Linear regression with repeated measures: Linear mixed models

Example

A research group carried out a study about the growth of children aged between 8 and 14 years. They used the distance between the pituitary gland and the pterygomaxillary fissure (in millimeters, mm) as growth measure. The researchers recruited 16 boys and 11 girls and measured this distance at ages 8, 10, 12 and 14. They aimed to answer the following questions:

- Does this distance change in time?
- Is the of pattern of change similar in boys and girls?

Let us import the data in `R`:

```
growth <- read.table("growth.txt", header=T, sep=" ")
head(growth)
```

```
  ID SEX AGE DIST J
1  1   F   8 21.0 1
2  1   F  10 20.0 2
3  1   F  12 21.5 3
4  1   F  14 23.0 4
5  2   F   8 21.0 1
6  2   F  10 21.5 2
```

The dataset contains the following variables:

- ID: identification of the subject
- SEX: gender of the subject, F=female, M=male
- AGE: age of the subject, in years
- DIST: growth measure (mm)
- J: order of the measurement

To answer the research questions first we need to model the distance in terms of sex and age. Thus, the first model that we need to estimate is:

$$\text{DIST}_{ij} = \beta_0 + \beta_1 \text{SEX}_{ij} + \beta_2 \text{AGE}_{ij} + \epsilon_{ij},$$

where DIST_{ij} refers to the jth measurement from the ith subject and ϵ_{ij} represents the random error, which is assumed to follow a Normal distribution with mean 0 and variance σ_e^2.

SEX is a categorical variable, so if we want to interpret it correctly, first we need to find out how `R` encodes it:

```r
contrasts(growth$SEX)
```

```
  M
F 0
M 1
```

This means that the reference category of SEX is Female, so

$$\text{SEX} = \begin{cases} 1 & \text{if SEX=M} \\ 0 & \text{if SEX=F} \end{cases}$$

Thus, β_1 is interpreted as the global difference, in mean, between boys and girls of a certain age.

Regarding β_2, given that AGE is a quantitative variable, it is interpreted as the change, in mean, of the distance per each year of age increase. However, we could think that this change in the distance is not the same in boys than in girls. In this situation, we should introduce an interaction effect between SEX and AGE in the model:

$$\text{DIST}_{ij} = \beta_0 + \beta_1 \text{SEX}_{ij} + \beta_2 \text{AGE}_{ij} + \beta_3 \text{SEX}_{ij} \cdot \text{AGE}_{ij} + \epsilon_{ij}$$

this is equivalent to performing two linear models between DIST and AGE, one per each sex:

- Boys:
$$\text{DIST}_{ij} = \beta_0 + \beta_1 + (\beta_2 + \beta_3)\, \text{AGE}_{ij} + \epsilon_{ij}$$
- Girls:
$$\text{DIST}_{ij} = \beta_0 + \beta_2 \text{AGE}_{ij} + \epsilon_{ij}$$

However, here we cannot estimate the model as we saw in multiple linear regression since one of the assumptions of the model is not met: the random errors are not independent. We have the same individuals measured more than once, and we cannot assume that the measurements from the same subject are independent. This situation is similar to the case of paired data we saw in previous topics, but now we have 4 measurements per individual instead of 2. A method to estimate regression models taking into account the dependence of the measurements from the same individual is **linear mixed regression**.

Definition of Linear mixed model

These models are called 'linear' because the response variable is modelled through a linear combination of parameters and covariates. The word 'mixed' indicates that we use two types of parameters: fixed and random. In previous topics we considered only fixed parameters. This means that they were the same for all individuals. However, random effects allow the

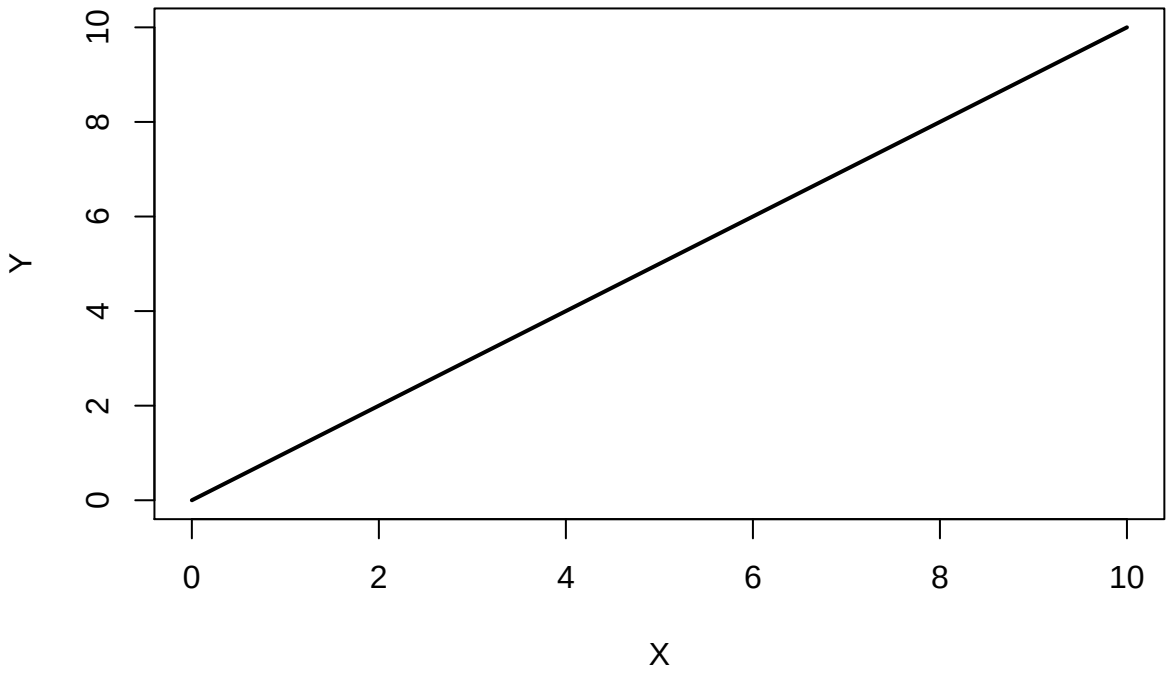
effects to change from one individual to another. For example, if we want the intercept β_0 to vary between subjects, the appropriate linear mixed model is

$$\text{DIST}_{ij} = (\beta_0 + b_{0i}) + \beta_1 \text{SEX}_{ij} + \beta_2 \text{AGE}_{ij} + \beta_3 \text{SEX}_{ij} \cdot \text{AGE}_{ij} + \epsilon_{ij},$$

where b_{0i} is assumed to randomly vary between individuals. In general, we assume that random effects follow a Normal distribution. Hence, in the case of this random intercept, we assume that it follows a Normal distribution with mean 0 and variance σ_0^2.

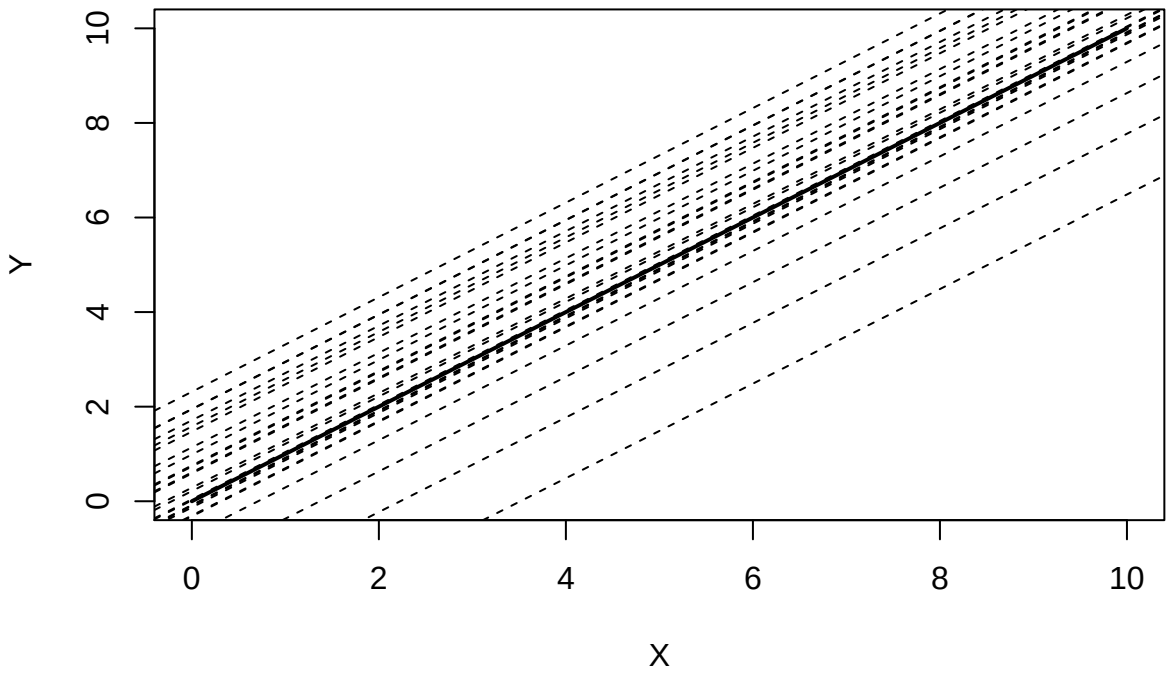
Graphically, in the linear regression model we adjusted a shared straight line for all the individuals:

```
plot(c(0, 10), c(0, 10), type="l", xlab="X", ylab="Y", lwd=2)
```



When we introduce the random intercept, we are adjusting a different regression for each individual with different intercepts and with a common slope:

```
plot(c(0,10), c(0,10), type="l", xlab="X", ylab="Y", lwd=2)
for (i in 1:25) abline(rnorm(1, 0, 2), 1, lty=2)
```



In fact, we are not correcting the lack of independence in the data, but we are forcing the model to take it into account. To see this, we can calculate the covariance between the measurements of a single individual; for example, between DIST_{i1} and DIST_{i2}. Using our model:

$$\text{DIST}_{i1} = (\beta_0 + b_{0i}) + \beta_1 \text{SEX}_{i1} + \beta_2 \text{AGE}_{i1} + \beta_3 \text{SEX}_{i1} \cdot \text{AGE}_{i1} + \epsilon_{i1},$$
$$\text{DIST}_{i2} = (\beta_0 + b_{0i}) + \beta_1 \text{SEX}_{i2} + \beta_2 \text{AGE}_{i2} + \beta_3 \text{SEX}_{i2} \cdot \text{AGE}_{i2} + \epsilon_{i2}.$$

Thus,

$$\text{Cov}\left(\text{DIST}_{i1}, \text{DIST}_{i2}\right) = \cdots = \text{Cov}\left(b_{0i} + \epsilon_{i1}, b_{0i} + \epsilon_{i2}\right).$$

Since random errors are assumed to be independent,

$$\text{Cov}\left(\text{DIST}_{i1}, \text{DIST}_{i2}\right) = \ldots = \text{Cov}\left(b_{0i}, b_{0i}\right) = \text{Var}\left(b_{0i}\right) = \sigma_0^2.$$

We can also show that two measurements from two different individuals have covariance equal to zero, that is,

$$\text{Cov}\left(\text{DIST}_{ij}, \text{DIST}_{i',j'}\right) = 0.$$

Hence, in linear mixed models we assume that measurements from the same individual are dependent but measurements from different individuals are independent.

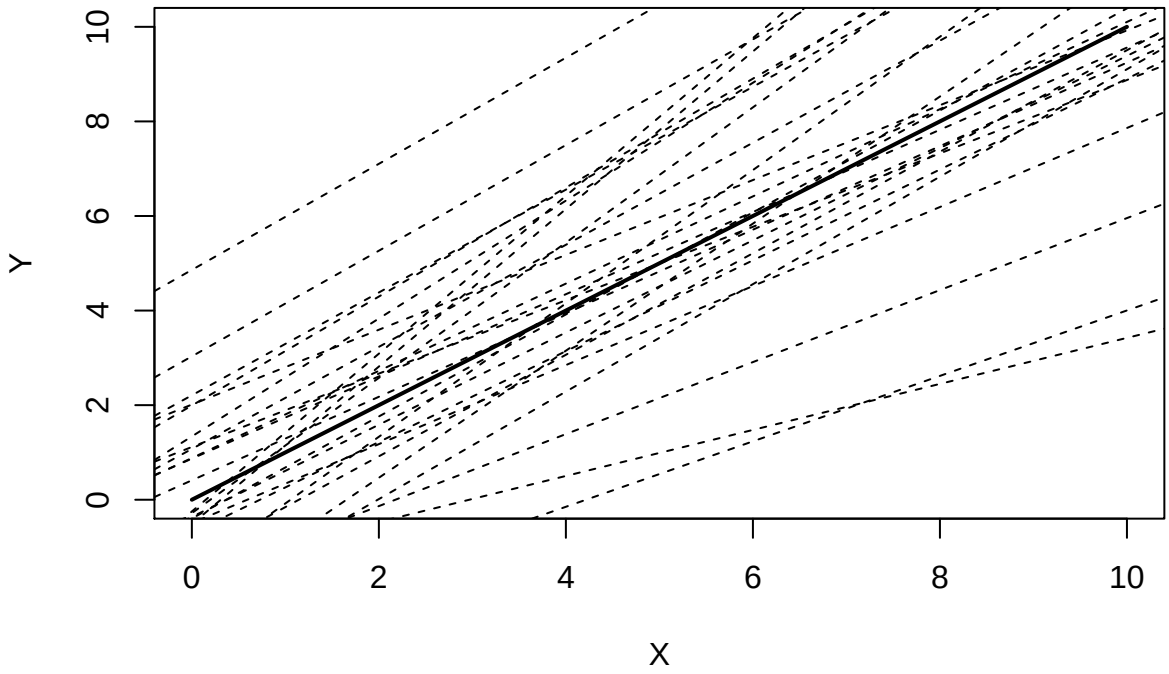
In our model, we could also allow that slopes change depending on the individual by introducing another random effect. For example,

$$\text{DIST}_{ij} = (\beta_0 + b_{0i}) + \beta_1 \text{SEX}_{ij} + (\beta_2 + b_{2i})\,\text{AGE}_{ij} + \beta_3 \text{SEX}_{ij} \cdot \text{AGE}_{ij} + \epsilon_{ij}.$$

where b_{2i} follows a Normal distribution with mean 0 and variance σ_2^2. In this model we are assuming that the effect of age is not the same for all the subjects.

Graphically,

```
plot(c(0, 10), c(0, 10), type="l", xlab="X", ylab="Y", lwd=2)
for (i in 1:25) abline(rnorm(1, 0, 2), rnorm(1, 1, 0.25), lty=2)
```



Notice that now the covariance between measurements from the same individual, DIST_{i1} and DIST_{i2}, is also a function of age. Thus, when we introduce a random effect in the slope, we are also imposing an structure to the covariance between measurements.

Model estimation

First of all, we will estimate the model including only a random effect in the intercept.

In `R`, linear mixed models are fitted using the `nlme` package:

```
library(nlme)
model.1=lme(DIST~SEX+AGE+SEX*AGE, data=growth, random=~1|ID)
summary(model.1)
```

```
Linear mixed-effects model fit by REML
 Data: growth
       AIC      BIC    logLik
  445.7572 461.6236 -216.8786

Random effects:
 Formula: ~1 | ID
        (Intercept) Residual
StdDev:    1.816214 1.386382

Fixed effects: DIST ~ SEX + AGE + SEX * AGE
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 1.1835071 79 14.679023  0.0000
SEXM        -1.032102 1.5374208 25 -0.671321  0.5082
AGE          0.479545 0.0934698 79  5.130483  0.0000
SEXM:AGE     0.304830 0.1214209 79  2.510520  0.0141
 Correlation:
         (Intr) SEXM   AGE
SEXM     -0.770
AGE      -0.869  0.669
SEXM:AGE  0.669 -0.869 -0.770

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.59804400 -0.45461690  0.01578365  0.50244658  3.68620792

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

In this output we can see, first of all, the standard deviations of the random effects. The importance of a random effect can be assessed by comparing its variance to that of the random error. In this case, the standard deviation of the random effect b_{0i} is approximately 1.31 times higher than than of the random error, thus this random effect is important.

More formally, we can evaluate the null hypothesis $\sigma_0 = 0$. If we do not reject it, we will conclude that it is not necessary to include this random effect in the model. First, we should estimate the model with no random effect using the function `gls`:

```
model.0=gls(DIST~SEX+AGE+SEX*AGE, data=growth)
summary(model.0)
```

```
Generalized least squares fit by REML
```

```
  Model: DIST ~ SEX + AGE + SEX * AGE
  Data: growth
       AIC      BIC    logLik
  493.5591 506.7811 -241.7796

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 17.372727 1.7080306 10.171204  0.0000
SEXM        -1.032102 2.2187969 -0.465163  0.6428
AGE          0.479545 0.1521635  3.151515  0.0021
SEXM:AGE     0.304830 0.1976661  1.542143  0.1261

 Correlation:
         (Intr) SEXM   AGE
SEXM     -0.770
AGE      -0.980  0.754
SEXM:AGE  0.754 -0.980 -0.770

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.48814895 -0.58569115 -0.07451734  0.58924709  2.32476465

Residual standard error: 2.256949
Degrees of freedom: 108 total; 104 residual
```

Once the model with no random effect is estimated, the hypothesis test should be performed with the function `anova`:

```
anova(model.0, model.1)
```

```
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model.0     1  5 493.5591 506.7811 -241.7796
model.1     2  6 445.7572 461.6236 -216.8786 1 vs 2 49.80187  <.0001
```

The p-value is lower than 0.05 (the usual type I error used in hypothesis tests), so we have enough evidence to say that the variance of the random effect is different from 0.

Then we think about adding a random effect in the age slope:

```
model.2=lme(DIST~SEX+AGE+SEX*AGE, data=growth, random=~AGE|ID)
summary(model.2)
```

```
Linear mixed-effects model fit by REML
 Data: growth
       AIC      BIC    logLik
  448.5817 469.7368 -216.2908

Random effects:
```

```
 Formula: ~AGE | ID
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2.4055009 (Intr)
AGE         0.1803455 -0.668
Residual    1.3100396

Fixed effects: DIST ~ SEX + AGE + SEX * AGE
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 1.2283958 79 14.142614  0.0000
SEXM        -1.032102 1.5957329 25 -0.646789  0.5237
AGE          0.479545 0.1037193 79  4.623492  0.0000
SEXM:AGE     0.304830 0.1347353 79  2.262432  0.0264
 Correlation:
         (Intr) SEXM   AGE
SEXM     -0.770
AGE      -0.880  0.678
SEXM:AGE  0.678 -0.880 -0.770

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.168077732 -0.385939009  0.007104087  0.445154545  3.849463576

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

The standard deviation of the random effect b_{2i} is very small compared to that of the random effect b_{0i}. If we test the null hypothesis of $\sigma_2 = 0$:

```
anova(model.1, model.2)
```

```
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model.1     1  6 445.7572 461.6236 -216.8786
model.2     2  8 448.5817 469.7368 -216.2908 1 vs 2 1.175588  0.5556
```

We do not reject the null hypothesis and so we should exclude this effect from our model.

In conclusion, our best model is the one that includes a random effect in the intercept and a fixed effect in the slope.

Now let us evaluate the fixed effects. First of all, we should test if there is an interaction between age and gender. Remember that the interaction effect implies the following: the slope between the distance (our response variable) and age is different depending on the gender. Thus, we need to test the null hypothesis $\beta_3 = 0$ versus the alternative $\beta_3 \neq 0$. To solve this contrast we may use the test shown in the summary table of the fixed effects:

```
summary(model.1)
```

```
Linear mixed-effects model fit by REML
 Data: growth
       AIC      BIC    logLik
  445.7572 461.6236 -216.8786

Random effects:
 Formula: ~1 | ID
        (Intercept) Residual
StdDev:    1.816214 1.386382

Fixed effects: DIST ~ SEX + AGE + SEX * AGE
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 1.1835071 79 14.679023  0.0000
SEXM        -1.032102 1.5374208 25 -0.671321  0.5082
AGE          0.479545 0.0934698 79  5.130483  0.0000
SEXM:AGE     0.304830 0.1214209 79  2.510520  0.0141
 Correlation:
         (Intr) SEXM   AGE
SEXM     -0.770
AGE      -0.869  0.669
SEXM:AGE  0.669 -0.869 -0.770

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.59804400 -0.45461690  0.01578365  0.50244658  3.68620792

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

The p-value corresponding to the interaction term is lower than 0.05, so we reject the null hypothesis. Thus, there is a significant interaction between age and gender.

The general model that we have obtained can be written as follows:

- Boys:

$$\text{DIST} = 17.37 - 1.03 + 0.48 \cdot \text{AGE} + 0.30 \cdot \text{AGE} = 16.34 + 0.78 \cdot \text{AGE}$$

- Girls:

$$\text{DIST} = 17.37 + 0.48 \cdot \text{AGE}$$

The slope for boys is higher than that for girls. That is, the growth rate in time is higher in boys than in girls. Moreover, the distance increases, in mean, 0.78 mm per year in boys and 0.48 mm per year in girls.

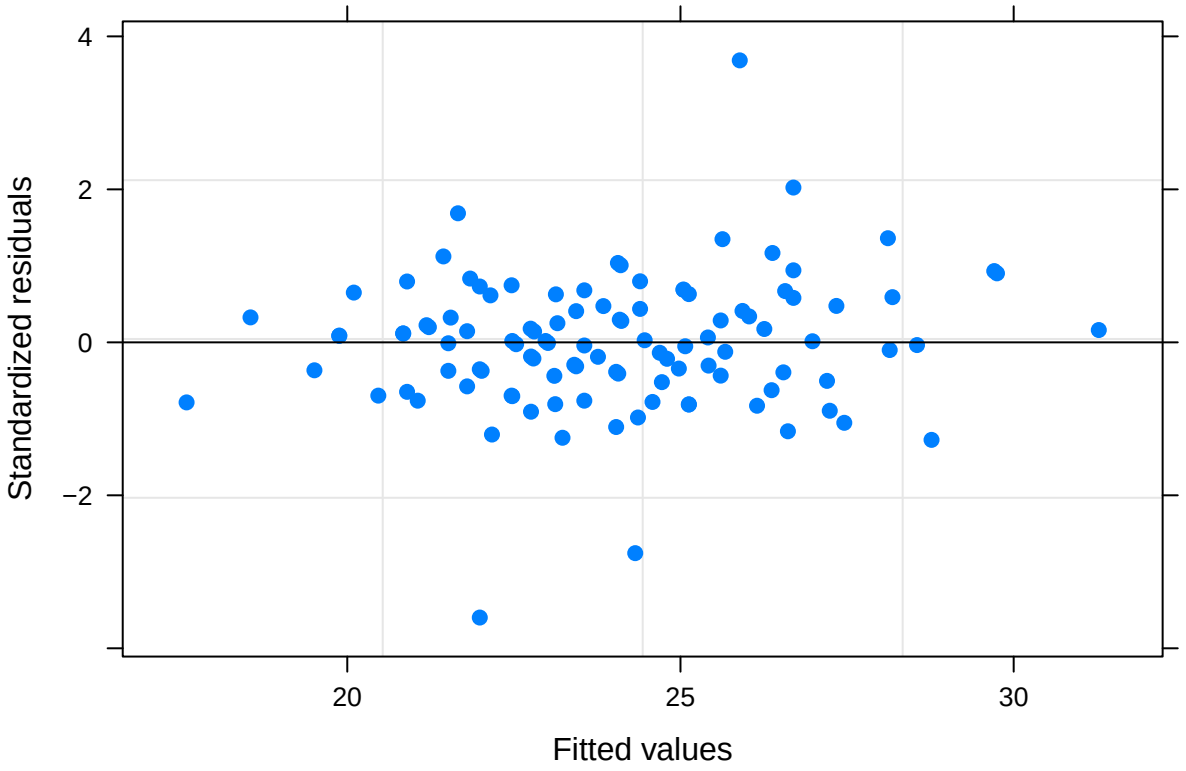
The random effect in the intercept can be interpreted as the variability of the initial distances between boys and girls. This variability is represented by a standard deviation of 1.81 mm.

Model validation

Once we have chosen a model for our data, the next step is to check if the assumptions of the model are met. We need to validate the following assumptions: independence of the random errors, normality of residuals and presence of outliers.

We can check the independence of the random errors using a dispersion plot of the standardized residuals versus the predicted values:

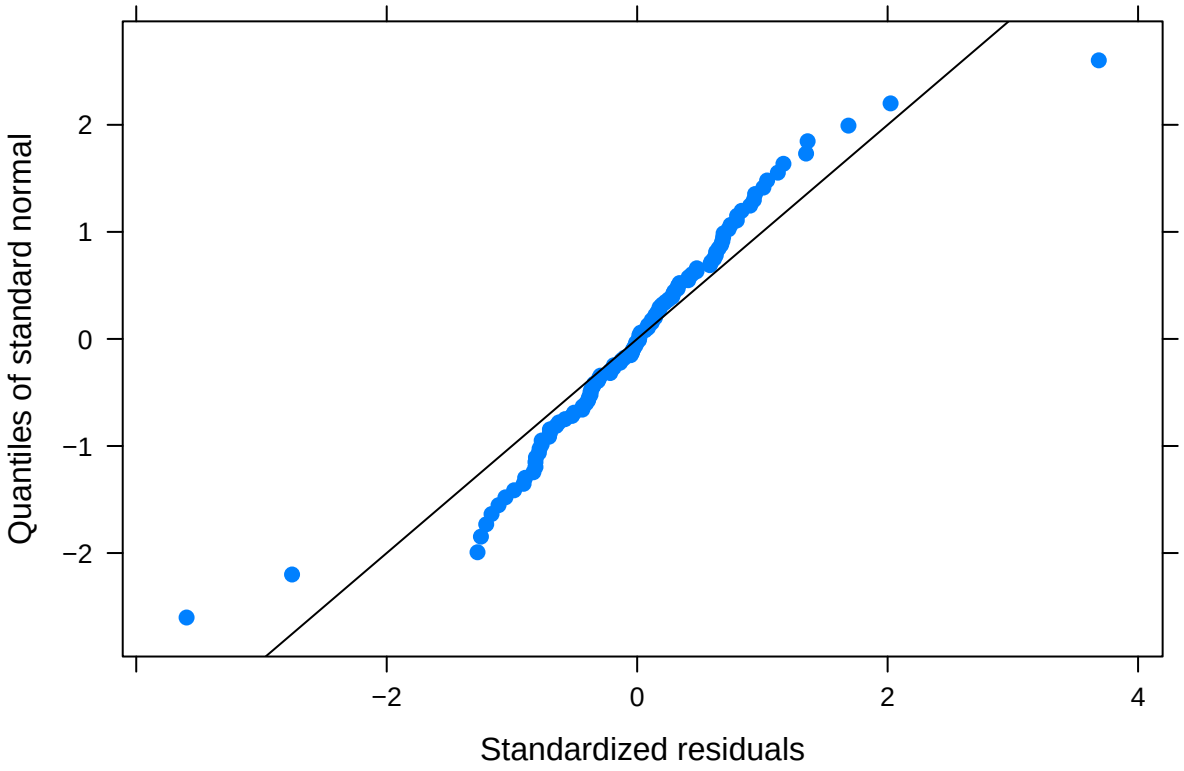
```
print(plot(model.1, pch=16, which=1))
```



In this plot we see that the behaviour of the residuals is, in general, right. Most of them take values between -2 and 2 and there are only 3 observations with a standardized residual (in absolute value) greater than 2. Moreover, we do not see any grouping of observations nor trends.

We can see the QQ-plot:

```
print(qqnorm(model.1, abline=c(0, 1), pch=16))
```



We continue to see these 3 outliers. Let us identify these observations in out database. First, let us obtain the standardized residuals:

```
res=resid(model.1, type="p")
```

Then we look for the observations with an absolute standardized residual greater than 2:

```
which(abs(res)>2)
```

```
20 20 24 24
78 79 93 96
```

We see that there are, in fact, 4 observations that lead to extreme residuals. Actually, these observations come from the individuals labelled 20 and 24. Let us inspect the individual with `ID=20`:

```
res[growth$ID==20]
```

```
        20          20          20          20
 0.1788771 -2.7559203  3.6862079 -1.0518444
```

The extreme residuals correspond to the second and third measurements, that is, the measurements at 10 and 12 years. Let us see these observations:

```
subset(growth, ID==20)
```

```
   ID SEX AGE DIST J
77 20   M   8 23.0 1
78 20   M  10 20.5 2
79 20   M  12 31.0 3
80 20   M  14 26.0 4
```

We see an strange behaviour in the distances of this individual: the distance decreased between 8 and 10 years, then increased again at age 12 and decreased once more at age 14. Since the model tells us that the distance increases with time, the oscillations of subject number 20 do not fit well in the model. In this situation, we could check with the researchers if these measurements are correct (the measurements can be affected by errors in the data collection and/or in the implementation of the database).

If we look at subject 24:

```
res[growth$ID==24]
```

```
        24         24         24         24
-3.5980440  0.6801783  0.6305888  2.0236032
```

```
subset(growth, ID==24)
```

```
   ID SEX AGE DIST J
93 24   M   8 17.0 1
94 24   M  10 24.5 2
95 24   M  12 26.0 3
96 24   M  14 29.5 4
```

The problem here is located in the first measurement, where this subject has a distance very far from its successive distances. Actually, if we calculate the mean distance of all 8-year boys:

```
mean(subset(growth, AGE==8 & SEX=="M")$DIST)
```

```
[1] 22.875
```

We conclude that the first measurement of the subject number 24 is quite far from this mean, so this is causing that this subject appears as an outlier in the residuals. The problem with outliers is that sometimes they have an important influence in the model estimates. We can check this by adjusting the model again, but excluding these 2 individuals. If the two models give similar estimates, we will conclude that the outliers are simple anomalies in the model. However, if they give very different estimates, we would consider invalid the model with outliers.

Let us adjust the model again but excluding individuals 20 and 24:

```
growth2=subset(growth, ID!=20 & ID!=24)
model.out=lme(DIST~SEX+AGE+SEX*AGE, data=growth2, random=~1|ID)
summary(model.out)
```

```
Linear mixed-effects model fit by REML
 Data: growth2
      AIC      BIC    logLik
  366.449 381.8351 -177.2245

Random effects:
 Formula: ~1 | ID
        (Intercept)  Residual
StdDev:      1.9588 0.9927077

Fixed effects: DIST ~ SEX + AGE + SEX * AGE
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 0.9556221 73 18.179496  0.0000
SEXM         0.073701 1.2770037 23  0.057714  0.9545
AGE          0.479545 0.0669283 73  7.165058  0.0000
SEXM:AGE     0.207955 0.0894368 73  2.325158  0.0228
 Correlation:
         (Intr) SEXM   AGE
SEXM     -0.748
AGE      -0.770  0.577
SEXM:AGE  0.577 -0.770 -0.748

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-1.91105116 -0.68010763 -0.01735606  0.59418206  2.13465001

Number of Observations: 100
Number of Groups: 25
```

The standard deviation estimate of the random effect is 1.96, whereas in the model with the complete dataset it was 1.82. Regarding the random error, now we have a standard deviation of 0.99, whereas in the preivous model it was 1.39. Thus, the change in these two variabilitys is small. If we look at the fixed effects, we conclude that the estimates barely change from one model to another, and the interaction term is still statistically significant. In conclusion, the two patients affecting the residuals do not have an important impact in the model estimates so we decide to keep the first model.