Päev 4
Lineaarsed mudelid: juhuslikud faktorid ja korduvad mõõtmised.

Avage *R*.

1.

Vaatame üht lihtsat katset – 4 erinevat sorti on kasvatatud erinevatel aastatel (2003-2007) uurimaks saagikuse erinevusi.

Lugege *R*-i vastav andmestik:

```
saagikus=read.csv("http://www.eau.ee/~ktanel/modelleerimise_koolitus_EMYs_2013/
saagikus.csv",sep=",",dec=".",header=TRUE)
```

1.1. Hinnake sordi mõju saagikusele tavalise lineaarse mudeli abil.

Sisestades (kopeerides) alljärgnevad käsud skriptiaknasse.

```
saak.model.1 <- lm(saak ~ sort, data=saagikus)
summary(saak.model.1)
```

Osa tulemusest:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3330.3      135.5  24.584  < 2e-16 ***
sortsort2     -556.0      191.6  -2.902  0.00413 **
sortsort3     -471.5      191.6  -2.461  0.01472 *
sortsort4     -152.5      191.6  -0.796  0.42686
```

Sordi 1 keskmise saagikuse hinnang on 3330,3 standardhälbega 135,5. Sordi 2 saagikus on 556,0 võrra madalam, sordi 3 saagikus 471,5 võrra madalam ja sordi 4 saagikus 152,5 võrra madalam.

95%-usaldusinetrvall 1. sordi saagikusele on leitav käsuga

```
predict(saak.model.1, data.frame(sort="sort1"), interval="confidence")
```

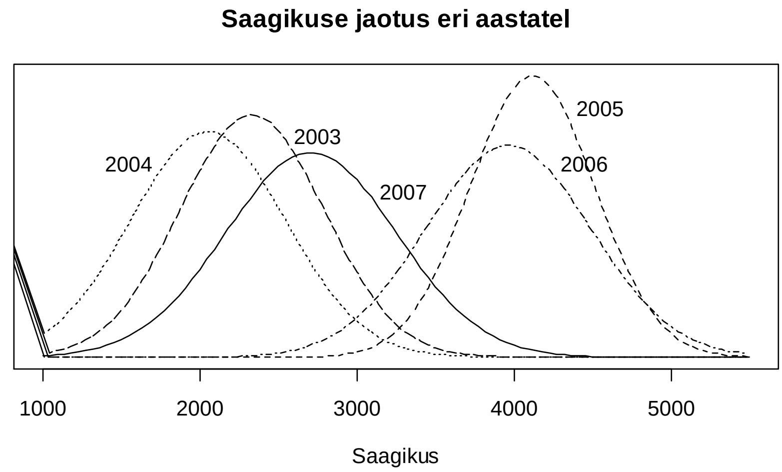
```
          fit      lwr      upr
[1,] 3330.275 3063.121 3597.429
```

95%-usaldusintervall 1. sordi keskväärtusele on 3063,1…3597,4.

1.2.

Küsimus: millal me võime teha ülaltoodud järeldusi erinevate sortide saagikuse kohta?

Vastus: kui saagikus ei sõltu aastast, siis kehtivad tehtud järeldused mistahes aasta kohta; kui aga saagikus aastati varieerub, siis vaid aastate 2003-2007 kohta.



Muideks, ülal toodud joonis, mis kirjeldab saagikuse jaotust eri aastatel eeldusel, et saagikus jaotub normaaljaotuse alusel, on saadud järgneva programmijupi abil.

```
attach(saagikus)
.x <- seq(1000, 5500, length=100)
plot(.x, dnorm(.x, mean=mean(saak[aasta==2005]),sd=sd(saak[aasta==2005])),
 xlab="Saagikus", ylab="", main="Saagikuse jaotus eri aastatel", lty=2, type="l",
 yaxt="n")
lines(.x, dnorm(.x, mean=mean(saak[aasta==2007]),sd=sd(saak[aasta==2007])),lty=1)
lines(.x, dnorm(.x, mean=mean(saak[aasta==2004]),sd=sd(saak[aasta==2004])),lty=3)
lines(.x, dnorm(.x, mean=mean(saak[aasta==2006]),sd=sd(saak[aasta==2006])),lty=4)
lines(.x, dnorm(.x, mean=mean(saak[aasta==2003]),sd=sd(saak[aasta==2003])),lty=5)
text(2900,0.0008, "2003", adj=c(1, 0.5))
text(1700,0.0007, "2004", adj=c(1, 0.5))
text(4700,0.0009, "2005", adj=c(1, 0.5))
text(4600,0.0007, "2006", adj=c(1, 0.5))
text(3450,0.0006, "2007", adj=c(1, 0.5))
remove(.x)
```

Uurimaks täpsemalt, kui suur on aastatevaheline erinevus, sobitame andmetele keerukama mudeli:

```
saak.model.2 <- lm(saak ~ sort + factor_aasta, data=saagikus)
summary(saak.model.2)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        2627.61      86.55  30.361  < 2e-16 ***
sortsort2          -556.03      86.55  -6.425 1.02e-09 ***
sortsort3          -471.47      86.55  -5.448 1.55e-07 ***
sortsort4          -152.54      86.55  -1.762 0.079578 .  
factor_aasta2004   -278.38      96.76  -2.877 0.004470 ** 
factor_aasta2005   1784.75      96.76  18.445  < 2e-16 ***
factor_aasta2006   1631.55      96.76  16.862  < 2e-16 ***
factor_aasta2007    375.39      96.76   3.880 0.000144 ***
```

Sordi 1 keskmine saagikus aastal 2003 on 2627,6. Samuti saame leida, et näiteks aastal 2007 on sordi 4 keskmine saagikus 2627,6 – 152,5 + 375,4 = 2850,5.

Seega saame hinnata saagikust mistahes sordi ja mistahes andmetes esindatud vaatlusaasta korral.

Kuigi aastatevahelise erinevuse statistiline olulisus ilmneb juba käsu `summary` tulemusena saadud efektide tabelist, võib lisaks käsuga `Anova` tellida ka testid faktorite mõjude omavahelise erinevuse kohta:

```
Anova(saak.model.2)
```

```
> Anova(saak.model.2)
Anova Table (Type II tests)

Response: saak
                Sum Sq  Df F value    Pr(>F)    
sort          10329848   3  18.388 1.576e-10 ***
factor_aasta 143881439   4 192.091 < 2.2e-16 ***
Residuals     35953332 192                      
```

Nii sordi kui ka aasta mõju on statistiliselt olulised.

Kui nüüd aga tahame prognoosida saagikust aastaks 2009, siis jääme jänni – prognoosida küll saab (võttes saagikuseks näiteks katseaastate keskmise), aga prognoosi täpsuse hindamiseks infot pole.

1.3.

Lahendus: käsitleme vaadeldud aastaid kui juhuslikku valimit kõikmõeldavatest aastatest, st et käsitleme tunnust 'aasta' juhusliku faktorina.

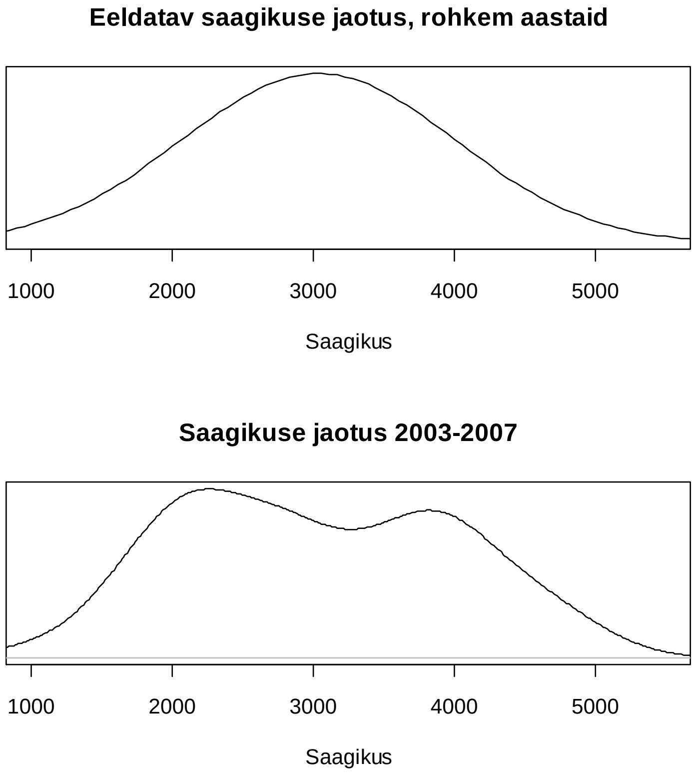
Sellisel juhul eeldatakse, et aastate mõjud A_j on normaaljaotusega suurused, $A_j \sim N(0, \sigma^2_{aasta})$, st et

– keskmine aasta mõju on 0 (keskmisest paremaid ja halvemaid aastaid on ühepalju);
– kas järgmine aasta on hea või halb, on juhuslik;
– σ_{aasta} on aasta mõjude standardhälve

(et normaaljaotuse puhul jääb ~95% väärtustest vahemikku ±2σ, siis saab ka järgmise aasta saagikuse prognoosimisi täpsuse osas konstateerida, et saagikus võib varieeruda ±2σ_{aasta} piires).

Kõrvaloleva pildi joonistanud programm:

```
par(mfrow=c(2,1))
.y <- seq(500, 6000, length=100)
plot(.y, dnorm(.y, mean=mean(saak),
 sd=sd(saak)), xlim=c(1000,5500),
 xlab="Saagikus", ylab="",
 main="Eeldatav saagikuse jaotus,
 rohkem aastaid", type="l",
 yaxt="n")
remove(.y)
plot(density(saagikus$saak),
 xlim=c(1000,5500), xlab="Saagikus",
 ylab="", main="Saagikuse jaotus
 2003-2007", yaxt="n")
par(mfrow=c(1,1))
```



Taolise mudeli sobitamiseks on *R*-s erinevate moodulite koosseisus olemas mitmeid sobivaid funktsioone.

Näiteks

```
library(nlme)
saak.model.3 <- lme(saak ~ sort, random=~1|factor_aasta, data=saagikus)
summary(saak.model.3)
```

või

```
library(lme4)
saak.model.4 <- lmer(saak ~ sort + (1|factor_aasta), data=saagikus)
summary(saak.model.4)
```

- Neist esimene, lisapakett `nlme` funktsiooniga `lme`, on *R*-i klassikaline vahend segamudelite hindamiseks;
- teine lisapakett, `lme4` funktsiooniga `lmer`, on aga uus ja arenev moodul, milles on hõlbus arvesse võtta suuremat arvu juhuslikke faktoreid ja mis võimaldab sobitada segamudeleid ka mittenormaaljaotusega tunnustele.

Samas ei pruugi keerulisemate mudelite parameetrite hindamisprotsess antud funktsiooniga koonduda (paketi arendamine on alles pooleli).

Juhusliku faktori kirjapilt kujul ’`1|factor_aasta`’ tähendab, et igale aastale hinnatakse oma vabaliige (jaotusest $A_j \sim N(0, \sigma^2_{aasta})$).

Tulemused on mõlemal juhul samad:

```
> library(nlme)
> saak.model.3 <- lme(saak ~ sort, random=~1|factor_aasta, data=saagikus)
> summary(saak.model.3)
Linear mixed-effects model fit by REML
 Data: saagikus
       AIC      BIC    logLik
  2984.390 3004.059 -1486.195

Random effects:
 Formula: ~1 | factor_aasta
        (Intercept) Residual
StdDev:   945.8211  432.7319

Fixed effects: saak ~ sort
               Value Std.Error  DF   t-value p-value
(Intercept) 3330.275  427.3882 192  7.792156  0.0000
sortsort2   -556.026   86.5464 192 -6.424605  0.0000
sortsort3   -471.470   86.5464 192 -5.447595  0.0000
sortsort4   -152.537   86.5464 192 -1.762490  0.0796
```

ja

```
> library(lme4)
> saak.model.4 <- lmer(saak ~ sort + (1|factor_aasta), data=saagikus)
> summary(saak.model.4)
Linear mixed-effects model fit by REML
Formula: saak ~ sort + (1 | factor_aasta)
   Data: saagikus
  AIC  BIC logLik MLdeviance REMLdeviance
 2982 2999  -1486       3018         2972
Random effects:
 Groups       Name        Variance Std.Dev.
 factor_aasta (Intercept) 894578   945.82
 Residual                 187257   432.73
number of obs: 200, groups: factor_aasta, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3330.28     427.39   7.792
sortsort2    -556.03      86.55  -6.425
sortsort3    -471.47      86.55  -5.448
sortsort4    -152.54      86.55  -1.762
```

- Sordi 1 saagikus on 3330,3, mis on sarnane eelnevalt leituga, küll on märgatavalt suurem saadud hinnangu standardhälve – 427,4 (vrd mudeli `saak.model.1` väljundiga lk. 1 lõpus). Suurema standardvea põhjus on selles, et püüame modelleerida märksa üldisemat olukorda (5 konkreetse aasta asemel mistahes aastate saagikust). St, et 3330,3 on sordi 1 keskmise saagikuse hinnanguks mistahes aastatel, mitte ainult aastail 2002-2007, nagu esimese mudeli korral.

- Aastate mõjude hinnanguid erinevalt aastat fikseeritud faktorina käsitlenud mudeli kokkuvõttest (vt `summary(saak.model.2)` lk. 3 alguses) enam vaikimisi välja ei trükita. Põhjuseks asjaolu, et ega meile nende konkreetsete katseaastate erinevus nii väga huvi pakugi, pigem on vaja hinnata, kuivõrd saagikus erinevatel aastatel üldse erineda võib (seda mõõdab aasta mõjude dispersioon (või standardhälve)).

Antud juhul on aasta mõjude standardhälbe hinnanguks 945,8 (vt eelmine lk).
Seega on halvematel aastatel saagikus keskmisest saagikusest umbes 2×945,8 = 1891,6 võrra väiksem, paremal aastatel suurem (vastavalt normaaljaotuse omadustele peaks ~2,5% kehvemaid aastaid olema keskmisest saagikusest $2\times\sigma_{aasta}$ võrra väiksema saagikusega, sama kehtib ka paremate aastate kohta).

Sedavõrd suur kõikumine on tingitud vaadeldud aastate väga suurest erinevusest.

- Siiski on funktsiooni `lmer` abil konstrueeritud mudelist lisakäsuga

```
ranef(saak.model.4)
```

tellitavad andmetes esindatud aastate juhuslike mõjude prognoosid (`ranef` tähendab „random effect“):

	(Intercept)
2003	-699.0053
2004	-975.9346
2005	1076.4546
2006	924.0540
2007	-325.5687

Kõrvutades saadud arve fikseeritud aastaefektide hinnangutega mudelist `saak.model.2` (vt käsu `summary` tulemust lk. 3 alguses), ilmneb, et aastate vaheline erinevus on pisut vähenenud.

Käsitledes aastat fikseeritud faktorina, on näiteks aastate 2005 ja 2004 vaheline erinevus 1784,75 – (–278,38) = 2063,13;
käsitledes aastat aga juhuslikuna, on sama erinevus 1076,46 – (–975,93) = 2052,39.
Põhjuseks jällegi aastat juhuslikuna käsitleva mudeli üldisem olemus; vähem tähelepanu pööratakse konkreetsete aastate vahelistele erinevustele, pigem käsitletakse neid erinevusi juhuslikena, mistap on tulemuseks ka väiksemad hinnangud.

- Käsu `lme` abil konstrueeritud mudelist saab aga lisakäsuga

```
predict(saak.model.3, data.frame(sort=c("sort1","sort1"),factor_aasta=2004:2005),level=0:1)
```

leida sordi 1 saagikuse prognoosid suvaliseks aastaks (`level=0`) ja hinnangud konkreetsete andmetes esindatud aastate (`level=1`) saagikustele:

	factor_aasta	predict.fixed	predict.factor_aasta
1	2004	3330.275	2354.341
2	2005	3330.275	4406.730

Keskmise saagikuse prognoos üle kõigi aastate on 3330,275, hinnang aasta 2004 saagikusele on 3330,275 – 975,9346 = 2354,341 (– 975,9346 on aasta 2004 mõju prognoos, vt eelmine tabel).

1.4. Tegelikult on eelnev mudel sama, kui **korduvmõõtmiste mudel**!

Vaikimisi eeldatakse üldiste linaarsete mudelite puhul, et kõik vaatlused on sõltumatud, st et erinevate vaatluste vaheline kovariatsioon (ja seega ka korrelatsioon) on null. Kui aga osa vaatlusi on teostatud samadel indiviididel/objektidel (põldudel/aastatel/jne), siis on ju loomulik eeldada, et need mõõtmised on omavahel seotud ja nende vaheline kovariatsioon ei ole null.

Korduvmõõtmiste mudelid püüavadki seda sarnaste vaatluste seotust arvesse võtta. Lähenemisi on siin mitmeid, tänapäeval on ehk levinuim püüd modelleerida teatud vaatluste vahelist sarnasust vaatluste sobivalt defineeritud kovariatsioonimaatriksi kaudu.

Lihtsaim variant sarnaste vaatluste vahelise seotuse modelleerimisest on eeldada, et kõigi samadel indiviididel/objektidel sooritatud mõõtmiste vahel on ühesugune korrelatsioon.

Antud ülesande kontekstis võiks siis eeldada, et kõik samal aastal sooritatud mõõtmised on tänu samal aastal valitsenud sarnastele tingimustele omavahel seotud, kusjuures korrelatsioon kõigi aastal 2003 sooritatud mõõtmiste vahel on ρ, nagu ka kõigi aastal 2004 sooritatud mõõtmiste vahel jne.

Taolise mudeli andmetele sobitamiseks on kasutatav paketis `nlme` sisalduv funktsioon `gls` kujul:

```
saak.model.1K = gls(saak ~ sort, data=saagikus, correlation=corCompSymm(form=~1|aasta))
summary(saak.model.1K)
```

```
> saak.model.1K = gls(saak ~ sort, data=saagikus, na.a
> summary(saak.model.1K)
Generalized least squares fit by REML
  Model: saak ~ sort
  Data: saagikus
      AIC      BIC    logLik
  2984.39 3004.059 -1486.195

Correlation Structure: Compound symmetry
 Formula: ~1 | aasta
 Parameter estimate(s):
      Rho
0.8269079

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 3330.275  427.3882  7.792157  0.0000
sortsort2   -556.026   86.5464 -6.424605  0.0000
sortsort3   -471.470   86.5464 -5.447595  0.0000
sortsort4   -152.537   86.5464 -1.762490  0.0795

 Correlation:
          (Intr) srtsr2 srtsr3
sortsort2 -0.101
sortsort3 -0.101  0.500
sortsort4 -0.101  0.500  0.500

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.8191693 -0.7512281 -0.1434760  0.8538539  1.8480468

Residual standard error: 1040.113
Degrees of freedom: 200 total; 196 residual
```

Korrelatsioon samal aastal sooritatud mõõtmiste vahel on 0,83.

Mudeli parameetrite hinnangud ja nende standardvead on identsed juhusliku aastamõjuga mudeleist hinnatutega.

Näitemaks, et juhuslikku aastamõju sisaldanud mudel on tegelikult identne samal aastal sooritatud mõõtmisi korduvate mõõtmistena modelleeerinud mudeliga, tuleb esmalt tõdeda, et aasta mõjude dispersioon $\mathrm{var}(A_j) = \sigma^2_{aasta}$ kujutab enesest ka samal aastal sooritatud mõõtmiste vahelist kovariatsiooni. Et iga üksiku mõõtmise koguvarieeruvus avaldub juhusliku aasta mõju korral dispersioonikomponentide summana kujul $\sigma^2_{saak} = \sigma^2_{aasta} + \sigma^2_{residual}$, peab samal aastal sooritatud mõõtmiste vaheline korrelatsioon esituma vastavalt korrelatsioonikordaja definitsioonile kujul

$$\rho = \frac{\sigma^2_{aasta}}{\sqrt{(\sigma^2_{aasta} + \sigma^2_{residual}) * (\sigma^2_{aasta} + \sigma^2_{residual})}}.$$

Pannes sellesse valemisse juhuslikku aastamõju sisaldanud mudelist (mudelid `saak.model.3` ja `saak.model.4`) hinnatud dispersioonikomponentide väärtused, saame

$$\rho = \frac{894578}{\sqrt{(894578 + 187257) * (894578 + 187257)}} = 0{,}827,$$

mis on identne viimatise korduvmõõtmiste mudelist hinnatud korrelatsiooniga.

1.5.

Aga põld? Nimelt oli katse jaoks välja valitud 10 põllulappi, kus erinevatel aastatel on erinevaid sorte proovitud. Järeldusi sortide saagikuse kohta oleks aga kena teha mitte ainult nende 10 konkreetse põllu tarvis … Loomulikult peaks siis ka põld olema juhuslik faktor.

```
saak.model.5 <- lmer(saak ~ sort + (1|factor_aasta) + (1|p6ld), data=saagikus)
summary(saak.model.5)
```

```
> saak.model.5 <- lmer(saak ~ sort + (1|factor_aasta) + (1|p6ld), data=saagikus)
> summary(saak.model.5)
Linear mixed-effects model fit by REML
Formula: saak ~ sort + (1 | factor_aasta) + (1 | p6ld)
   Data: saagikus
  AIC  BIC logLik MLdeviance REMLdeviance
 2899 2919  -1443       2931         2887
Random effects:
 Groups       Name        Variance Std.Dev.
 p6ld         (Intercept)  87962   296.58
 factor_aasta (Intercept) 896639   946.91
 Residual                 104793   323.72
number of obs: 200, groups: p6ld, 10; factor_aasta, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3330.28     436.14   7.636
sortsort2    -556.03      64.74  -8.588
sortsort3    -471.47      64.74  -7.282
sortsort4    -152.54      64.74  -2.356
```

Juhuslike faktorite mõjude dispersioonide suhe näitab konkreetse faktori osa uuritava tunnuse koguvarieeruvuses.

Hinnang uuritava tunnuse kogudispersioonile avaldub summana

$$\sigma^2_{saak} = \sigma^2_{p6ld} + \sigma^2_{aasta} + \sigma^2_{residual} = 87962 + 896639 + 104793 = 1089397.$$

Aasta mõju osakaal koguvarieeruvusest on

$$\frac{\sigma^2_{aasta}}{\sigma^2_{saak}} = \frac{896639}{1089397} = 0,823$$

ja põllu mõju osakaal koguvarieeruvusest

$$\frac{\sigma^2_{põld}}{\sigma^2_{saak}} = \frac{87962}{1089397} = 0,081.$$

Seega on antud juhul aasta mõju ligikaudu 10 korda suurem, kui põllu mõju.

Kui te vahepeal (näiteks leheküljel 2 paikneva joonise tegemisel) kasutasite käsku

```
attach(saagikus)
```

siis nüüd, antud andmestikuga töö lõpuks, tuleb rakendada ka käsku

```
detach(saagikus)
```

PS.

Tegelikult on analüüsitud andmete näol tegu etteantud skeemi alusel arvuti poolt genereeritud juhuslike andmetega. Soovi ja huvi korral võite lasta *R*-l genereerida uue samade tunnustega andmestiku (näiteks nimega ’saagikus2’), rakendada kasutatud mudeleid uuele andmestikule ja püüda aru saada analüüside tulemustest.

Andmeid genereeriv programm:

```
p6ld <- rep(c("p1","p2","p3","p4","p5","p6","p7","p8","p9","p10"),c(20,20,20,20,20,20,20,20,20,20))
aasta <- rep(c(rep(2003,4),rep(2004,4),rep(2005,4),rep(2006,4),rep(2007,4)),10)
sort <- rep(c("sort1","sort2","sort3","sort4"),50)

p6lluefekt <- c(50+300*rnorm(20),-50+300*rnorm(20),0+300*rnorm(20),100+300*rnorm(20),
 -100+300*rnorm(20),300+300*rnorm(20),-300+300*rnorm(20),500+300*rnorm(20),
 -500+300*rnorm(20),70+300*rnorm(20))
saagikus <- data.frame(p6ld,aasta,sort,p6lluefekt)

saagikus$saak <- 3250+saagikus$p6lluefekt
saagikus$aastaefekt <- 0
saagikus$aastaefekt[aasta==2003] <- -700+600*rnorm(1)
saagikus$aastaefekt[aasta==2004] <- 50+400*rnorm(1)
saagikus$aastaefekt[aasta==2005] <- 500+310*rnorm(1)
saagikus$aastaefekt[aasta==2006] <- 1300+450*rnorm(1)
saagikus$aastaefekt[aasta==2007] <- -300+170*rnorm(1)

saagikus$sordiefekt <- 0
saagikus$sordiefekt[sort=="sort2"] <- -500
saagikus$sordiefekt[sort=="sort3"] <- -350
saagikus$sordiefekt[sort=="sort4"] <- -75


saagikus$saak <- saagikus$saak + saagikus$aastaefekt + saagikus$sordiefekt

saagikus$factor_aasta <- as.factor(saagikus$aasta)
```

2.

Lugege *R*-i andmestik

```
vasikas=read.csv("http://www.eau.ee/~ktanel/modelleerimise_koolitus_EMYs_2013/
vasikas.csv",sep=",",dec=".",header=TRUE)
```

Tegu on 55 vasika kehamassidega, mis on määratud vanuses 0 kuni 857 päeva keskmise intervalliga 43 päeva. Vaja oleks hinnata vasikate kasvukõverad ning kehamassid 700 päeva vanuses.

Kui igat vasikat oleks kaalutud täpselt 100-päevaste vahedega (vanuses 0 päeva, 100 päeva jne) vähemalt 700.-nda päevani välja, võiks arvutada iga taolise ajamomendi tarvis vasikate kehamasside keskmise ning esitada saadud keskmisi ühendava joone kasvukõverana.

Aga mida teha siis, kui

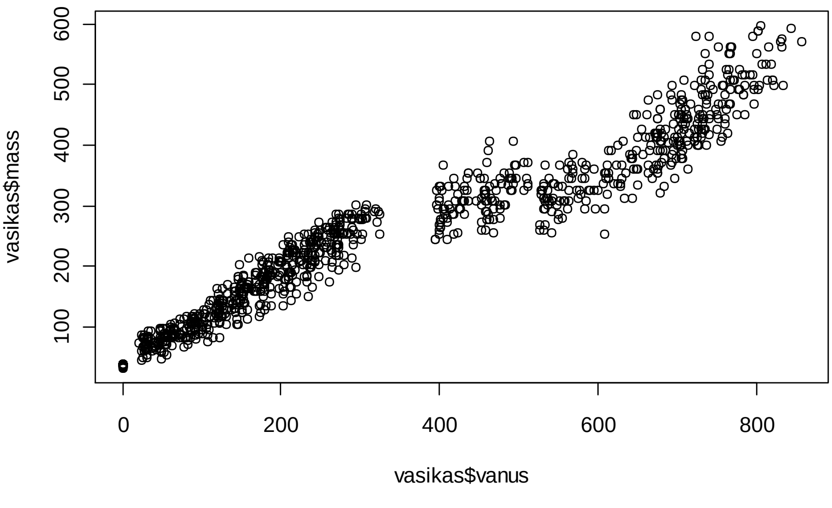
1. vasikaid ei ole kaalutud samadel vanustel;
2. osasid vasikaid on kaalutud pikema, osasid lühema aja jooksul;
3. soovime kasvukõverat interpoleerida mingi vanuse jaoks ka neile vasikatele, kellel sellest perioodist veel massi pole;
4. leidub üksikuid kahtlaseid väärtuseid, mis võivad vaid ühe looma tarvis välja joonistatud kõverale absurdse kuju anda?

Lahenduseks on hinnata kasvukõver iga vasika tarvis käsitledes vasika-spetsiifilisi efekte juhuslikena. St, et üle kõigi vasikate hinnatakse fikseeritud kõver ning sellele lisaks iga vasika kohta juhuslik kõrvalekalle fikseeritud kõverast. Inglise keeles nimetatakse taolist mudelit **’*random regression model*’** või ’*random coefficient model*’ või ….

2.1.

Andmetest esmase ülevaate saamiseks võib välja joonistada hajuvusdiagrammi. Näiteks käsuga

```
plot(vasikas$vanus, vasikas$mass)
```



Joonise alusel võiks proovida vasikate kehamassi muutumist modelleerida kuuppolünoomiga:

$$\text{mass}_i = \underbrace{b0 + b1\times\text{vanus}_i + b2\times(\text{vanus}_i)^2 + b3\times(\text{vanus}_i)^3}_{\text{fikseeritud kasvukõver üle kõigi vasikate}} + \underbrace{b0_i + b1_i\times\text{vanus}_i + b2_i\times(\text{vanus}_i)^2 + b3_i\times(\text{vanus}_i)^3}_{i\text{.-nda vasika spetsiifiline kasvukõver}}$$

Vasika-spetsiifilised juhuslikud regressioonikordajad eeldatakse jaotuvat vastavalt normaaljaotusele: $b0_i \sim N(0,\sigma^2_{b0})$, $b1_i \sim N(0,\sigma^2_{b1})$, $b2_i \sim N(0,\sigma^2_{b2})$ ja $b3_i \sim N(0,\sigma^2_{b3})$.

Taolist juhuslike regressioonikordajatega mudelit võib *R*’is andmetele sobitada nii käsuga `lme` kui ka käsuga `lmer`.
Nagu ikka, on nende käskude süntaksid pisut erinevad, kuid tunnuse nime `loom` kirjutamine püstkriipsu järele tähendab mõlemal juhul, et igale loomale tuleb hinnata erinevad regressioonikordajad ja seeläbi ka oma kasvukõver.

```
vasikas.model.1a <- lme(mass ~ vanus+I(vanus^2)+I(vanus^3),
 random=~1+vanus+I(vanus^2)+I(vanus^3)|loom, data=vasikas)
summary(vasikas.model.1a)
coef(vasikas.model.1a)
```

või

```
vasikas.model.1b <- lmer(mass ~ 1+vanus+I(vanus^2)+I(vanus^3) +
 (1+vanus+I(vanus^2)+I(vanus^3)|loom), data=vasikas)
summary(vasikas.model.1b)
```

Käsk `summary` väljastab mõlemal juhul olulisemad tulemused, lisaks võimaldab käsk `coef` välja trükkida kõigile vasikatele omased regressioonivõrrandi parameetrid.

- Käsu `lme` tulemustest hakkab silma, et juhuslike regressioonikordajate varieeruvus on väga väike, eriti polünoomi ruut- ja kuupliikme kordajail. Ühelt poolt võib see viidata antud parameetrite suhtelisele sarnasusele erinevatel loomadel. Teine ja praegusel juhul peapõhjus on aga hinnatavate regressioonikordajate eneste (vt fikseeritud regr. kordajate hinnanguid) väga väikesed väärtused (tuleb ju kuupliikmele vastavat kordajat korrutada kaalumisvanuse kuubiga, st et vastav argument võib omada väärtuseid 0-st 700^3=343000000-ni …

Hinnatud parameetrite tähendus.

Näiteks $\hat{\sigma}^2_{b3} = (0{,}0000000638)^2$, st et $b3_i \sim N[0;(0{,}0000000638)^2]$.

Fikseeritud regressioonikordajate hinnangud

```
Random effects:
 Formula: ~1 + vanus + I(vanus^2) + I(vanus^3) | loom
 Structure: General positive-definite, Log-Cholesky pa
              StdDev        Corr
(Intercept) 4.304048e+00 (Intr) vanus  I(v^2)
vanus       8.989510e-02  0.853
I(vanus^2)  9.714711e-05 -0.800 -0.860
I(vanus^3)  6.388967e-08 -0.087 -0.282 -0.032
Residual    2.091722e+01

Fixed effects: mass ~ vanus + I(vanus^2) + I(vanus^3)
                Value Std.Error  DF   t-value p-value
(Intercept) 22.093326 2.0785240 930  10.62933       0
vanus        1.246283 0.0267173 930  46.64697       0
I(vanus^2)  -0.002217 0.0000756 930 -29.31386       0
I(vanus^3)   0.000002 0.0000001 930  27.55055       0
```

- Käsu `lmer` rakendamine lõppeb aga hoopis veateatega:

```
Messages
[55] ERROR:  Downdated X'X is not positive definite, 4.
```

Või siis väljastab *R* küll mingid hinnangud, aga teadete aknas seisab info hindamisprotsessi mittekoondumisest, mistap ei ole ka väljastatud mudeli parameetrite hinnangud ilmselgelt andmetega sobivaimad.

```
Messages
[19] WARNING: Warning in mer_finalize(ans) : false convergence (8)
```

Ilmselt on sobitatud mudel antud andmete ja funktsiooni `lmer` tarvis pisut liiga keeruline.

2.2.

See, et käsu `lme` tulemusena saadud hinnang vasika-spetsiifilise kõvera kuupliikme varieeruvusele peaaegu 0 tuli, vihjab olukorrale, et erinevate vasikate kasvukõverate kuupliikme poolt määratud osad on peaaegu ühesugused ($b3_i \approx 0,\ \forall i$).

Sestap võiks järgnevalt sobitada andmetele ilma juhusliku vasika-spetsiifilise kuupliikmeta mudelit.

```
vasikas.model.2a <- lme(mass ~ vanus+I(vanus^2)+I(vanus^3),
 random=~1+vanus+I(vanus^2)|loom, data=vasikas)
summary(vasikas.model.2a)
```

```
Random effects:
 Formula: ~1 + vanus + I(vanus^2) | loom
 Structure: General positive-definite, Log-Cholesky pa
             StdDev       Corr
(Intercept)  4.047655e+00 (Intr) vanus
vanus        9.137866e-02  0.844
I(vanus^2)   1.105467e-04 -0.737 -0.900
Residual     2.092328e+01

Fixed effects: mass ~ vanus + I(vanus^2) + I(vanus^3)
                 Value Std.Error  DF   t-value p-value
(Intercept) 22.084254 2.0686171 930  10.67585       0
vanus        1.246706 0.0267685 930  46.57361       0
I(vanus^2)  -0.002219 0.0000757 930 -29.32490       0
I(vanus^3)   0.000002 0.0000001 930  28.00254       0
```

Sedakorda ei teki mudeli parameetrite hindamisega probleeme ei funktsioonil `lme` ega ka funktsioonil `lmer`.

Funktsiooni `lmer` võite ise kirja panna. Tulemus:

```
> vasikas.model.2b <- lmer(mass ~ 1+vanus+I(vanus^2)+I(vanus^3) + (1+vanus+I(vanus^2)|loom), data=vasikas)

> summary(vasikas.model.2b)
Linear mixed model fit by REML
Formula: mass ~ 1 + vanus + I(vanus^2) + I(vanus^3) + (1 + vanus + I(vanus^2) |      loom)
   Data: vasikas
  AIC  BIC logLik deviance REMLdev
 9117 9171  -4547     9038    9095
Random effects:
 Groups   Name        Variance   Std.Dev.   Corr
 loom     (Intercept) 1.4631e+01 3.8250e+00
          vanus       8.0842e-03 8.9912e-02  1.000
          I(vanus^2)  1.1587e-08 1.0764e-04 -0.901 -0.901
 Residual             4.3926e+02 2.0959e+01
Number of obs: 988, groups: loom, 55

Fixed effects:
              Estimate Std. Error t value
(Intercept)  2.209e+01  2.064e+00   10.70
vanus        1.247e+00  2.670e-02   46.69
I(vanus^2)  -2.219e-03  7.562e-05  -29.34
I(vanus^3)   1.798e-06  6.419e-08   28.01
```

Nagu näha, on juhuslike efektide varieeruvuse hinnagud pisut erinevad – näiteks loomaspetsiifiliste kõverate vabaliikmete varieeruvuseks hindas funktsioon `lme` $\hat{\sigma}^2_{b0,\mathrm{lme}} = (4{,}048)^2$ ja funktsioon `lmer` $\hat{\sigma}^2_{b0,\mathrm{lmer}} = (3{,}825)^2$. See, et erinevad funktsioonid ja/või arvutiprogrammid annavad tulemuseks pisut erinevad parameetrite hinnangud, on keerulisemate mudelite puhul loomulik, sest kasutatakse erinevaid hindamisalgoritme ja ei ole võimalik täpselt leida, millise parameerite kombinatsiooniga mudel andmetele kõige paremini vastab (*R*-i puhul loetakse funktsiooni `lmer` hinnanguid pisut täpsemaiks, samas, nagu oli eelnevaltki näha, ei pruugi funktsiooni `lmer` kasutatav algoritm alati koonduda).

- Igaks juhuks võib täiendavalt testida, ega uus mudel vasikate kehamassi muutusi kehvemini modelleeri, kui esimene, juhuslikku kuupliiget sisaldanud mudel:

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

```
> anova(vasikas.model.1a,vasikas.model.2a)
                  Model df      AIC      BIC    logLik   Test   L.Ratio p-value
vasikas.model.1a      1 15 9124.785 9198.160 -4547.393
vasikas.model.2a      2 11 9117.097 9170.905 -4547.549 1 vs 2 0.3117441   0.989
```

Järeldus: esimest mudelit ei ole mingit alust lugeda teisest paremaks (p = 0,989 > 0,05). Seega piisab, kui hinnata kõigi vasikate kasvukõveraile ühine kuupliikme kordaja.

Mudelite võrdlemise väljund sisaldab ka keerukamate mudelite võrdlemisel sageli kasutatavate kordajate *AIC* (Akaike informatsioonikriteerium) ja *BIC* (Bayesi informatsioonikriteerium) väärtuseid. Need kordajad ei testi mudelite vahelise erinevuse statistilist olulisust vaid üksnes kirjeldavad seda, sõltuvad nad ühelt poolt sellest, kui hästi mudel andmetele vastab, ja teiselt poolt mudeli keerukusest. Samas on nad kasutatavad ka siis, kui tavaline tõepärasuhte test (*LR-test*) seda ei ole (viimase puhul on eelduseks võrreldavate mudelite allutatus, samuti võivad probleeme tekitada juhuslikud faktorid). Andmetele vastab paremini see mudel, millele vastavad *AIC*-i ja *BIC*-i väärtused on väiksemad.

Antud juhul on nii *AIC* kui ka *BIC* väiksemad mudeli 2 (juhusliku kuupliikmeta mudeli) puhul, mis on veelkordne tõend teise mudeli paremusest.

2.3.

Aga, kas ruutliikme kordajat on mõtet igale vasikale eraldi hinnata. Ehk teisisõnu, kas erinevatele vasikatele hinnatud ruutliikme kordajate varieeruvust on põhjust 0-st erinevaks lugeda?

Vaatame järgi. Sobitame oma andmetele mudeli ka ilma juhusliku ruutliikmeta:

```
vasikas.model.3a <- lme(mass ~ vanus+I(vanus^2)+I(vanus^3),
 random=~1+vanus|loom, data=vasikas)
summary(vasikas.model.3a)
```

```
Random effects:
 Formula: ~1 + vanus | loom
 Structure: General positive-definite, Log-Cholesky pa
            StdDev      Corr
(Intercept) 11.60696171 (Intr)
vanus        0.03766962 0.228
Residual    21.71359004

Fixed effects: mass ~ vanus + I(vanus^2) + I(vanus^3)
                Value Std.Error  DF   t-value p-value
(Intercept) 21.649331 2.5935096 930   8.34750       0
vanus        1.255950 0.0249687 930  50.30091       0
I(vanus^2)  -0.002256 0.0000756 930 -29.85199       0
I(vanus^3)   0.000002 0.0000001 930  28.33958       0
```

Võrreldes mudelite 2 ja 3 väljundeid hakkab silma, et vasika-spetsiifilise ruutliikme mudelist välja jätmine tõi kaasa juhuslike vabaliikmete varieeruvuse peaaegu 3-kordse suurenemise (4,05 *vs* 11,61).

Seega hinnates kõigi vasikate kasvukõveratele ka ühise ruutliikme kordaja, püüab mudel vasikate erinevat kasvamist modelleerida, paigutades kõverate alguspunktid üksteisest kaugemale.

- Kas selline tegevus ka piisavalt edukas on? Testime.

```
> anova(vasikas.model.3a,vasikas.model.2a)
                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
vasikas.model.3a     1  8 9150.988 9190.121 -4567.494
vasikas.model.2a     2 11 9117.097 9170.905 -4547.549 1 vs 2 39.89047  <.0001
```

Ega ikka ei ole küll. Keerulisem mudel (mudel 2) modelleerib vasikate kasvamist statistiliselt oluliselt paremini, kui lihtsam mudel (mudel 3), $p < 0{,}0001$. Ka *AIC*-i ja *BIC*-i väätused on mudeli 2 puhul väiksemad.

- Seega võiks vasikate kasvukõveraid modelleerida 3. järku polünoomiga, kus igale vasikale *i* on hinnatud individuaalsed mudeli vabaliige ning lineaar- ja ruutliikme kordajad:

$$\text{mass}_i = \underbrace{(b0+b0_i)} + \underbrace{(b1+b1_i)}\times\text{vanus}_i + \underbrace{(b2+b2_i)}\times(\text{vanus}_i)^2 + b3\times(\text{vanus}_i)^3,$$

$$b0_i \sim N(0,\sigma^2_{b0}),\ b1_i \sim N(0,\sigma^2_{b1}),\ b2_i \sim N(0,\sigma^2_{b2})\ .$$

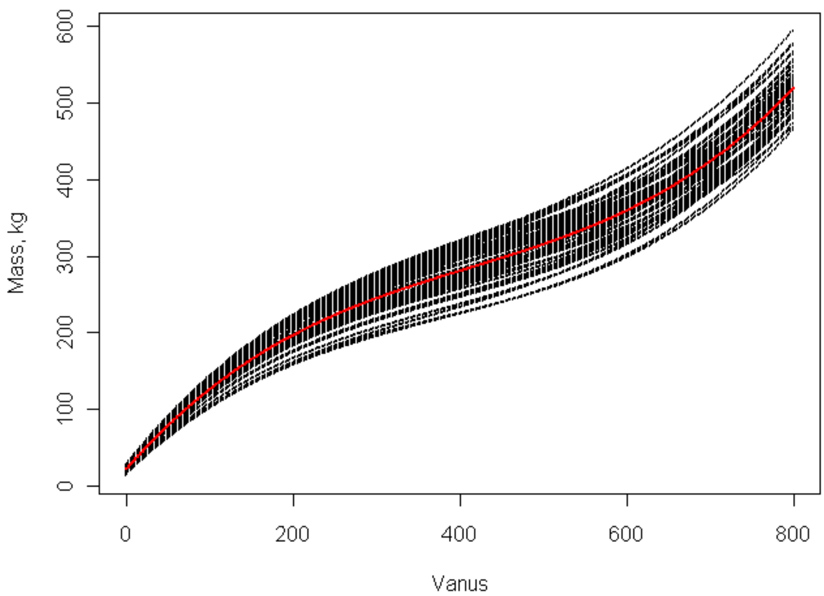
```
> coef(vasikas.model.2)
     (Intercept)    vanus  I(vanus^2)   I(vanus^3)
2684    19.22594 1.188462 -0.002115887 1.797621e-06
2685    17.18089 1.150321 -0.002152235 1.797621e-06
2686    22.67008 1.277953 -0.002287691 1.797621e-06
2687    24.80621 1.313947 -0.002217329 1.797621e-06
2688    21.22402 1.238332 -0.002287055 1.797621e-06
2689    23.64002 1.279134 -0.002217506 1.797621e-06
2690    25.88169 1.334806 -0.002291997 1.797621e-06
2691    24.04751 1.288894 -0.002279265 1.797621e-06
2693    19.93963 1.194271 -0.002166115 1.797621e-06
2695    17.62261 1.140073 -0.002120251 1.797621e-06
2696    26.56294 1.362341 -0.002379284 1.797621e-06
2697    14.78917 1.065995 -0.002068288 1.797621e-06
```

2.4.

Kuidas need kõverad välja näevad?

```
x=rep(seq(0,800,2), length(unique(vasikas$loom)))
y=predict(vasikas.model.2a, data.frame(vanus=x,loom=unique(vasikas$loom)),level=1)
plot(x, y, cex=0.1, xlab="Vanus", ylab="Mass, kg")
lines(rep(seq(0,800,1)),
 predict(vasikas.model.2a, data.frame(vanus=rep(seq(0,800,1))),level=0),lwd=2,col="red")
```

selle käsuga on mudelist hinnatud keskmised kehamassid vanuste 0-800 päeva tarvis



2.5.

Aga isa mõju?

Andmestikus on 8 erineva pulli järglaseid. Kas erinevate isade järglaste kasvukõverad on erinevad?

Arvestades, et isade mõju saab tähendada vaid järglastele pärandatud geenide mõju (miks?) ning viimane on iga järglase puhul erinev (iga järglane saab küll pooled isa geenidest, aga millised täpselt ja millistes kombinatsioonides, on juhuslik), on loomulik käsitleda ka isa mõju juhuslikuna.

Funktsioon `lme` ei võimalda sobitada andmetele mitme juhusliku faktoriga mudelit. Sestap on ainuke variant kasutada funktsiooni `lmer`.

- Esmalt võiks mudelile 2 lisada ka isa-spetsiifilise ruutpolünoomi (et üksikute loomade kasvukõverate eripära modelleerimiseks oli vaja ruutpolünoomi, võiks arvata, et ehk ilmnevad samalaadsed erinevused ka isade vahel). Andmetele on seega vaja sobitada järgnevat mudeli:

$$\text{mass}_{ij} = \left.\begin{array}{l} b0 + b1\times\text{vanus}_{ij} + b2\times(\text{vanus}_{ij})^2 + b3\times(\text{vanus}_{ij})^3 + \\ b0_i + b1_i\times\text{vanus}_{ij} + b2_i\times(\text{vanus}_{ij})^2 + \\ b0_{ij} + b1_{ij}\times\text{vanus}_{ij} + b2_{ij}\times(\text{vanus}_{ij})^2 \end{array}\right\}$$

isa *i* järglase *j* kehamass =
fikseeritud kasvukõver üle kõigi vasikate +
i.-nda isa spetsiifiline kasvukõver +
i.-nda isa *j*.-nda järglase spetsiifiline kasvukõver

Õieti tähendavad isa- ja vasika-spetsiifilised parameetrid toodud mudelis erinevusi keskmisest (fikseeritud) kõverast.

Vastavalt juhuslike efektide olemusele eeldatakse nii isa- kui ka vasika-spetsiifiliste juhuslike regressioonikordajate jaotumist vastavalt normaaljaotusele (mida see sisuliselt tähendab?):
$b0_i \sim N(0,\sigma^2_{b0,S}), b1_i \sim N(0,\sigma^2_{b1,S}), b2_i \sim N(0,\sigma^2_{b2,S}), b0_{ij} \sim N(0,\sigma^2_{b0}), b1_{ij} \sim N(0,\sigma^2_{b1})$ ja $b2_{ij} \sim N(0,\sigma^2_{b2})$.

Vastava mudeli *R*-s rakendamiseks:

```
vasikas.model.4 <- lmer(mass ~ 1+vanus+I(vanus^2)+I(vanus^3) +
    (1+vanus+I(vanus^2)|loom) + (1+vanus+I(vanus^2)|isa), data=vasikas)
```

```
summary(vasikas.model.4)
```

Tulemus:

```
Linear mixed model fit by REML
Formula: mass ~ 1 + vanus + I(vanus^2) + I(vanus^3) + (1 + vanus + I(vanus^2) | loom) + (1 + vanus + I(vanus^2) | isa)
   Data: vasikas
  AIC  BIC logLik deviance REMLdev
 9127 9210  -4546     9037    9093
Random effects:
 Groups   Name        Variance   Std.Dev.   Corr
 loom     (Intercept) 1.9397e+01 4.4042e+00
          vanus       7.8939e-03 8.8848e-02  0.860
          I(vanus^2)  1.2699e-08 1.1269e-04 -0.570 -0.909
 isa      (Intercept) 1.3258e+01 3.6412e+00
          vanus       1.6226e-03 4.0282e-02 -1.000
          I(vanus^2)  1.6091e-09 4.0114e-05  1.000 -1.000
 Residual             4.3642e+02 2.0891e+01
Number of obs: 988, groups: loom, 55; isa, 8

Fixed effects:
              Estimate Std. Error t value
(Intercept)  2.148e+01  2.639e+00    8.14
vanus        1.253e+00  3.206e-02   39.09
I(vanus^2)  -2.226e-03  7.784e-05  -28.60
I(vanus^3)   1.798e-06  6.424e-08   27.99
```

Juhuslike looma- ja isa-spetsiifiliste regressioonikordajate dispersioonid ja standardhälbed.

Üks asi, mis antud analüüsi väljundist silma hakkab, on juhuslike isa-spetsiifiliste kordajate maksimaalne korrelleeritus.

See, et erinevate parameetrite hinnangud on omavahel seotud, on loomulik, sest hinnatakse nad ju kõik koos kompleksselt. Aga absoluutväärtuselt ühega võrduvad isa-spetsiifiliste regressioonikordajate hinnangute vahelised korrelatsioonikordajad on enam kui kahtlased.

Absoluutväärtuselt ühega võrduvad korrelatsioonikordajad näitavad, et mistahes isa-spetsiifilise kordaja alusel on täpselt välja arvutatavad ka teised samale isale vastavad kordajad – ehk siis isa mõju kirjeldamiseks piisab vaid ühest parameetrist.

- Seega võiks järgnevalt püüda andmetele sobitada vaid isa-spetsiifilisi vabaliikmeid sisaldavat mudelit (st, et püüame kõigi loomade keskmise kasvukõvera hinnata kuuppolünoomina, modelleerida iga isa järglaste keskmist erinevust sellest lihtsalt vertikaalse nihkena üles- või allapoole ning viimaks modelleerida vasikaspetsiifilisi kasvukõveraid ruutpolünoomina):

```
vasikas.model.5 <- lmer(mass ~ 1+vanus+I(vanus^2)+I(vanus^3) + (1|isa) +
    (1+vanus+I(vanus^2)|loom), data=vasikas)

summary(vasikas.model.5)
```

Tulemus:

```
Linear mixed model fit by REML
Formula: mass ~ 1 + vanus + I(vanus^2) + I(vanus^3) + (1 | isa) + (1 + vanus + I(vanus^2) | loom)
   Data: vasikas
  AIC  BIC logLik deviance REMLdev
 9118 9177  -4547     9037    9094
Random effects:
 Groups   Name        Variance   Std.Dev.   Corr
 loom     (Intercept) 1.5866e+01 3.9833e+00
          vanus       8.4125e-03 9.1720e-02  0.869
          I(vanus^2)  1.3091e-08 1.1441e-04 -0.590 -0.912
 isa      (Intercept) 0.0000e+00 0.0000e+00
 Residual             4.3863e+02 2.0944e+01
Number of obs: 988, groups: loom, 55; isa, 8

Fixed effects:
             Estimate Std. Error t value
(Intercept)  2.208e+01  2.068e+00   10.68
vanus        1.247e+00  2.682e-02   46.48
I(vanus^2)  -2.219e-03  7.592e-05  -29.23
I(vanus^3)   1.798e-06  6.437e-08   27.93
```

Isa mõjude dispersioon on null. St, et mingeid isa-spetsiifilisi kasvukõveraid ei ole antud andmete alusel võimalik hinnata ja isa mõju vasikate kasvule puudub.

Soovi korral võib silmad üle lasta viimase mudeli ja ilma isa mõjuta mudeli (mudel 2, lk 11) parameetrite hinnangutest – erinevusi praktiliselt pole, seega isa mudelisse lisamine midagi juurde ei anna.

Samuti võib lasta *R*-l võrrelda isa mõjuga ja isa mõjuta mudeleid
NB! Isa mõjuta mudeli näol peab võrdlemisel kasutama funktsiooniga `lmer` hinnatud mudelit (vasikas.model.2b), sest erinevate funktsioonidega `lme` ja `lmer` hinnatud mudelid ei ole omavahel võrreldavad.

```
> anova(vasikas.model.5,vasikas.model.2b)
Data: vasikas
Models:
vasikas.model.2b: mass ~ 1 + vanus + I(vanus^2) + I(vanus^3) + (1 + vanus + I(vanus^2) | loom)
vasikas.model.5: mass ~ 1 + vanus + I(vanus^2) + I(vanus^3) + (1 | isa) + (1 + vanus + I(vanus^2) | loom)
                 Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
vasikas.model.2b 11 9060.2 9114.0 -4519.1
vasikas.model.5  12 9061.5 9120.2 -4518.7 0.7396      1     0.3898
```

Järeldus: keerukamat, juhuslikku isa efekti sisaldavat mudelit ei ole mingit alust lugeda teisest paremaks (p = 0,390 > 0,05).
Sama järelduse saab teha ka kordajate *AIC* ja *BIC* alusel – väiksemad väärtused vastavad mudelile 2.