Kazalo

1 LINEARNI MEŠANI MODELI

Linearni mešani modeli (LMM) predstavljajo široko paleto modelov, v katerih so poleg **fiksnih vplivov** vključeni tudi **slučajni vplivi**. Do sedaj smo se v linearnih modelih (LM) ukvarjali samo s fiksnimi vplivi, v katerih je za napake $\boldsymbol{\varepsilon}$ veljalo, da so slučajne, porazdeljene normalno $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Lambda})$. Ko dodamo v linearni model tudi slučajne vplive, nepojasnjeno variabilnost, ki je bila v LM pripisana napakam, razdelimo na dva dela: nepojasnjena variabilnost, ki jo lahko pripišemo znanim slučajnim vplivom, in preostanek nepojasnjene variabilnosti. V modelu poleg parametrov variančno-kovariančne matrike napak ocenjujemo še parametre **variančno-kovariančne matrike slučajnih vplivov**. S tem ne vplivamo na ocene parametrov modela za fiksne vplive, temveč samo na njihove standardne napake.

Slučajni vplivi so v LMM določeni glede na naravo zbiranja podatkov (načrt vzorčenja, poskusna zasnova). Za te podatke je značilno, da so na različne načine zbrani po skupinah. V poglavju o polinomski regresiji in regresiji zlepkov smo modelirali pridelek koruze v odvisnosti od gostote setve, v modelu smo sprva upoštevali tudi slučajni vpliv t. i. blokov, ker je bil poljski poskus zasnovan v poskusni zasnovi slučajni bloki. Izkazalo se je, da slučajni vpliv blokov ne pojasni statistično značilnega dela variabilnosti.

Denimo, da imamo S skupin, $i = 1, ..., S$, v vsaki skupini je n_i enot. Za podatke znotraj skupin velja, da so medsebojno odvisni, podatki med različnimi skupinami pa so medsebojno neodvisni. Trije najbolj pogosti načini zbiranja podatkov (vzorčenja) po skupinah so:

- **longitudinalne študije**: na istih enotah izvedemo meritve ob različnih časih. Slučajni vpliv je tu enota (npr. oseba, rastlina, predmet,...). Merimo/opazujemo, kako se vrednost odzivne spremenljivke spreminja s časom pod različnimi pogoji, ki jih opisujejo ostale napovedne spremenljivke v modelu;

- **hierarhično vzorčenje**: npr. po principu slučajnosti vzorčimo šole na danem območju, znotraj šol po principu slučajnosti izberemo razrede ter znotraj razredov spet po principu slučajnosti izberemo učence (vzorčenje v več ravneh). Enota vzorčenja je učenec. V takem primeru v LMM vključimo dva slučajna vpliva: slučajni vpliv `šola` in slučajni vpliv `razred`, ki je gnezden znotraj šole;

- **poskusi s ponovljajočimi meritvami** (*repeated measures design*): istim enotam vzorčenja priredimo več različnih obravnavanj oz. jih opazujemo pod različnimi pogoji. Slučajni vpliv tudi v tem primeru predstavlja enota vzorčenja.

1.1 Matematični model

Analiziramo k fiksnih vplivov (v model je vključenih k regresorjev) in l slučajnih vplivov. Linearni mešani model za i-to skupino, $i = 1, ..., S$, v kateri je n_i enot, zapišemo takole:

$$\mathbf{y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{u_i} + \boldsymbol{\varepsilon_i}, \tag{1}$$

$$\mathbf{u_i} \sim \mathbf{N}_l(\mathbf{0}, \boldsymbol{\Psi}),$$

$$\boldsymbol{\varepsilon_i} \sim \mathbf{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{\Lambda}_i).$$

- $\mathbf{y_i}$ je vektor vrednosti odzivne spremenljivke iz i-te skupine dimenzije $n_i \times 1$;
- $\mathbf{X_i}$ je **modelska matrika fiksnih vplivov** za opazovanja iz i-te skupine dimenzije $n_i \times (k+1)$;
- $\boldsymbol{\beta}$ je vektor parametrov modela za fiksne vplive dimenzije $(k+1) \times 1$;
- $\mathbf{Z_i}$ je **modelska matrika slučajnih vplivov** za opazovanja iz i-te skupine dimenzije $n_i \times l$;
- $\mathbf{u_i}$ je vektor slučajnih vplivov za i-to skupino dimenzije $l \times 1$;
- $\boldsymbol{\varepsilon_i}$ je vektor napak za i-to skupino dimenzije $n_i \times 1$ (*within-group error vector*);
- $\boldsymbol{\Psi}$ je variančno-kovariančna matrika slučajnih vplivov dimenzije $l \times l$;
- $\sigma^2\boldsymbol{\Lambda}_i$ je variančno-kovariančna matrika napak za i-to skupino dimenzije $n_i \times n_i$; v osnovnem linearnem mešanem modelu predpostavimo neodvisne napake s konstantno varianco, kar pomeni, da je $\boldsymbol{\Lambda}_i = \mathbf{I}$.

Predpostavimo, da so za vsako skupino i, $i = 1, ..., S$ slučajni vplivi u_i in napake ε_i medsebojno neodvisni, prav tako velja, da so neodvisni slučajni vplivi u_i med različnimi skupinami, enako velja za napake ε_i različnih skupin.

Slučajne vplive v LMM vključimo kot **slučajne spremenljivke**, za katere predpostavimo, da so porazdeljene po normalni porazdelitvi $\mathbf{u_i} \sim \mathbf{N}_l(\mathbf{0}, \boldsymbol{\Psi})$ in so neodvisne med sabo. V LMM ocenjujemo parametre variančno-kovariančne matrike slučajnih vplivov $\boldsymbol{\Psi}$ in tudi vrednosti slučajnih vplivov $\mathbf{u_i}$.

Za oceno parametrov LMM se standardno uporablja metoda REML (*REstricted Maximum Likelihood method*), ki v primeru majhnih vzorcev zmanjša pristranskost ML ocen za komponente variance slučajnih vplivov in za varianco napak. Če primerjamo modele z različnim številom parametrov $\boldsymbol{\beta}$ za fiksne vplive med sabo, moramo parametre modela oceniti po ML metodi.

2 LONGITUDINALNO VZORČENJE

Analizirali bomo podatke iz podatkovnega okvira `Orthodont` v paketu `nlme` (Pinheiro in Bates, 2000). Imamo podatke za 27 otrok (16 fantov in 11 deklet), ki so jih spremljali od starosti 8 do 14 let. Vsaki dve leti so jim izmerili razdaljo med dvema točkama lobanje, ki sta dobro določljivi na rentgenskem posnetku (`distance`), ta razdalja je ortodontsko zanimiva. Zanima nas `distance` v odvisnosti od starosti (`age`) in spola (`Sex`) otrok.

```
> library(nlme)
> head(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
5     21.5   8     M02 Male
6     22.5  10     M02 Male

> class(Orthodont)

[1] "nfnGroupedData" "nfGroupedData"  "groupedData"    "data.frame"
```

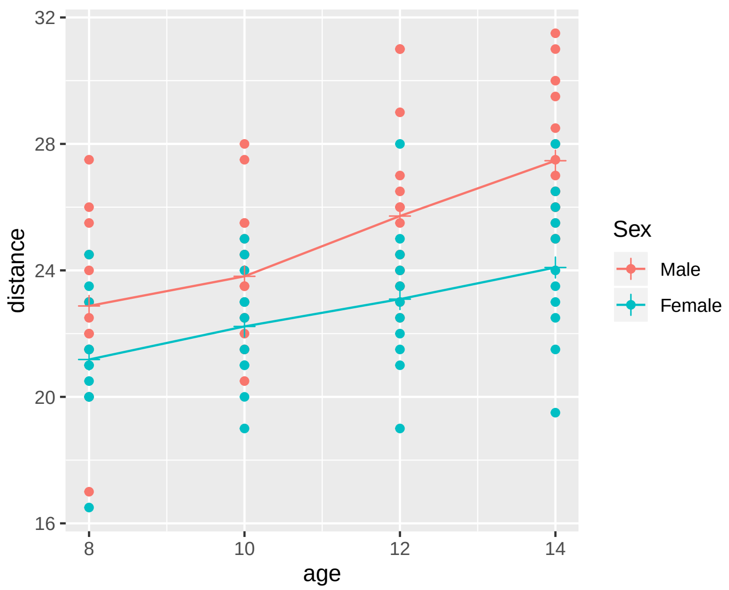
Podatkovni okvir `Orthodont` je vrste `data.frame` in hkrati vrste `groupedData`, kar pomeni, da je med podatki definirana struktura, ki poveže podatke za `distance` v odvisnosti od `age` za istega otroka: `distance` $\sim$ `age` | `Subject`. Ta podatkovni okvir je hkrati tudi vrste `nfGroupedData`, kar pomeni, da je odzivna spremenljivka `distance` številska, grupiranje podatkov je narejeno na podlagi ene opisne spremenljivke `Subject`; hkrati je tudi vrste `nfnGroupedData`, kar pomeni, da je napovedna spremenljivka `age` številska.

2.1 Linearni model s fiksnimi vplivi

Zanima nas vpliv spola, starosti in njune interakcije na `distance`.

Analizo podatkov bomo začeli tako, da dejstva, da so podatki v času izmerjeni na istem otroku, ne bomo upoštevali. Tako podatke grafično prikazuje Slika 1. Vidimo, da lahko za prvi približek privzamemo, da `distance` s časom linearno narašča, slika kaže tudi vpliv spola.

```
> library(ggplot2)
> library(dplyr)
> # pripravimo podatke za grafični prikaz povprečij po skupinah določenih glede na starost in s
> gd <- Orthodont %>% group_by(age,Sex) %>% summarise(distance = mean(distance))
> gd

# A tibble: 8 x 3
# Groups:   age [4]
    age Sex    distance
  <dbl> <fct>     <dbl>
1     8 Male       22.9
2     8 Female     21.2
3    10 Male       23.8
4    10 Female     22.2
5    12 Male       25.7
6    12 Female     23.1
7    14 Male       27.5
8    14 Female     24.1

> ggplot(data=as.data.frame(Orthodont), aes(x=age, y=distance, col=Sex)) +
+   geom_point() + geom_point(data = gd, size = 3, shape = 3) +
+   geom_line(data=gd)
```



Slika 1: Odvisnost `distance` od `age` glede na `Sex`, lomljeni črti povezujeta povprečja po `age`

Podatke za `age` bomo centrirali, od vseh starosti bomo odšteli povprečje, torej `age-11`. Razloga sta dva, interpretacija parametra β_0 na centriranih podatkih za `age` je smiselna, saj ocenjujemo povprečje za `distance` pri povprečni starosti; ta parameter ni koreliran z naklonom, kar pomeni, da se s centriranjem v splošnem znebimo tudi korelacije med ocenama parametrov β_0 in β_1.

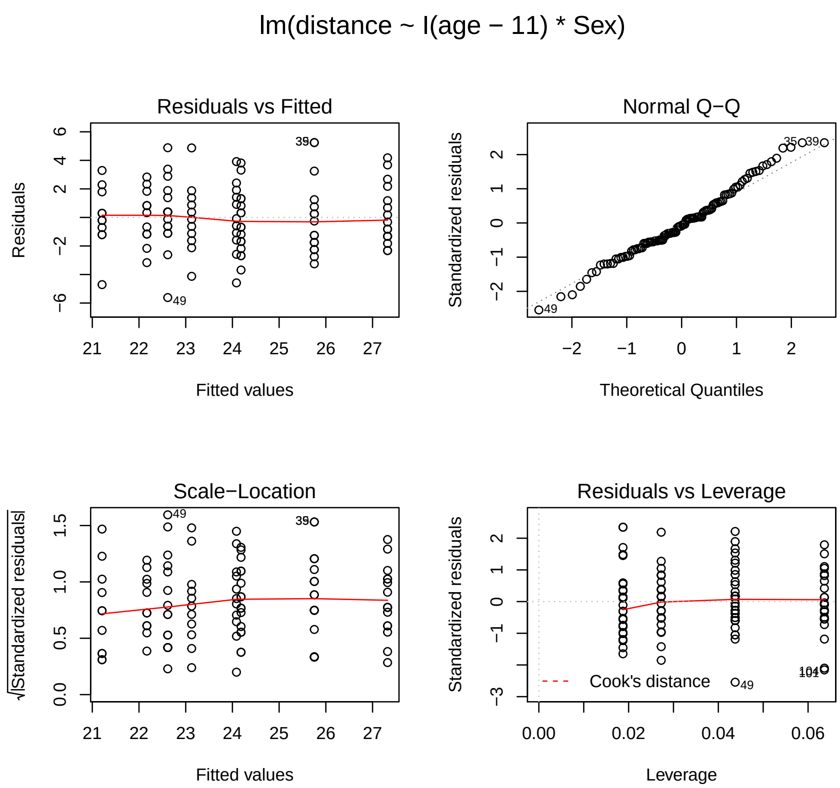
Za ilustracijo, kako centriranje vpliva na variančno-kovariančno matriko ocen parametrov, naredimo model na necentriranih in centriranih podatkih za `age`:

```
> mod.lm0<-lm(distance~age*Sex, data=Orthodont)
> round(vcov(mod.lm0), 2)
```

```
              (Intercept)   age SexFemale age:SexFemale
(Intercept)          2.01 -0.18     -2.01          0.18
age                 -0.18  0.02      0.18         -0.02
SexFemale           -2.01  0.18      4.92         -0.43
age:SexFemale        0.18 -0.02     -0.43          0.04
```

```
> mod.lm<-lm(distance~I(age-11)*Sex, data=Orthodont)
> round(vcov(mod.lm), 2)

                     (Intercept) I(age - 11) SexFemale I(age - 11):SexFemale
(Intercept)                 0.08        0.00     -0.08                  0.00
I(age - 11)                 0.00        0.02      0.00                 -0.02
SexFemale                  -0.08        0.00      0.20                  0.00
I(age - 11):SexFemale       0.00       -0.02      0.00                  0.04

> summary(mod.lm)

Call:
lm(formula = distance ~ I(age - 11) * Sex, 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)            24.9687     0.2821  88.504  < 2e-16 ***
I(age - 11)             0.7844     0.1262   6.217 1.07e-08 ***
SexFemale              -2.3210     0.4420  -5.251 8.05e-07 ***
I(age - 11):SexFemale  -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
```

Rezultati za `mod.lm` kažejo, da je interakcija med `age` in `Sex` neznačilna ($p = 0.126$), vpliva spola in starosti sta statistično značilna. Z modelom je pojasnjene 42.2 % variabilnosti odzivne spremenljivke.

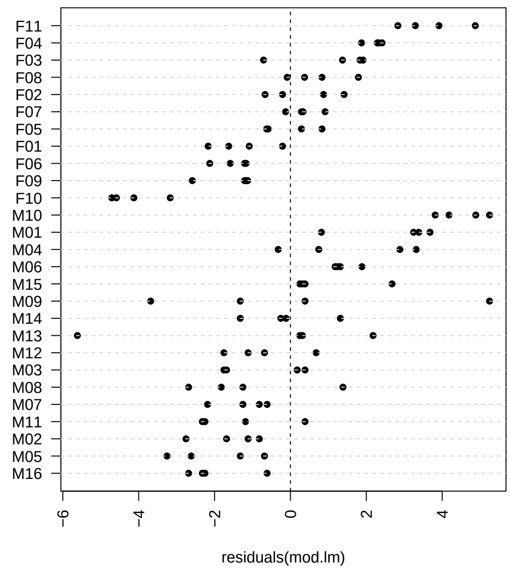


Slika 2: Ostanki za `mod.lm`

V ostankih modela `mod.lm` na Sliki 2 se pokaže rahlo odstopanje od normalne porazdelitve v zgornjem repu porazdelitve, sicer je model sprejemljiv. Poglejmo ostanke še drugače - narišimo ostanke za vsako osebo posebej. Najprej bomo uporabili funkcijo `stripchart` (Slika 3). Malo drugačen, vendar v tem primeru preglednejši grafični prikaz, naredimo s funkcijo `bwplot` iz paketa `lattice` (Slika 4). Opomba: okvirji z ročaji so narisani samo na štirih podatkih, vendar je Slika 4 bolj nazorna kot Slika 3.

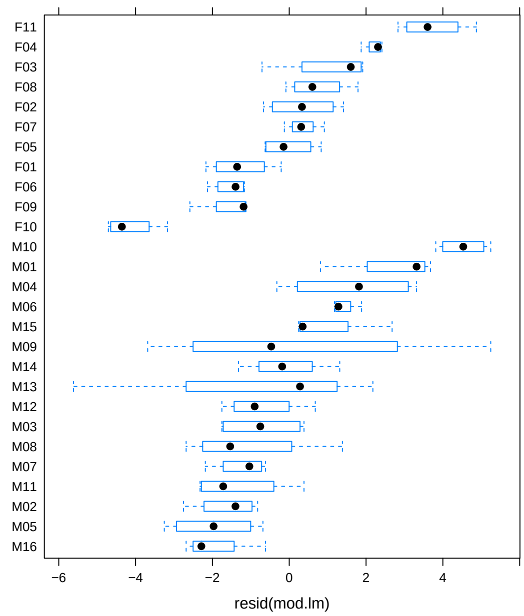
Na Slikah 3 in 4 vidimo, da so ostanki odvisni od osebe, kar pomeni, da `mod.lm` ni sprejemljiv. V okviru `lm` modelov odpravljanje te pomanjkljivosti ni mogoče.

```
> stripchart(residuals(mod.lm) ~ Subject, vertical = F, pch=16, las=2, data = Orthodont)
> abline(v=0, lty=2); abline(h=c(1:27), col="lightgrey", lty=2)
```



Slika 3: Ostanki za `mod.lm` razdeljeni v skupine po `Subject`

```
> library(lattice)
> bwplot(getGroups(Orthodont)~resid(mod.lm))
```

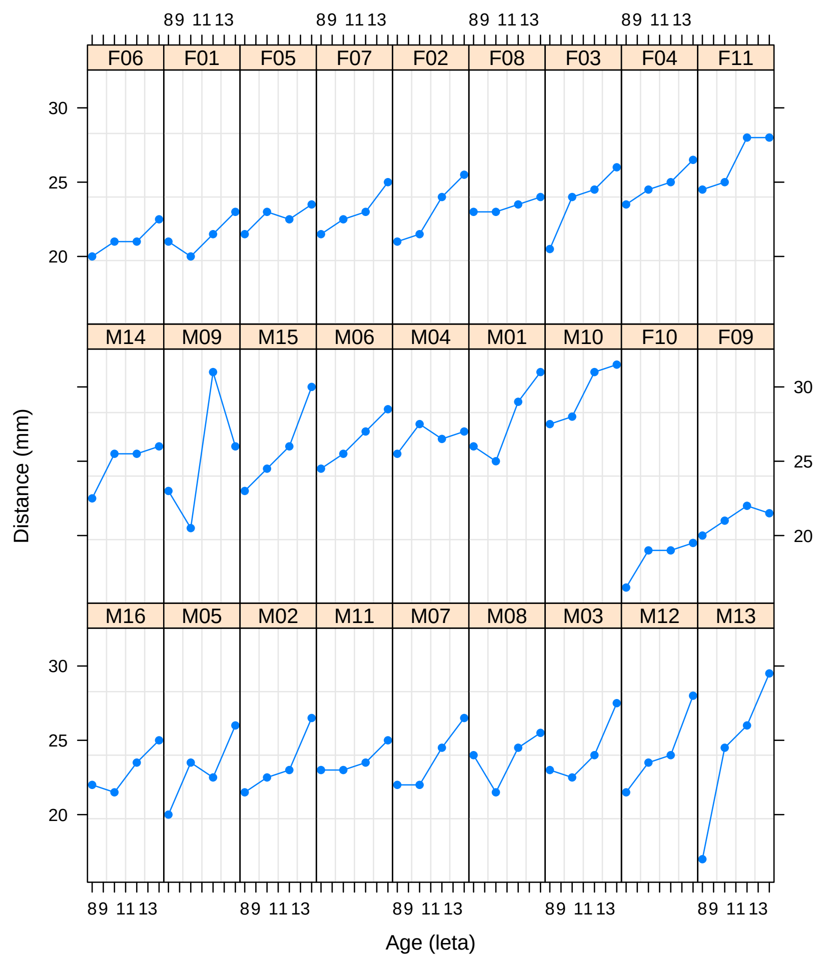


Slika 4: Okvirji z ročaji za ostanke za `mod.lm` razdeljeni v skupine po `Subject`

2.2 Linearni model za vsako osebo posebej

Narišimo odvisnost `distance` od `age` za vsako osebo posebej (Slika 5). Ker je podatkovni okvir `Orthodont` vrste `nfnGroupedData`, ustrezno sliko vrne kar funkcija `plot`.

```
> plot(Orthodont, ylab="Distance (mm)", xlab="Age (leta)", pch=16)
```



Slika 5: Odvisnost `distance` od `age` za vsako osebo (`Subject`) posebej

Grafi na Sliki 5 kažejo, da je linearna odvisnost `distance` od `age` sprejemljiva. Vizualno se presečišča in nakloni precej razlikujejo med osebami. Med podatki so za vsako osebo le štiri meritve v času, kar je za modeliranje odvisnosti od časa zelo malo.

Da bomo v nadaljevanju lažje razumeli vključevanje slučajnih vplivov v model, naredimo model za vsako osebo posebej. V paketu `nlme` uporabimo funkcijo `lmList`, ki za podatkovne okvire vrste `groupedData` izračuna modele za vsako skupino podatkov; v primeru `Orthodont` so skupine določene s spremenljivko `Subject`.

```
> mod.list<-lmList(distance~I(age-11), data=Orthodont)
> summary(mod.list)

Call:
  Model: distance ~ I(age - 11) | Subject
```

```
   Data: Orthodont

Coefficients:
   (Intercept)
    Estimate Std. Error  t value     Pr(>|t|)
M16   23.000  0.6550198 35.11344 7.229908e-39
M05   23.000  0.6550198 35.11344 7.229908e-39
M02   23.375  0.6550198 35.68594 3.127804e-39
M11   23.625  0.6550198 36.06761 1.801423e-39
M07   23.750  0.6550198 36.25845 1.369868e-39
M08   23.875  0.6550198 36.44928 1.043080e-39
M03   24.250  0.6550198 37.02178 4.641294e-40
M12   24.250  0.6550198 37.02178 4.641294e-40
M13   24.250  0.6550198 37.02178 4.641294e-40
M14   24.875  0.6550198 37.97595 1.234662e-40
M09   25.125  0.6550198 38.35762 7.332650e-41
M15   25.875  0.6550198 39.50262 1.580399e-41
M06   26.375  0.6550198 40.26596 5.812844e-42
M04   26.625  0.6550198 40.64763 3.548813e-42
M01   27.750  0.6550198 42.36513 4.059867e-43
M10   29.500  0.6550198 45.03681 1.633487e-44
F10   18.500  0.6550198 28.24342 5.063500e-34
F09   21.125  0.6550198 32.25093 5.809918e-37
F06   21.125  0.6550198 32.25093 5.809918e-37
F01   21.375  0.6550198 32.63260 3.173132e-37
F05   22.625  0.6550198 34.54094 1.692434e-38
F07   23.000  0.6550198 35.11344 7.229908e-39
F02   23.000  0.6550198 35.11344 7.229908e-39
F08   23.375  0.6550198 35.68594 3.127804e-39
F03   23.750  0.6550198 36.25845 1.369868e-39
F04   24.875  0.6550198 37.97595 1.234662e-40
F11   26.375  0.6550198 40.26596 5.812844e-42
   I(age - 11)
    Estimate Std. Error   t value     Pr(>|t|)
M16    0.550  0.2929338 1.8775576 6.584707e-02
M05    0.850  0.2929338 2.9016799 5.361639e-03
M02    0.775  0.2929338 2.6456493 1.065760e-02
M11    0.325  0.2929338 1.1094659 2.721458e-01
M07    0.800  0.2929338 2.7309929 8.511442e-03
M08    0.375  0.2929338 1.2801529 2.059634e-01
M03    0.750  0.2929338 2.5603058 1.328807e-02
M12    1.000  0.2929338 3.4137411 1.222240e-03
M13    1.950  0.2929338 6.6567951 1.485652e-08
M14    0.525  0.2929338 1.7922141 7.870160e-02
M09    0.975  0.2929338 3.3283976 1.577941e-03
M15    1.125  0.2929338 3.8404587 3.247135e-04
M06    0.675  0.2929338 2.3042752 2.508117e-02
M04    0.175  0.2929338 0.5974047 5.527342e-01
```

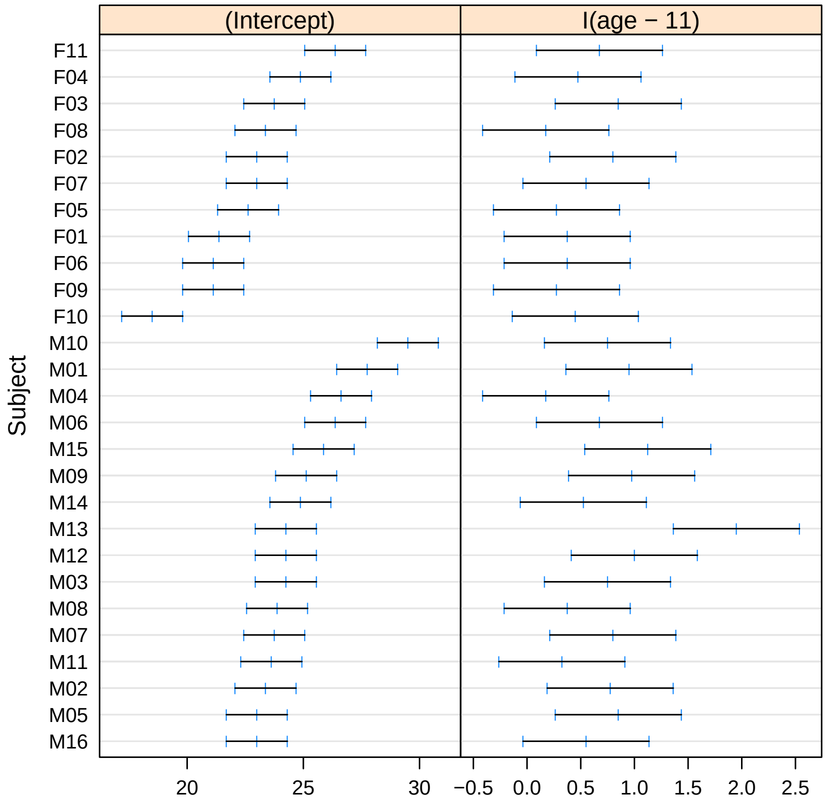
```
M01    0.950  0.2929338 3.2430540 2.030113e-03
M10    0.750  0.2929338 2.5603058 1.328807e-02
F10    0.450  0.2929338 1.5361835 1.303325e-01
F09    0.275  0.2929338 0.9387788 3.520246e-01
F06    0.375  0.2929338 1.2801529 2.059634e-01
F01    0.375  0.2929338 1.2801529 2.059634e-01
F05    0.275  0.2929338 0.9387788 3.520246e-01
F07    0.550  0.2929338 1.8775576 6.584707e-02
F02    0.800  0.2929338 2.7309929 8.511442e-03
F08    0.175  0.2929338 0.5974047 5.527342e-01
F03    0.850  0.2929338 2.9016799 5.361639e-03
F04    0.475  0.2929338 1.6215270 1.107298e-01
F11    0.675  0.2929338 2.3042752 2.508117e-02

Residual standard error: 1.31004 on 54 degrees of freedom
```

V `mod.list` je ocenjenih 54 parametrov za premice in skupna standardna napaka $\widehat{\sigma} = 1.310$ (`Residual standard error`). Ta standardna napaka je precej manjša kot za `mod.lm`, kjer je $\widehat{\sigma} = 2.257$. Intervale zaupanja za β_0 in β_1 za vsako osebo posebej vrne funkcija `intervals`, njihov grafični prikaz je na Sliki 6.

Slika 6 kaže velike razlike med ocenami parametra β_0 (presečišče) po osebah, torej velike razlike v pričakovani vrednosti `distance` pri starosti 11 let med osebami. Veliko intervalov zaupanja za parameter β_0 se ne prekriva. Taki rezultati nakazujejo smiselnost vključitve slučajnega vpliva, ki podaja variabilnost ocene parametra β_0 med osebami.

Intervali zaupanja za β_1 se za vse osebe, razen za M13, prekrivajo, kar kaže na to, da je razlika v naklonih manjša. Zato je vključitev slučajnega vpliva, ki podaja variabilnost ocene parametra β_1 med osebami, morda nepotrebna.

```
> plot(intervals(mod.list))
```



Slika 6: Intervali zaupanja za parametre modelov odvisnosti `distance` od `age` za vsakega od 16 fantov in 11 deklet

2.3 Mešani model s slučajnim vplivom na presečišče in naklon

Naredili bomo mešani model, v katerega bomo najprej vključili fiksna vpliva `Sex` in `age-11` ter njuno interakcijo, in dva slučajna vpliva: slučajni vpliv osebe (`Subject`) na presečišče in slučajni vpliv osebe na naklon. Za i-to osebo, $i = 1, ..., S = 27$, v času t, $t = 1, ..., 4$, zapišemo model takole:

$$distance_{it} = \beta_0 + \beta_1 Sex_i + \beta_2(age_{it} - 11) + \beta_3 Sex_i(age_{it} - 11) + u_{i0} + u_{i1}(age_{it} - 11) + \varepsilon_{it}$$

ali pa

$$distance_{it} = (\beta_0 + u_{i0}) + \beta_1 Sex_i + (\beta_2 + u_{i1})(age_{it} - 11) + \beta_3 Sex_i(age_{it} - 11) + \varepsilon_{it}, \qquad (2)$$

$$\mathbf{u_i} \sim \mathbf{N}_2(\mathbf{0}, \mathbf{\Psi}), \quad \boldsymbol{\varepsilon_i} \sim \mathbf{N}_4(\mathbf{0}, \sigma^2).$$

V enačbi (2) u_{i0} predstavlja slučajni vpliv osebe na presečišče, u_{i1} pa slučajni vpliv osebe na naklon. V tem primeru je variančno-kovariančna matrika slučajnih vplivov dimenzije 2×2:

$$\mathbf{\Psi} = \begin{pmatrix} \sigma^2_{u_0} & Cov(u_0, u_1) \\ Cov(u_0, u_1) & \sigma^2_{u_1} \end{pmatrix}. \qquad (3)$$

V LMM poleg parametrov $\boldsymbol{\beta}$ in σ ocenjujemo tudi varianci slučajnih vplivov $\sigma^2_{u_0}$ in $\sigma^2_{u_1}$, kovarianco slučajnih vplivov $Cov(u_0, u_1)$ ter slučajne vplive u_{i0} in u_{i1}, $i = 1, ..., S = 27$. Za napake najprej predpostavimo, da so neodvisne med sabo. Prav tako so medsebojno neodvisni slučajni vplivi med skupinami. Predpostavimo tudi, da so napake in slučajni vplivi skupin neodvisni.

Za izračun ocen parametrov LMM uporabimo funkcijo `lme` iz paketa `nlme`. Slučajna vpliva definiramo z argumentom `random=~I(age-11)|Subject`, kar pomeni vključitev dveh slučajnih vplivov, slučajni vpliv osebe na presečišče in slučajni vpliv osebe na naklon.

```
> mod.lme1<-lme(distance~I(age-11)*Sex, random=~I(age-11)|Subject,
+               data=Orthodont, method="REML")
> summary(mod.lme1)

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) 1.8303267 (Intr)
I(age - 11) 0.1803454 0.206
Residual    1.3100397
```

```
Fixed effects: distance ~ I(age - 11) * Sex
                          Value Std.Error DF  t-value p-value
(Intercept)           24.968750 0.4860007 79 51.37596  0.0000
I(age - 11)            0.784375 0.0859995 79  9.12069  0.0000
SexFemale             -2.321023 0.7614168 25 -3.04829  0.0054
I(age - 11):SexFemale -0.304830 0.1347353 79 -2.26243  0.0264
 Correlation:
                      (Intr) I(g-11) SexFml
I(age - 11)            0.102
SexFemale             -0.638 -0.065
I(age - 11):SexFemale -0.065 -0.638   0.102

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.168078485 -0.385939135  0.007103929  0.445154686  3.849463229

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

Ocene parametrov se ob osnovni nastavitvi funkcije izračunajo po metodi omejenega največjega verjetja (REML). Izpis za `lme` model v prvem delu poda vrednosti kriterijskih funkcij AIC in BIC ter logaritem verjetja. V drugem delu `Random effects` dobimo ocene za variančno-kovariančno matriko slučajnih vplivov $\hat{\boldsymbol{\Psi}}$, izražene s standardnimi odkloni: $\hat{\sigma}_{u_0} = 1.830$, $\hat{\sigma}_{u_1} = 0.180$ in $\widehat{Cor}(u_0, u_1) = 0.206$. Ocena za standardni odklon napak je $\hat{\sigma} = 1.31$, kar je isto kot pri modelu `mod.list`.

V tretjem delu izpisa so ocene parametrov fiksnih vplivov v taki obliki kot pri `lm` modelu. Sledi simetrična matrika korelacijskih koeficientov med ocenami parametrov fiksnih vplivov, izpisani so samo členi pod diagonalo.

V primerjavi z `mod.lm` se v `mod.lme1` pokaže statistično značilna tudi interakcija `(age-11)*Sex` ($p = 0.0264$). To pomeni, da se `distance` pri dekletih spreminja s starostjo drugače kot pri fantih, kar lahko opazimo tudi na Sliki 5.

2.4 Mešani model s slučajnim vplivom na presečišče

Ocena standardnega odklona za slučajni vpliv osebe na naklon premice je za en velikostni red manjša kot ocena standardnega odklona slučajnega vpliva na presečišče, zato bomo preverili, ali je slučajni vpliv osebe na naklon statistično pomemben. Naredimo model `mod.lme2` s samo enim slučajnim vplivom: oseba vpliva preko slučajnega vpliva samo na presečišče. V tem primeru bo argument za slučajni vpliv `random=` $\sim 1|$`Subject`. Fiksni vplivi v modelu so isti, zato parametre modela ponovno ocenimo po metodi omejenega največjega verjetja REML.

```
> mod.lme2<-lme(distance~I(age-11)*Sex, random=~1|Subject,data=Orthodont)
```

S primerjavo modelov `mod.lme1` in `mod.lme2` testiramo ničelno domnevo, da je varianca slučajnega vpliva osebe na naklon enaka 0.

```
> anova(mod.lme1, mod.lme2)

         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.lme1     1  8 448.5817 469.7368 -216.2908
mod.lme2     2  6 445.7572 461.6236 -216.8786 1 vs 2 1.175588  0.5556
```

Pokaže se, da slučajnega vpliva osebe na naklon v modelu ni potrebno upoštevati ($p = 0.5556$). Torej naprej analiziramo `mod.lme2`.

```
> summary(mod.lme2)

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

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

Fixed effects: distance ~ I(age - 11) * Sex
                          Value Std.Error DF  t-value p-value
(Intercept)           24.968750 0.4860008 79 51.37595  0.0000
I(age - 11)            0.784375 0.0775011 79 10.12082  0.0000
SexFemale             -2.321023 0.7614168 25 -3.04829  0.0054
I(age - 11):SexFemale -0.304830 0.1214209 79 -2.51052  0.0141
 Correlation:
                      (Intr) I(g-11) SexFml
I(age - 11)            0.000
SexFemale             -0.638  0.000
I(age - 11):SexFemale  0.000 -0.638   0.000

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
```

Poglejmo intervale zaupanja za ocenjene parametre in jih primerjajmo z intervali zaupanja za `mod.lm`, ki ne upošteva slučajnega vpliva osebe na presečišče.

```
> intervals(mod.lme2)

Approximate 95% confidence intervals
```

```
 Fixed effects:
                           lower       est.      upper
(Intercept)           24.0013897 24.9687500 25.9361103
I(age - 11)            0.6301129  0.7843750  0.9386371
SexFemale             -3.8891901 -2.3210227 -0.7528554
I(age - 11):SexFemale -0.5465118 -0.3048295 -0.0631473
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                   lower     est.    upper
sd((Intercept)) 1.321019 1.816214 2.497038

 Within-group standard error:
   lower     est.    upper
1.186236 1.386382 1.620298
```

```
> confint(mod.lm)
```

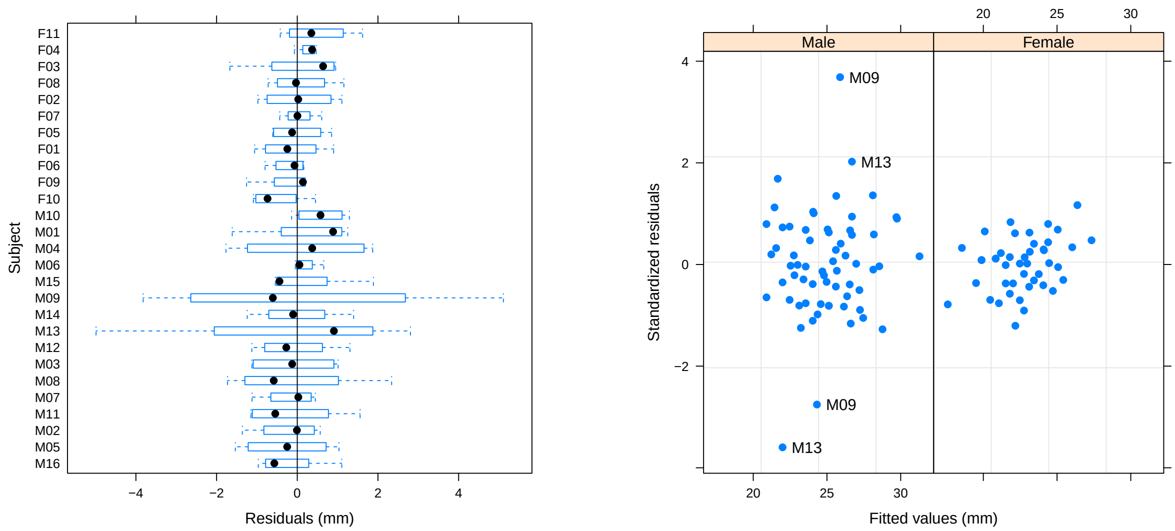
```
                           2.5 %      97.5 %
(Intercept)           24.4092982 25.52820179
I(age - 11)            0.5341806  1.03456945
SexFemale             -3.1975152 -1.44453022
I(age - 11):SexFemale -0.6968089  0.08714982
```

Ključna razlika je opazna pri `I(age - 11)`; interval zaupanja za naklon je za `mod.lme2` precej ožji kot za `mod.lm`.

Ostanke za `mod.lme2` po osebah prikazuje Slika 7. Ostanki za vse osebe so zdaj porazdeljeni okoli vrednosti 0, videti je tudi, da je njihova variabilnost pri fantih večja kot pri dekletih.

```
> plot(mod.lme2, Subject~resid(.), abline=0,pch=16)
```

```
> plot(mod.lme2, resid(.,type="pearson")~fitted(.)|Sex,id=0.05,adj=-0.3,pch=16)
```



(a) (b)

Slika 7: Ostanki za `model.lme2`

2.5 Modeliranje heteroskedastičnosti v LMM

Slika 7 kaže različno variabilnost med ostanki za dekleta in za fante. To heteroskedastičnost bomo poskusili odpraviti z modeliranjem variančne matrike napak tako, da uporabimo variančno strukturo `varIdent(form=∼1|Sex)`. Variančno-kovariančna matrika napak $\mathbf{\Lambda}$ je diagonalna matrika z dvema različnima vrednostma, ena je varianca za fante in druga za dekleta. Model dopolnimo z dodatnim argumentom `weights` na enak način kot pri `gls` modelih.

```
> mod.lme3<-lme(distance~I(age-11)*Sex, random=~1|Subject,
+                 weights=varIdent(form=~1|Sex), data=Orthodont)
> summary(mod.lme3)

Linear mixed-effects model fit by REML
 Data: Orthodont
       AIC      BIC    logLik
  429.2205 447.7312 -207.6102

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:     1.84757 1.669823

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | Sex
 Parameter estimates:
     Male    Female
1.0000000 0.4678944
Fixed effects: distance ~ I(age - 11) * Sex
```

```
                          Value Std.Error DF  t-value p-value
(Intercept)           24.968750 0.5068650 79 49.26115  0.0000
I(age - 11)            0.784375 0.0933459 79  8.40288  0.0000
SexFemale             -2.321023 0.7623026 25 -3.04475  0.0054
I(age - 11):SexFemale -0.304830 0.1071828 79 -2.84402  0.0057
 Correlation:
                      (Intr) I(g-11) SexFml
I(age - 11)            0.000
SexFemale             -0.665  0.000
I(age - 11):SexFemale  0.000 -0.871   0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.00556474 -0.63419474  0.01890475  0.55016878  3.06446971

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

```
> anova(mod.lme2,mod.lme3)
```

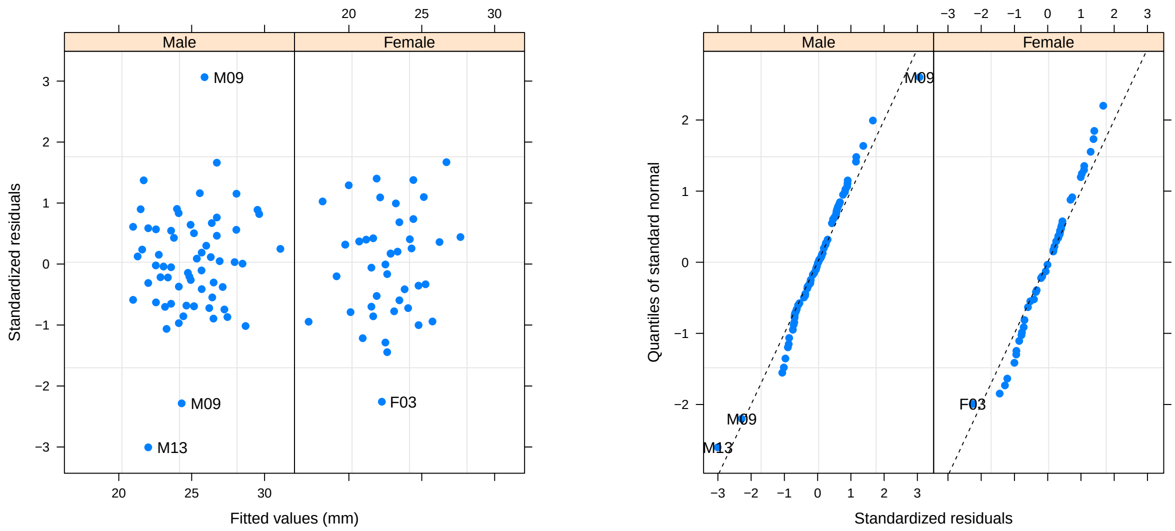
```
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.lme2     1  6 445.7572 461.6236 -216.8786
mod.lme3     2  7 429.2205 447.7312 -207.6102 1 vs 2 18.53677  <.0001
```

Ocenjeno razmerje med standardnim odklonom napak za dekleta in standardnim odklonom napak za fante je 0.468 in `mod.lme3` se pri testiranju pokaže za boljšega od `mod.lme2` ($p < 0.0001$).

Grafična prikaza na Sliki 8 kažeta ustrezno porazdelitev ostankov.

```
> plot(mod.lme3, resid(.,type="pearson")~fitted(.)|Sex,
+       id=0.05, adj=-0.3, pch=16)
```

```
> qqnorm(mod.lme3, ~resid(.,type="pearson")|Sex, abline=c(0, 1), lty=2,
+        grid=T, id=0.05, adj=0.5, pch=16)
```



Slika 8: Ostanki za `model.lme3` glede na `Sex`

```
> library(car)
> compareCoefs(mod.lme2,mod.lme3)

Calls:
1: lme.formula(fixed = distance ~ I(age - 11) * Sex, data = Orthodont,
  random = ~1 | Subject)
2: lme.formula(fixed = distance ~ I(age - 11) * Sex, data = Orthodont,
  random = ~1 | Subject, weights = varIdent(form = ~1 | Sex))

                      Model 1 Model 2
(Intercept)            24.969  24.969
SE                      0.486   0.507

I(age - 11)            0.7844  0.7844
SE                     0.0775  0.0933

SexFemale              -2.321  -2.321
SE                      0.761   0.762

I(age - 11):SexFemale  -0.305  -0.305
SE                      0.121   0.107
```

Ocene parametrov fiksnih vplivov se ne spremenijo, malo pa se povečajo njihove standardne napake.

Ocene slučajnega vpliva posameznika na vrednost presečišča $\mathbf{u_0}$ izpišemo s funkcijo `random.effects`.

```
> u0<-random.effects(mod.lme3)
> u0
```

```
    (Intercept)
M16 -1.63488808
M05 -1.63488808
M02 -1.32348083
M11 -1.11587599
M07 -1.01207357
M08 -0.90827115
M03 -0.59686390
M12 -0.59686390
M13 -0.59686390
M14 -0.07785181
M09  0.12975302
M15  0.75256753
M06  1.16777720
M04  1.37538203
M01  2.30960379
M10  3.76283764
F10 -3.97023056
F09 -1.45756410
F06 -1.45756410
F01 -1.21826253
F05 -0.02175469
F07  0.33719766
F02  0.33719766
F08  0.69615002
F03  1.05510237
F04  2.13195942
F11  3.56776883
```

```
> mean(u0[,1])
```

```
[1] -1.422136e-15
```

```
> sd(u0[,1])
```

```
[1] 1.700613
```

Povprečje slučajnih vplivov na presečišče $\mathbf{u_0}$ je 0 in njihov standardni odklon je 1.70, kar je zelo blizu oceni za standardni odklon tega slučajnega vpliva izračunani po metodi največjega verjetja (`Random effects, Residual = 1.67`).

2.6 Testiranje domnev

Obrazložimo vpliv starosti in spola na `distance` na podlagi končnega modela `mod.lme3`. Testiramo štiri ničelne domneve: H_0: povprečje distance pri starosti 11 let je pri moških in ženskah enako, H_0: naklon pri moških je 0, H_0: naklon pri ženskah je 0 in H_0: naklon je pri moških in ženskah enak.

```
> library(multcomp)
> names(coefficients(mod.lme3))

[1] "(Intercept)"           "I(age - 11)"           "SexFemale"
[4] "I(age - 11):SexFemale"

> C<-rbind(c(0,0,1,0),c(0,1,0,0),c(0,1,0,1),c(0,0,0,1))
> rownames(C)<-c("presec.Female-presec.Male", "naklon.Male",
+                "naklon.Female","naklon.Female-naklon.Male")
> test<-glht(mod.lme3, linfct=C)
> summary(test)

         Simultaneous Tests for General Linear Hypotheses

Fit: lme.formula(fixed = distance ~ I(age - 11) * Sex, data = Orthodont,
    random = ~1 | Subject, weights = varIdent(form = ~1 | Sex))

Linear Hypotheses:
                               Estimate Std. Error z value Pr(>|z|)
presec.Female-presec.Male == 0 -2.32102    0.76230  -3.045  0.00796 **
naklon.Male == 0                0.78438    0.09335   8.403  < 0.001 ***
naklon.Female == 0              0.47955    0.05268   9.104  < 0.001 ***
naklon.Female-naklon.Male == 0 -0.30483    0.10718  -2.844  0.01576 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> confint(test)

         Simultaneous Confidence Intervals

Fit: lme.formula(fixed = distance ~ I(age - 11) * Sex, data = Orthodont,
    random = ~1 | Subject, weights = varIdent(form = ~1 | Sex))

Quantile = 2.4363
95% family-wise confidence level


Linear Hypotheses:
                               Estimate lwr     upr
presec.Female-presec.Male == 0 -2.3210  -4.1782 -0.4638
naklon.Male == 0                0.7844   0.5570  1.0118
naklon.Female == 0              0.4795   0.3512  0.6079
naklon.Female-naklon.Male == 0 -0.3048  -0.5660 -0.0437
```

Pri starosti 11 let je pri ženskah povprečje `distance` za 2.32 enot manjše kot pri moških (95 % interval zaupanja je 0.46, 4.18). Na leto se pri moških povprečje `distance` poveča za 0.78 enot (95 % interval zaupanja je 0.56, 1.01), pri ženskah pa za 0.48 enot (95 % interval zaupanja je 0.35, 0.61), razlika v naklonih je statistično značilna.

2.7 Napovedane vrednosti

Z `lme` modelom dobimo dve vrsti napovedanih vrednosti, **napovedi za populacijo**:

$$\widehat{distance}_{it} = \hat{\beta}_0 + \hat{\beta}_1 Sex_i + \hat{\beta}_2(age_{it} - 11) + \hat{\beta}_3 Sex_i(age_{it} - 11) \tag{4}$$

in **napovedi za posamezno skupino** (v našem primeru za vsako osebo):

$$\widehat{distance}_{it} = (\hat{\beta}_0 + \hat{u}_{i0}) + \hat{\beta}_1 Sex_i + \hat{\beta}_2(age_{it} - 11) + \hat{\beta}_3 Sex_i(age_{it} - 11). \tag{5}$$

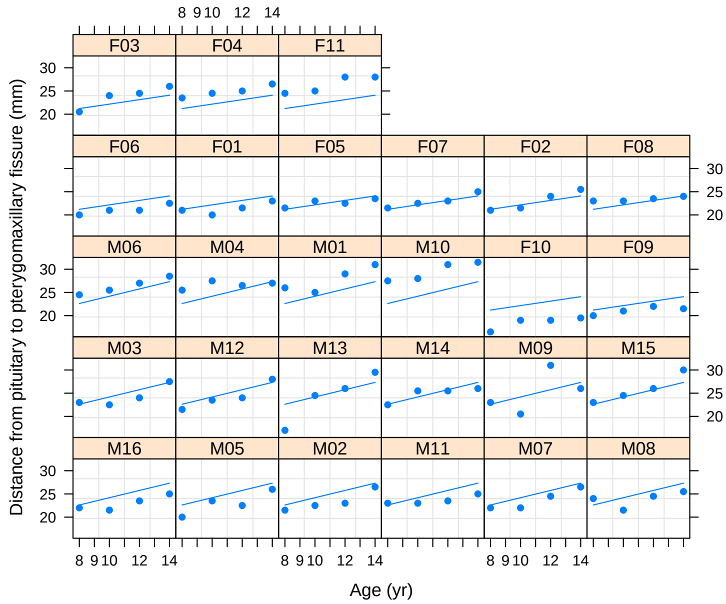
V funkciji `fitted` z argumentom `levels` določimo, za katere napovedi gre; `level=0` določi napovedi za populacijo in `level=1` za posamezne osebe. Poglejmo izpis za napovedi za prvi dve in zadnji dve osebi.

```
> cbind(Orthodont, round(fitted(mod.lme3, level=0:1),1))[1:8,]

  distance age Subject  Sex fixed Subject
1     26.0   8     M01 Male  22.6    24.9
2     25.0  10     M01 Male  24.2    26.5
3     29.0  12     M01 Male  25.8    28.1
4     31.0  14     M01 Male  27.3    29.6
5     21.5   8     M02 Male  22.6    21.3
6     22.5  10     M02 Male  24.2    22.9
7     23.0  12     M02 Male  25.8    24.4
8     26.5  14     M02 Male  27.3    26.0

> cbind(Orthodont, round(fitted(mod.lme3, level=0:1),1))[101:108,]

    distance age Subject    Sex fixed Subject
101     16.5   8     F10 Female  21.2    17.2
102     19.0  10     F10 Female  22.2    18.2
103     19.0  12     F10 Female  23.1    19.2
104     19.5  14     F10 Female  24.1    20.1
105     24.5   8     F11 Female  21.2    24.8
106     25.0  10     F11 Female  22.2    25.7
107     28.0  12     F11 Female  23.1    26.7
108     28.0  14     F11 Female  24.1    27.7
```

Napovedi za populacijo (`fixed`) so za vse fante enake, prav tako za vsa dekleta, je pa razlika med spoloma; napovedi za `Subject` pa se med osebami razlikujejo.

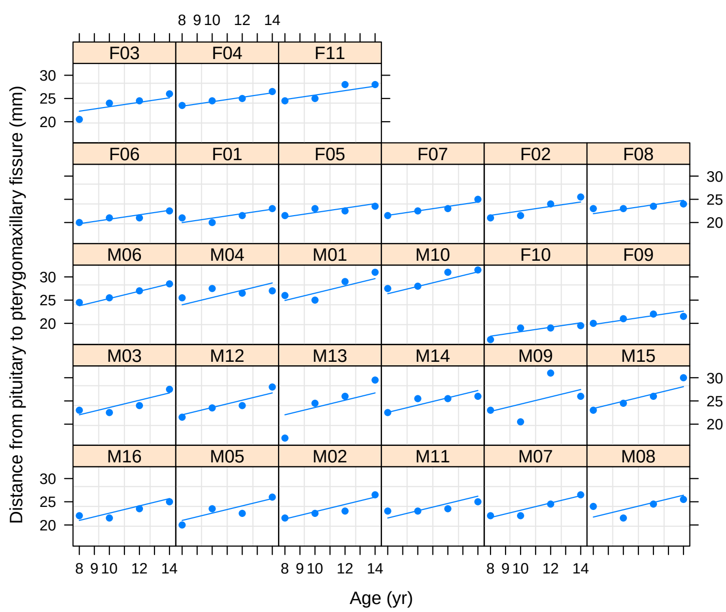
Napovedane vrednosti za `mod.lme3` za populacijo kaže Slika 9, za posamezno osebo pa Slika 10.

```
> plot(augPred(mod.lme3, level=0), grid=T, pch=16)
```



Slika 9: Napovedane vrednosti za populacijo za `mod.lme3`

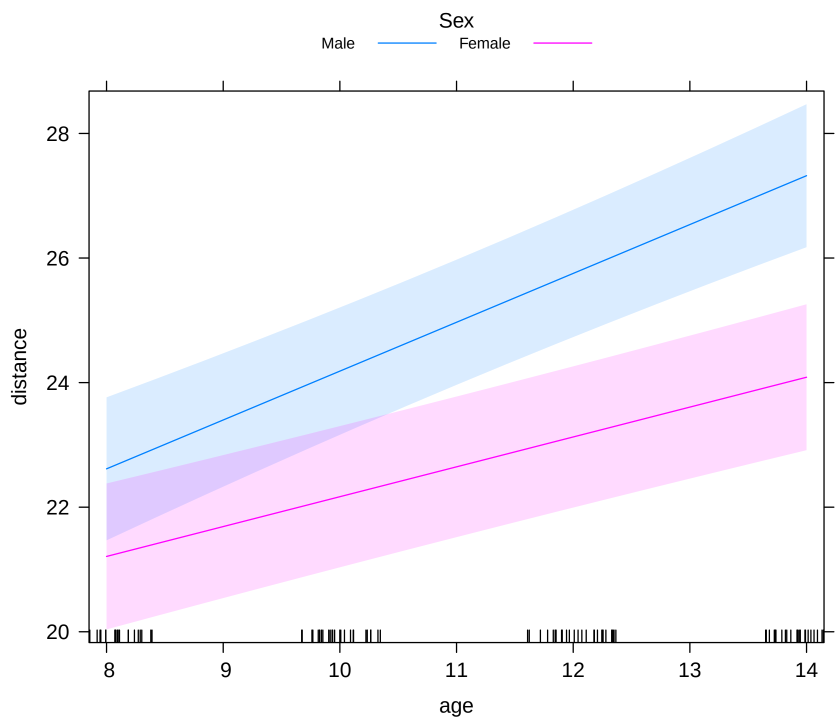
```
> plot(augPred(mod.lme3, level=1), pch=16, grid=T)
```



Slika 10: Napovedane vrednosti za posameznika za `mod.lme3`

Na Sliki 9 sta samo dve različni premici, ena je za vse fante in druga za vsa dekleta, vidna so odstopanja posameznika od populacije. Na Sliki 10 so nakloni za fante enaki, prav tako so enaki vsi nakloni za dekleta, presečišča pa so različna (vzporedne premice). Primerjava napovedi za populacijo fantov in deklet se bolj jasno vidi na Sliki 11.

```
> library(effects)
> plot(effect(c("I(age-11)","Sex"), mod.lme3), ci.style="bands", multiline=T, main="")
```



Slika 11: Napovedane vrednosti s pripadajočimi 95 % IZ za povprečno napoved za populacijo za `mod.lme3`

Dodatek: napovedi za `distance` za posameznike M10, M11 in F03 pri starosti `age=13` in `15` (za ti dve starosti nimamo meritev), pa tudi za populacijo toliko starih fantov in deklet, dobimo s funkcijo `predict`.

```
> novaOrth<-data.frame(Subject=rep(c("M10","M11","F03"),c(2,2,2)),
+                      Sex=rep(c("Male","Male","Female"),c(2,2,2)),
+                      age=rep(c(13,15),3))
```

```
> novaOrth

  Subject    Sex age
1     M10   Male  13
2     M10   Male  15
3     M11   Male  13
4     M11   Male  15
5     F03 Female  13
6     F03 Female  15

> predict(mod.lme3, newdata=novaOrth,level=0:1)

  Subject predict.fixed predict.Subject
1     M10      26.53750        30.30034
2     M10      28.10625        31.86909
3     M11      26.53750        25.42162
4     M11      28.10625        26.99037
5     F03      23.60682        24.66192
6     F03      24.56591        25.62101
```

2.8 Alternativni pristop z modeliranjem variančno-kovariančne matrike napak

Modeliranje slučajnega vpliva osebe na presečišče zamenjamo z modeliranjem variančno-kovariančne matrike napak v okviru `gls` modela:

$$distance_{it} = \beta_0 + \beta_1(age_{it} - 11) + \beta_2 Sex_i + \beta_3 Sex_i(age_{it} - 11) + \varepsilon_{it}, \quad i = 1, ..., S = 27, \quad t = 1, ..., 4. \tag{6}$$

$$\boldsymbol{\varepsilon_i} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}_i).$$

V kontekstu mešanih modelov zapišemo variančno-kovariančno matriko napak za i-to osebo takole:

$$\mathbf{\Sigma}_i = \sigma^2 \mathbf{\Lambda}_i = \sigma^2 \mathbf{V_i} \mathbf{C_i} \mathbf{V_i}.$$

V tem primeru imata matriki $\mathbf{V_i}$ in $\mathbf{C_i}$ za i-to osebo naslednjo strukturo:

$$\mathbf{V_i} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 \\ 0 & 0 & \delta_3 & 0 \\ 0 & 0 & 0 & \delta_4 \end{bmatrix}, \tag{7}$$

$$\mathbf{C_i} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}. \tag{8}$$

Predpostavimo, da je variančno-kovariančna matrika $\mathbf{\Lambda}_i$ enaka za vse osebe. Ocenjujemo torej štiri parametre, ki določajo variančno strukturo napak, in šest korelacijskih koeficientov, skupaj 10 parametrov. Taki strukturi korelacijske matrike napak $\mathbf{C_i}$ v kontekstu mešanih modelov pravimo **splošna korelacijska struktura** (*general correlation structure*). V paketu `nlme` tako strukturo povzame funkcija `corSymm`, ki jo uporabimo v argumentu `correlation` pri modeliranju s funkcijama `lme` ali `gls`.

Spremenljivko `distance` bomo s funkcijo `gls` modelirali v odvisnosti od `I(age-11)*Sex` s predpostavko različnih varianc pri različnih starostih in splošno strukturo korelacije napak.

```
> mod.gls1<-gls(distance~I(age-11)*Sex,
+               weights=varIdent(form=~1|age),
+               correlation=corSymm(form=~1|Subject), data=Orthodont)
> summary(mod.gls1)

Generalized least squares fit by REML
  Model: distance ~ I(age - 11) * Sex
  Data: Orthodont
       AIC      BIC    logLik
  452.5468 489.5683 -212.2734

Correlation Structure: General
 Formula: ~1 | Subject
 Parameter estimate(s):
 Correlation:
  1     2     3
2 0.568
3 0.659 0.581
4 0.522 0.725 0.740
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | age
 Parameter estimates:
        8        10        12        14
1.0000000 0.8788793 1.0744596 0.9586878

Coefficients:
                          Value Std.Error  t-value p-value
(Intercept)           24.937123 0.4728666 52.73606  0.0000
I(age - 11)            0.826804 0.0822177 10.05627  0.0000
SexFemale             -2.271743 0.7408396 -3.06644  0.0028
I(age - 11):SexFemale -0.350439 0.1288104 -2.72058  0.0076

 Correlation:
                      (Intr) I(g-11) SexFml
I(age - 11)            0.112
SexFemale             -0.638 -0.072
```

```
I(age - 11):SexFemale -0.072 -0.638   0.112

Standardized residuals:
        Min          Q1         Med          Q3         Max 
-2.34272826 -0.63481506 -0.07904148  0.63772264  2.16523296 

Residual standard error: 2.329213 
Degrees of freedom: 108 total; 104 residual
```

Ocenjeni korelacijski koeficienti napak so zelo podobni med sabo in tudi ocene razmerij med standardnimi odkloni napak ne odstopajo bistveno od 1. Poglejmo še intervale zaupanja za ocenjene parametre.

```
> intervals(mod.gls1)

Approximate 95% confidence intervals

 Coefficients:
                          lower       est.       upper
(Intercept)           23.9994110 24.9371232 25.87483543
I(age - 11)            0.6637629  0.8268037  0.98984450
SexFemale             -3.7408554 -2.2717427 -0.80262996
I(age - 11):SexFemale -0.6058749 -0.3504390 -0.09500314
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
             lower      est.     upper
cor(1,2) 0.2528700 0.5681970 0.7744035
cor(1,3) 0.3840801 0.6589493 0.8265260
cor(1,4) 0.1846617 0.5220393 0.7493509
cor(2,3) 0.2728723 0.5806064 0.7805551
cor(2,4) 0.4827690 0.7249210 0.8640956
cor(3,4) 0.5133557 0.7396207 0.8697391
attr(,"label")
[1] "Correlation structure:"

 Variance function:
       lower      est.    upper
10 0.6372783 0.8788793 1.212075
12 0.8032124 1.0744596 1.437308
14 0.6877138 0.9586878 1.336432
attr(,"label")
[1] "Variance function:"

 Residual standard error:
   lower     est.    upper 
1.766609 2.329213 3.070986 
```

Vsi intervali zaupanja za korelacijske koeficiente se prekrivajo, kar nakazuje na to, da bi v modelu lahko privzeli enostavnejšo korelacijsko strukturo, kjer je korelacija med vsemi časi/starostmi enaka ρ (*compound symmetry*):

$$\mathbf{C_i} = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} \tag{9}$$

Tako korelacijsko strukturo povzame funkcija `corCompSymm`. Tudi ocenjena razmerja standardnih odklonov pri posameznih starostih so blizu 1, vsi intervali zaupanja za razmerje standardnih odklonov vsebujejo vrednost 1, kar kaže, da heteroskedastičnosti po času ni potrebno vključiti v model. Model `mod.gls1` zato poenostavimo tako, da odstranimo modeliranje strukture varianc in korelacije napak modeliramo z `corCompSymm`:

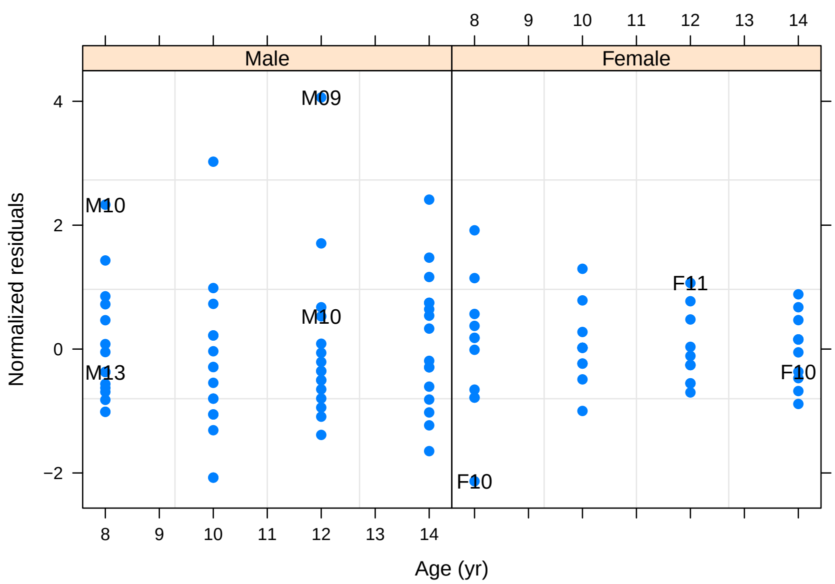
```
> mod.gls2<-gls(distance~I(age-11)*Sex,weights=NULL,
+               correlation=corCompSymm(form=~1|Subject), data=Orthodont)
> summary(mod.gls2)

Generalized least squares fit by REML
  Model: distance ~ I(age - 11) * Sex
  Data: Orthodont
       AIC      BIC    logLik
  445.7572 461.6236 -216.8786

Correlation Structure: Compound symmetry
 Formula: ~1 | Subject
 Parameter estimate(s):
      Rho
0.6318381

Coefficients:
                          Value Std.Error  t-value p-value
(Intercept)           24.968750 0.4860003 51.37600  0.0000
I(age - 11)            0.784375 0.0775011 10.12082  0.0000
SexFemale             -2.321023 0.7614161 -3.04830  0.0029
I(age - 11):SexFemale -0.304830 0.1214209 -2.51052  0.0136

 Correlation:
                      (Intr) I(g-11) SexFml
I(age - 11)            0.000
SexFemale             -0.638  0.000
I(age - 11):SexFemale  0.000 -0.638   0.000

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.45773173 -0.57853118 -0.07360637  0.58204364  2.29634479

Residual standard error: 2.284881
Degrees of freedom: 108 total; 104 residual
```

Primerjava modelov `mod.gls1` in `mod.gls2` pokaže, da je `mod.gls2` zaradi svoje enostavnejše strukture boljši, razlika v številu ocenjenih parametrov je velika, $14 - 6 = 8$.

```
> anova(mod.gls1,mod.gls2)

         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.gls1     1 14 452.5468 489.5683 -212.2734
mod.gls2     2  6 445.7572 461.6236 -216.8786 1 vs 2 9.210449  0.3249
```

Če je model, v katerem modeliramo variančno-kovariančno matriko napak sprejemljiv, velja, da so t. i. **normalizirani ostanki**, $r_i = \hat{\sigma}^{-1}(\hat{\Lambda}_i^{-1/2})^T(y_i - \hat{y}_i)$ porazdeljeni $N(0, \sigma^2 I)$. Zato za diagnostične grafične prikaze uporabimo normalizirane ostanke, kar pomeni, da ima argument `type` v funkciji `resid` vrednost `"n"`:

```
> plot.lme(mod.gls2, resid(.,type="n")~age|Sex,id=0.05,adj=0.5, pch=16)
```

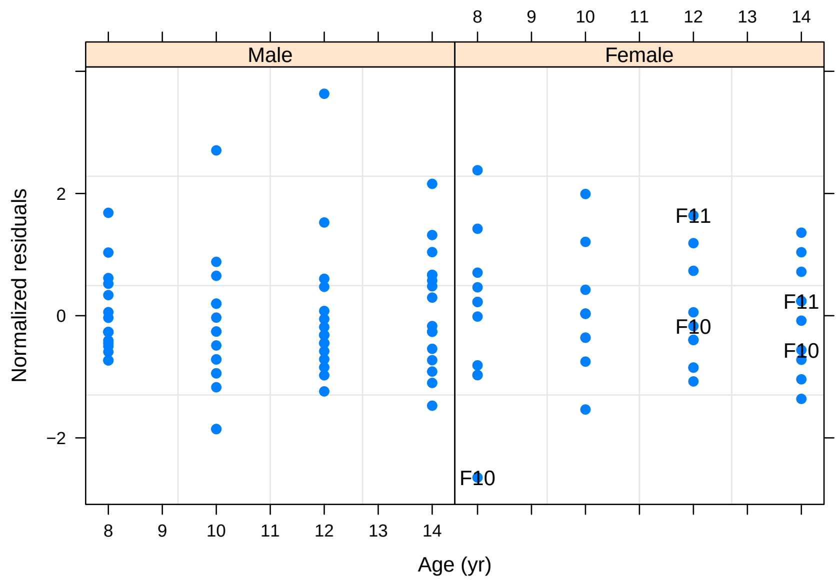


Slika 12: Ostanki za `mod.gls2` po `age` in `Sex`

Slika 12, ki prikazuje ostanke za `mod.gls2`, je zelo podobna Sliki 7, vidi se razliko v variabilnosti ostankov med fanti in dekleti, zato naredimo `mod.gls3`, kjer to heteroskedatičnost modeliramo s funkcijo `varIdent(form=` $\sim$ `1|Sex)`.

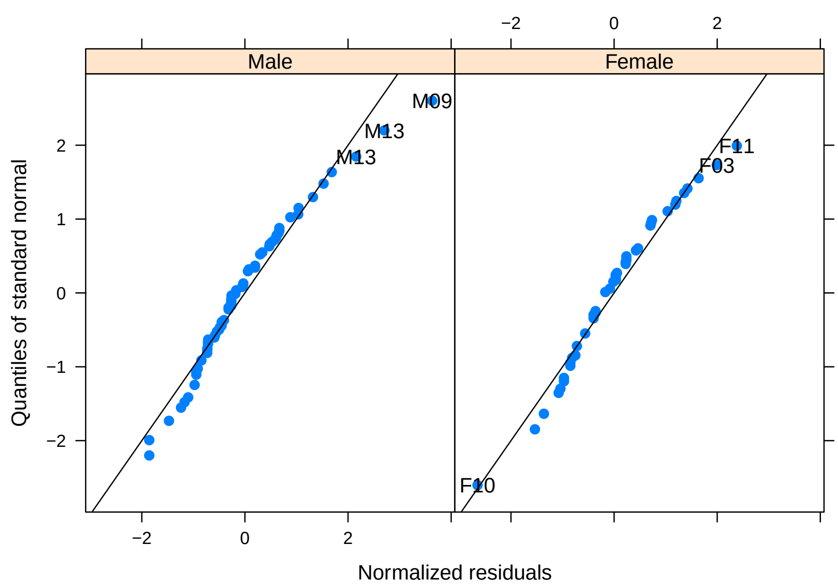
```
> mod.gls3<-gls(distance~I(age-11)*Sex,
+               weights=varIdent(form=~1|Sex),
+               correlation=corCompSymm(form=~1|Subject), data=Orthodont)
> anova(mod.gls2,mod.gls3)

         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.gls2     1  6 445.7572 461.6236 -216.8786
mod.gls3     2  7 436.1887 454.6994 -211.0943 1 vs 2 11.56859   7e-04
```

```
> plot(mod.gls3, resid(.,type="n")~age|Sex,id=0.05,adj=0.5,pch=16)
```



Slika 13: Ostanki za `mod.gls3` po `age` in `Sex`

```
> qqnorm(mod.gls3, ~resid(.,type="n")|Sex, abline=c(0, 1), id=0.05, adj=0.5, pch=16)
```



Slika 14: Ostanki za `mod.gls3` po `Sex`

Sliki 13 in 14 kažeta na ustrezno porazdelitev ostankov modela `mod.gls3`, ki je zelo podobna porazdelitvi ostankov modela `mod.lme3`.

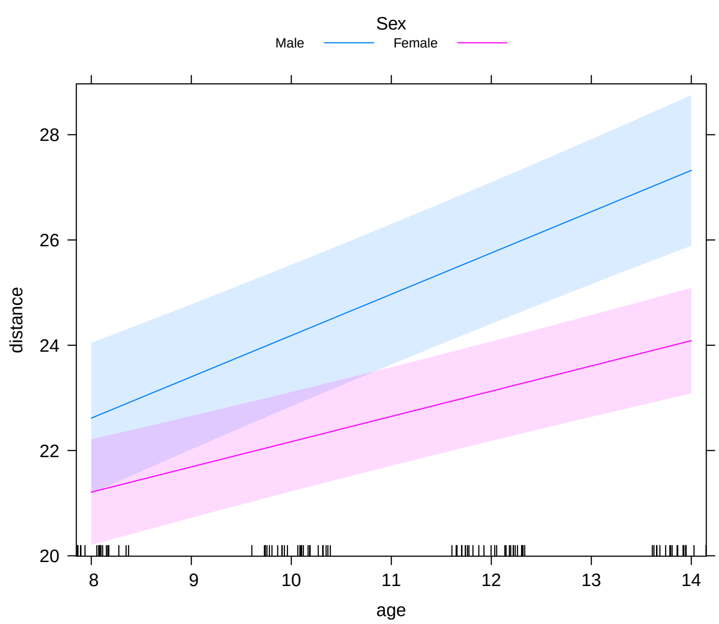
```
> summary(mod.gls3)

Generalized least squares fit by REML
  Model: distance ~ I(age - 11) * Sex
  Data: Orthodont
       AIC      BIC    logLik
  436.1887 454.6994 -211.0943

Correlation Structure: Compound symmetry
 Formula: ~1 | Subject
 Parameter estimate(s):
      Rho
0.7342506
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | Sex
 Parameter estimates:
     Male    Female
1.0000000 0.5818398

Coefficients:
                         Value Std.Error  t-value p-value
(Intercept)          24.968750 0.6727718 37.11326  0.0000
I(age - 11)           0.784375 0.0866677  9.05037  0.0000
SexFemale            -2.321023 0.8218888 -2.82401  0.0057
I(age - 11):SexFemale -0.304830 0.1058772 -2.87909  0.0048

 Correlation:
                      (Intr) I(g-11) SexFml
I(age - 11)            0.000
SexFemale             -0.819  0.000
I(age - 11):SexFemale  0.000 -0.819   0.000

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.69114771 -0.60955147 -0.07845008  0.50459020  2.78466249

Residual standard error: 3.007434
Degrees of freedom: 108 total; 104 residual
```

Primerjajmo še modela `mod.lme3` in `mod.gls3`, ki nista hierarhična, zato se test logaritma razmerja verjetij ne izvede:

```
> anova(mod.lme3,mod.gls3)

         Model df      AIC      BIC    logLik
mod.lme3     1  7 429.2205 447.7312 -207.6102
mod.gls3     2  7 436.1887 454.6994 -211.0943
```

Glede na AIC kriterij je boljši model `mod.lme3`. Primerjava standardnih napak ocen parametrov obeh modelov ne pokaže bistvenih razlik.

```
> compareCoefs(mod.lme3,mod.gls3)

Calls:
1: lme.formula(fixed = distance ~ I(age - 11) * Sex, data = Orthodont,
  random = ~1 | Subject, weights = varIdent(form = ~1 | Sex))
2: gls(model = distance ~ I(age - 11) * Sex, data = Orthodont, correlation =
   corCompSymm(form = ~1 | Subject), weights = varIdent(form = ~1 | Sex))

                      Model 1 Model 2
(Intercept)            24.969  24.969
SE                      0.507   0.673

I(age - 11)            0.7844  0.7844
SE                     0.0933  0.0867

SexFemale              -2.321  -2.321
SE                      0.762   0.822

I(age - 11):SexFemale  -0.305  -0.305
SE                      0.107   0.106

> plot(effect(c("I(age-11)","Sex"), mod.gls3), ci.style="bands", multiline=T, main="")
```



Slika 15: Napovedane vrednosti s pripadajočimi 95 % IZ za povprečno napoved za populacijo za `mod.gls3`

Kako se odločiti, ali `lme` ali `gls`? To je precej odvisno od narave vzorčenja. Če imamo opravka z vzorčenjem v skupinah, je bolj naravna pot `lme`. To velja tudi za kratke časovne vrste kot v našem primeru; sicer pa se pri analizi časovnih vrst in prostorskih podatkov v splošnem bolj ukvarjamo z modeliranjem v okviru `gls`.

3 VAJE

3.1 Število jajčnih foliklov

Jajčni folikel je izoblikovana tkivna struktura v jajčniku, ki vsebuje jajčece. Raziskovalci so opazovali število velikih jajčnih foliklov pri enajstih kobilah (Pierson in Ginther, 1987) v več časovnih točkah estrusnega cikla. Podatki so v podatkovnem okviru `Ovary` vrste `groupedData` v paketu `nlme`. Spremenljivka `follicles` je število jajčnih foliklov večjih od 10 mm v premeru, `Time` je čas v estrusnem ciklu. Opazovanja so bila narejena vsak dan od treh dni pred ovulacijo do treh dni po naslednji ovulaciji. Za vsako kobilo je `Time=0` pri prvi opazovani ovulaciji in `Time=1` pri naslednji ovulaciji; `Mare` je spremenljivka, ki označuje posamezno kobilo. Primer je povzet iz knjige Pinheiro J. C. in Bates D. M. (2001).

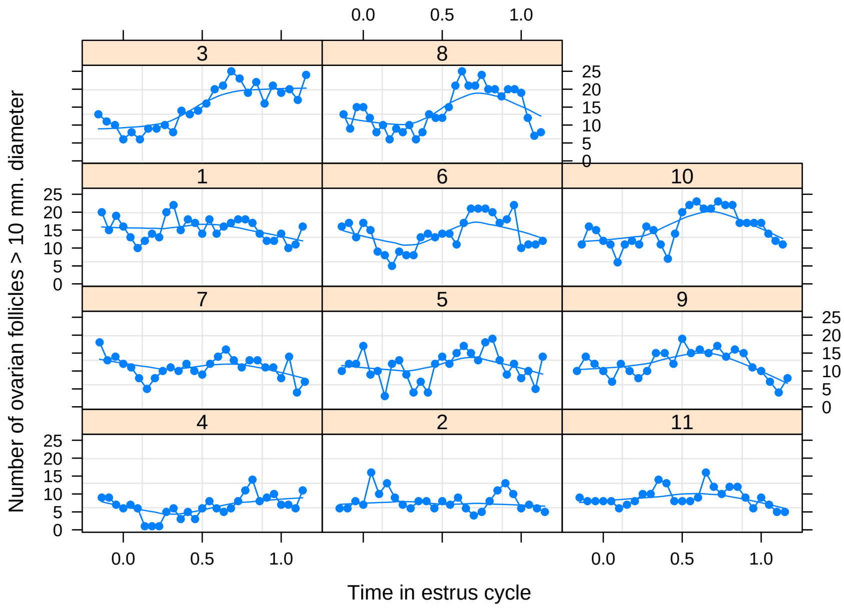
```
> head(Ovary)

Grouped Data: follicles ~ Time | Mare
  Mare        Time follicles
1    1 -0.13636360        20
2    1 -0.09090910        15
3    1 -0.04545455        19
4    1  0.00000000        16
5    1  0.04545455        13
6    1  0.09090910        10
```

```
> summary(Ovary)

      Mare          Time            follicles
 8       : 31   Min.   :-0.1667   Min.   : 1.00
 4       : 29   1st Qu.: 0.1667   1st Qu.: 8.00
 5       : 29   Median : 0.5000   Median :12.00
 1       : 29   Mean   : 0.5000   Mean   :12.04
 6       : 29   3rd Qu.: 0.8333   3rd Qu.:15.00
 10      : 29   Max.   : 1.1667   Max.   :25.00
 (Other):132
```

```
> str(Ovary)

Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':       308 obs. of
 $ Mare     : Ord.factor w/ 11 levels "4"<"2"<"11"<"7"<..: 7 7 7 7 7 7 7 7 7 7 ...
 $ Time     : num  -0.1364 -0.0909 -0.0455 0 0.0455 ...
 $ follicles: num  20 15 19 16 13 10 12 14 13 20 ...
 - attr(*, "formula")=Class 'formula'  language follicles ~ Time | Mare
  .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
```

```
 - attr(*, "labels")=List of 2
  ..$ x: chr "Time in estrus cycle"
  ..$ y: chr "Number of ovarian follicles > 10 mm. diameter"
```

```
> plot(Ovary, type=c("p","l","smooth"), pch=16)
```



Slika 16: Število jajčnih foliklov s premerom nad 10 mm v času estrusnega cikla za enajst kobil; za vsako kobilo je `Time=0` pri prvi ovulaciji in `Time=1` pri drugi ovulaciji

Slika 16 kaže, da se število jajčnih foliklov v času znotraj ciklusa periodično spreminja. V mešanem modelu bomo zato fiksni vpliv časa modelirali kot sinusno nihanje (harmonična regresija). Sinusno nihanje z amplitudo A, frekvenco f in faznim zamikom α lahko izrazimo z linearno kombinacijo sinusnega in kosinusnega člena:

$$Asin(2\pi ft + \alpha) = \beta_1 sin(2\pi ft) + \beta_2 cos(2\pi ft). \tag{10}$$

Do izraza na desni strani (10) pridemo ob upoštevanju zvez: $\sin^2(x) + cos^2(x) = 1$ ter $\sin(x + y) = sin(x)cox(y) + cos(x)sin(y)$. Velja: $Acos(\alpha) = \beta_1$ in $Asin(\alpha) = \beta_2$. Iz teh dve enačb sledi:

$$\begin{aligned} A &= \sqrt{\beta_1^2 + \beta_2^2}, \\ sin(\alpha) &= \frac{\beta_2}{\sqrt{\beta_1^2 + \beta_2^2}}. \end{aligned} \tag{11}$$

V kontekstu podatkov Ovary ima f vrednost 1 (opazovan je en estrusni cikel). Fazni zamik α pa predstavlja čas med ovulacijo ($t = 0$) in časom, ko nihanje doseže povprečno število jajčnih foliklov (β_0). Model harmonične regresije v tem primeru zapišemo:

$$y_{ij} = (\beta_0 +) + (\beta_1) sin(2\pi t_{ij}) + (\beta_2) cos(2\pi t_{ij}) + \varepsilon_{ij}. \tag{12}$$

V model bomo v prvem primeru vključili tri slučajne vplive kobile: na presečišče, na sinusni in na kosinusni člen. Model za i-to kobilo v j-tem času t_{ij} za ta primer zapišemo takole:

$$y_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i}) sin(2\pi t_{ij}) + (\beta_2 + u_{2i}) cos(2\pi t_{ij}) + \varepsilon_{ij}. \tag{13}$$

V modelu ocenjujemo tri parametre fiksnih vplivov $(\beta_0, \beta_1, \beta_2)$. Parameter β_0 predstavlja povprečno število jajčnih foliklov v estrusnem ciklu, na podlagi parametrov β_1 in β_2 pa izračunamo amplitudo A in fazni zamik sinusnega nihanja α. Poleg fiksnih vplivov v modelu ocenjujemo tri variance slučajnih vplivov in tri kovariance slučajnih vplivov ter varianco napak. V (13) u_{i0} predstavlja slučajni vpliv kobile na presečišče, u_{i1} slučajni vpliv kobile na sinusni člen in u_{i2} slučajni vpliv kobile na kosinusni člen.

$$\mathbf{u_i} = \begin{pmatrix} u_0 \\ u_1 \\ u_2 \end{pmatrix} \sim \mathbf{N}_3(\mathbf{0}, \mathbf{\Psi}), \quad \boldsymbol{\varepsilon_{ij}} \sim \mathbf{N}(\mathbf{0}, \sigma^2). \tag{14}$$

V tem primeru je splošna variančno-kovariančna matrika slučajnih vplivov dimenzije 3×3:

$$\mathbf{\Psi} = \begin{pmatrix} \sigma^2_{u_0} & Cov(u_0, u_1) & Cov(u_0, u_2) \\ Cov(u_0, u_1) & \sigma^2_{u_1} & Cov(u_1, u_2) \\ Cov(u_0, u_2) & Cov(u_1, u_2) & \sigma^2_{u_2} \end{pmatrix}. \tag{15}$$

```
> mod.o0.lme <- lme(follicles ~ sin(2*pi*Time) + cos(2*pi*Time),
+                   random=~sin(2*pi*Time) + cos(2*pi*Time)|Mare, data=Ovary)
> intervals(mod.o0.lme)

Approximate 95% confidence intervals

 Fixed effects:
                       lower        est.       upper
(Intercept)        10.237383 12.1859113 14.13443968
sin(2 * pi * Time) -4.637722 -3.2966775 -1.95563307
cos(2 * pi * Time) -1.664766 -0.8731382 -0.08151003
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Mare
                                            lower       est.      upper
sd((Intercept))                         1.9790090  3.2293322 5.26960016
sd(sin(2 * pi * Time))                  1.2554966  2.0928355 3.48862798
```

```
sd(cos(2 * pi * Time))                      0.3269650  1.0670096 3.48205292
cor((Intercept),sin(2 * pi * Time))        -0.8939344 -0.5697096 0.14594207
cor((Intercept),cos(2 * pi * Time))        -0.9793782 -0.8014465 0.07665652
cor(sin(2 * pi * Time),cos(2 * pi * Time)) -0.6619608  0.1780972 0.81984005

 Within-group standard error:
   lower     est.    upper
2.775573 3.019479 3.284819
```

Intervali zaupanja za korelacijske člene variančno-kovariančne matrike slučajnih vplivov so široki in vsebujejo vrednost 0, zato ocenimo, da so slučajni vplivi medsebojno neodvisni in v drugem koraku modeliramo variančno-kovariančno matriko slučajnih vplivov kot diagonalno matriko. V `lme` funkciji `random` argument definiramo s funkcijo `pdDiag`:

$$\boldsymbol{\Psi} = \begin{pmatrix} \sigma_{u_0}^2 & 0 & 0 \\ 0 & \sigma_{u_1}^2 & 0 \\ 0 & 0 & \sigma_{u_2}^2 \end{pmatrix}. \tag{16}$$

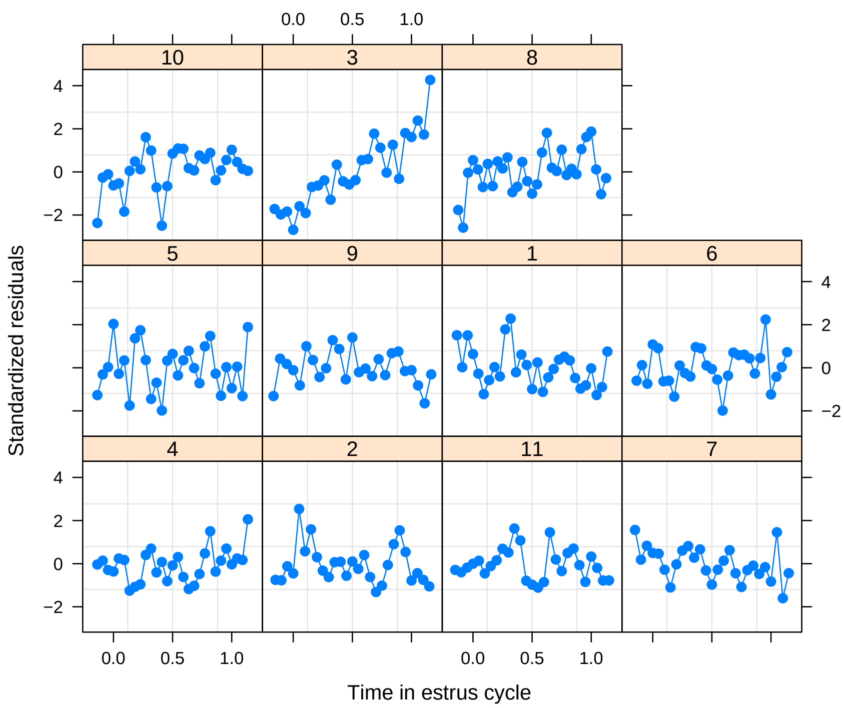
```
> mod.o1.lme <- lme(follicles ~ sin(2*pi*Time) + cos(2*pi*Time),
+                   random=pdDiag(~sin(2*pi*Time) + cos(2*pi*Time)),
+                   data=Ovary)
> intervals(mod.o1.lme)

Approximate 95% confidence intervals

 Fixed effects:
                       lower       est.       upper
(Intercept)        10.276652 12.1871657 14.09767953
sin(2 * pi * Time) -4.637521 -3.2981263 -1.95873172
cos(2 * pi * Time) -1.667829 -0.8820666 -0.09630429
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Mare
                           lower     est.    upper
sd((Intercept))        2.0105422 3.164144 4.979654
sd(sin(2 * pi * Time)) 1.2549259 2.089711 3.479801
sd(cos(2 * pi * Time)) 0.5461156 1.054054 2.034424

 Within-group standard error:
   lower     est.    upper
2.777786 3.020299 3.283984
```

Aproksimativni intervali zaupanja za `mod.o1.lme` za standardne odklone slučajnih vplivov kažejo, da je ocena standardnega odklona slučajnega vpliva kobile na kosinusni člen pol manjša kot pri sinusnem členu, kar pomeni, da morda ta slučajni vpliv v modelu ni potreben. Test logaritma razmerja verjetij tega ne pokaže:

```
> mod.o2.lme <- lme(follicles ~ sin(2*pi*Time) + cos(2*pi*Time),
+                   random=pdDiag(~sin(2*pi*Time)),
+                   data=Ovary)
> anova(mod.o1.lme, mod.o2.lme )

           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.o1.lme     1  7 1633.616 1659.658 -809.8078
mod.o2.lme     2  6 1638.082 1660.404 -813.0409 1 vs 2 6.466236   0.011
```

Modela nista enakovredna, zato nadaljujemo z analizo modela `mod.o1.lme`. Poglejmo časovne vrste ostankov za posamezno kobilo in izračunajmo in narišimo njihov avtokorelogram, ki ga za `lme` model naredi funkcija `ACF` iz paketa `nlme` (Sliki 17 in 18).

```
> plot(mod.o1.lme, resid(.,type="p")~Time|Mare, pch=16, type="b")
```

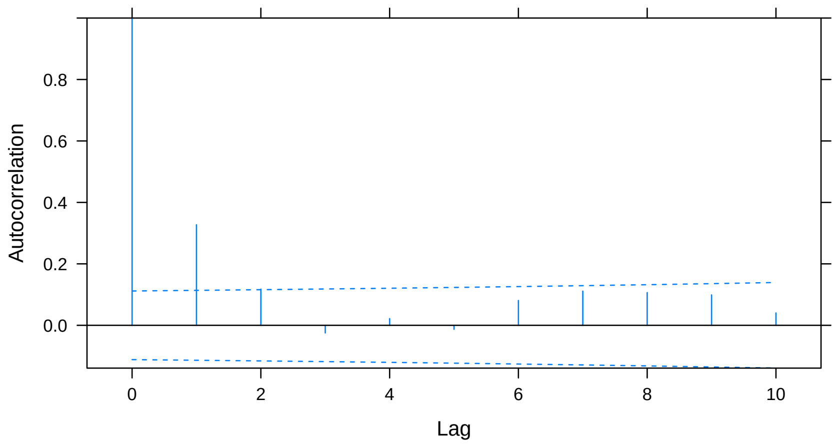


Slika 17: Časovne vrste ostnkov za `mod.o1.lme`

```
> ACF(mod.o1.lme, maxLag=10)

  lag         ACF
1   0  1.00000000
2   1  0.32719494
3   2  0.11764533
4   3 -0.02482731
5   4  0.02176065
6   5 -0.01305529
```

```
7     6  0.08078725
8     7  0.11135095
9     8  0.10651786
10    9  0.09906737
11   10  0.04033596
```

```
> plot(ACF(mod.o1.lme, maxLag=10, resType="n"), alpha=0.05)
```



Slika 18: Avtokorelogram za ostanke `mod.o1.lme`

Avtokorelogram ostankov za `mod.o1.lme` kaže, da med njimi obstaja avtokorelacija prvega ali drugega reda. Poskusimo v prvem primeru za modeliranje avtokorelacije napak uporabiti avtoregresijski model prvega reda AR(1).

```
> mod.o1.lme.cor <- update(mod.o1.lme, correlation=corAR1())
> anova(mod.o1.lme, mod.o1.lme.cor)

               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.o1.lme         1  7 1633.616 1659.658 -809.8078
mod.o1.lme.cor     2  8 1565.448 1595.210 -774.7240 1 vs 2 70.16759  <.0001
```

Test logaritma razmerja verjetij pokaže, da smo z modeliranje avtokorelacije bolje opisali podatke. Aproksimativni 95 % interval zaupanja za koeficient avtokorelacije napak prvega reda je (0.43, 0.69) in ne vsebuje vrednosti 0, kar dodatno kaže na smiselnost modeliranja avtokorelacije napak.

```
> intervals(mod.o1.lme.cor)

Approximate 95% confidence intervals

 Fixed effects:
                      lower       est.       upper
(Intercept)       10.330919 12.1880885 14.04525774
```

```
sin(2 * pi * Time) -4.177139 -2.9852978 -1.79345659
cos(2 * pi * Time) -1.818054 -0.8777618  0.06253073
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Mare
                              lower          est.         upper
sd((Intercept))        1.699262e+00 2.8584039583 4.808248e+00
sd(sin(2 * pi * Time)) 3.926131e-01 1.2579880575 4.030773e+00
sd(cos(2 * pi * Time)) 6.464180e-91 0.0003289627 1.674094e+83

 Correlation structure:
        lower      est.     upper
Phi 0.4314982 0.5721863 0.6857022
attr(,"label")
[1] "Correlation structure:"

 Within-group standard error:
   lower     est.    upper
3.023508 3.507051 4.067926
```

Pri analizi 95 % intervalov zaupanja za standardne odklone slučajnih vplivov vidimo, da je ocena standardnega odklona za slučajni vpliv kobile na kosinusni člen zelo blizu 0 in pripadajoč interval zaupanja je zaradi numerične nestabilnosti na robu definicijskega prostora parametra strašno širok. To kaže, da smo ob modeliranju avtokorelacije napak izničili ta slučanji vpliv, zato ga izločimo iz modela (`mod.o2.lme.cor`).

```
> mod.o2.lme.cor <- lme(follicles ~ sin(2*pi*Time) + cos(2*pi*Time),
+                    random=pdDiag(~sin(2*pi*Time)), correlation=corAR1(), data=Ovary)
> anova(mod.o1.lme.cor, mod.o2.lme.cor)

               Model df      AIC     BIC   logLik   Test      L.Ratio p-value
mod.o1.lme.cor     1  8 1565.448 1595.21 -774.724
mod.o2.lme.cor     2  7 1563.448 1589.49 -774.724 1 vs 2 1.833037e-07  0.9997
```

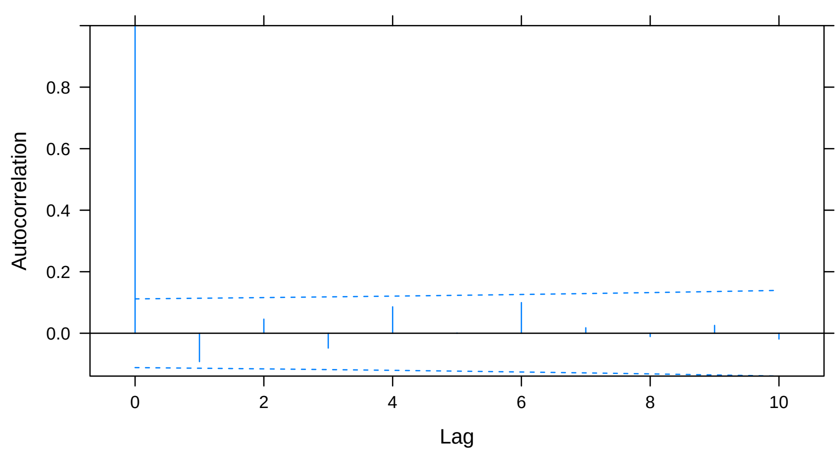
Poglejmo še, ali je potrebno upoštevati slučajni vpliv kobile na sinusni člen. Z modeliranjem avtokorelacije v ostankih se je tudi ocena za standardni odklon tega slučajnega vpliva zmanjšala, njen interval zaupanja pa povečal, njegova spodnja meja je blizu 0:

```
> mod.o3.lme.cor <- lme(follicles ~ sin(2*pi*Time) + cos(2*pi*Time),
+                    random=pdDiag(~1), correlation=corAR1(), data=Ovary)
> anova(mod.o2.lme.cor, mod.o3.lme.cor)

               Model df      AIC      BIC    logLik   Test   L.Ratio p-value
mod.o2.lme.cor     1  7 1563.448 1589.490 -774.7240
mod.o3.lme.cor     2  6 1562.447 1584.769 -775.2233 1 vs 2 0.9987813  0.3176

> intervals(mod.o3.lme.cor)
```

```
Approximate 95% confidence intervals

 Fixed effects:
                        lower      est.      upper
(Intercept)         10.328909 12.189583 14.0502567
sin(2 * pi * Time) -3.936398 -2.947283 -1.9581675
cos(2 * pi * Time) -1.892351 -0.880716  0.1309191
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Mare
                    lower     est.    upper
sd((Intercept)) 1.635785 2.807268 4.817719

 Correlation structure:
        lower      est.     upper
Phi 0.4878973 0.6074422 0.7046207
attr(,"label")
[1] "Correlation structure:"

 Within-group standard error:
   lower     est.    upper
3.195701 3.665450 4.204249
```

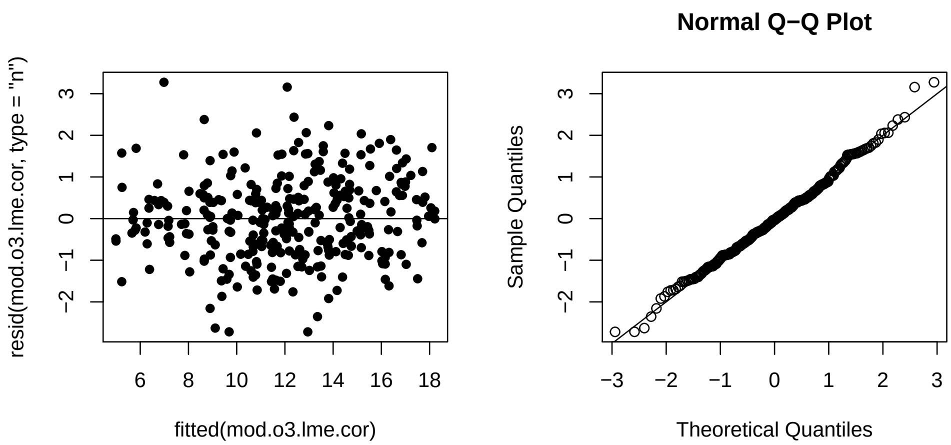
Tudi slučajni vpliv kobile na sinusni člen ni več statistično značilen. Slika 19 kaže, da ni več avtokorelacije v ostankih modela `mod.o3.lme.cor`.

```
> plot(ACF(mod.o3.lme.cor, maxLag=10, resType="n"), alpha=0.05)
```



Slika 19: Avtokorelogram za normalizirane ostanke za `mod.o3.lme.cor`

```
> par(mfrow=c(1,2))
> plot(fitted(mod.o3.lme.cor), resid(mod.o3.lme.cor, type="n"), pch=16)
> abline(h=0)
> qqnorm(resid(mod.o3.lme.cor, type="n"))
> abline(a=0,b=1)
```

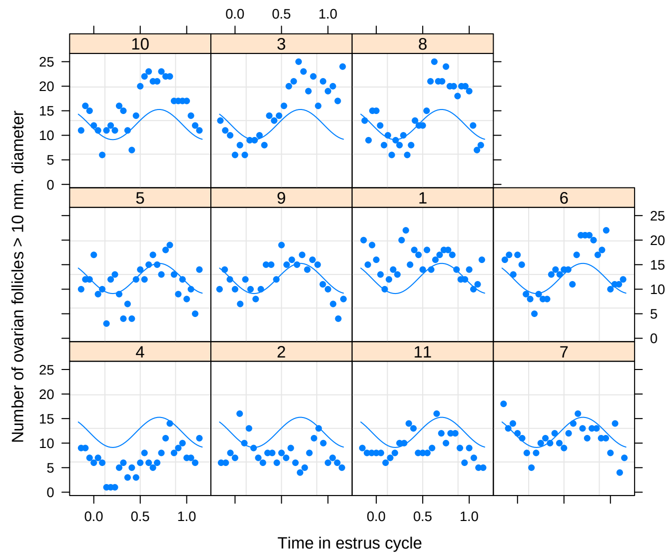


Slika 20: Ostanki za `mod.o3.lme.cor`

Slika 21 kaže populacijske napovedi izbranega modela `mod.o3.lme.cor` in Slika 22 kaže napovedi za posamezno kobilo. Končni model torej predpostavlja sinusno nihanje števila jajčnih foliklov v času estrusnega cikla. Upoštevan je slučajni vpliv kobile na presečišče in avtokorelacija napak AR(1). Ocena za povprečno število jajčnih foliklov v estrusnem ciklu je 12.2 s 95 % intervalom zaupanja (10.3, 14.1). Ocena za amplitudo je $\sqrt{(2.95^2 + 0.88^2)} = 3.08$ in za fazni zamik $asin(0.88/3.08) = 0.29$.

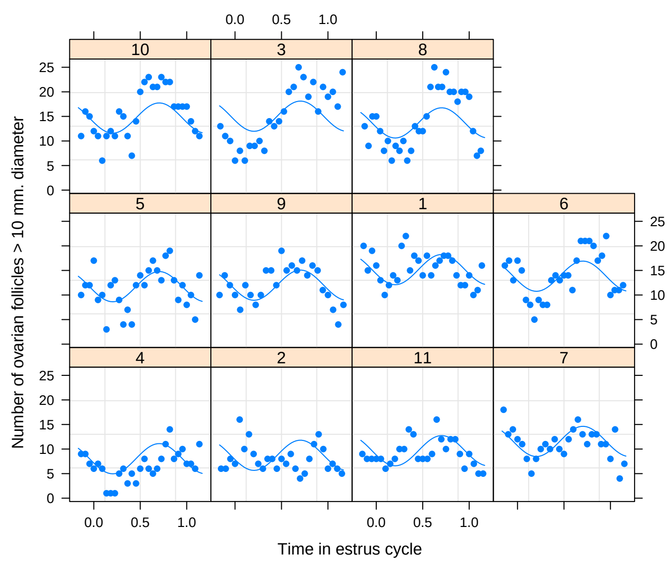
Kako bi ocenili interval zaupanja za ti dve oceni?

```
> plot(augPred(mod.o3.lme.cor, level=0), grid=T, pch=16)
```



Slika 21: Napovedane vrednosti za populacijo za `mod.o3.lme.cor`

```
> plot(augPred(mod.o3.lme.cor, level=1), grid=T, pch=16)
```



Slika 22: Napovedane vrednosti za kobile `mod.o3.lme.cor`

Modelirajte avtokorelacijo napak še z modeloma MA(2) ali ARMA(1,1) in primerjajte rezultate.