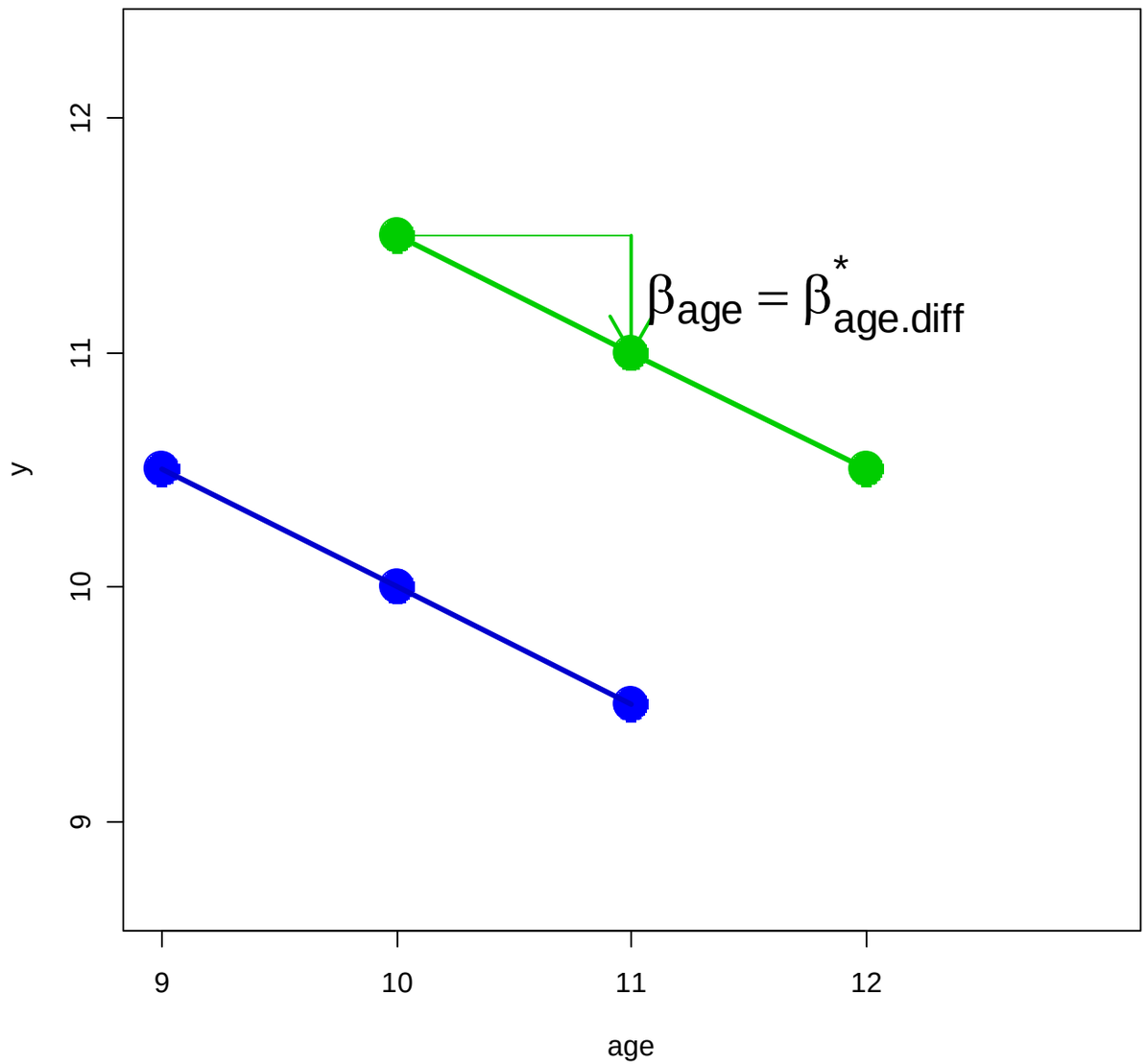
but

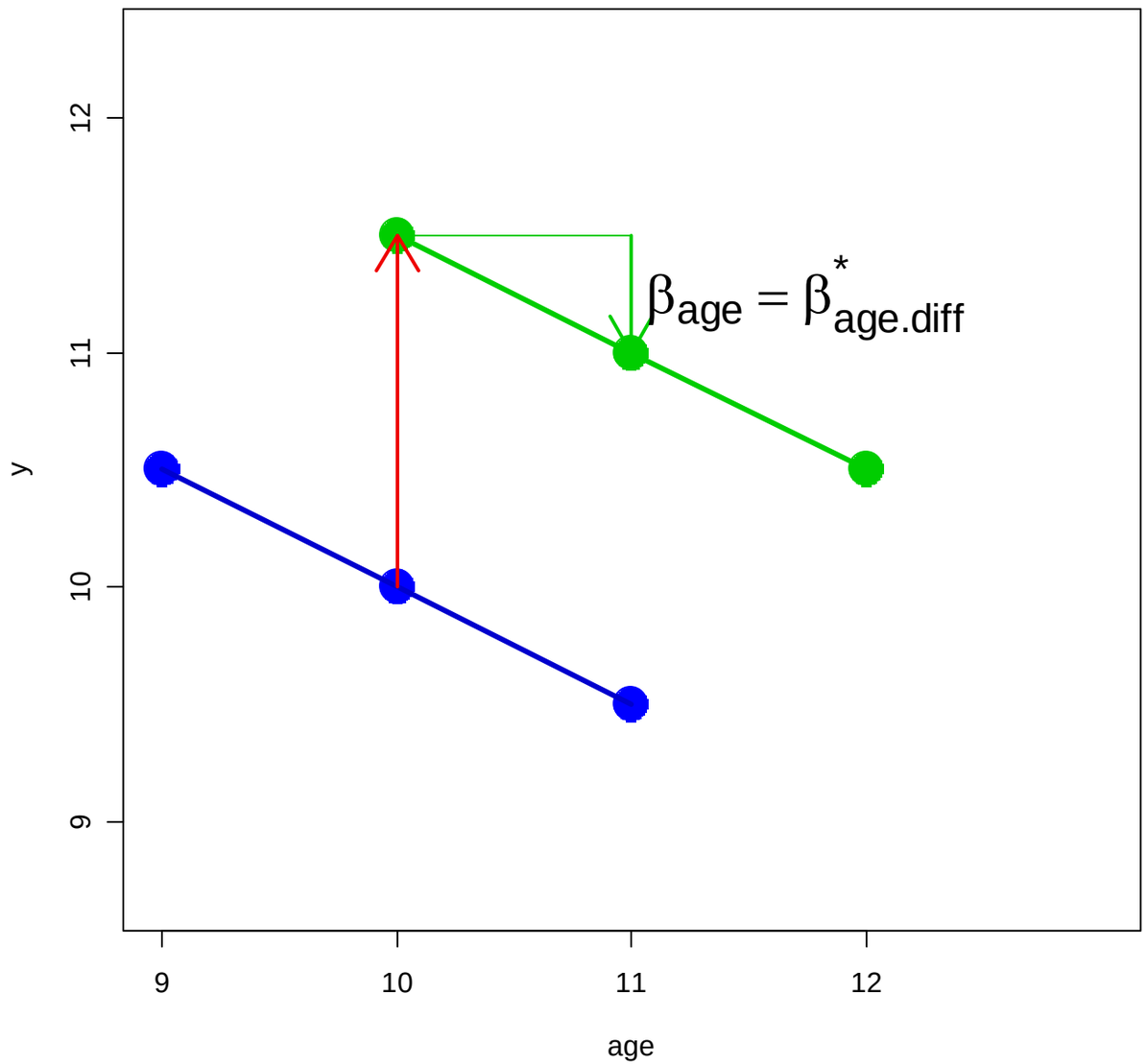
$$\begin{aligned}\beta_{age.mean}^* &= \beta_{age.mean} + \beta_{age.diff}^* \\ &= \beta_{age.mean} + \beta_{age}\end{aligned}$$

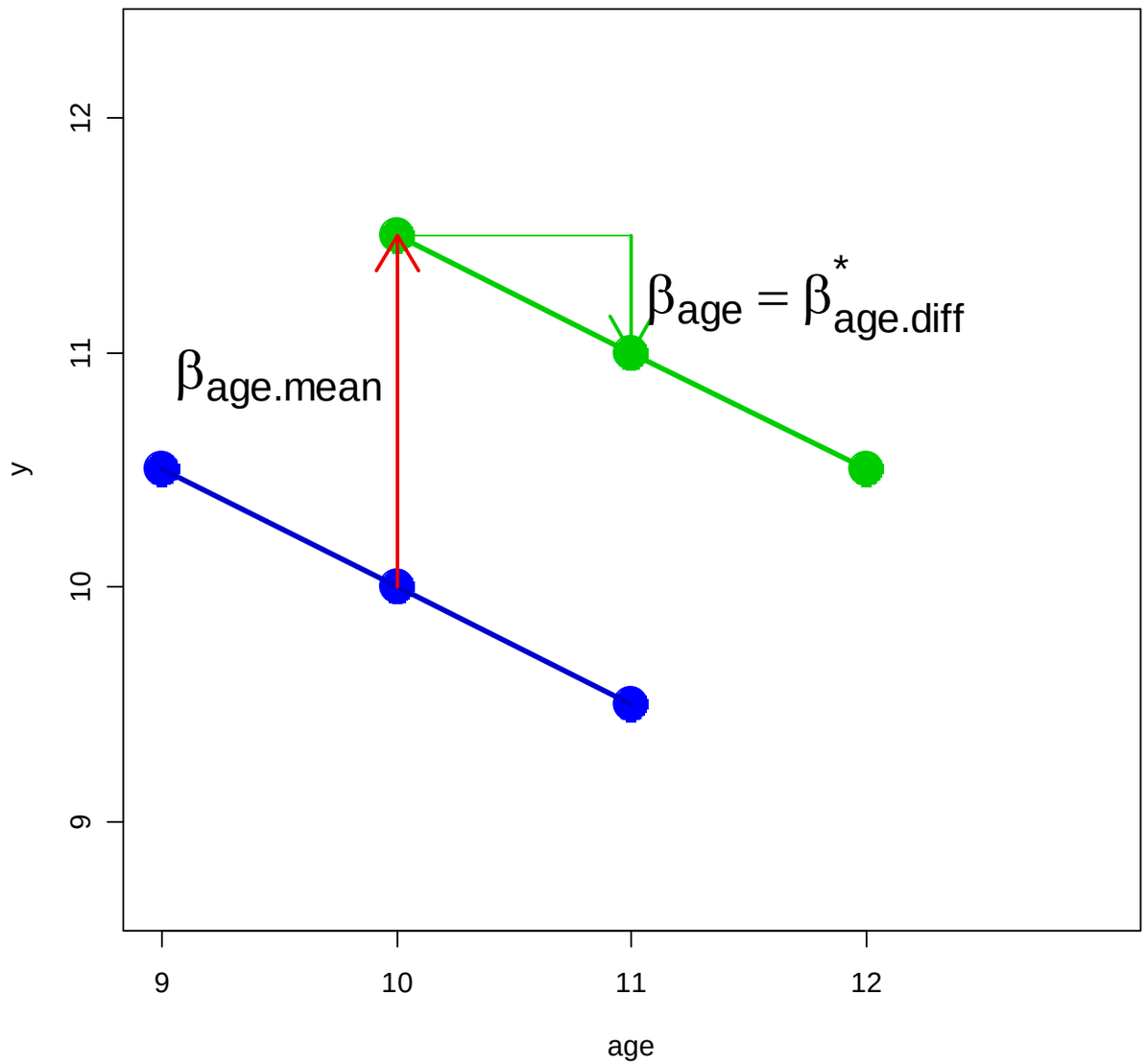

y
12
11
10
9
9
10
11
12
age

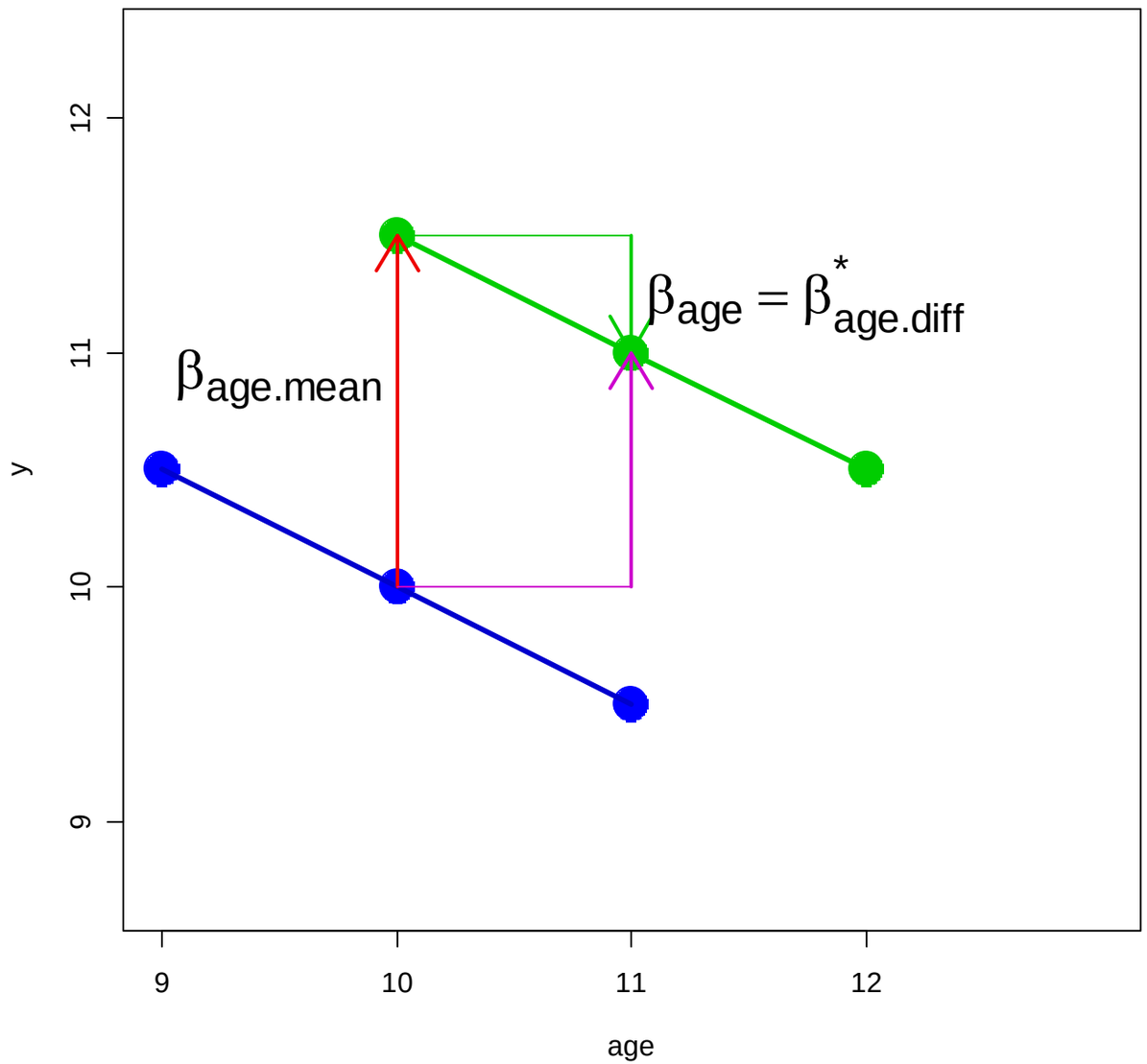

12
11
10
9
y
9
10
11
12
age

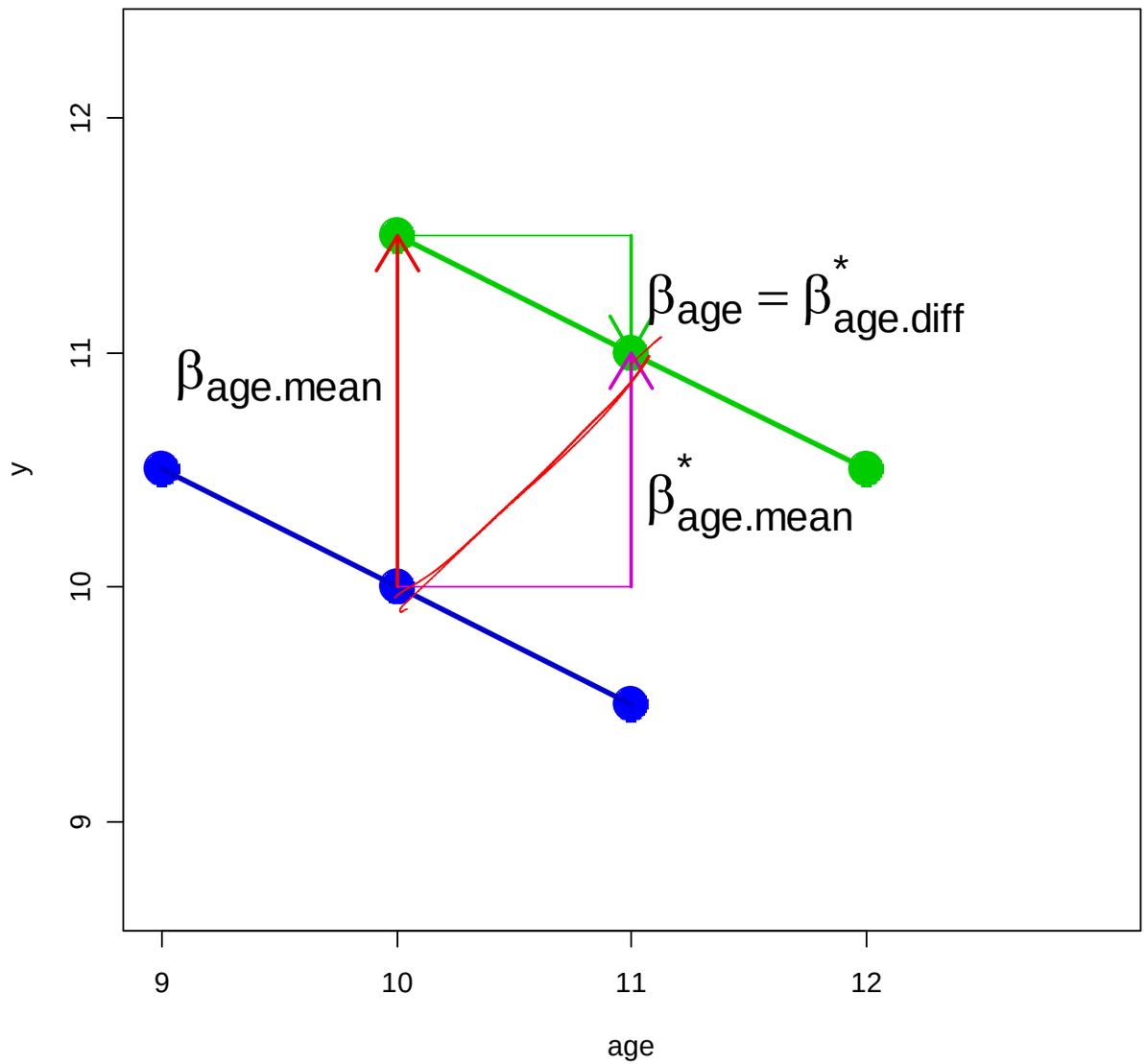

12
11
10
9
y
9
10
11
12
age
β_{age}


12
11
10
9
y
9
10
11
12
age
$\beta_{age} = \beta^*_{age.diff}$


12
11
10
9
y
9
10
11
12
age
$\beta_{age} = \beta^*_{age.diff}$


12
11
10
9
y
9
10
11
12
age
$\beta_{age.mean}$
$\beta_{age} = \beta^{*}_{age.diff}$


12
11
10
9
y
9
10
11
12
age
$\beta_{age.mean}$
$\beta_{age} = \beta^{*}_{age.diff}$



$$\beta^*_{age.mean} = \beta_{age.mean} + \beta_{age}$$

$\beta^*_{age.mean}$ keeps age.diff constant

$\beta_{age.mean}$ keeps age constant

Compositional effect
= Contextual effect
+ Within-subject effect

Using 'lme' with a contextual mean

```
> fit.contextual <- lme(
+           y ~ (age + cvar(age,Subject) ) * Sex,
+           du,
+           random = ~ 1 + age | Subject)
```

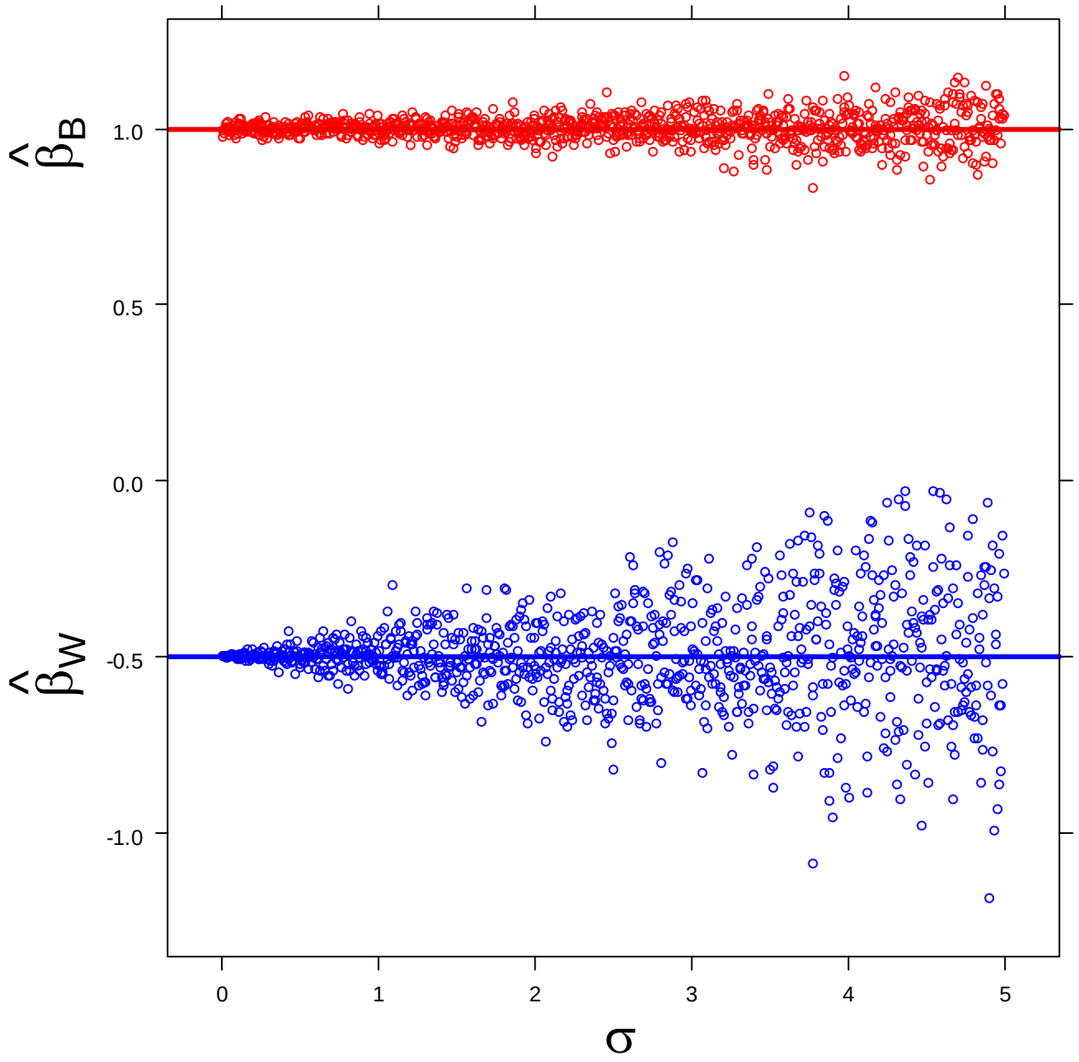
```
> summary(fit.contextual)
Linear mixed-effects model fit by REML
 Data: du
       AIC      BIC    logLik
  296.8729 323.1227 -138.4365

Random effects:
 Formula: ~1 + age | Subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 1.53161007 (Intr)
age         0.03287630 0.024
Residual    0.51263884
```

```
Fixed effects: y ~ (age + cvar(age, Subject)) * Sex
                                Value Std.Error DF    t-value
(Intercept)                 -3.681624 1.6963039 79  -2.170380
age                         -0.493880 0.0343672 79 -14.370670
cvar(age, Subject)           1.628584 0.0695822 23  23.405165
SexFemale                    6.000170 2.5050694 23   2.395211
age:SexFemale                0.060143 0.0538431 79   1.116996
cvar(age, Subject):SexFemale -0.313087 0.1266960 23  -2.471167
                               p-value
(Intercept)                     0.0330
age                             0.0000
cvar(age, Subject)              0.0000
SexFemale                       0.0251
age:SexFemale                   0.2674
cvar(age, Subject):SexFemale    0.0213

. . . . .

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-1.871139553 -0.502221634 -0.006447848  0.552360837  2.428148053

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

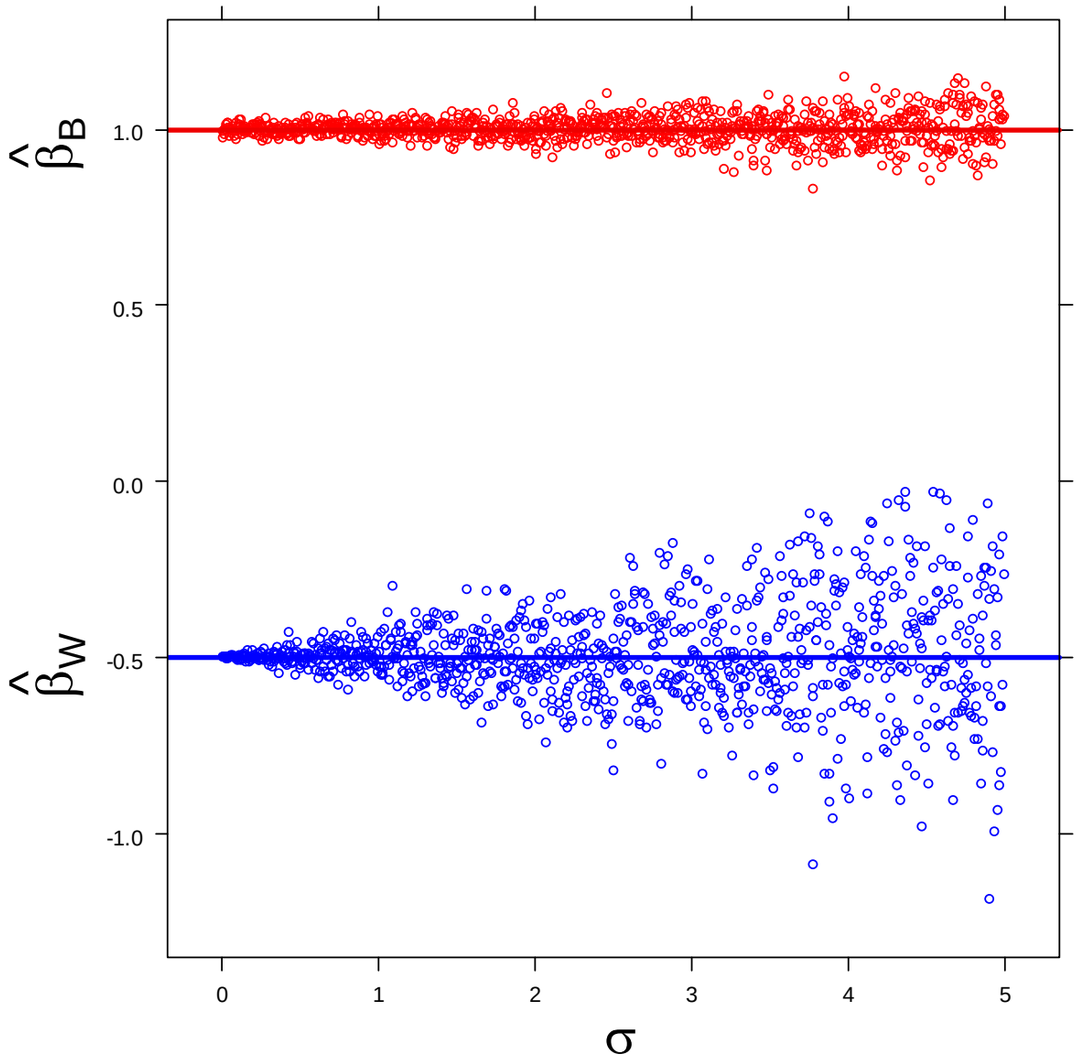
```
> fit.compositional <- lme( y ~ (dvar(age,Subject) +
+        cvar(age,Subject) ) * Sex, du,
+        random = ~ 1 + age | Subject)
> summary(fit.compositional)
Linear mixed-effects model fit by REML
 Data: du
       AIC      BIC    logLik
  296.8729 323.1227 -138.4365

Random effects:
 Formula: ~1 + age | Subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 1.53161006 (Intr)
age         0.03287629 0.024
Residual    0.51263884

Fixed effects: y ~ (dvar(age, Subject) + cvar(age, Subject)) * Sex
                                  Value Std.Error DF    t-value
(Intercept)                  -3.681624 1.6963039 79  -2.170380
dvar(age, Subject)           -0.493880 0.0343672 79 -14.370670
cvar(age, Subject)            1.134704 0.0616092 23  18.417778
SexFemale                     6.000170 2.5050694 23   2.395211
dvar(age, Subject):SexFemale  0.060143 0.0538431 79   1.116996
cvar(age, Subject):SexFemale -0.252945 0.1161225 23  -2.178257
```

```
                               p-value
(Intercept)                     0.0330
dvar(age, Subject)              0.0000
cvar(age, Subject)              0.0000
SexFemale                       0.0251
dvar(age, Subject):SexFemale    0.2674
cvar(age, Subject):SexFemale    0.0399

. . . . .

Standardized Within-Group Residuals:
          Min            Q1           Med            Q3           Max
-1.871139550 -0.502221640 -0.006447847  0.552360836  2.428148063

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

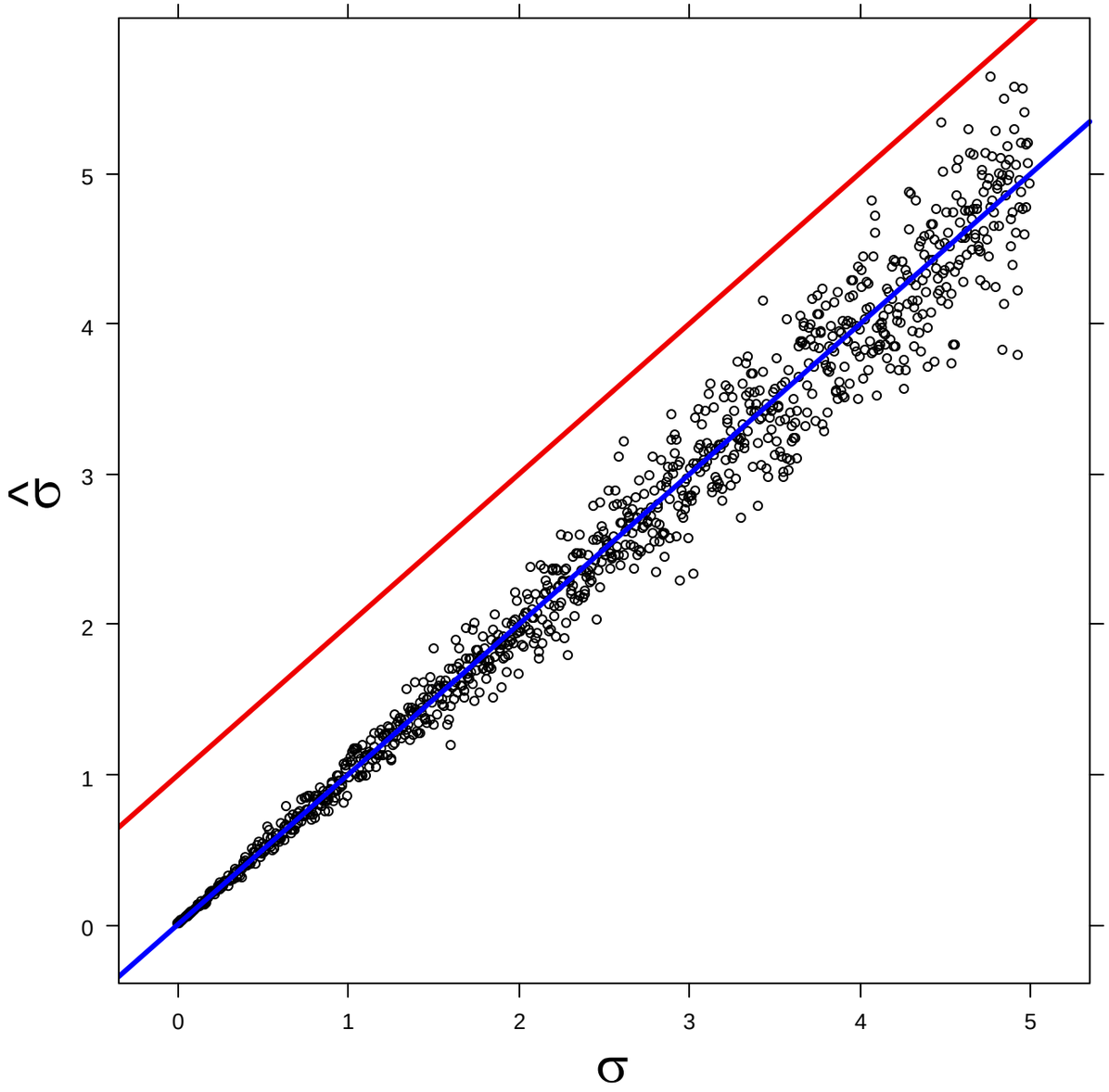
Simulation Revisited



1,000 simulations using the same models as the earlier simulation, i.e. the same configuration of Xs with $\beta_W = -1/2$ and $\beta_B = 1$, keeping $g_{00} = 1/2$ and allowing σ^2 to vary from 0.005 to 5



Here a mixed model is used with mean age by subject and the within-subject residual of age from mean age.



Including the contextual variable gives better estimates of variance components. The estimate of σ does not eventually include g_{00}

Power

The best way to carry out power calculations is to simulate. You end up learning about a lot more than power.

Nevertheless, Stephen Raudenbush and colleagues have a nice graphical package available at Optimal Design Software.

Some links

- There is a very good current bibliography as well as many other resources at the UCLA Academic Technology Services site. Start your visit at http://www.ats.ucla.edu/stat/sas/topics/repeated_measures.htm

- Another important site is the Centre for Multilevel Modeling, currently at the University of Bristol: http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-support/news.shtml

A few books

- Pinheiro, Jose C. and Bates, Douglas M. (2000) *Mixed-Effects Models in S and S-PLUS.* Springer

- Fitzmaurice, Garrett M., Laird, Nan M., Ware, James H. (2004) *Applied Longitudinal Analysis*, Wiley.

- Allison, Paul D. (2005) *Fixed Effects Regression Methods for Longitudinal Data Using SAS*, SAS Institute.

- Littell, Ramon C. et al. (2006) *SAS for Mixed Models* (2nd ed.), SAS Institute.

- Singer, Judith D. and Willett, John B. (2003) *Applied Longitudinal Data Analysis : Modeling Change and Event Occurrence*. Oxford University Press.

Appendix: Reinterpreting weights

The mixed model estimate using a random intercept model can be seen either as a weighted combination of $\hat{\beta}_{\mathrm{B}}$ and $\hat{\beta}_{\mathrm{W}}$ or of $\hat{\beta}_{\mathrm{P}}$ and $\hat{\beta}_{\mathrm{W}}$

$$\hat{\beta}_{\mathrm{MM}} = \left(\frac{\sigma^2/T}{\sigma^2/T + g_{00}} W_{\mathrm{B}} + W_{\mathrm{W}} \right)^{-1} \left(\frac{\sigma^2/T}{\sigma^2/T + g_{00}} W_{\mathrm{B}} \hat{\beta}_{\mathrm{B}} + W_{\mathrm{W}} \hat{\beta}_{\mathrm{W}} \right)$$

$$= \left(\left(\frac{\sigma^2}{T} + g_{00} \right)^{-1} W_{\mathrm{B}} + \left(\frac{\sigma^2}{T} \right)^{-1} W_{\mathrm{W}} \right)^{-1} \left(\left(\frac{\sigma^2}{T} + g_{00} \right)^{-1} W_{\mathrm{B}} \hat{\beta}_{\mathrm{B}} + \left(\frac{\sigma^2}{T} \right)^{-1} W_{\mathrm{W}} \hat{\beta}_{\mathrm{W}} \right)$$

$$= \left((g_{00})^{-1} (W_{\mathrm{B}} + W_{\mathrm{W}}) + \left(\frac{\sigma^2}{T} \right)^{-1} W_{\mathrm{W}} \right)^{-1} \left((g_{00})^{-1} (W_{\mathrm{B}} + W_{\mathrm{W}}) \hat{\beta}_{\mathrm{P}} + \left(\frac{\sigma^2}{T} \right)^{-1} W_{\mathrm{W}} \hat{\beta}_{\mathrm{W}} \right)$$