

How to understand and apply mixed-effect models

Deming Mi

Department of Biostatistics

4/27/2007

José C. Pinheiro
Douglas M. Bates

Mixed-Effects Models in S and S-PLUS



 Springer

Copyrighted Material

Longitudinal data

Orthodontic growth data, Example from Pinheiro and Bates (2000)

Investigators at the University of North Carolina Dental School followed the growth of 27 children (16 males, 11 females) from age 8 until age 14. Every two years they measured the distance between the pituitary and the pterygomaxillary fissure, two points that are easily identified on x-ray exposures of the side of the head.

The data

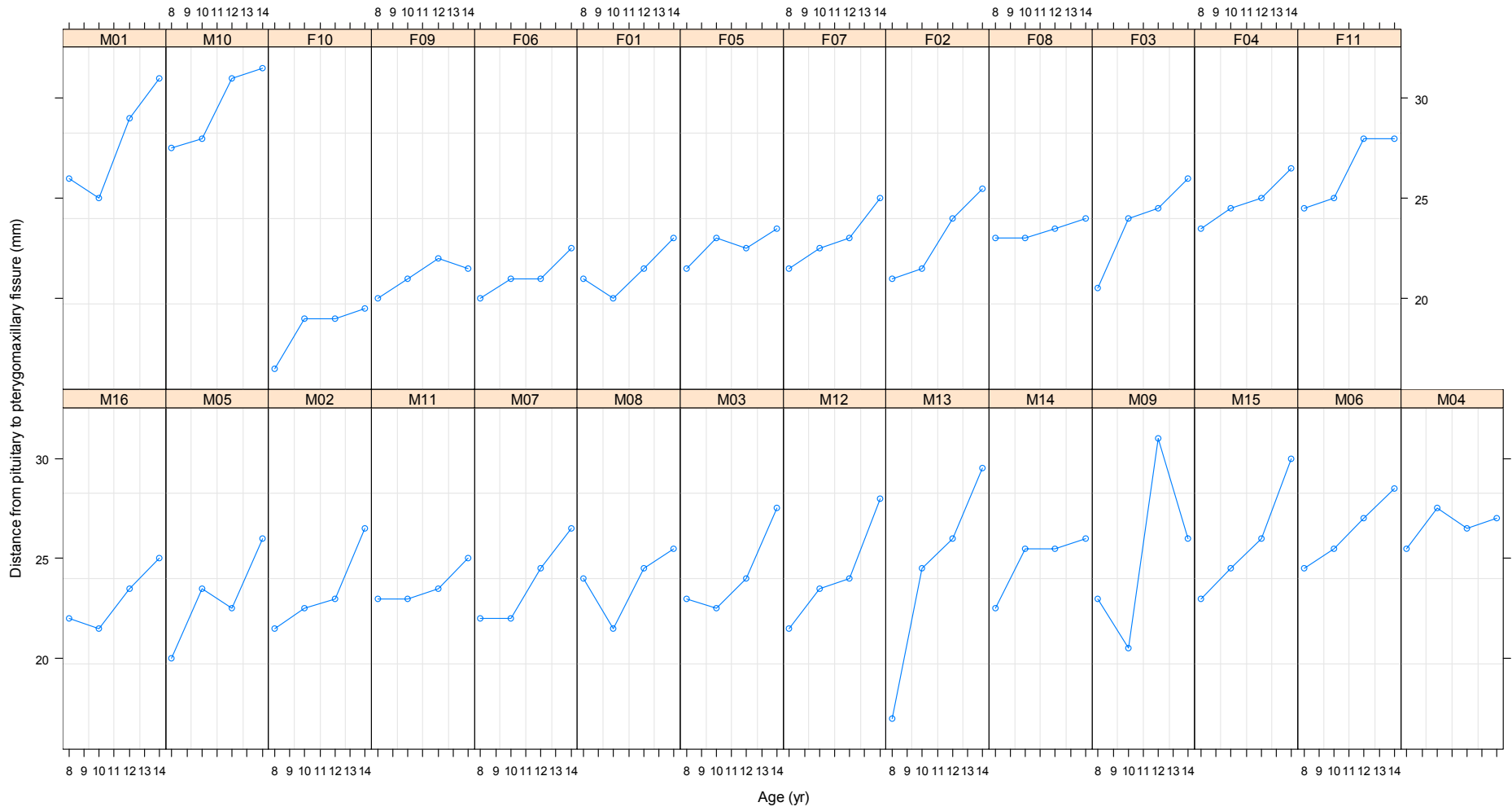
```
library(nlme)
```

```
Data(Orthodont)
```

```
Grouped Data: distance ~ age | Subject
```

```
  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
.....

105  24.5   8    F11  Female
106  25.0  10    F11  Female
107  28.0  12    F11  Female
108  28.0  14    F11  Female
```



Orthodontic growth data

Questions

- Is there an age effect on growth?
- Is there a gender difference?
- Is growth different in both sexes (Is there an interaction)?
- Is an ordinary linear regression model adequate?

The simple linear regression model

M subjects and n_i measurements per subject

$$y_{ij} = \beta_1 + \beta_2 x_{ij} + \varepsilon_{ij}$$
$$i = 1, \dots, M \quad (M = 27)$$
$$j = 1, \dots, n_i \quad (n_i = 4)$$
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

R call:

```
fm2Orth.lm <- lm(distance~Sex*I(age-11), Orthodont)
```

Data centering

- All measurements were made between age 8 and 14.
- Intercept is the extrapolated response at age 0.
- There will be a high negative correlation between the intercept and slope in the linear model.
- Such a high correlation in a regression model will cause multicollinearity problem
- A guard against this is to use centered data

Result

Call:

```
lm(formula = distance ~ Sex * I(age - 11), data = Orthodont)
```

Residuals:

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	24.9688	0.2821	88.504	< 2e-16	***
SexFemale	-2.3210	0.4420	-5.251	8.05e-07	***
I(age - 11)	0.7844	0.1262	6.217	1.07e-08	***
SexFemale:I(age - 11)	-0.3048	0.1977	-1.542	0.126	

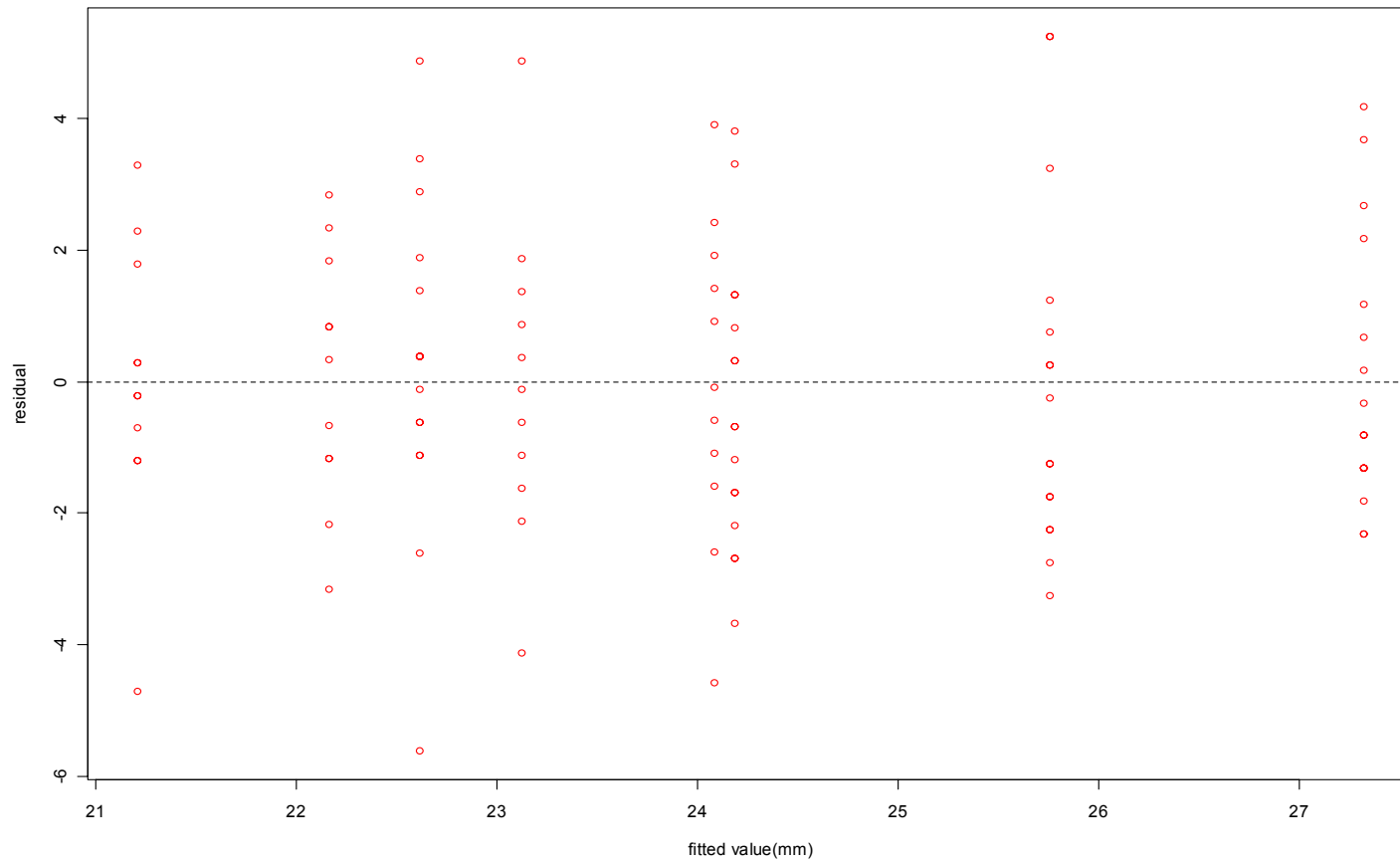
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.257 on 104 degrees of freedom

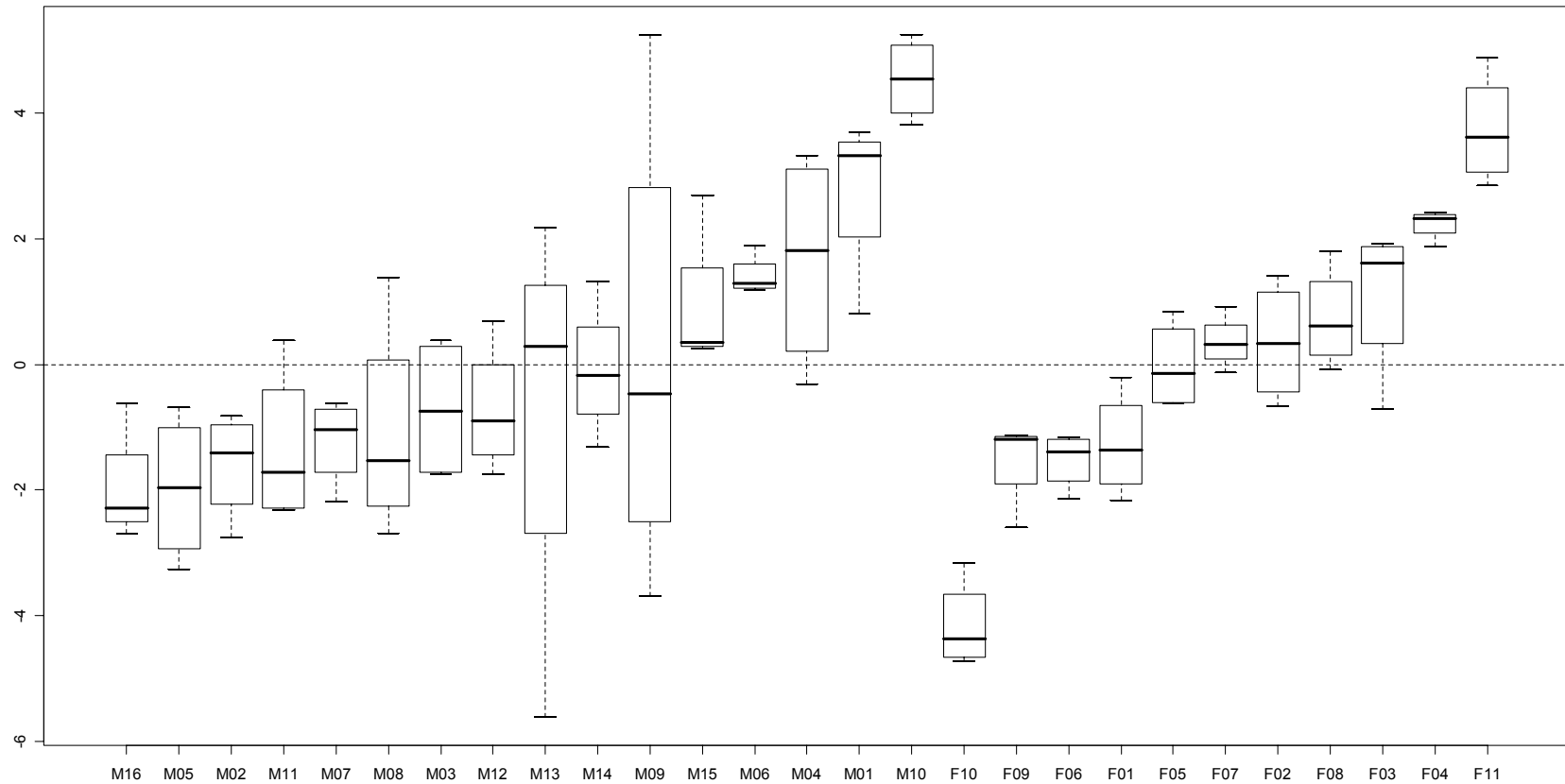
Multiple R-Squared: 0.4227, Adjusted R-squared: 0.4061

F-statistic: 25.39 on 3 and 104 DF, p-value: 2.108e-12

Residual plot of simple linear regression model



Residual plots by subject (simple linear regression model)

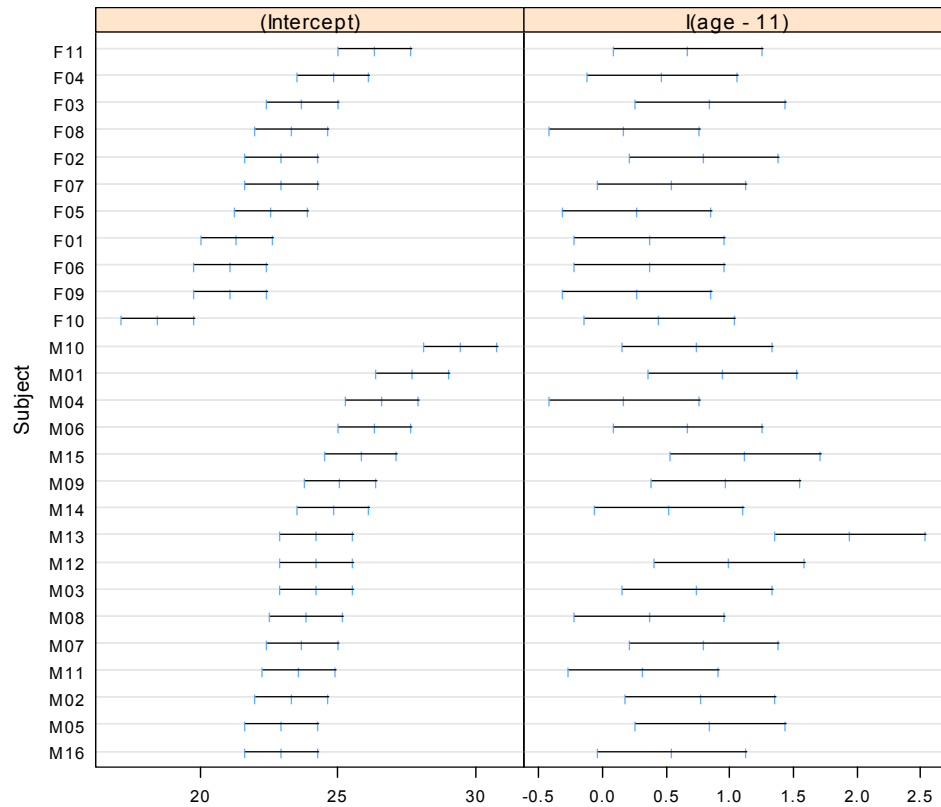


Fit a simple linear regression model per subject
with 'age' as the only covariate

R call:

```
fm2Orth.lis <- lmList(distance ~ I(age-11) | Subject, Orthodont)
```

95% confidence intervals on intercept and slope for each subject



Comments

- Residuals corresponding to the same subject tend to have the same sign.
- There is a significant subject-to-subject variability for intercept.
- Need to incorporate a “subject effect” in the model to account for between-subject variability.

What if using dummy variables for “subject effect”?

- Your inference about “subject effect” will not be applicable to the whole population
- You will need $(M - 1)$ dummy variables for the M subjects

Fit a single-level mixed effect model

$$y_{ij} = \beta_1 + \beta_2 x_{ij} + b_{i,1} + b_{i,2} x_{ij} + \varepsilon_{ij}$$

$$i = 1, \dots, M, \quad j = 1, \dots, n_i$$

$$\begin{bmatrix} b_{i,1} \\ b_{i,2} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{\Psi}), \quad \mathbf{\Psi} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

y_{ij} is the j th observation in the i th group of data and x_{ij} is the corresponding value of the covariate.

Between - subject variability will be accounted for by random effects $b_{i,1}$ and $b_{i,2}$

Parameters to estimate : $\beta_1, \beta_2, \sigma_1, \sigma_2, \sigma_{12}, \sigma$

Comment about Ψ

- Variance-covariance matrix for the random effects
- Customized variance-covariance matrix

pdSymm(default)	general positive-definite matrix
pdBlocked	block-diagonal
pdCompSymm	compound-symmetry structure
pdDiag	diagonal
pdIdent	multiple of an identity

General symmetric positive-definite var-cov matrix (default)

$$\begin{bmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

Diagonal var-cov matrix

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

Multiple of an identity var-cov matrix

$$\begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

Block-diagonal var-cov matrix

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix}$$

Compound-symmetry var-cov matrix

$$\begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{bmatrix}$$

Model matrix for subject M01

$$\begin{bmatrix} y_{M01,1} \\ y_{M01,2} \\ y_{M01,3} \\ y_{M01,4} \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} b_{M01,1} \\ b_{M01,2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{M01,1} \\ \varepsilon_{M01,2} \\ \varepsilon_{M01,3} \\ \varepsilon_{M01,4} \end{bmatrix}$$

Note: centered age are -3, -1, 1, 3

Fit a single-level mixed effect model with random effects associated with both intercept and slope

R Call

```
fm1Orth.lme <- lme(distance~l(age-11), data=Orthodont, random=~l(age-11)|Subject)
```

Note: at this point, we omit the “sex” effect for simplicity.

Result

Linear mixed-effects model fit by REML

Data: Orthodont

Log-restricted-likelihood: -221.3183

Fixed: distance ~ I(age - 11)

(Intercept) I(age - 11)

24.0231481 0.6601852

Random effects:

Formula: ~I(age - 11) | Subject

Structure: General positive-definite, Log-Cholesky parametrization

StdDev Corr

(Intercept) 2.1343289 (Intr)

I(age - 11) 0.2264278 0.503

Residual 1.3100402

Number of Observations: 108

Number of Groups: 27

$$\beta_1 = 24.02, \beta_2 = 0.66$$

$$\sigma_1 = 2.13, \sigma_2 = 0.22, \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} = 0.5, \sigma = 1.31$$

Check the confidence intervals of the fit

```
> intervals(fm1Orth.lme)
```

```
Approximate 95% confidence intervals
```

```
Fixed effects:
```

	lower	est.	upper
(Intercept)	23.1680973	24.0231481	24.8781990
I(age - 11)	0.5183866	0.6601852	0.8019837

```
attr(,"label")
```

```
[1] "Fixed effects:"
```

```
Random Effects:
```

```
Level: Subject
```

	lower	est.	upper
sd((Intercept))	1.5844025	2.1343289	2.8751278
sd(I(age - 11))	0.1025110	0.2264278	0.5001372
cor((Intercept), I(age - 11))	-0.2672333	0.5026262	0.8808415

```
Within-group standard error:
```

lower	est.	upper
1.084830	1.310040	1.582004

Confidence interval (-0.27, 0.88) suggests a diagonal variance-covariance structure for random effects

Fit a single-level mixed effect model with diagonal var-cov structure for random effects

R call:

```
fm1Orth.diagonal.lme <- lme(distance~I(age-11), data=Orthodont,  
random=list(Subject=pdDiag(~I(age-11))))
```

Result

```
> fm1Orth.diagonal.lme
```

```
Linear mixed-effects model fit by REML
```

```
Data: Orthodont
```

```
Log-restricted-likelihood: -222.4924
```

```
Fixed: distance ~ I(age - 11)
```

```
(Intercept) I(age - 11)
```

```
24.0231481    0.6601852
```

```
Random effects:
```

```
Formula: ~I(age - 11) | Subject
```

```
Structure: Diagonal
```

```
(Intercept) I(age - 11) Residual
```

```
StdDev:      2.134331    0.2264279 1.310039
```

```
Number of Observations: 108
```

```
Number of Groups: 27
```

$$\beta_1 = 24.02, \beta_2 = 0.66$$

$$\sigma_1 = 2.13, \sigma_2 = 0.22, \sigma = 1.31$$

How to compare two mixed effect models

$$AIC = -2\log Lik + 2n_{par}$$

$$BIC = -2\log Lik + n_{par} \log(N)$$

"Smaller is better" based on this definition of *AIC* and *BIC*

(different statistical computing software may have different definition for *AIC* and *BIC*)

Likelihood - ratio test : $2\log(L_2/L_1) = 2[\log(L_2) - \log(L_1)] \sim \chi^2(k_2 - k_1)$

(k_1 and k_2 are number of parameters in the two models to be compared)

`anova(modelA, modelB)`

Note : model A and modelB have to have the same fixed effects structure

Compare models with different var-cov structures for random effects

```
fm1Orth.diagonal.lme <- lme(distance~I(age-11), data=Orthodont,  
random=list(Subject=pdDiag(~I(age-11))))
```

```
fm1Orth.lme <- lme(distance~I(age-11), data=Orthodont, random=~I(age-  
11)|Subject)
```

```
> anova(fm1Orth.diagonal.lme, fm1Orth.lme)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
fm1Orth.diagonal.lme	1	5	454.9848	468.3020	-222.4924			
fm1Orth.lme	2	6	454.6367	470.6173	-221.3183	1 vs 2	2.348112	0.1254

There is no significant difference between the two models. Keep the model with general positive-definite matrix (fm1Orth.lme) since this is a more general model with less assumption.

Fit a single-level mixed effect model with random effects associated with slope only

R Call

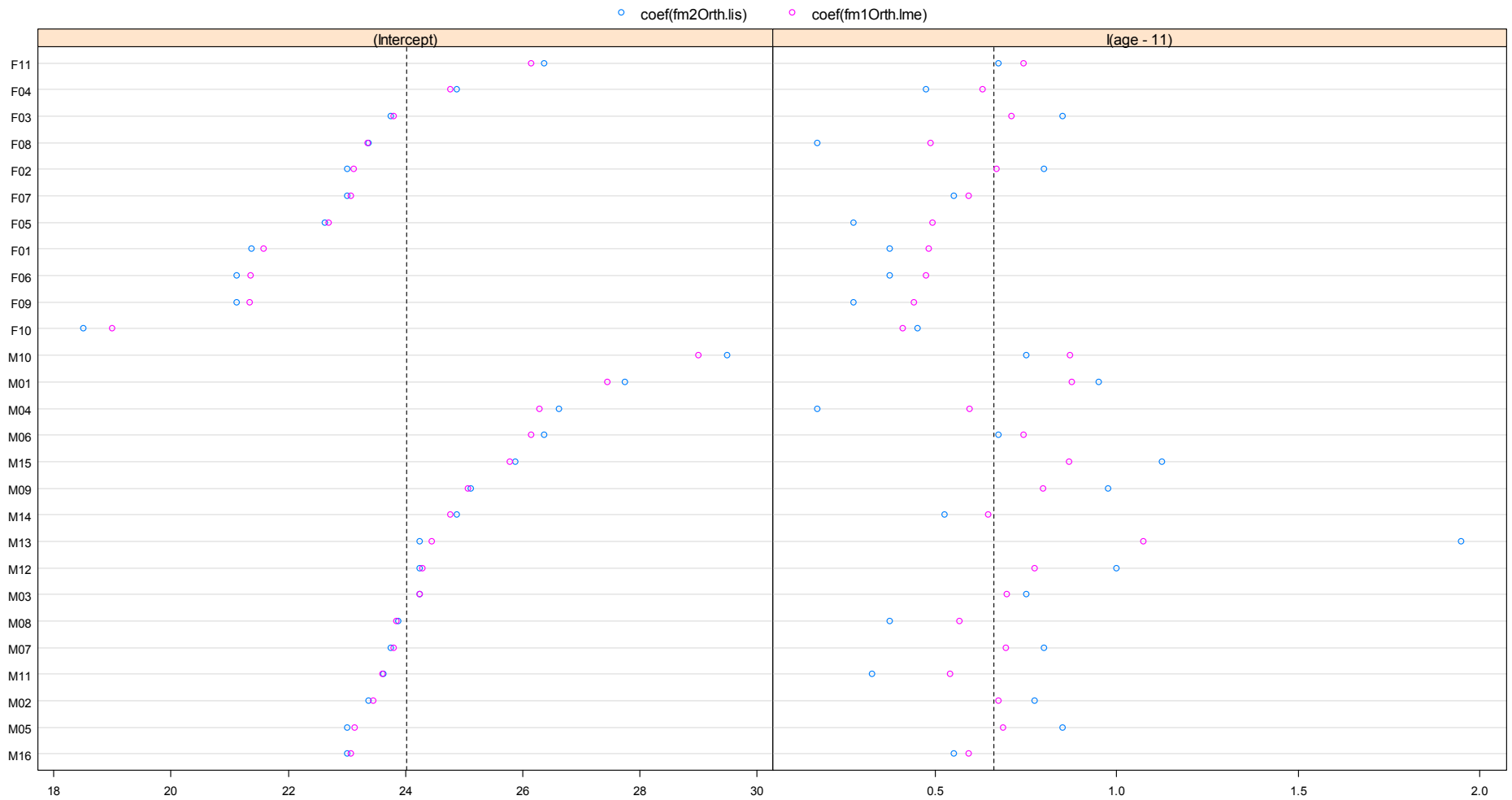
```
fm1Orth.slope.lme <- lme(distance~l(age-11), data=Orthodont, random=~l(age-11)-1|Subject)
```

Compare models with random effect associated with slope only against the model with random effect associated with both intercept and slope

```
> anova(fm1Orth.slope.lme, fm1Orth.lme)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
fm1Orth.slope.lme	1	4	517.1695	527.8233	-254.5848			
fm1Orth.lme	2	6	454.6367	470.6173	-221.3183	1 vs 2	66.53281	<.0001

Based on p-value and AIC or BIC criteria, model with random effects associated with both intercept and slope (fm1Orth.lme) is much better than model with random effects associated with slope only.



Fixed model: $y_{ij} = \beta_1 + \beta_2 x_{ij} + \varepsilon_{ij}$

Mixed model: $y_{ij} = (\beta_1 + b_{i,1}) + (\beta_2 + b_{i,2})x_{ij} + \varepsilon_{ij}$

$i = 1, \dots, M, \quad j = 1, \dots, n_i$

Consider a possible gender difference

R call:

```
fm2Orth.lme <-
```

```
lme(distance~Sex*I(age-11), data=Orthodont, random=~I(age-11)|Subject)
```

Result

Linear mixed-effects model fit by REML

Data: Orthodont

	AIC	BIC	logLik
	448.5817	469.7368	-216.2908

Random effects:

Formula: ~I(age - 11) | Subject

Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	1.8303268	(Intr)
I(age - 11)	0.1803454	0.206
Residual	1.3100397	

Fixed effects: distance ~ Sex * I(age - 11)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	24.968750	0.4860007	79	51.37595	0.0000
SexFemale	-2.321023	0.7614168	25	-3.04829	0.0054
I(age - 11)	0.784375	0.0859995	79	9.12069	0.0000
SexFemale:I(age - 11)	-0.304830	0.1347353	79	-2.26243	0.0264

Correlation:

	(Intr)	SexFml	I(-11)
SexFemale	-0.638		
I(age - 11)	0.102	-0.065	
SexFemale:I(age - 11)	-0.065	0.102	-0.638

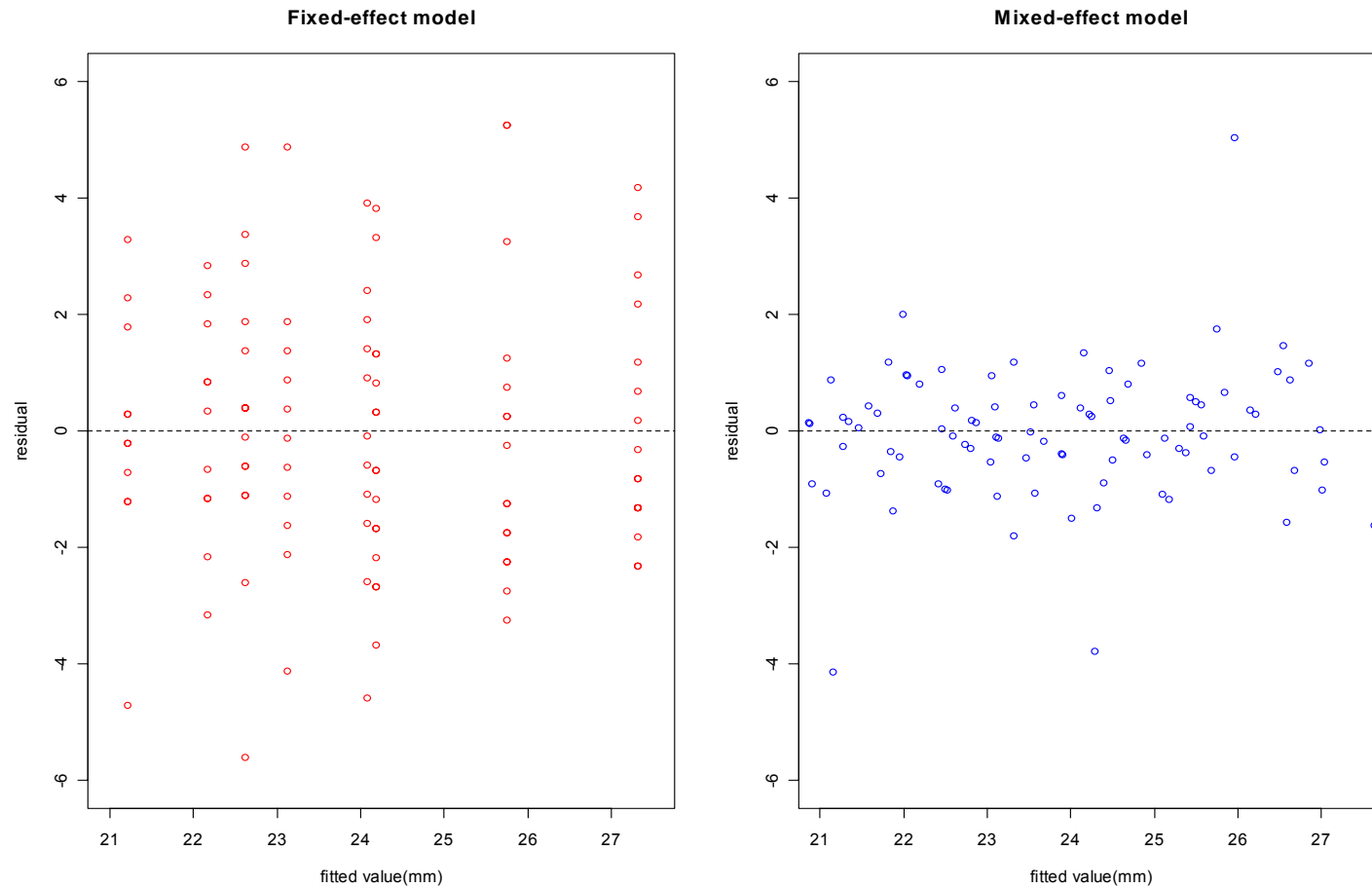
Standardized Within-Group Residuals:

Min	Q1	Med	Q3	Max
-3.168078325	-0.385939104	0.007103934	0.445154637	3.849463326

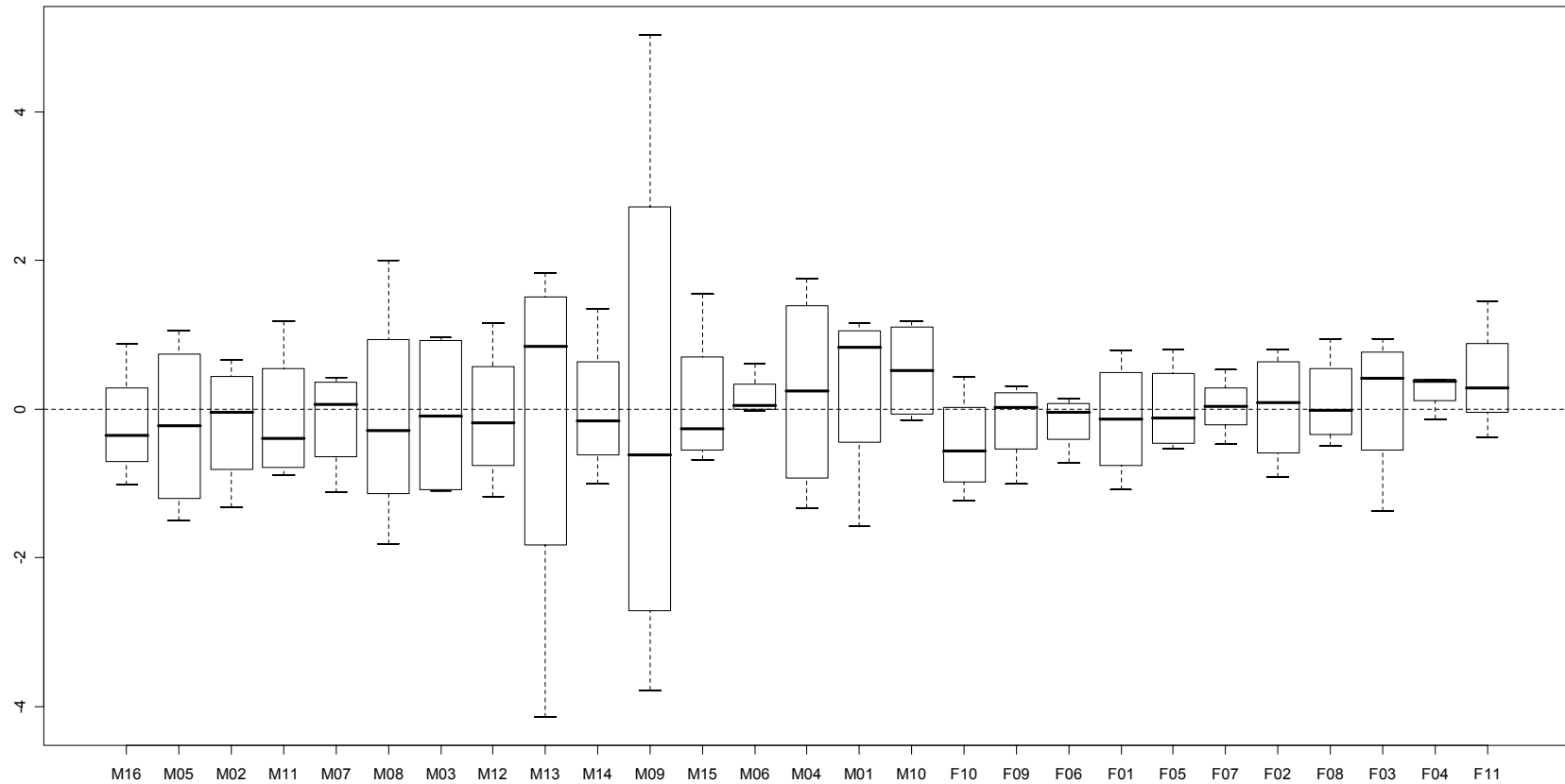
Number of Observations: 108

Number of Groups: 27

Compare the residual plots from fixed-effect model and mixed-effect model



Residual plots by subject (mixed-effect model)



Compare mixed effect model and fixed effect model

```
fm2Orth.lme <- lme(distance~Sex*I(age-11), data=Orthodont, random=~I(age-11)|Subject)
```

```
fm2Orth.lm <- lm(distance~Sex*I(age-11), Orthodont)
```

```
> anova(fm2Orth.lme, fm2Orth.lm)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
fm2Orth.lme	1	8	448.5817	469.7368	-216.2908			
fm2Orth.lm	2	5	493.5591	506.7811	-241.7796	1 vs 2	50.97746	<.0001

Take a look at the fixed-effect and random-effect estimators

```
> fixed.effects(fm2Orth.lme)
```

(Intercept)	SexFemale	I(age - 11)	SexFemale:I(age - 11)
24.9687500	-2.3210227	0.7843750	-0.3048295

```
> coef(fm2Orth.lme)
```

	(Intercept)	SexFemale	I(age - 11)	SexFemale:I(age - 11)
M16	23.21038	-2.321023	0.6957279	-0.3048295
M05	23.23019	-2.321023	0.7759009	-0.3048295
M02	23.55722	-2.321023	0.7608123	-0.3048295
M11	23.74881	-2.321023	0.6438560	-0.3048295

```
> random.effects(fm2Orth.lme)
```

	(Intercept)	I(age - 11)
M16	-1.75837369	-0.088647071
M05	-1.73855508	-0.008474092
M02	-1.41153061	-0.023562683
M11	-1.21993911	-0.140519049
M07	-1.07789993	-0.011926948
M08	-0.99531659	-0.123853783
M03	-0.63856420	-0.018682906
M12	-0.62204869	0.048127909
M13	-0.55928975	0.302009008
M14	-0.10012962	-0.070554884
M09	0.15091771	0.053007686
M15	0.82478527	0.103003482
M06	1.23769618	-0.010649781
M04	1.42598457	-0.140968310
M01	2.47312002	0.081009179
M10	4.00914352	0.050682243
F10	-3.67384214	-0.062697295
F09	-1.36154913	-0.074782291
F06	-1.35494293	-0.048057965
F01	-1.13362351	-0.044754862
F05	-0.03363264	-0.054963676
F07	0.31651355	0.023482874
F02	0.33302906	0.090293690
F08	0.62371940	-0.071778696
F03	1.00029041	0.113565160
F04	1.97145451	0.028212897
F11	3.31258341	0.101480164

Revisit the mixed effect model with single level grouping

The n_i - dimensional response vector y_i for the i th group

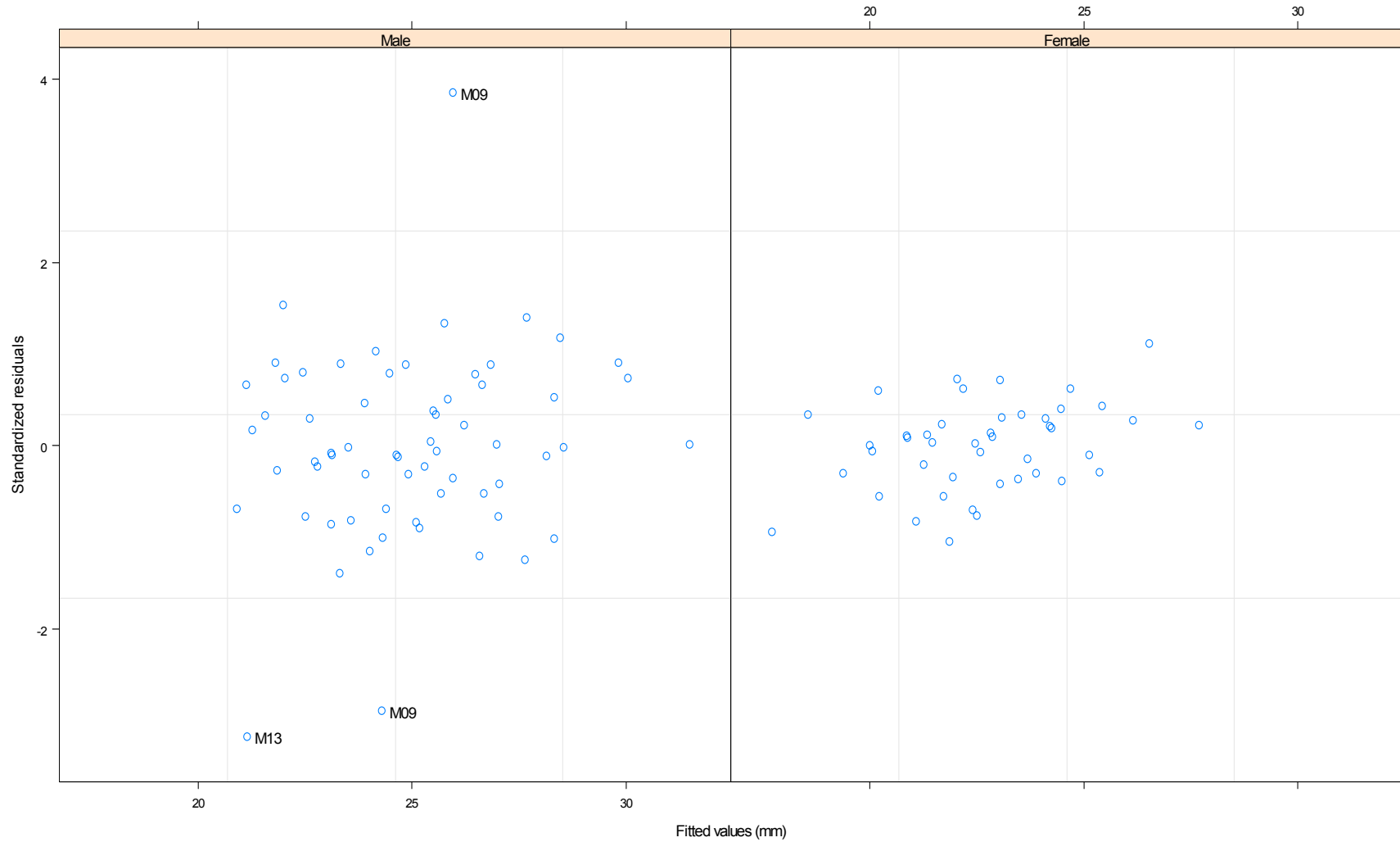
$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \dots, M$$

$$\mathbf{b}_i \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\varepsilon}_i \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

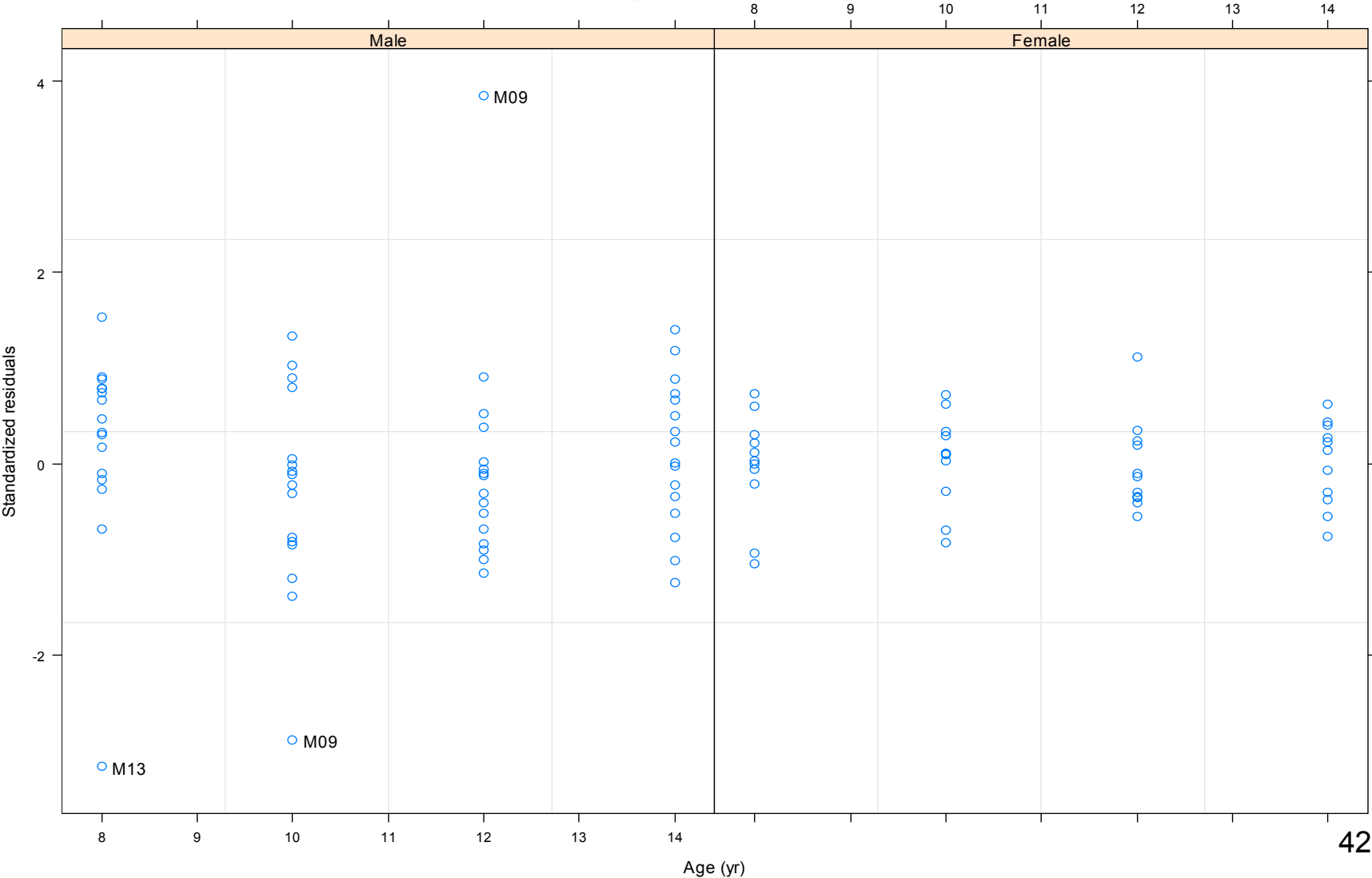
The within - group errors $\boldsymbol{\varepsilon}_i$ are assumed to :

- 1) be independent for different groups of different i ;
- 2) be independent for repeated measurements within the same groups;
- 3) be independent of random effects \mathbf{b}_i ;
- 4) have constant variances for both different group and repeated measurements within the same group, that is, homoscedastic.

Scatter plots of standardized residuals versus fitted values by gender (homoscedastic model)



Scatter plots of standardized residuals versus age by gender



Extending the basic linear mixed-effects model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \dots, M$$

$$\mathbf{b}_i \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Psi})$$

Instead of assuming a homoscedastic $\boldsymbol{\varepsilon}_i \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, the extended single - level linear mixed - effects model relaxes the assumption by allowing heteroscedastic and correlated within - group errors $\boldsymbol{\varepsilon}_i \sim \mathbf{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$, $i = 1, \dots, M$ where $\boldsymbol{\Lambda}_i$ is a symmetric positive definite matrix

$\boldsymbol{\varepsilon}_i$ is now assumed to :

- 1) independent for different groups of different i ;
- 2) independent of the random effects \mathbf{b}_i ;
- 3) have a variance structure as a function of some covariates (e.g. age, sex etc.)
- 4) have a covariance structure among repeated measurements within the same group (e.g. measurements from different time points for the same subject), that is, $\text{corr}(\boldsymbol{\varepsilon}_{ij}, \boldsymbol{\varepsilon}_{ij'}) \neq 0$.

Extending the basic linear mixed-effects model (continued)

The symmetric positive definite matrices Λ_i can always be decomposed into a product of simpler matrices

$$\Lambda_i = \mathbf{V}_i \mathbf{C}_i \mathbf{V}_i$$

Where \mathbf{V}_i is diagonal and \mathbf{C}_i is a correlation matrix with all diagonal elements equal to one.

Easy to verify :

1. Take the square root of the diagonal elements of variance - covariance matrix Λ_i to be the diagonal elements of \mathbf{V}_i
2. Solve for the corresponding correlation coefficients from the variance - covariance

matrix Λ_i by $\rho_{AB} = \frac{\text{Cov}(A, B)}{\sqrt{\text{Var}(A)\text{Var}(B)}}$

3. There is an R function `decompose.cov` for this purpose

e.g.

$$\begin{bmatrix} 4 & 4 & 1.2 \\ 4 & 16 & 3.6 \\ 1.2 & 3.6 & 9 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0.5 & 0.2 \\ 0.5 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Extending the basic linear mixed-effects model (continued)

$$\boldsymbol{\varepsilon}_i \sim \mathbf{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i), i = 1, \dots, M$$

$$\boldsymbol{\Lambda}_i = \mathbf{V}_i \mathbf{C}_i \mathbf{V}_i$$

It is easy to verify that

$$\text{Var}(\varepsilon_{ij}) = \sigma^2 [\mathbf{V}_i]_{jj}^2 \quad \text{cor}(\varepsilon_{ij}, \varepsilon_{jk}) = [\mathbf{C}_i]_{jk}$$

Hence variance - covariance matrix $\boldsymbol{\Lambda}_i$ for within - group errors $\boldsymbol{\varepsilon}_i$ is decomposed into two components :

\mathbf{V}_i : variance structure component

\mathbf{C}_i : correlation structure component

Such a decomposition allow us to model the variance and correlaton structure of within - group errors separately.

Model variance component of within-group errors

Variance for the within-group errors in a single-level linear mixed-effects model as a function of fitted values and variance covariates (e.g. age, sex, etc.)

$$\text{Var}(\varepsilon_{ij}) \approx \sigma^2 g^2(\hat{\mu}_{ij}, \mathbf{v}_{ij}, \boldsymbol{\delta}), \quad i = 1, \dots, M \quad j = 1, \dots, n_i$$

$g(\cdot)$: the variance function

$\hat{\mu}_{ij}$: the fitted values

\mathbf{v}_{ij} : vector of variance covariates (e.g. age, sex, etc.)

$\boldsymbol{\delta}$: vector of parameters for variance covariates

Model variance component of within-group errors: variance functions for modeling heteroscedasticity

Fixed variance	$Var(\varepsilon_{ij}) = \sigma^2 \nu_{ij}$	varFixed(~age)	Variance ↑ or ↓ linearly with variance covariate (e.g. age)
Different variances across stratum (s=1, 2...,S, e.g Sex)	$Var(\varepsilon_{ij}) = \sigma^2 \delta_{sij}^2$	varIdent(form=~1 sex)	Constant variance within stratum. Delta represents the ratios between the variances and a reference variance
Power of covariate	$Var(\varepsilon_{ij}) = \sigma^2 \nu_{ij} ^{2\delta}$	varPower(form=~fitted(.) Sex, fixed=1)	Variance ↑ or ↓ proportionally to a power of absolute value of variance covariate (e.g. fitted value)
Exponential of covariate	$Var(\varepsilon_{ij}) = \sigma^2 \exp(2\delta\nu_{ij})$	varExp(form=~age sex, fixed=c(Female=0))	Variance ↑ or ↓ proportionally to the exponential of the variance covariate (e.g. age)
Constant plus power of covariate	$Var(\varepsilon_{ij}) = \sigma^2 (\delta_1 + \nu_{ij} ^{\delta_2})^2$	varConstPower(power=0.5, fixed=list(const=1))	Similar to power of covariate (varPower)
varComb	Multiplying any two variance functions from above	(varFunc1)(varFunc2)	For more flexible while more complicated variance structure

Refit a heteroscedastic model for the orthodontic growth data

R call

```
fm3Orth.lme <- lme(distance~Sex*I(age-11), data=Orthodont, random=~I(age-11)|Subject,  
weights=varIdent(form=~1|Sex))
```


Result of fitting a mixed effect model with heteroscedastic within-group errors

Linear mixed-effects model fit by REML

Data: Orthodont

Log-restricted-likelihood: -205.7612

Fixed: distance ~ Sex * I(age - 11)

(Intercept)	SexFemale	I(age - 11)	SexFemale:I(age - 11)
24.9687500	-2.3210227	0.7843750	-0.3048295

Random effects:

Formula: ~I(age - 11) | Subject

Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	1.8549792	(Intr)
I(age - 11)	0.1565170	0.394
Residual	1.6295849	

Variance function:

Structure: Different standard deviations per stratum

Formula: ~1 | Sex

Parameter estimates:

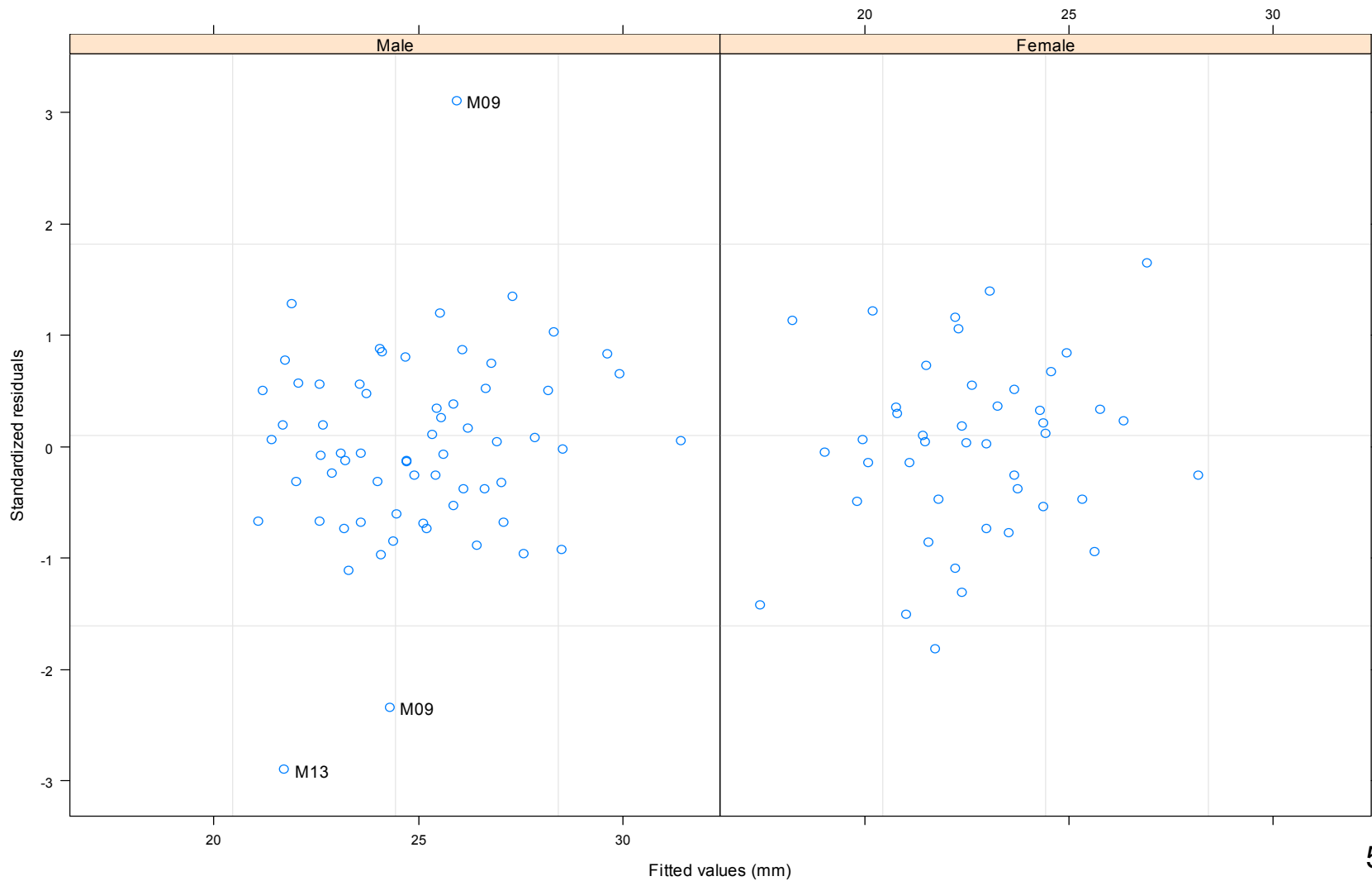
	Male	Female
	1.0000000	0.4088464

Number of Observations: 108
Number of Groups: 27

$$\text{Std}_{\text{male}} = 1.63$$

$$\text{Std}_{\text{female}} = 0.41 \text{Std}_{\text{male}}$$

Scatter plots of standardized residuals versus fitted values by gender (heteroscedastic model)



Model correlation component of within-group errors

- General idea: two within-group errors ε_{ij} and $\varepsilon_{ij'}$ are correlated with each other, and such a correlation is assumed to depend on the spatial or temporal distance between them.
- Correlation structure is isotropic: correlation only depends on relative distance, NOT on particular spatial or temporal positions of two measurements

$$\text{cor}(\varepsilon_{ij}, \varepsilon_{ij'}) = h[d(\mathbf{p}_{ij}, \mathbf{p}_{ij'}), \boldsymbol{\rho}] \quad i = 1, \dots, M \quad j, j' = 1, \dots, n_i$$

$\mathbf{p}_{ij}, \mathbf{p}_{ij'}$: spatial or temporal position vectors for $\varepsilon_{ij}, \varepsilon_{ij'}$

$d(\cdot)$: distance function

$h(\cdot)$: correlation functions taking values between -1 and 1

$\boldsymbol{\rho}$: vectors of correlation parameters

Common correlation structures

Compound symmetry	$h = \rho$ if $j \neq j'$, 1 if $j = j'$ where $0 \leq \rho \leq 1$	corCompSymm
Autoregressive-moving average lag 1	$h = \rho^s$ where $0 \leq \rho \leq 1$	corCAR1
Exponential	$h = \exp(-s/\rho)$	corExp
Gaussian	$h = \exp(-(s/\rho)^2)$	corGaus
Linear	$h = (1 - s/\rho)I(s < \rho)$	corLin
Rational quadratic	$h = 1 - (s/\rho)^2 / [1 + (s/\rho)^2]$	corRatio
Spherical	$h = [1 - 1.5(s/\rho) + 0.5(s/\rho)^3]I(s < \rho)$	corSpher

$s = d(\mathbf{p}_{ij}, \mathbf{p}_{ij'})$ distance between two response measurements

Refit a heteroscedastic model with dependent within-group errors for the orthodontic growth data

R call:

```
fm4Orth.lme <- lme(distance~Sex*I(age-11), data=Orthodont, random=~I(age-11)|Subject,  
weights=varIdent(form=~1|Sex), correlation=corCAR1(form=~I(age-11)|Subject))
```

Result of fitting a mixed effect model with corCAR1 correlation structure for the within-group errors

Linear mixed-effects model fit by REML

Data: Orthodont

Log-restricted-likelihood: -205.7612

Fixed: distance ~ Sex * I(age - 11)

(Intercept)

24.9687500

SexFemale

-2.3210227

I(age - 11) SexFemale:I(age - 11)

0.7843750

-0.3048295

Random effects:

Formula: ~I(age - 11) | Subject

Structure: General positive-definite, Log-Cholesky parametrization

StdDev Corr

(Intercept) 1.8549782 (Intr)

I(age - 11) 0.1565193 0.394

Residual 1.6295832

Correlation Structure: Continuous AR(1)

Formula: ~I(age - 11) | Subject

Parameter estimate(s):

Phi

8.511159e-05

Variance function:

Structure: Different standard deviations per stratum

Formula: ~1 | Sex

Parameter estimates:

Male Female

1.0000000 0.4088455

Number of Observations: 108

Number of Groups: 27

The estimated correlation between two measurements taken 2 year apart is $(8.5E-5)^2$

Compare the independent errors model with dependent errors model

R call:

```
fm3Orth.lme <- lme(distance~Sex*I(age-11), data=Orthodont, random=~I(age-11)|Subject, weights=varIdent(form=~1|Sex))
```

```
fm4Orth.lme <- lme(distance~Sex*I(age-11), data=Orthodont, random=~I(age-11)|Subject, weights=varIdent(form=~1|Sex), correlation=corCAR1(form=~I(age-11)|Subject))
```

```
anova(fm3Orth.lme, fm4Orth.lme)
```

Result of model comparison

```
> anova(fm3Orth.lme, fm4Orth.lme)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
fm3Orth.lme	1	9	429.5225	453.3220	-205.7612			
fm4Orth.lme	2	10	431.5225	457.9664	-205.7612	1 vs 2	6.156108e-08	0.9998

Large p-value suggests simpler independent errors model (fm3Orth.lme) is enough

Comments

- Mixed-effect model provides a flexible tool for modeling difference sources of variability among grouped data
- Readily available in open-source software (R library(nlme))
- Requires extensive assumptions about the distribution of random effects