Sekamallit

Tapio Nummi

11. elokuuta 2005

1 Taustaa

1.1 Kiinteiden vaikutusten malli (Fixed Effects Model)

Yksisuuntainen varianssianalyysin malli on populaation osaryhmien vertailun yksinkertaisin mallityyppi. Perusoletus on, että mukana ovat kaikki kiinnostavat käsittelyt. Täydellisesti satunnaistetussa kokeessa ainoana vaihtelun aiheuttajana pidetään ryhmän saamaa käsittelyä. Kaikki muu vaihtelu johtuu tuntemattomista (kontrolloimattomista) tekijöistä. Malliyhtälö voidaan kirjoittaa muotoon

$$E(y_{ij}) = \mu_i, \tag{1}$$

missä $i = 1, \ldots, k$ (käsittelyt) ja $j = 1, \ldots, n_i$ (käsittelyn i havainnot) tai

$$E(y_{ij}) = \mu + \alpha_i, \tag{2}$$

missä μ on yleiskeskiarvo ja α_i kuvaa käsittelyn i aikaansaamaa poikkeamaa yleiskeskiarvosta μ.

Esimerkki 1. Puutarhuri kokeilee palstoillaan (24 kpl) erilaisten tomaattilajikkeiden (4 kpl) satoisuutta. Palstat valittiin satunnaisesti ja kutakin lajiketta viljeltiin kuudelle palstalle. Jos y_{ij} on nyt lajikkeen i tuotto palstalta j, niin

$$E(y_{ij}) = \mu_i,$$

missä $i = 1, \ldots, 4$ ja $j = 1, \ldots, 6$. Tässä mallissa odotusarvo μ_i on tuntematon kiinteä vakio, joka halutaan estimoida. Mielenkiinto kohdistuu nyt nimenomaan neljään tiettyyn tomaattilajikkeeseen. Mitään johtopäätöksiä muiden tomaattilajikkeiden omanaisuuksista ei tehdä. Nämä neljä vaikutusta ovat nyt kyseisen mallin *kiinteitä* vaikutuksia.

Kiinteiden vaikutusten mallissa virhe ϵ_{ij} määritellään havaintojen y_{ij} poikkeamana odotusarvosta $E(y_{ij})$, jolloin siis

$$\epsilon_{ij} = y_{ij} - E(y_{ij}) = y_{ij} - (\mu + \alpha_i).$$

Tästä saadaan

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad (KVM)$$

tai yhtäpitävästi

$$y_{ij} = \mu_i + \epsilon_{ij}.$$

Varianssianalyysin mallissa parametrit μ ja α_i (tai μ_i) ajatellaan kiinteiksi. Virheet ϵ_{ij} ovat sensijaan riippumattomia satunnaismuuttujia siten, että

$$E(\epsilon_{ij}) = E[y_{ij} - E(y_{ij})] = 0$$

ja

$$Var(\epsilon_{ij}) = \sigma_\epsilon^2, \ \forall \ i, j.$$

Jos oletetaan, että riippumattomat satunnaisvirheet noudattavat normaalijakaumaa

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2),$$

niin havainnot y_{ij} ovat riippumattomia ja noudattavat normaalijakaumaa

$$y_{ij} \sim N(\mu + \alpha_i, \sigma_\epsilon^2), \ \forall \ i, j.$$

1.2 Satunnaisvaikutusten malli (Random Effects Model)

Satunnaisvaikutusten mallissa oletetaan, että "käsittelyt"α_i ovat otos "käsittelyjen"populaatiosta. Satunnaisvaikutusten mallissa satunnaisvaikutuksia käsitellään riippumattomina satunnaismuuttujina. Oletamme, että

$$E(\alpha_i) = 0$$

ja

$$Var(\alpha_i) = E[\alpha_i - E(\alpha_i)]^2 = E(\alpha_i)^2 = \sigma_\alpha^2.$$

Olkoon satunnaismuuttuja α^* ja datasta saatu satunnaismuuttujan realisaatio α_i, jolloin saadaan

$$E(y_{ij} \mid \alpha^* = \alpha_i) = \mu + \alpha_i.$$

Virheeksi saadaan

$$\epsilon_{ij} = y_{ij} - E(y_{ij} \mid \alpha^* = \alpha_i) = y_{ij} - (\mu + \alpha_i)$$

ja malliyhtälöksi

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad (SVM)$$

joka "muistuttaa" kiinteiden vaikutusten mallia KVM. Mallin (SVM) ominaisuudet poikkeavat kuitenkin huomattavasti kiinteiden vaikutusten mallista. Satunnaisvaikutusten mallissa oletetaan (kuten kiinteiden vaikutusten mallissakin), että ϵ_{ij}:t ovat riippumattomia ja niillä on sama varianssi. Lisäksi oletetaan, että satunnaisvaikutukset α_i:t ovat riippumattomia ja niillä on sama varianssi σ_α^2 ja että ϵ_{ij}:t ja α_i:t ovat keskenään riippumattomia. Havaintojen y_{ij} varianssi on nyt

$$Var(y_{ij}) = Var(\mu + \alpha_i + \epsilon_{ij}),$$

mikä on

$$\sigma_y^2 = \sigma_\alpha^2 + \sigma_\epsilon^2.$$

Taulukko 1: Antibioottipitoisuudet.

Erä	1	2	3	4	5	6	7	8
1. mittaus	40	33	46	55	63	35	56	34
2. mittaus	42	34	47	52	59	38	56	29

Havaintojen y varianssi σ_y^2 muodostuu nyt kahdesta komponentista σ_α^2 ja σ_ϵ^2. Näitä variansseja kutsutaan *varianssikomponenteiksi*. Huomataan erikseen, että vaikka α_i ja ϵ_{ij} ovat korreloimattomia, havainnot y_{ij} eivät ole, koska

$$Cov(y_{ij}, y_{ij'}) = \sigma_\alpha^2,$$

kun $j \neq j'$.

Esimerkki 2. Tutkitaan erään antibiootin tehoa kahden vuoden varastoinnin jälkeen. Kokeen perusteella halutaan estimoida lääkkeen keskimääräinen antibioottipitoisuus. Lisäksi halutaan tutkia onko antibioottierällä merkitsevää vaikutusta mitattujen antibioottitasojen vaihteluun. Aineistoon on valittu satunnaisesti kahdeksan antibioottierää (Taulukko 1). Kustakin erästä on tehty kahden alkion otos. Malli on nyt muotoa

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

missä $i = 1, \ldots, 8$ ja $j = 1, 2$. Mallissa siis μ on keskimääräinen antibioottipitoisuus, α_i on erään i liittyvä satunnaisvaikutus ja ϵ_{ij} on satunnaisvirhe. Oletukset:

$$\alpha_i \sim IN(0, \sigma_\alpha^2), \quad \epsilon_{ij} \sim IN(0, \sigma_\epsilon^2)$$

sekä α_i ja ϵ_{ij} riippumattomia. Nyt siis

$$E(y_{ij}) = \mu$$

ja

$$Var(y_{ij}) = \sigma_\alpha^2 + \sigma_\epsilon^2.$$

Kiinnostuksen kohteena tässä aineistossa ovat siis keskimääräinen antibioottitaso μ ja satunnaisvaikutusten α_i varianssi σ^2_α.

Jos nyt oletettaisiinkin, että kahdeksan antibioottierää muodostaisivat koko varaston (eivät olisi satunnaisotos), niin vaikutukset α_i olisivatkin kiinteitä, jolloin malli muodollisesti olisi samannäköinen, mutta

$$E(y_{ij}) = \mu + \alpha_i$$

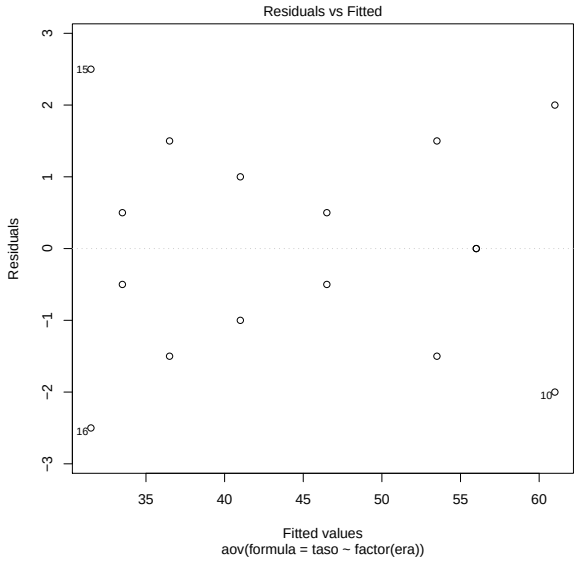
ja

$$Var(y_{ij}) = \sigma^2_\epsilon.$$

Saadaan siten varsin erilainen malli kuin SVM. Nyt siis jokaisella antibioottierällä on oma kiinteä taso α_i ja havaintojen y_{ij} vaihtelu aiheutuu ainoastaan virhevaihtelusta σ^2_ϵ. Tilastollinen malli, jolla aineistoa analysoidaan riippuu siten siitä tulkitaanko antibioottierät satunnaisotokseksi vaiko ei.

Esimerkki 3a. Seuraavassa R-esimerkissä antibioottiaineistoon on sovitettu tavallinen yksisuuntainen varianssianalyysin malli.

```
>  era<-scan() # luetaan muuttujan era arvot
1: 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8
17:
Read 16 items
>  taso<-scan() # luetaan antibiottiarvot
1: 40 42 33 34 46 47 55 52 63 59 35 38 56 56 34 29
17:
Read 16 items
>  ls() # tulostetaan työtilan sisaltö
[1] "era"  "taso"
>  plot(era,taso) # piirretään muuttujien yhteisjakauma
# Ajetaan varianssianalyysi
```



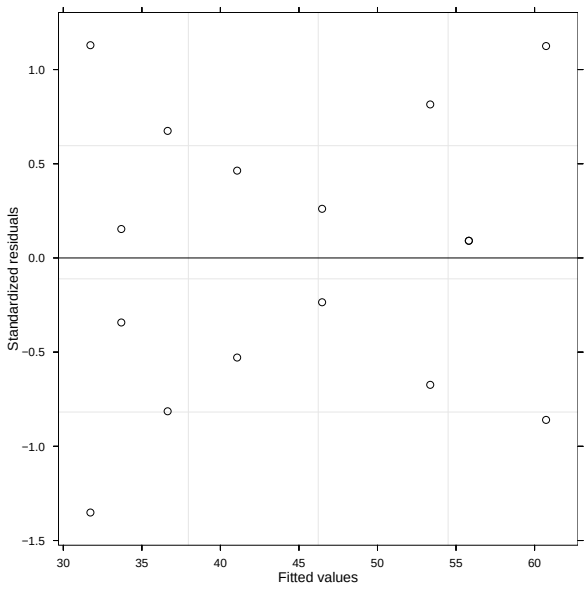
Kuva 1: R-plottaus varianssianalyysin mallista

```
# Huomaa, että muuttujaan era käytetään funktiota factor
>  antib.kiint<-aov(taso~factor(era))
>  summary(antib.kiint)
             Df  Sum Sq Mean Sq F value    Pr(>F)
factor(era)  7 1708.44  244.06  60.077 2.744e-06 ***
Residuals    8   32.50    4.06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>library(lattice)
#Tässä monistetta varten tausta on vaihdettu valkoiseksi ja
#teksti mustaksi
>trellis.device(device=getOption("device"),color=FALSE,bg="white")
>  plot(antib.kiint) # Varianssianalyysiin liittyvää grafiikkaa
```

Esimerkki 3b. Tässä esimerkissä antibioottiaineistoon on sovitettu satunnaisvaikutusten malli.

```
#  Muodostetaan muuttujista taso ja era havaintomatriisi
>  antib<-data.frame(taso=taso, era=factor(era))
>  antib # Tulostetaan aineiston sisalto
    taso era
1    40   1
2    42   1
3    33   2
4    34   2
5    46   3
    ...

>  library(nlme) # Ladataan ohjelmisto NLME
Attaching package: 'nlme'

        The following object(s) are masked from package:stats :

         contr.SAS
# Estimoidaan satunnaisvaikutusten malli.
>  antib.satv<-lme(taso~1, data=antib, random=~1|era)
>  summary(antib.satv)
Linear mixed-effects model fit by REML
  Data: antib
        AIC      BIC    logLik
   101.0371 103.1613 -47.51855

Random effects:
```

```
  Formula: ~1 | era
        (Intercept) Residual
StdDev:    10.95445 2.015564

Fixed effects: taso ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 44.9375  3.905626  8 11.50584  <.0001

Standardized Within-Group Residuals:
         Min          Q1         Med          Q3         Max
-1.35131971 -0.56486636  0.09135843  0.51634833  1.12937550

Number of Observations: 16
Number of Groups: 8
# Testataan tasoparametri
>  anova(antib.satv)
             numDF denDF  F-value p-value
(Intercept)     1     8 132.3843  <.0001
# Lasketaan mallin parametreille luottamusvälit
>  intervals(antib.satv)
Approximate 95% confidence intervals

  Fixed effects:
               lower    est.    upper
(Intercept) 35.93111 44.9375 53.94389

  Random Effects:
   Level: era
                   lower     est.    upper
```



Kuva 2: R-plottaus satunnaisvaikutusten mallista

```
sd((Intercept)) 6.430086 10.95445 18.66228


  Within-group standard error:
    lower     est.    upper
1.234752 2.015564 3.290134
# piirretään mallin residuaalikuvio
>  plot(antib.satv)
>
```

1.3 Sekamalli (Mixed Model)

Mallia, jossa on sekä kiinteitä että satunnaisia vaikutuksia kutsutaan sekamalliksi. Malliyhtälö voidaan kirjoittaa muotoon

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

missä μ ja ϵ_{ij} ovat kuten edellä, α_i on kiinteä vaikutus ($i = 1, \ldots, k$) ja β_j on satunnaisvaikutus ($j = 1, \ldots, n$).

Esimerkki 4. Alustavissa testeissä uusi verenpainelääke oli osoittautunut tehokkaaksi. Seuraavassa vaiheessa vaikutusta haluttiin tutkia tarkemmin suuremmalla aineistolla. Kokeiluun otti osaa 9 eri maata, jotka tässä ajatellaan satunnaisotokseksi kaikista maista. Joka maasta valittiin ryhmä ihmisiä jotka arvottiin joko lääkeryhmään tai placeboryhmään. Vastemuuttuja on *difbp* (=**dif**ference in **b**lood **p**ressure before and after treatment). Tutkimuksen päätavoite on selvitää voidaanko alustavan tutkimuksen tulokset vahvistaa isommalla aineistolla. Mallissa ryhmä (lääke c. placebo) on kiinteä vaikutus ja maa satunnaisvaikutus. Aineisto on annettu liitteessä 1. Jos aineistoon sovitetaan sekamalli, niin malli voisi olla

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

missä $k = 1, \ldots, n_{ij};\ j = 1, \ldots, 9;\ i = 1, 2$, y_{ijk} on henkilön k $difbp$ käsittelylle i ja maana j, μ on yleiskeskiarvo, α_i on käsittelyn i (lääke c. placebo) vaikutus β_j on maan j satunnaisvaikutus ja ϵ_{ijk} on satunnaisvirhe. Mallin oletukset olisivat

$$\beta_j \sim IN(0, \sigma^2_\beta), \epsilon_{ijk} \sim IN(0, \sigma^2_\epsilon) \ \text{ ja } \ Cov(\beta_j, \epsilon_{ijk}) = 0,\ \forall i, j, k.$$

Mielenkiinnon kohteena olisivat nyt esimerkiksi hypoteesit

$$H_0 : \alpha_1 = \alpha_2 \text{ ja } H_0 : \sigma^2_\beta = 0.$$

Voitaisiin siis olla kiinnostuneita siitä mikä on lääkkeen vaikutus verrattuna placeboon ja siitä vaihteleeko vaikutusten taso maittain.

1.4 Kiinteä vs. satunnainen vaikutus

Aina ei ole itsestään selvää se, onko vaikutus kiinteä vai satunnainen. Jos esimerkiksi tarkastellaan vuoden vaikutusta vehnäsatoon, onko vuoden vaikutus vehnäsatoon kiinteä vai satunnainen vaikutus? Peräkkäisten vuosien arvoja ei voida pitää satunnaisina, mutta yksittäisen vuoden vaikutus voidaan ajatella satunnaiseksi. Kuitenkin, jos ollaan kiinnostuttu esimerkiksi vertaamaan tiettyjä vuosia toisiinsa, on vaikutusta parempi käsitellä kiinteänä.

Kun pohditaan onko vaikutus kiinteä vai satunnainen on syytä miettiä suoritetaanko analyysit juuri kyseisillä faktoritasoilla vai voidaanko tasot ajatella otokseksi jostakin tasojen populaatiosta. Ensinmainitussa tapauksessa vaikutuksia tulisi pitää kiinteinä ja jälkimmäisessä satunnaisina.

Huomautus. Satunnaisefektit eivät välttämättä ole normaalisti jakautuneita. Normaalijakaumaoletus kuitenkin usein tehdään, kun tarkastellaan estimaattoreiden jakaumaominaisuuksia

2 Yleinen sekamalli (Mixed Model)

2.1 Taustaa

Lähtökohtana on tavallinen lineaarinen malli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

missä $\mathbf{y}$ on $n \times 1$ havaintovektori, $\mathbf{X}$ on annettu $n \times p$ suunnittelumatriisi, $\boldsymbol{\beta}$ on $p \times 1$ parametrivektori ja $\boldsymbol{\epsilon}$ on $n \times 1$ satunnaisvirheiden muodostama vektori.

Taulukko 2:

Luokka	y_{i1}	y_{i2}	y_{i3}	y_{i4}
$i = 1$	3	3	12	2
$i = 2$	11	13	17	7
$i = 3$	4	2	1	33

Esimerkki 5.

Malli

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

missä $i = 1, 2, 3$ ja $j = 1, 2, 3, 4$. Voidaan esittää matriisiyhtälönä

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 12 \\ 2 \\ 11 \\ 13 \\ 17 \\ 7 \\ 4 \\ 2 \\ 1 \\ 33 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \boldsymbol{\epsilon}.$$

eli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

missä havainto- ja virhevektoreita merkitään $\mathbf{y}$ ja $\boldsymbol{\epsilon}$ sekä

$$\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3)'.$$

Merkinnöissä voidaan käyttää summausvektoria $\mathbf{1}_k = (1, 1, \ldots, 1)'$. Tätä merkintää käyttäen

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{1}_{12} \, \mathbf{0} & \mathbf{1}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_4 \end{pmatrix}$$

ja

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{1}_{12}\mu + \begin{pmatrix} \mathbf{1}_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_4 \end{pmatrix} \boldsymbol{\alpha},$$

missä $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)'$. Malli voidaan lausua kätevästi myös Kroneckerin tulon avulla

$$\mathbf{y} = \left(\mathbf{1}_3 \otimes \mathbf{1}_4\right) \mu + \left(\mathbf{I}_3 \otimes \mathbf{1}_4\right) \boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

missä $\mathbf{I}_3$ on 3×3 yksikkömatriisi. *Huom.* Edelläoleva malli on yliparametrisoitu, koska $\mathbf{X}$:n sarakeaste ei ole täysi.

Esimerkki 5. (jatkoa, yliparametrisoitu malli).
Usean muuttujan regressiomalli

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

määritellään R:ssä tyyliin

$$y \sim x1 + x2 + x3.$$

Vastaava mallimatriisi voidaan osittaa seuraavasti

$$\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3).$$

Vakiotermi tulee R-ohjelmassa oletusarvona mukaan. Jos vakiotermi halutaan jättää pois, niin annetaan

$$y \sim x1 + x2 + x3 - 1.$$

Tarkastellaan seuraavaksi yksinkertaista varianssianalyysin mallia

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij},\ i = 1, \ldots, n_j;\ j = 1, \ldots, k.$$

Vastaava mallimatriisi on nyt

$$\mathbf{X} = (\mathbf{1}, \mathbf{X}_a),$$

missä $\mathbf{X}_a$ sisältää sopivasti ykkösiä ja nollia. Matriisin $\mathbf{X}$ aste ei ole täysi (malli on yliparametrisoitu, koska esimerkiksi matriisin $\mathbf{X}_a$ sarakkeet summautuvat vektoriksi $\mathbf{1}$ (ks. esim. 5). Nyt esimerkiksi tavallista PNS estimaattoria

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

ei voida laskea, koska matriisi $\mathbf{X}'\mathbf{X}$ on singulaarinen. Eräs mahdollisuus on estimoida malli

$$y_{ij} = \alpha_j + \epsilon_{ij},\ i = 1, \ldots, n_j;\ j = 1, \ldots, k$$

eli jättää vakiotermi pois. Usein menetellään kuitenkin niin, että käytetään alkuperäisen tilalla mallimatriisia

$$\mathbf{X}^* = (\mathbf{1}, \mathbf{X}_a\mathbf{C}_a),$$

missä $\mathbf{C}_a$ (*contrast matrix*) on valittu siten, että matriisin $\mathbf{X}^*$ sarakeaste on täysi. Parametrit liittyvät toisiinsa seuraavasti

$$\boldsymbol{\alpha} = \mathbf{C}_a\boldsymbol{\alpha}^*,$$

missä vektori $\boldsymbol{\alpha}$ sisältää alkuperäisen mallin parametrit ja vektori $\boldsymbol{\alpha}^*$ muunnetun mallin parametrit. Jos nyt R-ohjelmistossa sovitetaan malli $y\tilde{}\ a$ niin saadaan estimaatit parametreille μ ja α^*.

```
# Muodostetaan havaintomatriisi
>  esim5<-data.frame(luokka=factor(c(1,1,1,1,2,2,2,2,3,3,3,3)),
+ y=c(3,3,12,2,11,13,17,7,4,2,1,33))
# Sovitetaan aineistoon KVM (yliparametrisoitu)
>  kvm<-lm(y ~ luokka, data=esim5)
>  summary(kvm)
    ...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.000      4.796   1.043    0.324
luokka2        7.000      6.782   1.032    0.329
luokka3        5.000      6.782   0.737    0.480

    ...
>  model.matrix(kvm)
    (Intercept) luokka2 luokka3
1             1       0       0
2             1       0       0
3             1       0       0
4             1       0       0
5             1       1       0
    ...
>  contrasts(esim5$luokka)
  2 3
1 0 0
2 1 0
3 0 1
>  alf.star<-coef(kvm)
>  alf.star
```

```
(Intercept)     luokka2     luokka3
          5           7           5


>  Ca<-contrasts(esim5$luokka)
>  Ca %*% alf.star[-1]
     [,1]
1    0
2    7
3    5
>  dummy.coef(kvm)
>  kvmb<-update(kvm, y~luokka-1) # Malli ei yliparametrisoitu
>  summary(kvmb)
    ...
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
luokka1    5.000      4.796   1.043   0.3243
luokka2   12.000      4.796   2.502   0.0337 *
luokka3   10.000      4.796   2.085   0.0667 .
   ...
>  model.matrix(kvmb)
    luokka1 luokka2 luokka3
1        1       0       0
2        1       0       0
3        1       0       0
4        1       0       0
5        0       1       0
    ...
```

2.2 Sekamalli

Sekamallissa osa mallin parametreista ajatellaan satunnaismuuttujiksi. Mallin yleinen muoto on

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

missä $\mathbf{Z}$ on $n \times c$ satunnaisvaikutusten suunnitelumatriisi ja $\mathbf{u}$ on $c \times 1$ satunnaisvaikutusten vektori. Sekamallissa määritellään

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \qquad \text{ja} \qquad E(\mathbf{y} \mid \mathbf{u}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

sekä

$$\boldsymbol{\epsilon} = \mathbf{y} - E(\mathbf{y} \mid \mathbf{u}),$$

missä

$$E(\mathbf{u}) = \mathbf{0} \qquad \text{ja} \qquad E(\boldsymbol{\epsilon}) = \mathbf{0}.$$

Sekamallissa oletetaan lisäksi, että *satunnaisvirheet ja satunnaisvaikutukset ovat riippumattomia,* jolloin

$$Cov(\mathbf{u}, \boldsymbol{\epsilon}) = \mathbf{O}.$$

Sekamallin yleisessä muodossa oletetaan, että $Var(\mathbf{u}) = \mathbf{D}$ ja $Var(\boldsymbol{\epsilon}) = \mathbf{R}$, missä $\mathbf{D}$ ja $\mathbf{R}$ ovat kovarianssimatriiseja. Havaintojen $\mathbf{y}$ kovarianssimatriisi on nyt muotoa

$$Var(\mathbf{y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}.$$

Huomautus. Merkinnöissä käytetään $\mathbf{y}$*:tä ja* $\mathbf{u}$*:ta merkitsemään satunnaisvektoreita* $\mathbf{y}$ *ja* $\mathbf{u}$ *sekä näiden arvoja.*

Esimerkki 6. Osaruutukoe (Split-Plot).

Tarkastellaan lannoituksen vaikutusta satoon. Sovelletaan kolmea lannoitustasoa (*main plots*) neljään peltoalueeseen ja alueiden sisällä kokeillaan kahta eri viljalajiketta (*sub-plots*). Maaperän erilaisuus peltojen välillä saattaa aiheuttaa lisävaihtelua tuloksiin (*block-effect*). Voidaan ajatella, että peltojen "tasot"ovat otos äärettömästä populaatiosta. Pelto-efekti on ns. satunnaisefekti.

Kiinteät vaikutukset

lannoite (tasot: $\alpha_1 \;\; \alpha_2 \;\; \alpha_3$)

laji (tasot: $\beta_1 \;\; \beta_2$)

Satunnaisvaikutukset

Blockin eli peltojen "tasot"on otos äärettömästä populaatiosta.

"Otoskoko "4: $B_1 \;\; B_2 \;\; B_3 \; B_4$

Lannoitteen ja maaperän yhdysvaikutus on satunnaisefekti.

"Otoskoko "12: $\quad \gamma_{ij} = \alpha_i * B_j, \quad i = 1, 2, 3; j = 1, 2, 3, 4$

Sekamallissa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

kiinteät vaikutukset ovat nyt $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_\alpha, \boldsymbol{\beta}'_\beta)$ ja vastaava suunnittelumatriisi on

$$\mathbf{X} = (\mathbf{X}_\alpha, \mathbf{X}_\beta)$$

sekä satunnaisvaikutukset $\mathbf{u}' = (\mathbf{u}'_B, \mathbf{u}'_{\alpha * B})$ ja satunnaisvaikutusten suunnittelumatriisi on

$$\mathbf{Z} = (\mathbf{Z}_B, \mathbf{Z}_{B*\alpha}).$$

Jos lisäksi oletetaan, että

$$Var(\boldsymbol{\epsilon}) = \sigma^2_\epsilon \mathbf{I}_n$$

ja

$$Var\begin{pmatrix}\mathbf{u}_B \\ \mathbf{u}_{B*\alpha}\end{pmatrix} = \begin{pmatrix}\sigma_B^2\mathbf{I} & \mathbf{O} \\ \mathbf{O} & \sigma_{B*\alpha}^2\mathbf{I}\end{pmatrix},$$

niin saadaan

$$Var(\mathbf{y}) = \mathbf{Z}Var(\mathbf{u})\mathbf{Z}' + Var(\boldsymbol{\epsilon})$$

$$= \sigma_B^2\mathbf{Z}_B\mathbf{Z}_B' + \sigma_{B*\alpha}^2\mathbf{Z}_{B*\alpha}\mathbf{Z}_{B*\alpha}' + \sigma_\epsilon^2\mathbf{I}_n,$$

missä siis σ_ϵ^2, σ_B^2 ja $\sigma_{B*\alpha}^2$ ovat varianssikomponentteja.

Esimerkki 7. Satunnaistettujen lohkojen koe (Randomized Block Design). Esimerkkinä käsitellään aineistoa ErgoStool (kirjastossa *nlme*). Tässä aineistossa on yhdeksän koehenkilöä joilta on mitattu neljälle eri tyyppiselle jakkaralle nousemiseen liittyvä ponnistus. Tarkoituksena on vertailla eri jakkaroiden saamia tuloksia. Muuttuja Type (*experimental factor*) on kiinteä vaikutus. Yhdeksän koehenkilöä edustaa otosta populaatiosta johon tuloksia halutaan yleistää, joten muuttuja Subject (*blocking factor*) on satunnaisvaikutus.

```
>  library(nlme)
Loading required package: nls
>  data(ergoStool)
>  contrasts(ergoStool$Type) # Muuttujaan Type liittyvät "kontrastit"
    T2 T3 T4
T1  0  0  0
T2  1  0  0
T3  0  1  0
T4  0  0  1

# Poimitaan aineistosta Subjektiin numero 1 liittyvä osa
>  ergo1<-ergoStool[ergoStool$Subject == "1",]
# Subjektiin numero 1 liittyva mallimatriisi
```

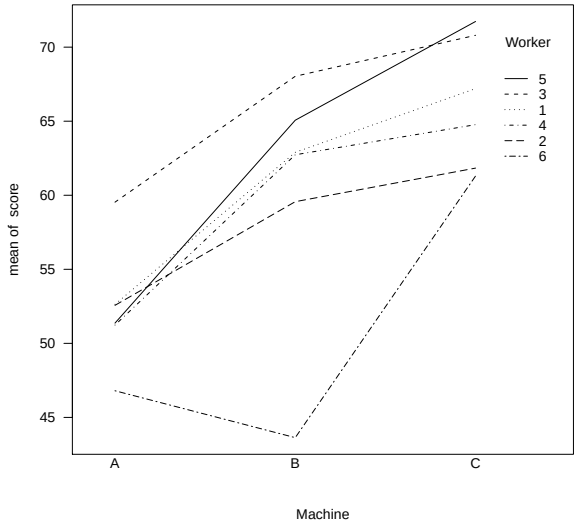
```
>  model.matrix(ergoStool~Type, ergo1)
   (Intercept) TypeT2 TypeT3 TypeT4
1           1      0      0      0
2           1      1      0      0
3           1      0      1      0
4           1      0      0      1
    ...

>  f1<-lme(effort~Type, data=ergoStool, random=~1|Subject)
>  summary(f1)
Linear mixed-effects model fit by REML
  Data: ergoStool
        AIC      BIC   logLik
   133.1308 141.9252 -60.5654

Random effects:
  Formula: ~1 | Subject
         (Intercept) Residual
StdDev:    1.332465 1.100295

Fixed effects: effort ~ Type
               Value Std.Error DF   t-value p-value
(Intercept) 8.555556 0.5760122 24 14.853079  <.0001
TypeT2      3.888889 0.5186838 24  7.497609  <.0001
TypeT3      2.222222 0.5186838 24  4.284348  0.0003
TypeT4      0.666667 0.5186838 24  1.285304  0.2110
    ...
```

```
>  f2<-update(f1, effort~Type-1)
>  summary(f2)
Linear mixed-effects model fit by REML
  Data: ergoStool
        AIC      BIC   logLik
   133.1308 141.9252 -60.5654
Random effects:
  Formula: ~1 | Subject
         (Intercept) Residual
StdDev:    1.332465 1.100295

Fixed effects: effort ~ Type - 1
            Value Std.Error DF  t-value p-value
TypeT1  8.555556 0.5760123 24 14.85308  <.0001
TypeT2 12.444444 0.5760123 24 21.60448  <.0001
TypeT3 10.777778 0.5760123 24 18.71102  <.0001
TypeT4  9.222222 0.5760123 24 16.01046  <.0001
    ...
```

Ajossa *f2* saadaan estimaatit kaikille faktorin *Type* tasoille. Ajossa *f1* on käytetty *treatment* "kontrasteja"(oletus R-ohjelmistossa). Estimaatit tulkitaan nyt siten, että *T1* asetetaan perustasoksi ja muut estimaatit ovat poikkeamia tästä perustasosta.

Esimerkki 8. Toistomittaukset.

Jakkaraesimerkissä kukin henkilö kokeili kutakin jakkaraa kerran. Usein kuitenkin suoritetaan kokeita, joissa on järkevää tehdä monia mittauksia. Käsitellään esimerkkinä aineistoa Machines (kirjastossa nlme). Komennolla *?Machines* saadaan aineiston tarkempi kuvaus. Seuraavaan esimerkkiin on valittu osa tulos-



Kuva 3: R-ohjelmiston "interaction plot"aineistosta Machines.

tuksesta:

```
Data on an experiment to compare three brands of machines used in
an industrial process are presented in Milliken and Johnson (p. 285,
1992). Six workers were chosen randomly among the employees of a fac-
tory to operate each machine three times. The response is an overall
productivity  score taking into account the number and quality of
components produced.

>  data(Machines)
>  Machines
Grouped Data: score ~ Machine | Worker
    Worker Machine score
1       1       A  52.0
```

```
2       1       A  52.8
3       1       A  53.1
4       2       A  51.8
5       2       A  52.8
6       2       A  53.1
7       3       A  60.0
8       3       A  60.2
            ...
```

```
>  attach(Machines)
>  interaction.plot(Machine, Worker, score, las=1)
>  detach()
```

Jos nyt kuviossa esitetyt käyrät olisivat yhdensuuntaisia, ei koneen ja ihmisen välillä olisi yhdysvaikutusta. Estimoidaan aluksi malli

```
>  fm1<-lme(score~Machine, data=Machines, random=~1 | Worker)
>  fm1
    ...
   Fixed: score ~ Machine
(Intercept)    Machine1    Machine2
   59.650000    3.983333    3.311111

Random effects:
  Formula: ~1 | Worker
         (Intercept) Residual
StdDev:    5.146552 3.161647
Number of Observations: 54
Number of Groups: 6
```

Koska muuttujan Worker arvot ovat satunnaisotos kiinnostuksen kohteena olevasta populaatiosta, ovat myös muuttujan Worker arvojen erot eri koneilla satunnaisvaikutuksia.

```
>  fm2<-update(fm1,  random=~1 | Worker/Machine)
>  fm2
    ...
Random effects:
  Formula: ~1 | Worker
         (Intercept)
StdDev:    4.781049
  Formula: ~1 | Machine %in% Worker
         (Intercept)  Residual
StdDev:    3.729536 0.9615768
Number of Observations: 54
Number of Groups:
              Worker Machine %in% Worker
                   6                  18
>  anova(fm1,fm2)
     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fm1      1  5 310.1209 310.1209 -145.2309
fm2      2  6 242.8620 242.8620 -109.6355 1 vs 2 71.19063  <.0001
```

Testauksessa malli *fm2* osoittautui paremmaksi. Lasketaan seuraavaksi luottamusvälit mallit parametreille.

```
>  intervals(fm2)
Approximate 95% confidence intervals

  Fixed effects:
```

```
                lower      est.    upper
(Intercept) 47.314060 52.355556 57.39705
MachineB     3.116066  7.966667 12.81727
MachineC     9.066066 13.916667 18.76727
attr(,"label")
[1] "Fixed effects:"

  Random Effects:
   Level: Worker
                   lower     est.    upper
sd((Intercept)) 2.249827 4.781049 10.16008
   Level: Machine
                   lower     est.    upper
sd((Intercept)) 2.382854 3.729536 5.837302

  Within-group standard error:
     lower      est.     upper
0.7635818 0.9615768 1.2109116
# poimitaan osa aineistoa muuttujaan Machine1
>  Machine1<-Machines[Machines$Worker=="1",]
>  model.matrix(score~Machine, Machine1)
    (Intercept) MachineB MachineC
1             1        0        0
2             1        0        0
3             1        0        0
19            1        1        0
20            1        1        0
    ...
>  model.matrix(~Machines$Machine-1,Machine1)
```

```
    MachineA MachineB MachineC
1          1        0        0
2          1        0        0
3          1        0        0
19         0        1        0
20         0        1        0
    ...
>  fm3<-update(fm2, random=~Machine-1| Worker)
>  summary(fm3)
Linear mixed-effects model fit by REML
  Data: Machines
        AIC      BIC    logLik
   247.6295 247.6295 -104.1556

Random effects:
  Formula: ~Machine - 1 | Worker
  Structure: General positive-definite, Log-Cholesky parametrization
          StdDev    Corr
MachineA 4.0792807 MachnA MachnB
MachineB 8.6252908 0.803
MachineC 4.3894795 0.623  0.771
Residual 0.9615766

Fixed effects: score ~ Machine
                Value Std.Error DF   t-value p-value
(Intercept) 52.35556  1.680711 46 31.150834  <.0001
MachineB     7.96667  2.420851 46  3.290854  0.0019
MachineC    13.91667  1.540100 46  9.036211  <.0001
    ...
```

```
>  anova(fm2,fm3)
     Model df      AIC      BIC    logLik   Test L.Ratio p-value
fm2      1  6 239.2785 239.2785 -107.8438
fm3      2 10 247.6295 247.6295 -104.1556 1 vs 2 7.37635  0.1173
```

Malli 3 ei selvästikään ole parempi kuin malli 2 (vrt. esim. BIC ja LRT), joten malli 2 jää voimaan.

3 Kiinteän osan estimointi ja hypoteesien testaus

Pienimmän neliösumman (PNS) estimaattori kiinteälle osalle on

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}.$$

PNS-estimaattori ei kuitenkaan ota huomioon havaintojen korreloituneisuutta, joten se ei ole kovin mielenkiintoinen käytännön analyyseissa. Paras lineaarinen ja harhaton estimaattori kiinteälle osalle on

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

ja

$$Var(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-},$$

jos $\mathbf{V}$ on ei-singulaarinen. Jos oletetaan normaalijakauma, niin $\tilde{\boldsymbol{\beta}}$ on myös suurimman uskottavuuden estimaattori. Edellä oletettiin matriisi $\mathbf{V}$ tunnetuksi. Käytännössä matriisia $\mathbf{V}$ ei yleensä kuitenkaan tunneta, vaan se korvataan estimaatilla $\hat{\mathbf{V}}$. Nyt kuitenkaan $\tilde{\boldsymbol{\beta}}$ ei ole enää paras lineaarinen ja harhaton estimaattori, mutta $\tilde{\boldsymbol{\beta}}$ on suurimman uskottavuuden estimaattori, jos $\hat{\mathbf{V}}$ on $\mathbf{V}$:n SU-estimaatti. Samoin

$$(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}$$

on $Var(\tilde{\boldsymbol{\beta}})$:n alaspäin harhainen estimaattori, koska estimaattorina $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^-$ ei ota huomioon sitä, että $\mathbf{V}$:n parametrit on estimoitu.

Tarkastellaan nollahypoteesia

$$H_0 : \mathbf{H}'\boldsymbol{\beta} = \mathbf{0}$$

ja vaihtoehtoista hypoteesia

$$H_1 : \mathbf{H}'\boldsymbol{\beta} \neq \mathbf{0},$$

missä $\boldsymbol{\beta}$ on $p \times 1$ vektori ja $\mathbf{H}$ on annettu q-asteinen $p \times q$ matriisi ($q \leq p$). Merkitään, että $\mathbf{h}_j$ on $\mathbf{H}$:n j. sarake ja oletetaan, että $\mathbf{h}_j'\boldsymbol{\beta}$ on estimoituva. Estimaattorina lineaarikombinaatioille $\mathbf{H}'\boldsymbol{\beta}$ käytämme estimaattoria $\mathbf{H}'\tilde{\boldsymbol{\beta}}$ ja koska

$$Var(\mathbf{H}'\tilde{\boldsymbol{\beta}}) = \mathbf{H}'Var(\tilde{\boldsymbol{\beta}})\mathbf{H} = \mathbf{H}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{H},$$

varianssin estimaattiksi saadaan

$$\mathbf{H}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^-\mathbf{H}.$$

Testisuureen

$$F = \frac{(\mathbf{H}'\tilde{\boldsymbol{\beta}})'[\mathbf{H}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{H}]^-(\mathbf{H}'\tilde{\boldsymbol{\beta}})}{rank(\mathbf{H})}$$

jakaumaa voidaan approksimoida F-jakaumalla, jonka osoittajan vapausaste on $\mathbf{H}$:n aste $rank(\mathbf{H})$ ja nimittäjän vapausaste on yleisessä tapauksessa approksimoitava. SAS:n MIXED-proceduurissa on käytössä useita menetelmiä vapausasteiden approksimointiin. Oletuksena on optio *contain*, joka antaa vapausasteiksi havaintojen lukumäärän vähennettynä kiinteiden ja satunnaisten vaikutusten vapausasteilla, jos mikään satunnaisvaikutus ei sisällä testattavaa kiinteätä vaikutusta.

Jos nyt $\mathbf{H} = \mathbf{h}$ ja $rank(\mathbf{h}) = 1$ sekä $T = \sqrt{F}$, niin testisuureen

$$T = \frac{\mathbf{h}'\tilde{\boldsymbol{\beta}}}{\sqrt{\mathbf{h}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^-\mathbf{h}}}$$

jakaumaa voidaan approksimoida t-jakaumalla. Jos esimerkissä 4 halutaan testata nollahypoteesi

$$H_0 : \alpha_1 = \alpha_2,$$

niin $\mathbf{h}' = (1, -1)$ ja testisuureen T osoittaja on $\mathbf{h}'\tilde{\boldsymbol{\beta}} = (1, -1)\begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{pmatrix} = \tilde{\alpha}_1 - \tilde{\alpha}_2$. Tällöin $(1 - \alpha)\times$ 100 %:n luottamusväliksi saadaan

$$\mathbf{h}'\tilde{\boldsymbol{\beta}} \pm t_{\alpha/2}\sqrt{\mathbf{h}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{h}}.$$

Vapausasteet saadaan varianssin $\mathbf{h}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{h}$ vapausasteista. Vapausasteiden approksimointiin voidaan käyttää samantyyppisiä menetelmiä kuin F-statistiikassa.

4 Satunnaisvaikutusten ennustaminen

Tarkastellaan esimerkin 2 mallia

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

missä $j = 1, \ldots, n$ ja $i = 1, \ldots, a$. Parametrien α_i prediktorina voidaan käyttää ehdollista odotusarvoa

$$\hat{\alpha}_i = E(\alpha_i \mid \bar{y}_{i.}).$$

Jos oletetaan normaalijakauma

$$\begin{pmatrix} \alpha_i \\ \bar{y}_{i.} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n} \end{pmatrix} \right),$$

niin

$$\begin{aligned} \hat{\alpha}_i &= E(\alpha_i \mid \bar{y}_{i.}) \\ &= \frac{n\sigma_\alpha^2}{n\sigma_\alpha^2 + \sigma_\epsilon^2}(\bar{y}_{i.} - \mu). \end{aligned}$$

Prediktori $\hat{\alpha}_i$ ei kuitenkaan ole vielä käyttökelpoinen käytännön tarkasteluissa, koska se sisältää tuntemattomat parametrit σ_α^2, σ_ϵ^2 ja μ. Ennusteelle saadaan numeerinen arvo, kun tuntemattomat parametrit korvataan estimaateillaan. Saadaan siis

$$\hat{\alpha}_i = \frac{n\hat{\sigma}_\alpha^2}{n\hat{\sigma}_\alpha^2 + \hat{\sigma}_\epsilon^2}(\bar{y}_{i.} - \bar{y}_{..}).$$

Jos esimerkiksi $\hat{\sigma}_\alpha^2 = 120$, $\hat{\sigma}_\epsilon^2 = 4.06$, $\bar{y}_{i.} = 41$ ja $\bar{y}_{..} = 44.94$, niin saadaan

$$\hat{\alpha}_i = \frac{2 \times 120}{2 \times 120 + 4.06}(41 - 44.94) = -3.87.$$

Yleisesti $\boldsymbol{\beta}$ ja $\mathbf{u}$ ratkaistaan ns. sekamalliyhtälöistä, jotka johdetaan vektoreiden $\mathbf{y}$ ja $\mathbf{u}$ yhteisjakaumasta

$$f(\mathbf{y}, \mathbf{u}) = g(\mathbf{y} \mid \mathbf{u})h(\mathbf{u})$$

$$= C \exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})] \exp[-\frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}].$$

Herd	Sire	Yield
1	A	110
1	D	100
2	B	110
2	D	100
2	D	100
3	C	110
3	C	110
3	D	100
3	D	100

Sekamalliyhtälöt voidaan lausua seuraavasti:

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

Ratkaisuiksi saadaan

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

ja

$$\tilde{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}).$$

Esimerkki 9. (Maidontuotanto)

Vaikutukset:

- isän geneettinen arvo (SIRE), satunnainen

- lauma (HERD), kiinteä

Oletukset:

$$Cov(\boldsymbol{\epsilon}) = \mathbf{I},\ Cov(\mathbf{u}) = 0.1\mathbf{I} \text{ ja } Cov(\boldsymbol{\epsilon}, \mathbf{u}) = \mathbf{0}.$$

Malli:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

missä

$$\mathbf{y} = (110, 100, 110, 100, 100, 110, 110, 100, 100)'$$

ja

$$\boldsymbol{\beta} = (h_1, h_2, h_3)',$$

missä h_i on i. lauman kiinteä vaikutus. Satunnaisvaikutusten vektori on

$$\mathbf{u} = (S_A, S_B, S_C, S_D)',$$

missä S_j on j. isän vaikutus tyttären maidontuotantoon.

Mallissa

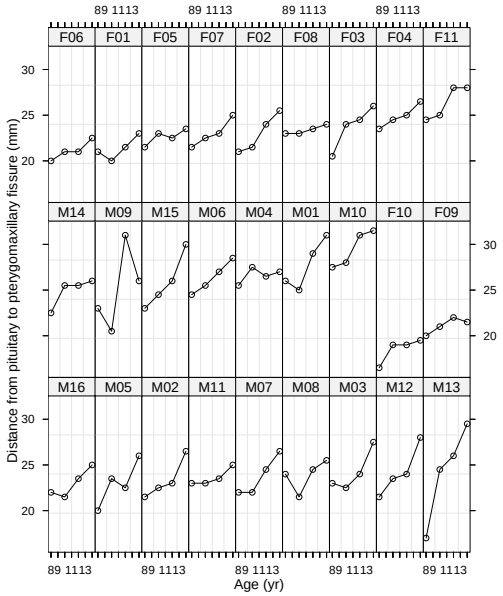
$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{ ja } \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Ratkaisuiksi saadaan

$$\tilde{\boldsymbol{\beta}} = (105.64, 104.28, 105.46)'$$

ja

$$\tilde{\mathbf{u}} = (0.4, 0.52, 0.76, -1.67)'.$$



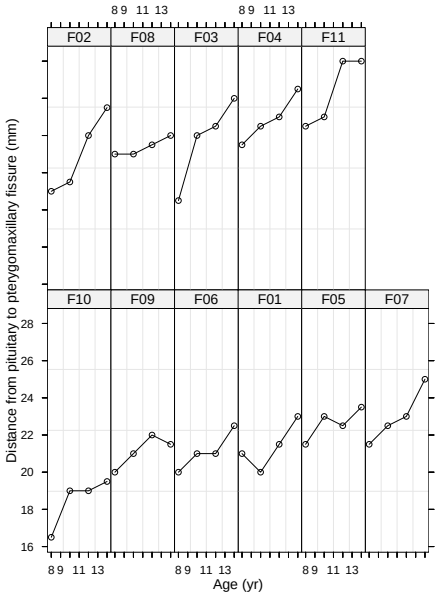
Kuva 4: Dental aineisto

Huom. Lauman 2 lehmillä on huonoin maidontuotannon estimaattiarvo 104.28. Isien A, B ja C tyttärillä on sama maidontuotannon keskiarvo 110 ja siten tässä mielessä isät A, B ja C ovat yhtä hyviä. Kuitenkin ennustetut satunnaisvaikutukset ovat eri suuria. Isän C vaikutus on suotuisin. Tämä johtunee siitä, että isältä C on kaksi tytärtä, mutta isiltä A ja B vain yksi, joten isästä C on käytössä enemmän informaatiota. Isän B tytär taas on peräisin hieman suuremmasta laumasta (3 lehmää) kuin isän A tytär (2 lehmää), joten isästä B on hieman enemmän informaatiota.

Esimerkki 10. Dental-aineiston tytöt.

Aineistossa yhdeltätoista tytöltä on mitattu tietty hampaistoon liittyvä etäisyys 8-, 10-, 12- ja 14-vuoden iässä (ks luku 7). Oletetaan, että etäisyys kasvaa lineaarisesti ajan funktiona.

```
> library(nlme)
```



Kuva 5: Dental aineiston tyttöjen mittaukset.

```
Loading required package: nls
>  data(Orthodont)
>  ?Orthodont
    ...
      Investigators at the University of North Carolina Dental School
      followed the growth of 27 children (16 males, 11 females) from age
      8 until age 14.  Every two years they measured the distance
      between the pituitary and the pterygomaxillary fissure, two points
      that are easily identified on x-ray exposures of the side of the
      head.
    ...
>  names(Orthodont)
[1] "distance" "age"      "Subject"  "Sex"
>  plot(Orthodont)
```

```
>  Orthofem<-Orthodont[Orthodont$Sex=="Female",]
>  plot(Orthofem)
>  femlin<-lmList(distance~age, data=Orthofem)
>  coef(femlin)
     (Intercept)   age
F10       13.55 0.450
F09       18.10 0.275
F06       17.00 0.375
F01       17.25 0.375
F05       19.60 0.275
F07       16.95 0.550
F02       14.20 0.800
F08       21.45 0.175
F03       14.40 0.850
F04       19.65 0.475
F11       18.95 0.675
>  femlinb<-lme(distance~age,data=Orthofem, random=~1+age | Subject)
>  summary(femlinb)
Linear mixed-effects model fit by REML
  Data: Orthofem
        AIC      BIC    logLik
   149.4287 159.8547 -68.71435

Random effects:
  Formula: ~1 + age | Subject
  Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 1.8839778 (Intr)
age         0.1609163 -0.354
```

```
Residual    0.6682885

Fixed effects: distance ~ age
                Value Std.Error DF  t-value p-value
(Intercept) 17.372727 0.7605627 32 22.84194  <.0001
age          0.479545 0.0662121 32  7.24256  <.0001
  Correlation:
     (Intr)
age -0.637

Standardized Within-Group Residuals:
         Min          Q1         Med          Q3         Max
-1.85445530 -0.46784363  0.06783482  0.42980708  1.59212526

Number of Observations: 44
Number of Groups: 11
>  random.effects(femlinb)
     (Intercept)         age
F10 -2.88783075 -0.10368479
F09 -0.10760472 -0.12657811
F06 -0.59939199 -0.08088186
F01 -0.41657310 -0.07544613
F05  0.98930860 -0.09396375
F07 -0.08887805  0.03985451
F02 -1.31834622  0.15409513
F08  2.02955253 -0.12335282
F03 -1.01578320  0.19325043
F04  1.65110405  0.04635027
F11  1.76444284  0.17035712
```

```
>  coef(femlinb)
     (Intercept)       age
F10    14.48490 0.3758607
F09    17.26512 0.3529673
F06    16.77334 0.3986636
F01    16.95615 0.4040993
F05    18.36204 0.3855817
F07    17.28385 0.5194000
F02    16.05438 0.6336406
F08    19.40228 0.3561926
F03    16.35694 0.6727959
F04    19.02383 0.5258957
F11    19.13717 0.6499026
>  femlinc<-lme(distance~age,data=Orthofem, random=~1| Subject)
>  anova(femlinc,femlinb)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
femlinc     1  4 149.2183 156.1690 -70.60916
femlinb     2  6 149.4287 159.8547 -68.71435 1 vs 2 3.789622  0.1503
```

Malli

$$y = \beta_0 + b + \beta_1 \times t + \epsilon,$$

jossa on vain yksi satunnaisvaikutus b näyttäisi siis parammalta kuin malli, jossa on kaksi satunnaisvaikutusta.

5 Kovarianssimatriisin estimointi

5.1 Suurimman uskottavuuden menetelmä (Maximum Likelihood)

Mikäli kovarianssimatriisin $\mathbf{V}$ parametrit on estimoitu, saadaan helposti laskettua estimaatit mallin kiinteälle osalle sekä tarvittaessa ennusteet satunnaisvaikutuksille. Käytössä on useita menetelmiä. Tällä kurssilla käytetään ns. uskottavuusfunktiopohjaisia menetelmiä.

Jos oletetaan normaalijakauma

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix} \right),$$

saadaan estimaatit parametreille $\mathbf{D}$ ja $\mathbf{R}$ maksimoimalla logaritmoitu uskottavuusfunktio

$$2l(\mathbf{D}, \mathbf{R}) = -\log \mid \mathbf{V} \mid -\mathbf{r}'\mathbf{V}^{-1}\mathbf{r},$$

missä $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. Maksimointi voidaan suorittaa esimerkiksi Newton-Raphson menetelmällä tai EM-algoritmilla. Olkoon $\boldsymbol{\theta}$ vektori, joka sisältää estimoitavat kovarianssimatriisien $\mathbf{D}$ ja $\mathbf{R}$ parametrit ja olkoon $\hat{\boldsymbol{\theta}}$ näiden parametrien SU-estimaattori. Eräs tärkeä SU-estimaattorin ominaisuus on, että asymptoottisesti

$$Var(\hat{\boldsymbol{\theta}}) \approx [\mathbf{I}(\theta)]^{-1},$$

missä

$$\mathbf{I}(\theta) = E[\frac{\partial l}{\partial \boldsymbol{\theta}} \frac{\partial l}{\partial \boldsymbol{\theta}'}]$$

on ns. Fisherin informaatiomatriisi. Lisäksi voidaan näyttää, että hyvin yleisten ehtojen vallitessa SUE on konsistentti ja asymptoottisesti normaalisti jakautunut odotusarvona $\boldsymbol{\theta}$ ja kovarianssimatriisina Fisherin informatiomatriisin käänteismatriisi

$$\hat{\boldsymbol{\theta}} \sim AN(\boldsymbol{\theta}, [\mathbf{I}(\theta)]^{-1}).$$

5.2 REML-menetelmä (Restricted Maximum Likelihood)

Suurimman uskottavuuden menetelmässä maksimodaan uskottavuusfunktio, kun havainnot $\mathbf{y}$ on annettu. REML- menetelmässä maksimoidaan uskottavuusfunktio, joka saadaan muunnoksesta $\mathbf{K}'\mathbf{y}$, missä $\mathbf{K}$ on annettu täysiasteinen (sarakeaste) matriisi siten, että $\mathbf{K}'\mathbf{X} = \mathbf{O}$. REML-menetelmässä maksimoidaan funktio

$$2l(\mathbf{D}, \mathbf{R})_R = 2l(\mathbf{D}, \mathbf{R}) - \log \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid .$$

Tämä funktio voidaan johtaa $\mathbf{K}'\mathbf{y}$:n tiheysfunktiosta $N(\mathbf{0}, \mathbf{K}'\mathbf{V}\mathbf{K})$.

Esimerkki 11.

Olkoon nyt $y_1, \ldots, y_n$ otos normaalijakaumasta $N(\mu, \sigma^2)$. Suurimman uskottavuuden estimaattori parametrille σ^2 on

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Tiedetään, että $\hat{\sigma}^2_{ML}$ on parametrin σ^2 alaspäin harhainen estimaattori. Harhaton estimaattori on

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Nyt siis $n-1$ ottaa huomioon yhden vapausasteen menetyksen parametrin μ estimoinnissa. Sen sijaan, jos μ olisi tunnettu, olisi estimaattori

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

harhaton. Siksi harha estimaattorissa $\hat{\sigma}^2_{ML}$ johtuu siitä, että μ pitää estimoida. Jos yleisessä sekamallissa valittaisiin $\mathbf{V} = \sigma^2\mathbf{I}$, $\mathbf{X} = \mathbf{1}$ ja $\boldsymbol{\beta} = \mu$, kiinteän osan estimaattoriksi saadaan

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{n}\mathbf{1}'\mathbf{y} = \bar{y}.$$

Olkoon $\mathbf{K}$ $n \times (n-1)$ matriisi, jonka sarakkeet $\mathbf{k}_i$, $i = 1, \ldots, n-1$, ovat lineaarisesti riippumattomia. Lisäksi oletetaan, että vektorit $\mathbf{X}'\mathbf{y}$ ja $\mathbf{K}'\mathbf{y}$ ovat

tilastollisesti riippumattomia, eli $Cov(\mathbf{X}'\mathbf{y}, \mathbf{K}'\mathbf{y}) = \mathbf{0}$ tai $\mathbf{K}'\mathbf{X} = \mathbf{0}$. Nyt vektori $\mathbf{K}'\mathbf{y}$ noudattaa normaalijakaumaa

$$N(\mathbf{0}, \sigma^2\mathbf{K}'\mathbf{K}),$$

jonka logaritmoitu uskottavuusfunktio ($\times 2$) on

$$2l_{REML}(\sigma^2) = c - (n-1)\log\sigma^2 - \sigma^{-2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}.$$

Ratkaisu saadaan, kun asetetaan uskottavuusfunktion $2l_{REML}(\sigma^2)$ derivaatta nollaksi:

$$\frac{\partial 2l_{REML}(\sigma^2)}{\partial\sigma^2} = -\frac{n-1}{\sigma^2} + \frac{\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}}{\sigma^4} = 0.$$

REML-estimaattori on nyt

$$\hat{\sigma}^2_{REML} = \frac{1}{n-1}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}.$$

Teoreema

Jos $\mathbf{K}'\mathbf{X} = \mathbf{0}$, missä matriisilla $\mathbf{K}$ on täysi sarakeaste ja $\mathbf{V}$ on positiivisesti definiitti, niin

$$\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}' = \mathbf{P},$$

missä

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}.$$

Nyt jos $\mathbf{V} = \mathbf{I}$, niin

$$\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}' = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$$

ja kun $\mathbf{X} = \mathbf{1}$, niin

$$\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{J},$$

missä $\mathbf{J} = \frac{1}{n}\mathbf{1}\mathbf{1}'$. Nyt saadaan

$$\hat{\sigma}^2_{REML} = \frac{1}{n-1}\mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Huomataan, että $\hat{\sigma}^2_{REML} = s^2$ ja lisäksi estimaattori on riippumaton matriisin $\mathbf{K}$ valinnasta. Lisäksi REML-menetelmä ei anna estimaattia mallin kiinteälle osalle.

Huomautus 1. Suurimman uskottavuuden menetelmällä saadut varianssiestimaatit ovat (alaspäin) harhaisia.

Huomautus 2. REML-tekniikalla saadaan yleensä vähemmän harhaisia estimaatteja.

*Huomautus 3. REML-menetelmää käytettäessä on huomioitava, että uskottavuusfunktio ei pysy samana, kun muutetaan mallin kiinteää osaa (**X**-matriisia). Tästä seuraa esimerkiksi se, että kun käytetään uskottavuusfunktiopohjaisia menetelmiä mallin valintaan, on mallin kiinteän osan pysyttävä samana.*

Jatkoa esimerkkiin 10.

```
> library(nlme)
> data(Orthodont)
> Orthofem<-Orthodont[Orthodont$Sex=="Female",]
> femlinc<-lme(distance~age,data=Orthofem, random=~1| Subject)
> femlinc2<-lme(distance~age+I(age^2),data=Orthofem, random=~1| Subject)
> anova(femlinc,femlinc2)
          Model df      AIC      BIC    logLik   Test  L.Ratio
femlinc       1  4 149.2183 156.1690 -70.60916
femlinc2      2  5 156.4092 164.9771 -73.20461 1 vs 2 5.190901
```

```
          p-value
femlinc
femlinc2  0.0227
Warning message:
Fitted objects with different fixed effects. REML comparisons are not
meaningful. in: anova.lme(femlinc, femlinc2)
>  femlinc<-lme(distance~age,data=Orthofem, random=~1| Subject, method="ML")
>  femlinc2<-lme(distance~age+I(age^2),data=Orthofem, random=~1| Subject,
method="ML")
> anova(femlinc,femlinc2)
          Model df      AIC      BIC    logLik   Test     L.Ratio p-value
femlinc       1  4 146.0304 153.1671 -69.01520
femlinc2      2  5 148.0208 156.9417 -69.01038 1 vs 2 0.009631381  0.9218
```

6 Kovarianssirakenteista

Sekamallin oletuksista seuraa, että havaintojen $\mathbf{y}$ kovarianssimatriisi on muotoa

$$Var(\mathbf{y}) = \mathbf{ZDZ}' + \mathbf{R} = \mathbf{V}.$$

Kovarianssimatriisit $\mathbf{D}$ ja $\mathbf{R}$ voidaan nykyisissä ohjelmistoissa (esim. R ja SAS) spesifioida varsin yleisesti.

Yksinkertaisin rakennne syntyy, kun oletetaan riippumattomuus ja sama varianssi:

$$\sigma^2\mathbf{I} = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{pmatrix}.$$

Tasakorrelaatiorakenne *(uniform, compound symmetry)* syntyy, kun

$$Cov(\epsilon_i, \epsilon_{i'}) = \begin{cases} \sigma_u^2 & , \text{kun } i \neq i' \\ \sigma_u^2 + \sigma^2 & , \text{kun } \ i = i'. \end{cases}$$

Tällöin

$$Var(\boldsymbol{\epsilon}) = \begin{pmatrix} \sigma_u^2 + \sigma^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma^2 & & \\ \vdots & & \ddots & \vdots \\ \sigma_u^2 & & \dots & \sigma_u^2 + \sigma^2 \end{pmatrix},$$

Autokorrelatiorakenne (AR(1)) syntyy, kun

$$Cov(\epsilon_i, \epsilon_{i'}) = \sigma^2 \rho^{|i-i'|}, \forall i \neq i'.$$

Tällöin kovarianssimatriisi näyttää seuraavanlaiselta

$$Var(\boldsymbol{\epsilon}) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ & 1 & \rho & \dots & \rho^{n-2} \\ & & 1 & \rho & \\ & & & \ddots & \rho \\ & \# & & & 1 \end{pmatrix}.$$

Rakenteettomassa kovarianssimatriisissa *(unstructured)* kovarianssimatriisille ei anneta mitään erityistä rakennetta.

7 Kovarianssirakenteen valinta

7.1 Informaatiokriteerit

Kovarianssirakenteiden paremmuutta voidaan tutkia esim. informaatiokriteerien AIC ja BIC avulla. Informaatiokriteeri AIC määritellään seuraavasti

$$AIC = -p(\boldsymbol{\theta}) + q,$$

missä $p(\boldsymbol{\theta})$ on logaritmoidun uskottavuusfunktion arvo (Log likelihood) ja q on estimoitujen kovarianssimatriisin parametrien lukumäärä. Rakenne, joka tuottaa pienimmän AIC-arvon on suositeltavampi. Informaatiokriteeri BIC määritellään seuraavasti

$$BIC = -p(\boldsymbol{\theta}) + \frac{1}{2}q \times logN^*,$$

missä N^* on havaintojen lukumäärä. Jälleen rakenne, joka tuottaa pienimmän arvon on suositeltavampi.

7.2 Uskottavuussuhteeseen perustuva testaus (Likelihood Ratio=LR)

Olkoon $\hat{\boldsymbol{\theta}}$ parametrivektorin $\boldsymbol{\theta}$ rajoittamaton estimaatti ja $\boldsymbol{\theta}^*$ rajoitteen vallitessa estimoitu $\boldsymbol{\theta}$ (esim. jotkin komponentit pakotettu nollaksi). LR-testi perustuu suureeseen

$$\lambda = \frac{L(\boldsymbol{\theta}^*)}{L(\hat{\boldsymbol{\theta}})},$$

missä L viittaa uskottavuusfunktioon. Jos H_0 on tosi (l. rajoitteet voimassa), niin $-2\log\lambda$ noudattaa likimain χ^2_k-jakaumaa, missä k on rajoitteiden lukumäärä. H_0 hylätään, jos

$$-2\log\lambda = 2\log L(\hat{\boldsymbol{\theta}}) - 2\log L(\boldsymbol{\theta}^*),$$

ylittää kriittisen arvon.

Esimerkki. Halutaan testata, onko kovarianssimatriisi muotoa (esim. toistomittaustilanteessa 3 toistomittausta tilastoyksikköä kohden)

$$H_0 : \mathbf{\Sigma} = \sigma^2 \mathbf{I}.$$

Koska rajoittamaton

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ & & \hat{\sigma}_3^2 \end{pmatrix},$$

niin $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_3^2, \hat{\sigma}_{12}, \hat{\sigma}_{13}, \hat{\sigma}_{23})'$, $\boldsymbol{\theta}^* = \sigma^{*2}$ ja $k = 5$.

Huomautus. Uskottavuussuhdetestiä voidaan käyttää, kun testataan alirakennetta. Siten esimerkiksi testausta AR(1)-rakenne c. tasakorrelaatirakenne ei voida uskottavuussuhdetestillä suorittaa. Lisäksi jos käytetään REML- menetelmää, niin mallin kiinteätä osaa ei voida testata uskottavuussuhdetestillä.

7.3 Waldin testi

Olkoon $\hat{\boldsymbol{\theta}}$ parametrin $\boldsymbol{\theta}$ estimaatti ja olkoon $\mathbf{I}(\boldsymbol{\theta})$ estimaattiin $\hat{\boldsymbol{\theta}}$ liittyä informaatiomatriisi. Hypoteesia $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^*$ voidaan nyt testata Waldin testillä

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'[\mathbf{I}(\boldsymbol{\theta}^*)]^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

joka tietyin edellytyksin on likimain χ^2-jakautunut vapausastein p, missä p on vektorin $\boldsymbol{\theta}$ alkioiden lukumäärä. Jos $p = 1$, niin likimain

$$\frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\hat{var}_\infty(\hat{\beta}_i)}} \sim AN(0, 1).$$

8 Satunnaisvaikutusten kovarianssirakenteen mallintaminen R-ohjelmistossa

Yleisimmät rakenteet satunnaisvaikutusten kovarianssimatriisille R-ohjelmistossa ovat:

pdBlocked (blokki-diagonaali)

pdCompsym (tasakorrelaatio)

pdDiag (diagonaali)

pdIdent (identiteetti)

pdSymm (yleinen)

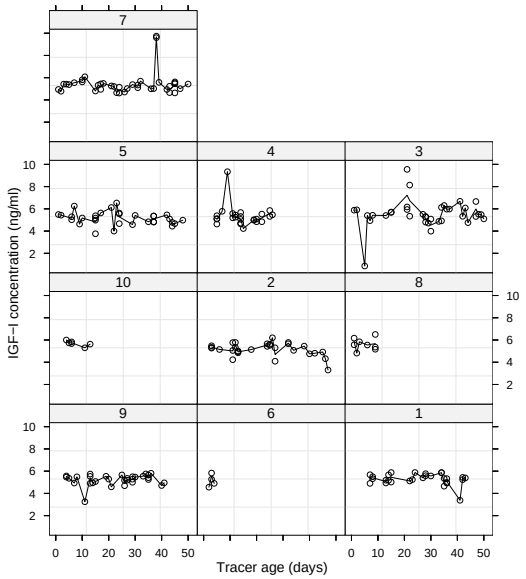
Esimerkki 12. Dental-aineiston tyttöjen muodostaman aineiston mallin satunnaisvaikutusosan kovarianssirakenteen mallintaminen.

```
>  data(Orthodont)
>  Orthofem<-Orthodont[Orthodont$Sex=="Female",]
>  la<-lme(distance~age,data=Orthofem,random=~age)
>  summary(la)
Linear mixed-effects model fit by REML
  Data: Orthofem
        AIC      BIC    logLik
   149.4287 159.8547 -68.71435
Random effects:
  Formula: ~age | Subject
  Structure: General positive-definite, Log-Cholesky parametrization
             StdDev    Corr
(Intercept) 1.8839778 (Intr)
age         0.1609163 -0.354
Residual    0.6682885
```

```
Fixed effects: distance ~ age
                Value Std.Error DF  t-value p-value
(Intercept) 17.372727 0.7605627 32 22.84194  <.0001
age          0.479545 0.0662121 32  7.24256  <.0001
  Correlation:
     (Intr)
age -0.637

Standardized Within-Group Residuals:
         Min          Q1         Med          Q3         Max
-1.85445530 -0.46784363  0.06783482  0.42980708  1.59212526

Number of Observations: 44
Number of Groups: 11
>  lb<-lme(distance~age,data=Orthofem,random=pdDiag(~age))
>  summary(lb)
Linear mixed-effects model fit by REML
  Data: Orthofem
        AIC      BIC    logLik
   147.7202 156.4085 -68.86009

Random effects:
  Formula: ~age | Subject
  Structure: Diagonal
        (Intercept)       age  Residual
StdDev:    1.572479 0.1328720 0.6934114
```

```
Fixed effects: distance ~ age
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 0.7072259 32 24.564609  <.0001
age          0.479545 0.0615674 32  7.788956  <.0001
  Correlation:
     (Intr)
age -0.552

Standardized Within-Group Residuals:
         Min          Q1         Med          Q3         Max
-2.04079146 -0.50460965  0.08842214  0.52868713  1.54130271

Number of Observations: 44
Number of Groups: 11
>  lc<-lme(distance~age,data=Orthofem,random=pdIdent(~age))
>  summary(lc)
Linear mixed-effects model fit by REML
  Data: Orthofem
        AIC      BIC   logLik
   149.7944 156.7451 -70.8972

Random effects:
  Formula: ~age | Subject
  Structure: Multiple of an Identity
         (Intercept)       age Residual
StdDev:   0.1838304 0.1838304 0.785641
```

```
Fixed effects: distance ~ age
                Value Std.Error DF   t-value p-value
(Intercept) 17.372727 0.5971412 32 29.093163  <.0001
age          0.479545 0.0766665 32  6.254958  <.0001
  Correlation:
    (Intr)
age -0.674

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.3980960 -0.5886844  0.1061709  0.5774736  1.7581531

Number of Observations: 44
Number of Groups: 11
>  anova(lc,lb,la)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
lc      1  4 149.7944 156.7451 -70.89720
lb      2  5 147.7202 156.4085 -68.86009 1 vs 2 4.074233  0.0435
la      3  6 149.4287 159.8547 -68.71435 2 vs 3 0.291473  0.5893
>  intervals(lb)
    ...
  Random Effects:
   Level: Subject
                      lower      est.     upper
sd((Intercept)) 0.78411447 1.5724791 3.1534816
sd(age)         0.06333797 0.1328720 0.2787423
  Within-group standard error:
    lower      est.     upper
0.5263491 0.6934114 0.9134990
```



Kuva 6: IGF-aineisto

Esimerkki 13. Satunnaisvaikutusten kovarianssirakenteen mallintaminen.

```
> library(nlme)
Loading required package: nls
> data(IGF)
> ?IGF
    ...
Format:
      This data frame contains the following columns:
      Lot an ordered factor giving the radioactive tracer lot.
      age a numeric vector giving the age (in days) of the
      radioactive tracer.

      conc a numeric vector giving the estimated concentration of IGF-I
```

```
      protein (ng/ml)
Details:
      Davidian and Giltinan (1995) describe data obtained
      during quality control radioimmunoassays for ten different
      lots of radioactive tracer used to calibrate the
      Insulin-like Growth Factor (IGF-I) protein concentration
      measurements.
    ...
>  plot(IGF)
>  IGF
Grouped Data: conc ~ age | Lot
     Lot age conc
1     1   7 4.90
2     1   7 5.68
3     1   8 5.32
4     1   8 5.50
5     1  13 4.94
...
234  10   6 5.68
235  10   6 5.83
236  10  11 5.30
237  10  13 5.63
>  lm1<-lm(cons ~ age, data=IGF)
>  summary(lm1)
    ...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.3510594  0.1037341   51.58   <2e-16 ***
age         -0.0006692  0.0039425   -0.17    0.865
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8327 on 235 degrees of freedom
Multiple R-Squared: 0.0001226,  Adjusted R-squared: -0.004132
F-statistic: 0.02882 on 1 and 235 DF,  p-value: 0.8654
```

Tassä mallissa kulmakerroin ei ole merkitsevä. Kokeillaan seuraavaksi satunnaiskertoimista regressiomallia

$$y = (\beta_0 + u_0) + (\beta_1 + u_1) \times x + \epsilon,$$

missä kertoimet β_0 ja β_1 ovat kiinteitä vaikutuksia ja kertoimet $\mathbf{u} = (u_0, u_1)'$ ovat satunnaisia vaikutuksia. Satunnaisvaikutusten kovarianssimatriisi $Var(\mathbf{u}) = \mathbf{D}$ on tässä ajossa rakenteeton.

```
>  sm1<-lme(IGF)
>  summary(sm1)
    ...
  Random effects:
  Formula: ~age | Lot
  Structure: General positive-definite
              StdDev      Corr
(Intercept) 0.081807079 (Intr)
age         0.008061817 -0.998
Residual    0.820634439
    ...
>  intervals(sm1)
Approximate 95% confidence intervals
  Fixed effects:
                 lower         est.       upper
(Intercept)  5.1632202  5.374910593 5.586601004
```

```
age         -0.0124569 -0.002528088 0.007400723

  Random Effects:

   Level: Lot

                              lower         est.      upper

sd((Intercept))       0.003887843  0.081807079 1.72136534

sd(age)               0.002211865  0.008061817 0.02938375

cor((Intercept),age) -1.000000000 -0.997677534 1.00000000


  Within-group standard error:

    lower      est.     upper

0.7489972 0.8206344 0.8991234
```

Näyttäisi siltä, että kovarianssimatriisissa muuttujien välinen kovarianssi (korrelaatio) ei olisi merkitsevä. Kokeillaan seuraavaksi erilaisia rakenteita satunnaisvaikutusten kovarianssimatriisille $\mathbf{D}$.

```
>  sm2<-update(sm1, random=pdDiag(~age))

>  sm3<-update(sm1, random=pdIdent(~age))

>  anova(sm3,sm2,sm1)

     Model df      AIC      BIC    logLik   Test   L.Ratio p-value

sm3      1  4 602.8038 616.6422 -297.4019

sm2      2  5 604.8017 622.0996 -297.4008 1 vs 2 0.0021268  0.9632

sm1      3  6 606.3700 627.1276 -297.1850 2 vs 3 0.4316558  0.5112
```

Testatuista rakenteista paras olisi siis $\mathbf{D} = \sigma^2\mathbf{I}$.

```
>  summary(sm3)

Linear mixed-effects model fit by REML

  Data: IGF

        AIC      BIC    logLik
```

```
   602.8038 616.6422 -297.4019

Random effects:
  Formula: ~age | Lot
  Structure: Multiple of an Identity
         (Intercept)         age Residual
StdDev: 0.005365771 0.005365771 0.821811
    ...
>  intervals(sm3)
Approximate 95% confidence intervals

  Fixed effects:
                   lower         est.       upper
(Intercept)  5.16447617  5.369021045 5.573565921
age         -0.01093177 -0.001928529 0.007074709

  Random Effects:
   Level: Lot
               lower        est.      upper
sd(age) 0.001840994 0.005365771 0.01563911

  Within-group standard error:
     lower      est.     upper
0.7500596 0.8218111 0.9004263
```

Mallissa muuttujan *age* kerroin ei selvästikään ole merkitsevä. Seuraavassa ajossa *age* pudotetaankin pois.

```
>  sm4<-update(sm3,conc~1)
>  summary(sm4)
```

```
Linear mixed-effects model fit by REML
  Data: IGF
       AIC     BIC    logLik
   592.0155 602.407 -293.0078

Random effects:
  Formula: ~age | Lot
  Structure: Multiple of an Identity
        (Intercept)         age  Residual
StdDev: 0.004895685 0.004895685 0.8212692

Fixed effects: conc ~ 1
               Value  Std.Error  DF  t-value p-value
(Intercept) 5.334441 0.06489369 227 82.20276  <.0001
    ...
>  intervals(sm4)
Approximate 95% confidence intervals
  Fixed effects:
              lower     est.    upper
(Intercept) 5.20657 5.334441 5.462312
  Random Effects:
   Level: Lot
              lower        est.      upper
sd(age) 0.001581789 0.004895685 0.01515229

  Within-group standard error:
    lower      est.     upper
0.7496691 0.8212692 0.8997077
```

Testataan mallin kiinteätä osaa. Vaihdetaan testattavissa malleissa estimointimenetelmäksi suurimman uskottavuuden menetelmä.

```
>  sm3b<-update(sm3, method="ML")
>  sm4b<-update(sm4, method="ML")
>  anova(sm4b,sm3b)
      Model df      AIC      BIC    logLik   Test   L.Ratio p-value
sm4b      1  3 588.3617 598.7659 -291.1808
sm3b      2  4 590.2020 604.0743 -291.1010 1 vs 2 0.1596393  0.6895
```

Tässä malliksi saatiin

$$y = \mu + u_0 + u_1 \times x + \epsilon,$$

missä $Var(\mathbf{u}) = \sigma^2\mathbf{I}$. Huomatettakoon, että kyseessä on nyt sekamalli, jossa satunnaisosan dimensio on suurempi kuin kiinteän osan dimensio.

9 Mallin oletusten tutkiminen graafisesti

Edellisessä luvussa käsiteltiin satunnaisosan kovarianssirakenteen mallintamista, kun satunnaisvirheillä $\boldsymbol{\epsilon}$ on kovarianssirakenne $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. Tutkitaan nyt sekamallin oletuksia tässä tilanteessa tarkemmin. Mallilla on nyt kaksi keskeistä oletusta:

1. Satunnaisvirheet ovat riippumattomia, normaalisti jakautuneita odotusarvona nolla ja niillä on sama varianssi, eli

   $$\boldsymbol{\epsilon} \sim IN(\mathbf{0}, \sigma^2\mathbf{I}).$$

2. Satunnaisvaikutukset ovat normaalisti jakautuneita odotusarvona nolla ja kovarianssimatriisina $\mathbf{D}$. Satunnaisvaikutukset ovat keskenään

riippumattomia, eli

$$\mathbf{u} \sim IN(\mathbf{0}, \mathbf{D}).$$

Oletuksia voidaan tutkia R-komennolla *plot*, jonka kutsu on muotoa:

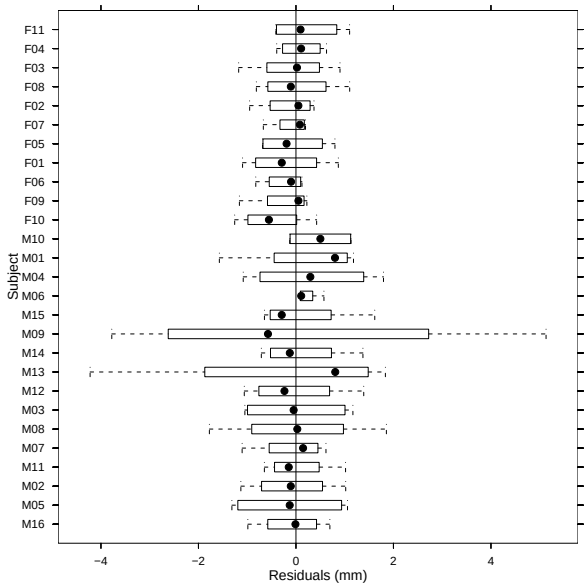
plot(tulos, muoto),

missä tulos on *lme*-funktion tulos ja muoto on lauseke, jossa määritellään muuttujat.

9.1 Satunnaisvirheiden jakauman tutkiminen

Esimerkki 14. Residuaalien jakauman tarkastelu yksilöittäin.
Esimerkissä tarkastellaan satunnaisvirheiden residuaalien jakaumaa. Residuaalit saadaan helposti myös yksilöittäin.

```
>  library(nlme)
Loading required package: nls
>  library(lattice)
>  trellis.device(device = getOption("device"),color = FALSE,
   bg ="white")
>  data(Orthodont)
>  formula(Orthodont)
distance ~ age | Subject
>  f1<-lme(Orthodont)
```



Kuva 7: Residuaalikuvio Dental-aineistosta

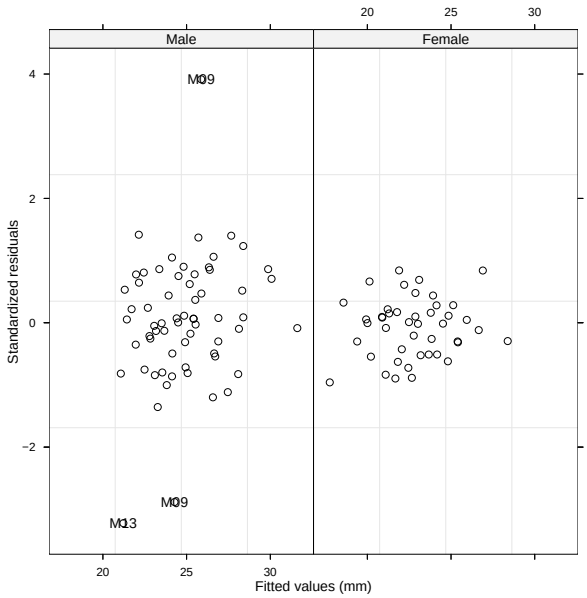
```
> plot(f1, Subject~resid(.), abline=0)
```

Argumentti *abline=0* lisää kuvioon pystysuoran viivan. Kuviosta nähdään, että residuaalien varianssi on suurempi pojilla. Lisäksi nähdään, että pojalla numero 13 (M13) on "vieras"havainto ja että havainnon M09 residuaalien vaihtelu on poikkeuksellisen suurta.

Esimerkki 15. Residuaalien ja sovitettujen arvojen yhteisjakauma.
Seuraavassa kuviossa on standartisoidut residuaalit ja sovitetut arvot sukupuolille erikseen:

```
> plot(f1, resid(., type="p")~fitted(.) | Sex, id=0.05, adj=-0.3)
```

Argumentilla $type = "p"$ saadaan standartisoidut residuaalit. Argumentti *id* liittyy aineistolle vieraiden havaintojen tunnistamiseen (ne standartisoidut resi-

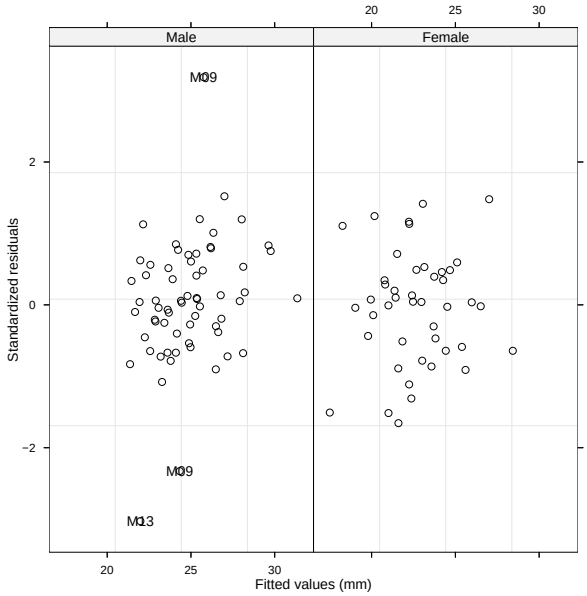


Kuva 8: Residuaalit ja sovitetut arvot erikseen tytöille ja pojille

duaalit joiden itseisarvo on suurempi kuin normaalijakauman kvantiili $1 - id/2$ merkitään kuvioon). Argumenttij *adj* liittyy vieraiden havaintojen tunnisteiden sijoitteluun.

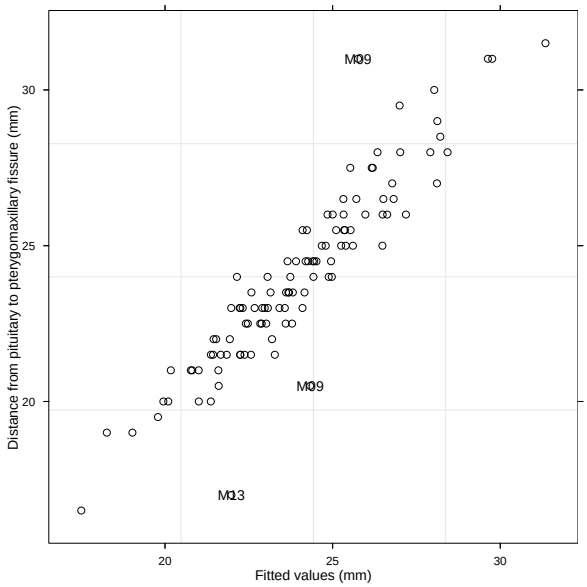
Saatu kuvio tukee selvästi käsitystä siitä, että residuaalien vaihtelu olisi suurempaa pojilla. Seuraavaksi sovitetaankin malli, jossa sukupuolille on annettu eri virhevarianssit.

```
>  f2<-update(f1, weights=varIdent(form=~1|Sex))
>  summary(f2)
    ...
             StdDev    Corr
(Intercept) 1.8992591 (Intr)
age         0.1980515 -0.446
Residual    1.6452956
```



Kuva 9: Residuaalit ja sovitetut arvot erikseen tytöille ja pojille

```
Variance function:
  Structure: Different standard deviations per stratum
  Formula: ~1 | Sex
  Parameter estimates:
     Male    Female
1.0000000 0.4040915
    ...
>  plot(f2, resid(., type="p")~fitted(.) | Sex, id=0.05, adj=-0.3)
>  anova(f1,f2)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
f1      1  6 454.6367 470.6173 -221.3183
f2      2  7 435.6466 454.2907 -210.8233 1 vs 2 20.99004  <.0001
```

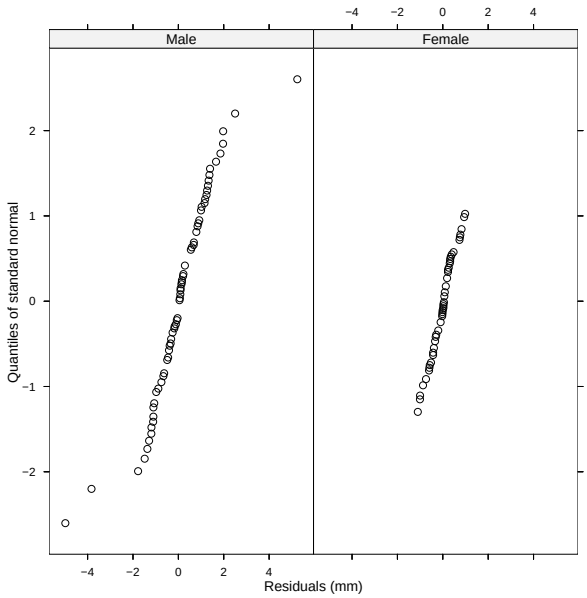


Kuva 10: Havaitut ja sovitetut arvot

Residuaalikuviot 3 ja 4 eivät selvästi poikkea toisistaan kummankaan mallin eduksi. Kuitenkin malli, jossa varianssit on mallinnettu erikseen tytöille ja pojille, osoittautuu paremmaksi testauksessa.

Esimerkki 16. Havaittujen ja sovitettujen arvojen yhteisjakauma.
Seuraavassa kuviossa tutkitaan havaittujen ja sovitettujen arvojen yhteisjakaumaa. Huomataan, että muutamaa poikkeusta lukuunottamatta valittu malli sopii varsin hyvin havaintoihin.

```
>  plot(f2, distance~fitted(.), id=0.05, adj=-0.3)
```



Kuva 11: Residuaalit ja sovitetut arvot erikseen tytöille ja pojille

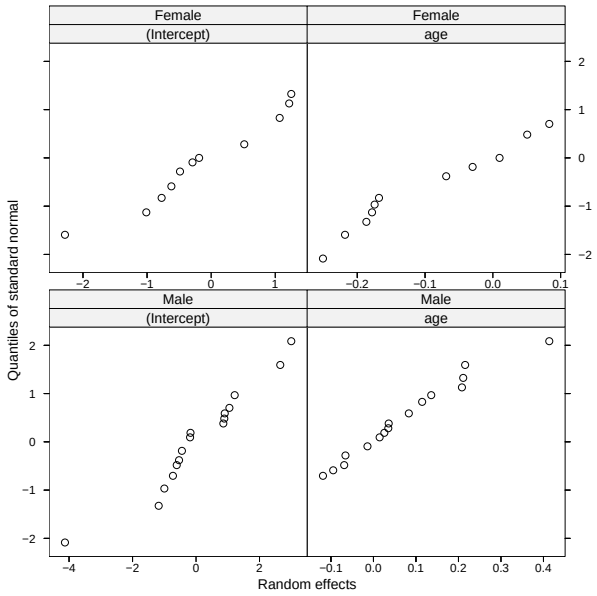
Esimerkki 17. Residuaalien normaalisuuden tutkiminen

Seuraavassa kuviossa tutkitaan jäännösten normaalisuutta. Kuvio tulkitaan siten, että tarkasteltava muuttuja on normaalisti jakautunut, jos pisteet ovat samalla suoralla. Kuviosta nähdään, että poikien aineistossa on muutama poikkeava arvo.

```
> qqnorm(f2, ~resid(.) | Sex)
```

9.2 Satunnaisvaikutusten jakauman tutkiminen

Seuraavassa tarkastellaan satunnaisvaikutuksien oletusten tutkimista. Keskeinen työkalu tarkasteluissa on funktio *ranef*, jolla saadaan estimoitua satunnaisvaikutusten *BLUP*-arvot *lme*- funktion tuloksesta.



Kuva 12: Satunnaisvaikutusten jakauman tutkiminen erikseen tytöille ja pojille

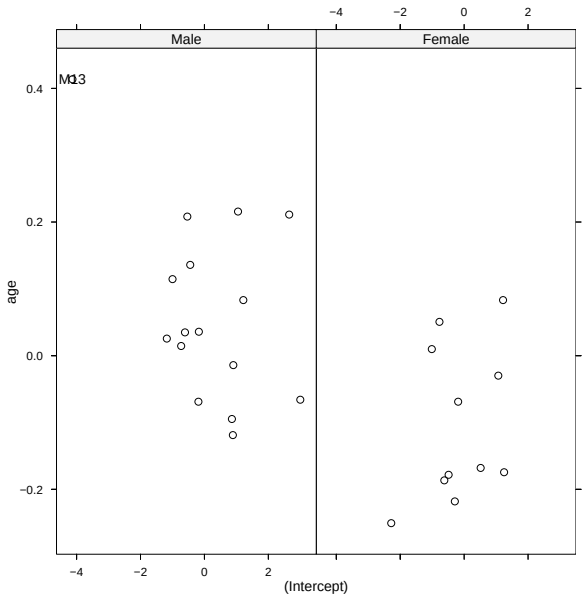
Tässä tarkastelussa on käytetty kahden tyyppisiä kuvioita:

1. *qqnorm* (normaalisuuden testaus)

2. *pairs* (satunnaisvaikutusten yhteisjakauma).

Esimerkki 18. Satunnaisvaikutusten normaalisuuden tutkiminen.
Esimerkissä tutkitaan satunnaisvaikutusten yhteisjakaumia erikseen tytöille ja pojille. Kuviot eivät kovin hyvin tue oletusta normaalijakaumasta.

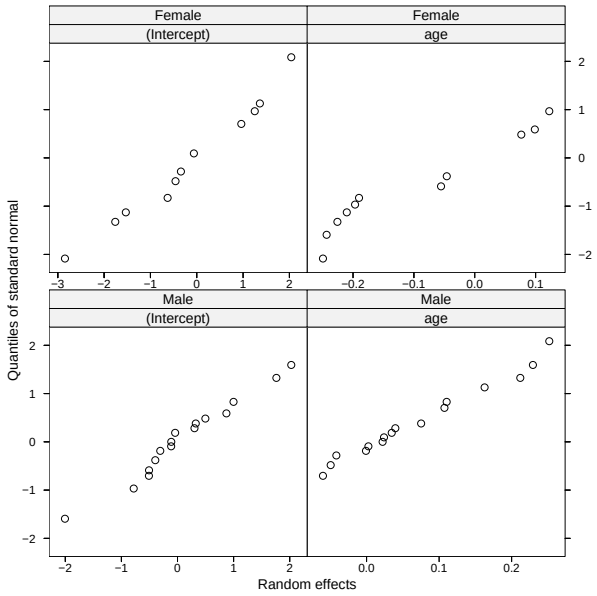
```
> qqnorm(f1, ~ranef(.)|Sex)
```



Kuva 13: Satunnaisvaikutusten yhteisjakauman tutkiminen erikseen tytöille ja pojille

Esimerkki 19. Satunnaisvaikutusten yhteisjauma.
Kuviolla voidaan tutkia satunnaisvaikutusten yhteisjakaumia tytöille ja pojille erikseen. Myöskaan tämä kuvio ei tue oletusta normaalijakaumasta.

```
>  pairs(f1, ~ranef(.)|Sex, id=~Subject=="M13", adj=-0.3)
```



Kuva 14: Normaalisuuden testaaminen erikseen tytöille ja pojille

Esimerkki 20. Normaalisuuden tutkiminen tytöille ja pojille.

Normaalisuuden testaaminen mallissa, jossa varianssit on mallinnettu erikseen tytöille ja pojille.

```
> qqnorm(f2, ~ranef(.)|Sex)
```

10 Satunnaisvirheiden kovarianssimatriisin mallintaminen R-ohjelmistossa

Kirjasto *nlme* tarjoaa monenlaisia rakenteita satunnaisvirheiden $\boldsymbol{\epsilon}$ kovariansimatriisin $Var(\boldsymbol{\epsilon}) = \mathbf{R}$ mallintamiseen. Esimerkiksi seuraavia rakenteita voidaan käyttää:

- *corCompSymm* (tasakorrelaatio)
- *corSymm* (yleinen)
- *corAR1* (ar(1)-rakenne)
- *corARMA* (ARMA-prosessista syntyvä rakenne)
- *corLin* (lineaarinen rakenne)

Funktiolle annetaan kaksi argumenttia: arvo ja muoto. Esimerkiksi

$$form = \tilde{} \mid Subject$$

Argumenttia *fixed* voidaan käyttää, jos halutaan kiinteä kovarianssimatriisi. Tällöin määritellään *FIXED=TRUE*.

Esimerkki 21. Sovitetaan Dental-aineiston satunnaisvirheille AR(1)- rakenne.

```
> library(nlme)
> data(Orthodont)
> m1<-lme(Orthodont, corr=corAR1(0.6))
> summary(m1)
    ...
Random effects:
```

```
  Formula: ~age | Subject
  Structure: General positive-definite
             StdDev    Corr
(Intercept) 3.3656038 (Intr)
age         0.3279253 -0.803
Residual    1.0900404

Correlation Structure: AR(1)
  Formula: ~1 | Subject
  Parameter estimate(s):
       Phi
-0.4899409
    ...
>  intervals(m1)
Approximate 95% confidence intervals
  Fixed effects:
                 lower       est.      upper
(Intercept) 15.1444247 16.6645551 18.1846855
age          0.5220068  0.6668926  0.8117785

  Random Effects:
   Level: Subject
                           lower       est.      upper
sd((Intercept))       2.2254703  3.3656038  5.0898407
sd(age)               0.2198051  0.3279253  0.4892290
cor((Intercept),age) -0.9233223 -0.8030639 -0.5393682

  Correlation structure:
         lower       est.       upper
```

```
Phi -0.7616757 -0.4899409 -0.07164797
    ...
```

Esimerkki 22. Sovitetaan Dental-aineistossa kovarianssimatriisille strukturoimaton rakenne.

```
m2<-update(m1, corr=corSymm())
>  summary(m2)
    ...
Random effects:
  Formula: ~age | Subject
  Structure: General positive-definite
             StdDev    Corr
(Intercept) 2.8828424 (Intr)
age         0.2742304 -0.75
Residual    1.3450149

Correlation Structure: General
  Formula: ~1 | Subject
  Parameter estimate(s):

  Correlation:
   1      2      3
2 -0.088
3  0.275 -0.167
4  0.147  0.442  0.182
    ...
```

Jos halutaan malli, jossa ei määritellä lainkaan satunnaisosaa, niin käytetään funktiota *gls (generalized least squares)*, jonka syntaksi on sama kuin *lme*-funktion

ilman satunnaisosaa.

Tarvittaessa kovarianssimatriiseille voidaan antaa alkuarvot. Jos esimerkiksi käytetään strukturoimatonta rakennetta, niin muodostettavan matriisin "ala-kolmiomatriisi"annetaan vektoreina sarakkeittain.

```
>  s1<-corSymm(value=c(0.2,0.1,-0.1,0,0.2,0), form = ~1 | Subject)
>  s1<-Initialize(s1, data=Orthodont)
>  s1
Correlation structure of class corSymm representing
  Correlation:
   1    2    3
2  0.2
3  0.1  0.0
4 -0.1  0.2  0.0
>  m2<-update(m1,corr=s1)
```

Jos käytetään tasakorrelaatiorakennetta, niin

```
>  cs<-corCompSymm(value=0.3, form=~1|Subject)
>  cs<-Initialize(cs, Orthodont)
```

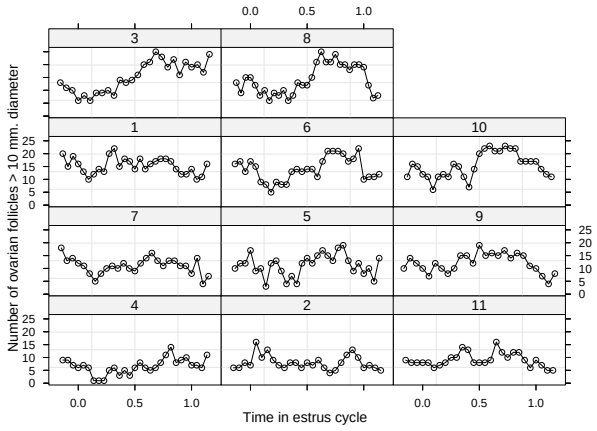
Kun tarkastellaan valitun korrelaatiomallin hyvyyttä, voidaan käytää standartoituja residuaaleja

$$\mathbf{r}_i = \hat{\mathbf{R}}^{-1/2}(\mathbf{y}_i - \hat{\mathbf{y}}_i),$$

missä $\hat{\mathbf{R}}$ on satunnaisvirheiden kovarianssimatriisin $\mathbf{R}$ estimaatti ja

$$\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{u}}_i.$$

Jos valittu kovarianssirakenne on kunnossa, niin residuaalien tulisi noudattaa likimain jakaumaa $IN(\mathbf{0}, \mathbf{I})$.



Kuva 15: Munarakkula-aineisto

Esimerkki 23. Aikasarja-aineiston mallintaminen. Aineistoon *Ovary* sovitetaan malli

$$y = (\beta_0 + b_0) + (\beta_1 + b_1)sin(2\pi t) + \beta_2 cos(2\pi t) + \epsilon,$$

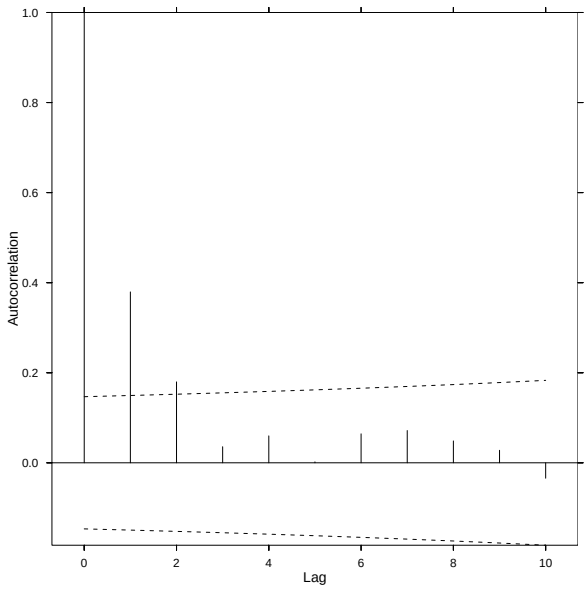
missä

$$(b_o, b_1)' \sim N(\mathbf{0}, diag(\sigma_0^2, \sigma_1^2))$$

ja $\epsilon \sim N(0, \sigma^2)$ sekä $\mathbf{b}$ ja ϵ ovat riippumattomia.

```
> data(Ovary)
> ?Ovary
    ...
Description:

     The 'Ovary' data frame has 308 rows and 3 columns.
```

```
Format:

      This data frame contains the following columns:

      Mare an ordered factor indicating the mare on which the
           measurement is made.

      Time time in the estrus cycle.  The data were recorded daily from
           3 days before ovulation until 3 days after the next
           ovulation. The measurement times for each mare are scaled so
           that the ovulations for each mare occur at times 0 and 1.

      follicles the number of ovarian follicles greater than 10 mm in
           diameter.
    ...
>  formula(Ovary)
follicles ~ Time | Mare
>  m1<-lme(follicles~sin(2*pi*Time)+cos(2*pi*Time),
+ data=Ovary,
+ random=pdDiag(~sin(2*pi*Time)))
>  m1
    ...
   Fixed: follicles ~ sin(2 * pi * Time) + cos(2 * pi * Time)
        (Intercept) sin(2 * pi * Time) cos(2 * pi * Time)
         12.1820241         -3.2985373         -0.8623725


Random effects:
  Formula: ~sin(2 * pi * Time) | Mare
```



Kuva 16: Empiirinen autokorrelaatio

```
  Structure: Diagonal
         (Intercept) sin(2 * pi * Time) Residual
StdDev:    3.052136           2.079312 3.112854
    ...

>  ACF(m1)
    lag          ACF
1    0  1.000000000
2    1  0.379480128
3    2  0.179722024
4    3  0.035692748
    ...
>  plot(ACF(m1, maxLag=10), alpha=0.01)
>  m2<-update(m1, corr=corAR1())
```

```
>  m2
Linear mixed-effects model fit by REML
   Data: Ovary
   Log-restricted-likelihood: -775.2224
   Fixed: follicles ~ sin(2 * pi * Time) + cos(2 * pi * Time)
        (Intercept) sin(2 * pi * Time) cos(2 * pi * Time)
         12.1895808         -2.9473432         -0.8807113

Random effects:
  Formula: ~sin(2 * pi * Time) | Mare
  Structure: Diagonal
         (Intercept) sin(2 * pi * Time) Residual
StdDev:    2.807293         0.03630784 3.665217

Correlation Structure: AR(1)
  Formula: ~1 | Mare
  Parameter estimate(s):
      Phi
0.6073908
Number of Observations: 308
Number of Groups: 11
>  anova(m1,m2)
    Model df      AIC      BIC    logLik   Test L.Ratio p-value
m1      1  6 1638.082 1660.404 -813.0409
m2      2  7 1564.445 1590.487 -775.2224 1 vs 2  75.637  <.0001
>  m3<-update(m1, corr=corARMA(q=2))
```

```
>  anova(m2,m3, test=F)
    Model df      AIC      BIC    logLik
m2      1  7 1564.445 1590.487 -775.2224
m3      2  8 1571.231 1600.994 -777.6156
>  ACF(m2)
    lag         ACF
1    0  1.00000000
2    1  0.52947658
3    2  0.36460864
4    3  0.23136531
5    4  0.21071815
6    5  0.13441680
7    6  0.13663687
8    7  0.08269753
    ...
>  m4<-update(m2, corr=corARMA(p=1,q=1))
Error in "coef<-.corARMA"(*tmp*, value = c(77.5763469424646,
128.313828938872,  :   Coefficient matrix not invertible
```

Saatu AR(1)-malli ei vielä ole tyydyttävä (ks. estimoitu autokorrelaatio). ARMA(1,1)-malli olisi voinut sopia aineistoon paremmin, mutta kyseisen mallin estimointi ei onnistunut.

11 Varianssin heteroskedastisuus

11.1 Taustaa

Sekamallin oletuksista seuraa kovarianssirakenne

$$Var(\mathbf{y}) = \mathbf{ZDZ}' + \mathbf{R},$$

missä $\mathbf{D}$ on satunnaisvaikutusten- ja $\mathbf{R}$ satunnaisvirheiden kovarianssimatriisi. Kovarianssimatriisi mallinnetaan antamalla matriiseille $\mathbf{D}$ ja $\mathbf{R}$ erilaisia rakenteita. Esimerkiksi $\mathbf{ZDZ}' = \sigma_u^2\mathbf{11}'$, jolloin havaintojen kovarianssirakennetta voitaisiin yrittää mallintaa antamalla jokin monimutkaisempi rakenne matriisille $\mathbf{R}$. Jos $\mathbf{R} = \sigma^2\mathbf{I}$, niin havaintojen kovarianssirakennetta voitaisiin yrittää mallintaa satunnaisosalla $\mathbf{ZDZ}'$. Käytännössä joudutaan usein toimimaan näiden tilanteiden välimaastossa.

Tarkastellaan rakennetta

$$\mathbf{R} = \sigma^2\mathbf{\Lambda}.$$

Jos $\mathbf{\Lambda}$ on positiivisesti definiitti, niin se voidaan aina esittää muodossa

$$\mathbf{\Lambda} = (\mathbf{\Lambda}^{1/2})'\mathbf{\Lambda}^{1/2}$$

ja käänteismatriisille saadaan

$$\mathbf{\Lambda}^{-1} = (\mathbf{\Lambda}^{-1/2})(\mathbf{\Lambda}^{-1/2})'.$$

Tehdään sekamallille

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}$$

muunnos $(\mathbf{\Lambda}^{-1/2})'\mathbf{y}$, niin saadaan

$$(\mathbf{\Lambda}^{-1/2})'\mathbf{y} = (\mathbf{\Lambda}^{-1/2})'\mathbf{X}\boldsymbol{\beta} + (\mathbf{\Lambda}^{-1/2})'\mathbf{Zu} + (\mathbf{\Lambda}^{-1/2})'\boldsymbol{\epsilon}.$$

Jos merkitään

$$\mathbf{y}^* = (\mathbf{\Lambda}^{-1/2})'\mathbf{y},\ \mathbf{X}^* = (\mathbf{\Lambda}^{-1/2})'\mathbf{X},\ \mathbf{Z}^* = (\mathbf{\Lambda}^{-1/2})'\mathbf{Z}\ \text{ja}\ \boldsymbol{\epsilon}^* = (\mathbf{\Lambda}^{-1/2})'\boldsymbol{\epsilon},$$

niin malliksi saadaan

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u} + \boldsymbol{\epsilon}^*.$$

Nyt esimerkiksi

$$Var(\boldsymbol{\epsilon}^*) = Var((\boldsymbol{\Lambda}^{-1/2})'\boldsymbol{\epsilon}) = (\boldsymbol{\Lambda}^{-1/2})'Var(\boldsymbol{\epsilon})\boldsymbol{\Lambda}^{-1/2}$$

ja koska $Var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{\Lambda}$, saadaan

$$Var(\boldsymbol{\epsilon}^*) = \sigma^2(\boldsymbol{\Lambda}^{-1/2})'\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} = \sigma^2(\boldsymbol{\Lambda}^{-1/2})'(\boldsymbol{\Lambda}^{1/2})'\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{-1/2} = \sigma^2\mathbf{I}.$$

Muunnetun mallin residuaalit ovat riippumattomia ja noudattavat normaalijakaumaa

$$\boldsymbol{\epsilon}^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

Jos nyt tarkastellaan mallia, jossa ei olisi satunnaisosaa, niin muunnetun mallin parametrit voitaisiin estimoida tavallisella PNS-menetelmällä. Tällöin

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= ((\mathbf{X}^*)'\mathbf{X}^*)^{-1}(\mathbf{X}^*)'\mathbf{y} \\
&= (\mathbf{X}'\boldsymbol{\Lambda}^{-1/2}(\boldsymbol{\Lambda}^{-1/2})'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Lambda}^{-1/2}(\boldsymbol{\Lambda}^{-1/2})'\mathbf{y} \\
&= (\mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Lambda}^{-1}\mathbf{y}.
\end{aligned}$$

"Palautettu"estimaattori on siten sama kuin alkuperäisen mallin *GLS*-estimaattori.

11.2 Yleinen varianssifunktio

Eräs tapa, jolla varianssin heteroskedastisuutta voidaan yrittää mallintaa on muodostaa varianssille malli, jonka arvot riippuvat esimerkiksi vasteen tasosta tai joistakin muista tekijöistä. Jos esimerkiksi varianssi on suhteessa vasteen tason neliöön, niin

$$Var(y) = \sigma^2\{f(\mathbf{x}, \boldsymbol{\beta})\}^2.$$

Sovelluksesta riippuen käytetään tietenkin erilaisia varianssifunktioita ja erilaisia taustamuuttujia varianssin mallintamiseen.

Yleinen varianssifunktio määritellään odotusarvon μ_i, kovariaattien $\boldsymbol{\nu}_i$ ja parametrien $\boldsymbol{\delta}$ funktiona. Varianssifunktion yleinen muoto on

$$Var(y_i) = \sigma^2 g^2(\mu_i, \boldsymbol{\nu}_i, \boldsymbol{\delta}),$$

missä $\mu_i = f(\mathbf{x}_i, \boldsymbol{\beta})$. Huomaa, että varianssifunktio riippuu parametreistä $\boldsymbol{\beta}$ funktion f kautta. Usein esimerkiksi biologisissa kokeissa varianssi on suhteessa tasoon, jolloin esimerkiksi

$$g(\mu_i, \boldsymbol{\nu}_i, \boldsymbol{\delta}) = \mu_i^{\delta},$$

missä $\delta > 0$ tai

$$g(\mu_i, \boldsymbol{\nu}_i, \boldsymbol{\delta}) = e^{\mu_i \delta}.$$

11.3 Varianssifunktio sekamallissa

Tarkastellaan satunnaisvirheiden kovarianssimatriisia $\mathbf{R}$, joka on muotoa $\mathbf{R} = \sigma^2 \boldsymbol{\Lambda}$. Matriisi $\boldsymbol{\Lambda}$ voidaan aina esittää muodossa

$$\boldsymbol{\Lambda} = \mathbf{TCT}, \qquad (HAJ1)$$

missä $\mathbf{C}$ on korrelaatiomatriisi ja $\mathbf{T}$ on sellainen diagonaalimatriisi, jonka alkiot muodostuvat muuttujien hajonnoista. Kertomalla $(HAJ1)$ puolittain käänteismatriisilla $\boldsymbol{T}^{-1}$ saadaan tavallinen korrelaatiomatriisi

$$\mathbf{C} = \mathbf{T}^{-1} \boldsymbol{\Lambda} \mathbf{T}^{-1}.$$

Helposti nähdään, että

$$Var(\epsilon_i) = \sigma^2 t_{ii}^2 \quad \text{ja} \quad cor(\epsilon_i, \epsilon_j) = c_{ij}.$$

Yleisillä varianssifunktioilla mallinnetaan nyt alkioita t_{ii}^2 ja edellisen luvun rakenteilla mallinnetaan korrelaatiomatriisia $\mathbf{C}$.

Varianssin mallintamiseen käytetään yleistä varianssifunktiota muodossa

$$Var(\epsilon_i \mid \mathbf{u}) = \sigma^2 g^2(\mu_i, \boldsymbol{\nu}_i, \boldsymbol{\delta}),$$

missä $\mu_i = E(y_i \mid \mathbf{u})$, $\boldsymbol{\nu}_i$ on kovariaattien vektori ja $\boldsymbol{\delta}$ sisältää varianssiparametrit ja $g(.)$ on varianssifunktio, joka oletetaan jatkuvaksi parametriensa suhteen. Jos esimerkiksi varianssin uskotaan kasvavan kovariaatin ν_i itseisarvon potenssina, niin

$$Var(\epsilon_i \mid \mathbf{u}) = \sigma^2 \mid \nu_i \mid^{2\delta} .$$

Varianssifunktio on nyt muotoa $g(x, y) =\mid x \mid^y$ ja kovariaatti ν_i voi olla odotusarvo μ_i.

Varianssifuntio $Var(\epsilon_i \mid \mathbf{u}) = \sigma^2 g^2(\mu_i, \boldsymbol{\nu}_i, \boldsymbol{\delta})$ tarjoaa varsin joustavan työkalun varianssin mallintamiseen, koska se mahdollistaa varianssin mallintamisen kiinteiden vaikutusten ja satunnaisten vaikutusten funktiona. Käytännössä ongelmaksi voi kuitenkin muodostua se, että satunnaisvaikutukset $\mathbf{u}$ ja satunnaisvirheet $\boldsymbol{\epsilon}$ eivät enää ole riippumattomia. Tällöin yksi mahdollisuus on käyttää approksimaatiota

$$Var(\epsilon_i) \approx \sigma^2 g^2(\hat{\mu}_i, \boldsymbol{\nu}_i, \boldsymbol{\delta}),$$

missä

$$\hat{\mu}_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\hat{\mathbf{u}}.$$

Approksimaatiossa satunnaisvirheet oletetaan riipumattomiksi satunnaisvaikutuksista.

R-ohjelmistossa voidaan käyttää seuraavia varianssifunktioita:

- *varFixed* (kiinteä varianssi)
- *varIdent* (havainnoille eri varianssit)
- *varPower* (kovariaatin potenssi)
- *varExp* (luku e potenssiin kovariaatti)
- *varConstPower* (vakio + kovariaatin pot.)
- *varComb* (varianssifunktioiden yhdistelmä)

Funktioita käytetään samaan tapaan kuin kovarianssirakenteen muodostavia funktioita. Argumentteina annetaan arvo ja muoto. Ensimmäisessä argumentissa annetaan δ ja toisessa annetaan varianssin kovariaatti ν ja mahdollinen ryhmittely. Jos esimerkiksi halutaan määritellä muuttuja *age* kovariaatiksi, niin että molemmille sukupuolille estimoidaan omat parametrit, niin voitaisiin määritellä

$$form = \tilde{} age \mid Sex$$

tai jos kovariaattina halutaan käyttää sovitettuja arvoja, niin

$$form = \tilde{} fitted(.)$$

Esimerkki 24. Kiinteä varianssi *(varFixed).*
Tämä varianssifunktio on yhden selittäjän funktio. Tässä varianssi ajatellaan tunnetuksi lukuunottamatta vakiota σ^2. Jos esimerkiksi oletetaan, että varianssi kasvaa lineaarisesti ajan *age* funktiona, niin

$$Var(\epsilon_i) = \sigma^2 age_i$$

ja vastaava varianssifunktio on

$$g(age_i) = \sqrt{age_i}.$$

```
> data(Orthodont)
> vf1fixed<-varFixed(~age)
> vf1fixed
Variance function structure of class varFixed with no parameters, or
uninitialized
> vf1fixed<-Initialize(vf1fixed, data=Orthodont)
> varWeights(vf1fixed)
   [1] 0.3535534 0.3162278 0.2886751 0.2672612 ...
```

Viimeisellä rivillä on käytetty funktiota *varWeights* joka laskee varianssifunktion käänteisarvot.

Esimerkki 25. Varianssit ositteille *(varIdent)*.

Tässä luokassa ositteille voidaan määritellä erisuuruiset varianssit. Funktiolla *varIdent* on varianssien suhteelle asetettu alkuarvo 0.5, jos tämä suhde halutaan laskennassa pitää vakiona, niin käytetään argumenttia fixed. Esimerkiksi

```
> vfident<-varIdent(c(Female=0.5), ~1 | Sex, fixed=c(Female=0.5))
> vfident<-Initialize(vfident, Orthodont)
> varWeights(vfident)
   Male   Male   Male
      1      1      1    ...
Female Female Female
      2      2      2    ...
```

Nyt siis varianssi on tytöille

$$Var(\epsilon) = \sigma^2\delta_1^2$$

ja pojille

$$Var(\epsilon) = \sigma^2\delta_2^2$$

Lisäksi

$$\frac{\delta_1}{\delta_2} = 0.5.$$

Vastaava varianssifunktio on muotoa

$$g(j, \boldsymbol{\delta}) = \delta_j, \ j = 1, 2.$$

Esimerkki 26. Potenssifunktio *(varPower)*.

Tässä luokassa varianssimalli on muotoa

$$Var(\epsilon_i) = \sigma^2 \mid \nu_i \mid^{2\delta}$$

ja vastaava varianssifunktio on muotoa

$$g(\nu_i, \delta) = \sigma^2 \mid \nu_i \mid^{\delta} .$$

Tässä parametri δ on rajoittamaton (poislukien 0), joten sillä voidaan mallintaa sekä varianssin kasvua, että vähenemistä. Esimerkiksi

```
> vf1power<-varPower(1,form=~fitted(.), fixed=0.5)
> vf1power<-Initialize(vf1power,Orthodont)
> varWeights(vf1power)
  [1] 1 1 1 ...
```

Varianssi kasvaa nyt lineaarisesti sovitettujen arvojen funktiona. Varianssi on nyt siis

$$Var(\epsilon) = \sigma^2 \mid \mu \mid$$

ja vastaava varianssifunktio on

$$g(\mu) = \mid \mu \mid^{0.5} .$$

Esimerkki 27. Eksponettifunktio *(varExp)*.

Tässä luokassa varianssi on muotoa

$$Var(\epsilon_i) = \sigma^2 e^{2\delta\nu_i},$$

ja vastaava varianssifunktio on

$$g(\nu_i, \delta) = e^{\delta \nu_i}.$$

Parametri δ on nyt rajoittamaton, joten varianssi voi nyt kasvaa tai vähetä kovariaatin ν_i funktiona. Muuten syntaksi on samantyyppinen kuin potenssifunktiolla. Jos esimerkiksi halutaan, että varianssi kasvaa pojille eksponentiaalisesti, mutta tytöille varianssi pidetään vakiona, niin

```
> vf1Exp<-varExp(form=~age|Sex, fixed=c(female=0))
```

Esimerkki 28. Vakio ja potenssi *(varConstPower)*.
Varianssimalli on nyt muotoa

$$Var(\epsilon_i) = \sigma^2(\delta_1 + \mid \nu_i \mid^{\delta_2})^2$$

ja vastaava varianssifunktio on

$$g(\nu_i, \boldsymbol{\delta}) = \delta_1 + \mid \nu_i \mid^{\delta_2},$$

missä $\delta_1 > 0$ ja δ_2 on rajoittamaton. Jos kovariaatti μ_i on nolla tai lähellä nollaa, niin varianssifunktio on vakio ja lähellä vakiota δ_1. Jos taas kovariaatin μ_i itseisarvo poikkeaa nollasta, niin varianssifunktio kasvaa kovariaatin itseisarvon potenssina. Näin saadaan usein realistisempi malli varianssille kuin pelkästään funktiolla *varPower*, kun kovariaatin arvo on nolla tai lähellä nolla. Esimerkiksi

```
> vf1ConstPower<-varConstPower(power=0.5, fixed=list(const=1))
```

asettaa parametrin δ_1 arvoksi 1 ja parametrin δ_2 alkuarvoksi 0.5. Kovariaattina käytetään sovitettuja arvoja μ_i.

Esimerkki 29. Varianssien yhdistelmä *(varComb)*.

Funktiolla *varComb* saadaan useamman varianssifunktioiden kombionaatio. Jos esimerkiksi halutaan kombinoida vakiovarianssi ja iän suhteen eksponentiaalinen varianssin kasvu, niin

$$Var(\epsilon_i) = \sigma^2 \delta_1^2 e^{2\delta_2 \nu_i}.$$

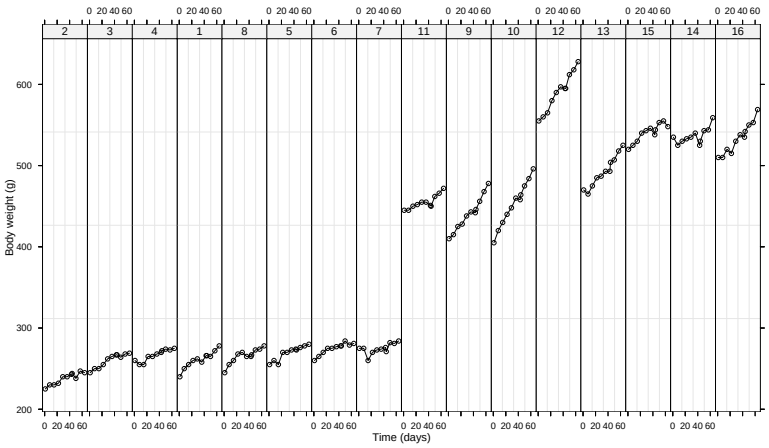
Esimerkiksi

```
> vf1comb<-varComb(varIdent(c(Female=0.5),~1 | Sex), varExp(1,~age))
```

Esimerkki 30. Varianssin mallintaminen rotta-aineistossa.

```
> data(BodyWeight)
> ?BodyWeight
    ...
      This data frame contains the following columns:
      weight a numeric vector giving the body weight of the rat (grams).
      Time a numeric vector giving the time at which the measurement is
           made (days).
      Rat an ordered factor with levels ‘2’ < ‘3’ < ‘4’ < ‘1’ < ‘8’ <
           ‘5’ < ‘6’ < ‘7’ < ‘11’ < ‘9’ < ‘10’ < ‘12’ < ‘13’ < ‘15’ <
           ‘14’ < ‘16’ identifying the rat whose weight is measured.
      Diet a factor with levels ‘1’ to ‘3’ indicating the diet that the
           rat receives.
    ...
> formula(BodyWeight)
weight ~ Time | Rat
> plot(BodyWeight)
```

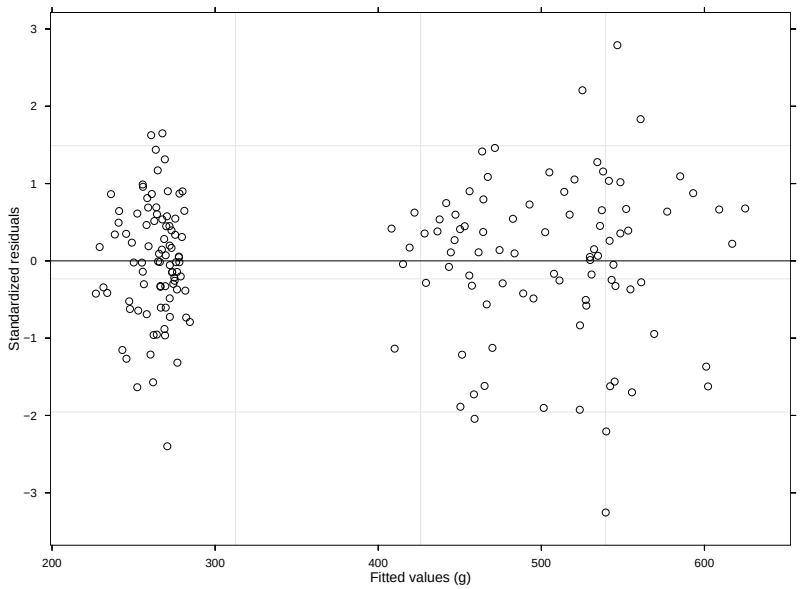
Aineistossa rottien paino on mitattu 64 päivän ajan, kun rotille on annettu kolmea erilaista ruokavaliota (3 ryhmää). Oletetaan, että paino kasvaa lineaa-



Kuva 17: Rottien painot

risesti ajan funktiona ja että sekä vakiotermiin, että kulmakertoimeen liittyy satunnaisvaikutus.

```
>  fm1bw.lme<-lme(weight~Time*Diet, data=BodyWeight, random=~Time|Rat)
>  summary(fm1bw.lme)
    ...
Random effects:
  Formula: ~Time | Rat
  Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 36.9390723 (Intr)
Time         0.2484113 -0.149
Residual     4.4436052
```



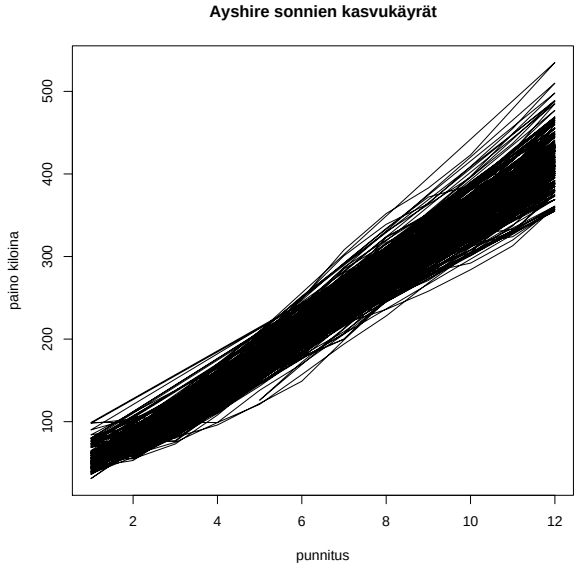
Kuva 18: Standartisoidut residuaalit ja sovitetut arvot rotta-aineistossa.

```
Fixed effects: weight ~ Time * Diet
                 Value Std.Error  DF   t-value p-value
(Intercept) 251.65165 13.094025 157 19.218816  <.0001
Time          0.35964  0.091140 157  3.946019  0.0001
Diet2       200.66549 22.679516  13  8.847873  <.0001
Diet3       252.07168 22.679516  13 11.114509  <.0001
Time:Diet2    0.60584  0.157859 157  3.837858  0.0002
Time:Diet3    0.29834  0.157859 157  1.889903  0.0606
    ...
>  plot(fm1bw.lme)
```

Kuvio paljastaa, että standartoiduissa residuaaleissa on heteroskedastisuutta. Sovitetaan sellainen varianssimalli, jossa varianssi kasvaa sovitettujen arvojen potenssina.

```
>  fm2bw.lme<-update(fm1bw.lme, weights=varPower())
>  summary(fm2bw.lme)
    ...
Variance function:
  Structure: Power of variance covariate
  Formula: ~fitted(.)
  Parameter estimates:
     power
0.5428088
Fixed effects: weight ~ Time * Diet
                 Value Std.Error  DF   t-value p-value
(Intercept) 251.60215 13.067404 157 19.254180  <.0001
Time          0.36109  0.088377 157  4.085746  0.0001
Diet2       200.77697 22.656366  13  8.861835  <.0001
Diet3       252.17016 22.661720  13 11.127583  <.0001
Time:Diet2    0.60183  0.155423 157  3.872210  0.0002
Time:Diet3    0.29524  0.155891 157  1.893863  0.0601
    ...
>  anova(fm1bw.lme,fm2bw.lme)
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
fm1bw.lme      1 10 1171.720 1203.078 -575.8599
fm2bw.lme      2 11 1163.921 1198.415 -570.9607 1 vs 2 9.798326  0.0017
```

Tässä siis varianssimalli, jossa varianssi kasvaa sovitettujen arvojen potenssina osoittautuu paremmaksi.

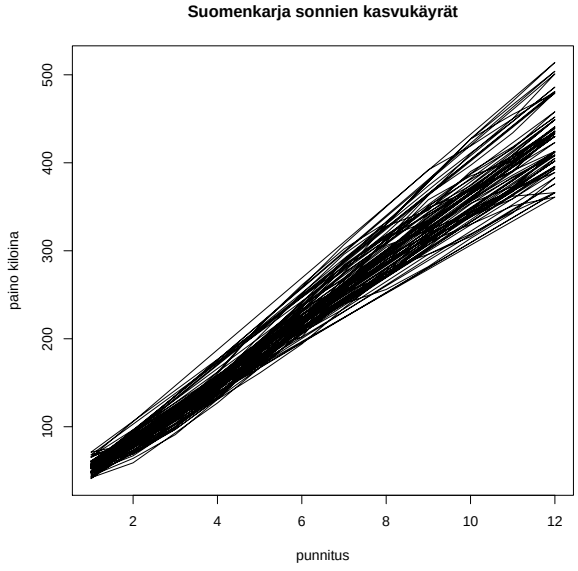


Kuva 19: Ayrshire-sonnien painon kehitys vuoden aikana.

Esimerkki 31. Varianssin mallintaminen sonni-aineistossa.

Esimerkkiaineistossa on tarkasteltu lihasonnien painon kehittymistä vuoden ajan. Punnituksia on tehty 12 kappaletta 30:n päivän välein. Aineistossa on 168 Ayrshire-sonnia ja 40 Suomenkarja- sonnia.

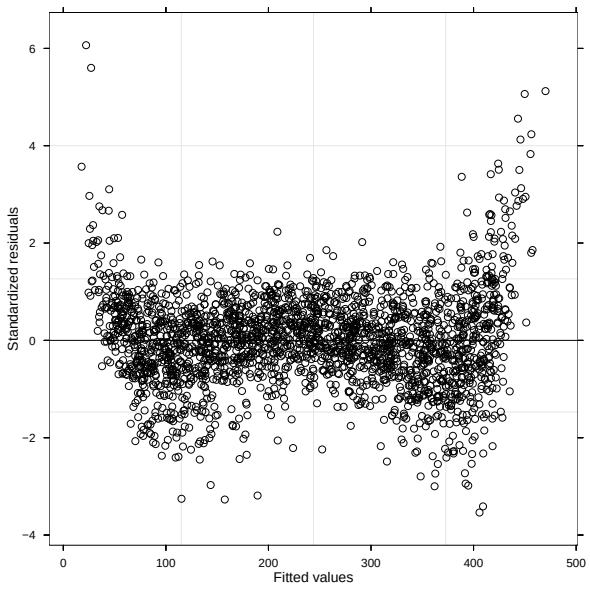
```
>  sonni<-read.table("C:/Omat tiedostot/opetus/mulli66.txt", header=TRUE)
>  library(nlme)
>  sonni<-groupedData(y~t|nro, data=sonni)
>  sonnia<-sonni[sonni$rotu==1,]
>  sonniy<-sonni[sonni$rotu==2,]
>  plot(sonnia$y~sonnia$t, type="l", xlab=c("punnitus"),
ylab=c("paino kiloina"),  + main=c("Ayrshire sonnien kasvukäyrät"))
>  plot(sonniy$y~sonniy$t, type="l", xlab=c("punnitus"),
ylab=c("paino kiloina"), main=c("Suomenkarja sonnien kasvukäyrät"))
```



Kuva 20: Suomenkarja-sonnien painon kehitys vuoden aikana.

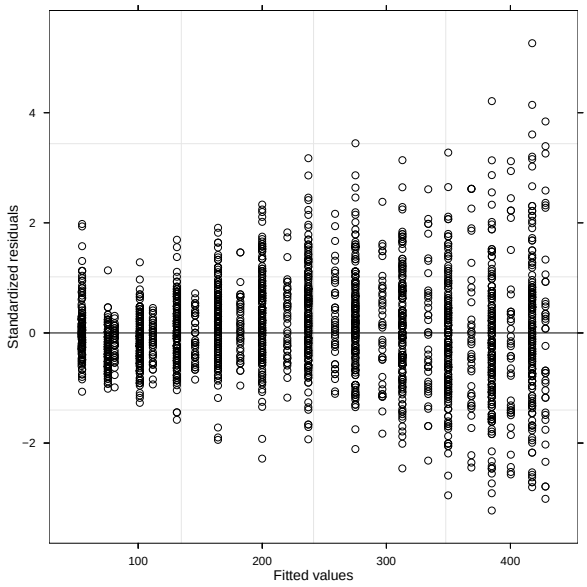
```
>sonni[1:10,]
Grouped Data: y ~ t | nro
    nro   y  t rotu
1  169  55  1    2
2  169  71  2    2
3  169  97  3    2
4  169 132  4    2
5  169 175  5    2
6  169 218  6    2
    ...
```

```
f1<-lme(y~factor(rotu)+factor(rotu):t+factor(rotu):I(t^2)+
factor(rotu):I(t^3)-1, data=sonni, random=~1)
```



Kuva 21: Mallin f1 residuaalit vs. sovitetut arvot.

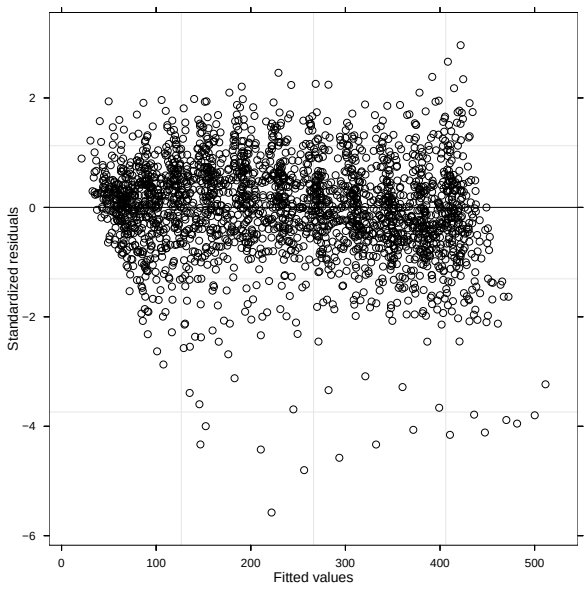
```
>  summary(f1)

    ...

Random effects:

  Formula: ~1 | nro

         (Intercept) Residual

StdDev:    16.14515 12.66613

Fixed effects: y ~ factor(rotu) + factor(rotu):t + ...

                          Value Std.Error   DF    t-value p-value

factor(rotu)1        29.866443  2.390765  206  12.492423  <.0001

factor(rotu)2        27.951263  4.153619  206   6.729376  <.0001

factor(rotu)1:t      15.933626  1.198731 2164  13.292076  <.0001

factor(rotu)2:t      20.646031  2.095210 2164   9.853921  <.0001

factor(rotu)1:I(t^2)  3.037233  0.200688 2164  15.134072  <.0001

factor(rotu)2:I(t^2)  2.876104  0.366977 2164   7.837299  <.0001
```



Kuva 22: Mallin f2 residuaalit vs. sovitetut arvot.

```
factor(rotu)1:I(t^3) -0.141519  0.009916 2164 -14.271961  <.0001
factor(rotu)2:I(t^3) -0.153435  0.018608 2164  -8.245577  <.0001
    ...

>trellis.device(device=getOption("device"),color=FALSE, bg="white")
>  plot(f1)
>  f2<-update(f1, corr=corAR1())
>  anova(f1,f2)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
f1     1 10 19467.24 19524.94 -9723.620
f2     2 11 17452.75 17516.22 -8715.373 1 vs 2 2016.495  <.0001
>  plot(f2)
```

Mallin *f2* residuaalikuviossa on selvää heteroskedastisuutta. Kokeilaan seuraavaksi varianssi mallintamista ajan (mittauskerta) funktiona.

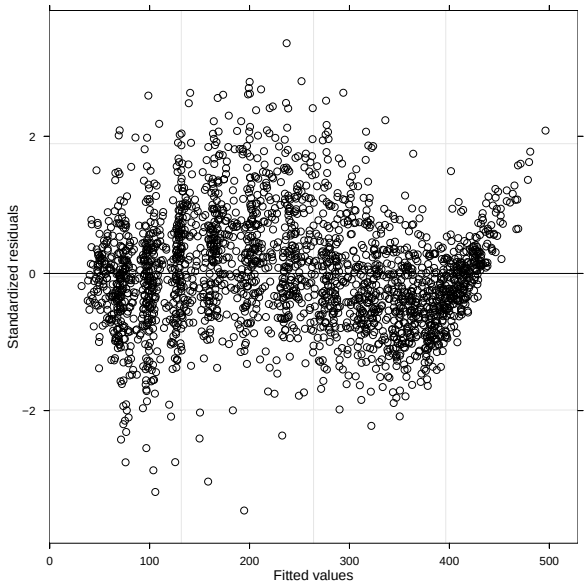


Kuva 23: Mallin f3 residuaalit vs. sovitetut arvot.

```
>  f3<-update(f2,weights=varFixed(~t))
>  anova(f2,f3)
    Model df      AIC      BIC    logLik
f2      1 11 17452.75 17516.22 -8715.373
f3      2 11 17423.85 17487.33 -8700.927
>plot(f3)
```

Mallin residuaalikuviossa ei ole enää selvää heteroskedastisuutta, mutta residuaalikuvio ei ole symmetrinen. Kokeillaan vielä mallin satunnaisosan mallintamista.

```
>  f4<-update(f3, random=~1+t+I(t^2))


>  anova(f3,f4)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
```



Kuva 24: Mallin f4 residuaalit vs. sovitetut arvot.

```
f3     1 11 17423.85 17487.33 -8700.927
f4     2 16 17120.31 17212.63 -8544.154 1 vs 2 313.5453  <.0001
>  plot(f4)
>  summary(f4)   ...
Random effects:
  Formula: ~1 + t + I(t^2) | nro
  Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
(Intercept) 17.3757038 (Intr) t
t            6.0829693 -0.976
I(t^2)       0.3839562  0.950 -0.921
Residual     5.6171172

Correlation Structure: AR(1)
```

```
  Formula: ~1 | nro
  Parameter estimate(s):
      Phi
0.8344463
Variance function:
  Structure: fixed weights
  Formula: ~t
Fixed effects: y ~ factor(rotu) + factor(rotu):t + ...
                        Value Std.Error   DF    t-value p-value
factor(rotu)1        41.24694 1.6643750  206  24.782241  <.0001
factor(rotu)2        36.95346 2.9822818  206  12.391001  <.0001
factor(rotu)1:t       8.37381 0.8899011 2164   9.409817  <.0001
factor(rotu)2:t      14.26612 1.5348475 2164   9.294809  <.0001
factor(rotu)1:I(t^2)  4.17368 0.1467409 2164  28.442554  <.0001
factor(rotu)2:I(t^2)  3.88177 0.2560731 2164  15.158851  <.0001
factor(rotu)1:I(t^3) -0.18788 0.0075962 2164 -24.732859  <.0001
factor(rotu)2:I(t^3) -0.19632 0.0137738 2164 -14.253183  <.0001
    ...
```

Mallin *f4* residuaalikuvio on selvästi "siistimpi"kuin edellisissä malleissa. Kuitenkin residuaaleissa on vielä havaittavissa selvää syklistä vaihtelua.

12 Toistomittaukset sekamallin avulla

Oletetaan jokaiselle yksilölle sekamalli

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, N,$$

missä $\mathbf{y}_i$ on $n_i \times 1$ havaintovektori ja $\mathbf{X}_i$ ja $\mathbf{Z}_i$ ovat annettuja matriiseja. Lisäksi oletetaan, että

$$\begin{cases} \mathbf{u}_i \sim N(\mathbf{0}, \mathbf{D}) \\ \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i) \\ \mathbf{u}_1, \ldots, \mathbf{u}_N, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N \text{ ovat riippumattomia,} \end{cases}$$

kun $i = 1, \ldots, N$. Oletuksista seuraa, että

$$\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$$

ja

$$Cov(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{O}, i \neq j,$$

missä

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i,$$

kun $i, j = 1, \ldots, N$. Usein oletetaan, että $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$, jolloin

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \sigma^2\mathbf{I}_{n_i}.$$

Matriisiesityksenä

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{Z}_1 & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{X}_2 & \mathbf{O} & \mathbf{Z}_2 & \ldots & \mathbf{O} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \mathbf{X}_N & \mathbf{O} & \mathbf{O} & \ldots & \mathbf{Z}_N \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{pmatrix} + \boldsymbol{\epsilon}$$

$$\mathbf{y} = (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}.$$

Jos tunnetaan

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{V}_2 & & \vdots \\ \vdots & & \cdots & \\ \mathbf{O} & & \cdots & \mathbf{V}_N \end{pmatrix},$$

niin parametrien $\boldsymbol{\beta}$ BLUE on

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= (\sum_{i=1}^{N}\mathbf{X}_i\mathbf{V}_i^{-1}\mathbf{X}_i)^{-}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i \end{aligned}$$

ja havainnolle i $\mathbf{u}_i$:n BLUP on

$$\hat{\mathbf{u}}_i = \mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}).$$

Esimerkki 32. Dental-aineiston mallintaminen.

- 4 mittauskertaa: mittaukset 8, 10, 12 ja 14-vuoden iässä
- 11 tyttöä ja 16 poikaa

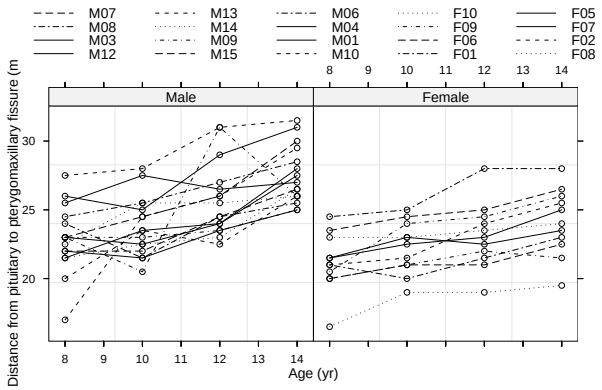
Jos etäisyys (y) kasvaa lineaarisesti ajan (t) funktiona, niin

$$y = \beta_0 + \beta_1 t + \epsilon.$$

Voidaan ajatella, että tasoparametriin β_0 liittyy satunnaisvaikutus u. Tästä saadaan sekamalli

$$y = (\beta_0 + u) + \beta_1 t + \epsilon,$$

missä u ja ϵ ovat riippumattomia sekä $u \sim N(0, d^2)$ ja $\epsilon \sim N(0, \sigma^2)$.



Kuva 25: Kasvukäyrät Dental-aineistossa

Matriisiesityksenä

- tytöt: $\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{Z}u_i + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, 11$
- pojat: $\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_2 + \mathbf{Z}u_i + \boldsymbol{\epsilon}_i,\ i = 12, \ldots, 27$

missä siis

$$\mathbf{X} = \begin{pmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{pmatrix} \quad \text{ja} \quad \mathbf{Z} = \mathbf{1}.$$

Jos nyt oletetaan, että $\mathbf{R} = \sigma^2\mathbf{I}$, niin

$$Var(\mathbf{y}_i) = d^2\mathbf{1}\mathbf{1}' + \sigma^2\mathbf{I}.$$

Edellä olevalla kovarianssimatriisilla on ns. *tasakorrelaatiorakenne (uniform struc-*

ture, compound symmetry). Yhtenä sekamallina lausuttuna saadaan

$$\mathbf{y} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{O} \\ \vdots & \vdots \\ \mathbf{X}_{11} & \mathbf{O} \\ \mathbf{O} & \mathbf{X}_{12} \\ \vdots & \vdots \\ \mathbf{O} & \mathbf{X}_{27} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{Z}_2 & & \vdots \\ \vdots & & \ddots & \\ \mathbf{O} & \cdots & & \mathbf{Z}_{27} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{27} \end{pmatrix} + \boldsymbol{\epsilon}$$

eli

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u} + \boldsymbol{\epsilon}.$$

Huom. Koska nyt $\mathbf{X}_1 = \cdots = \mathbf{X}_{27}$ ja $\mathbf{Z}_1 = \cdots \mathbf{Z}_{27}$, voidaan aineisto esittää myös ns. kasvukäyrämallina (GMANOVA). Nyt saadaan

$$\mathbf{X}^* = \left(\begin{pmatrix} \mathbf{1}_{11} & \mathbf{0}_{11} \\ \mathbf{0}_{16} & \mathbf{1}_{16} \end{pmatrix} \otimes \mathbf{X} \right) \text{ ja } \mathbf{Z}^* = (\mathbf{I}_{27} \otimes \mathbf{Z}),$$

jolloin malli on

$$\begin{aligned} \mathbf{y} &= \left(\begin{pmatrix} \mathbf{1}_{11} & \mathbf{0}_{11} \\ \mathbf{0}_{16} & \mathbf{1}_{16} \end{pmatrix} \otimes \mathbf{X} \right) \boldsymbol{\beta} + (\mathbf{I}_{27} \otimes \mathbf{Z})\mathbf{u} + \boldsymbol{\epsilon} \\ &= (\mathbf{A} \otimes \mathbf{X})\boldsymbol{\beta} + (\mathbf{I}_{27} \otimes \mathbf{Z})\mathbf{u} + \boldsymbol{\epsilon}, \end{aligned}$$

missä

$$(\mathbf{A} \otimes \mathbf{X}) = \left(\begin{pmatrix} \mathbf{1}_{11} & \mathbf{0}_{11} \\ \mathbf{0}_{16} & \mathbf{1}_{16} \end{pmatrix} \otimes \mathbf{X} \right).$$

Tällöin

$$E(\mathbf{y}) = (\mathbf{A} \otimes \mathbf{X})\boldsymbol{\beta}$$

ja

$$\begin{aligned} Var(\mathbf{y}) &= (\mathbf{I}_{27} \otimes \mathbf{Z})(\mathbf{I} \otimes \mathbf{D})(\mathbf{I}_{27} \otimes \mathbf{Z}') + \sigma^2\mathbf{I}_{27} \\ &= (\mathbf{I}_{27} \otimes \mathbf{Z}\mathbf{D}\mathbf{Z}') + \sigma^2\mathbf{I}_{27}. \end{aligned}$$

Koska $\mathbf{Z} = \mathbf{1}_4$, saadaan

$$Var(\mathbf{y}) = (\mathbf{I}_{27} \otimes d^2 \mathbf{1}_4 \mathbf{1}_4') + \sigma^2 \mathbf{I}_{27}.$$

Jos sovelletaan tulosta

$$vec(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A}) vec\mathbf{B},$$

niin saadaan

$$(\mathbf{A} \otimes \mathbf{X})\boldsymbol{\beta} = vec(\mathbf{XBA}'),$$

missä $\boldsymbol{\beta} = vec(\mathbf{B})$. Malli $\mathbf{XBA}'$ on ns. kasvukäyrämalli (GMANOVA, Generalized Multivariate Analysis of Variance). Perinteinen kasvukäyrämallien teoria perustuu kuitenkin siihen, että jäännöksille ei oleteta mitään erityistä rakennetta. Kasvukäyrämallissa lisäksi oletetaan, että yksilöt $\mathbf{y}_i$ on mitattu samoissa aikapisteissä eikä puuttuvaa tietoa sallita. Yleisessä sekamallissa ei näitä rajoituksia ole.

```
>  library(nlme)
Loading required package: nls
>  data(Orthodont)
>  l1<-lme(distance~Sex+Sex:age-1, data=Orthodont, random=~1, weights=
+ varIdent(c(Female=0.5), ~1|Sex))
>  summary(l1)
    ...
Random effects:
  Formula: ~1 | Subject
         (Intercept) Residual
StdDev:    1.847574 1.669823

Variance function:
  Structure: Different standard deviations per stratum
```

```
  Formula: ~1 | Sex
  Parameter estimates:
     Male    Female
1.0000000 0.4678937
Fixed effects: distance ~ Sex + Sex:age - 1
                  Value Std.Error DF   t-value p-value
SexMale       16.340625 1.1450949 25 14.270105  <.0001
SexFemale     17.372727 0.8123610 25 21.385478  <.0001
SexMale:age    0.784375 0.0933460 80  8.402882  <.0001
SexFemale:age  0.479545 0.0526752 80  9.103815  <.0001
    ...
>  l2<-lme(distance~Sex+Sex:age-1, data=Orthodont, random=~1)
>  anova(l2,l1)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
l2      1  6 445.7572 461.6236 -216.8786
l1      2  7 429.2205 447.7312 -207.6102 1 vs 2 18.53677  <.0001
```

Selvästi malli, jossa satunnaisvirheiden varianssi mallinnetaan erikseen tytöille ja pojille osoittautuu paremmaksi. Estimoidut malli tytöille ja pojille ovat:

$$y = (17.372727 + u_0) + 0.479545 \times t + \epsilon,$$

missä $Var(u_0) = 1.847574^2$ ja $Var(\epsilon) = (0.4678937 * 1.669823)^2$ sekä

$$y = (16.340625 + u_0) + 0.784375 \times t + \epsilon,$$

missä $Var(u_0) = 1.847574^2$ ja $Var(\epsilon) = 1.669823^2$.

13 Kiinteiden vaikutusten testaamisesta

Jos mallit ovat sisäkkäisiä, voidaan kiinteää osaa testata esimerkiksi uskottavuussuhdetestillä, kun estimointimenetelmänä käytetään suurimman uskottavuuden menetelmää. Testataan identtisyyshypoteesi Dental-aineistossa. Nollahypoteesi on siis

$$H_0 : \beta_{0t} = \beta_{0p} \text{ ja } \beta_{1t} = \beta_{1p}.$$

```
# ks. esim. 32
>  l1<-update(l1, method="ML")
>  l0<-update(l1, distance~age)
>  anova(l0,l1)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
l0      1  5 435.6933 449.1040 -212.8467
l1      2  7 423.3524 442.1273 -204.6762 1 vs 2 16.34097   3e-04
```

Tytöille ja pojille estimoidut suorat eivät siis selvästikään olisi identtisiä.

Yleisemmin testaamiseen käytetään yleistä lineaarista hypoteesia

$$H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$$

ja testisuureena käytetään F-testisuuretta (ks. luku 3). Identtisyyshypoteesi olisi nyt:

$$H_0 : \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_{0t} \\ \beta_{0p} \\ \beta_{1t} \\ \beta_{1p} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Leftrightarrow H_0 \begin{pmatrix} \beta_{0t} - \beta_{0p} \\ \beta_{1t} - \beta_{1p} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Jos testataan suorien yhdensuuntaisuus, niin

$$H_0 : \begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_{0t} \\ \beta_{0p} \\ \beta_{1t} \\ \beta_{1p} \end{pmatrix} = 0 \Leftrightarrow H_0 : \beta_{1t} - \beta_{1p} = 0$$

Yhdensuuntaisuus voitaisiin R-ohjelmistossa testata seuraavasti:

```
>  anova(l1, L=c(0,0,1,-1))
F-test for linear combination(s)
   SexMale:age SexFemale:age
             1             -1
   numDF denDF  F-value p-value
1      1    80 7.976532   0.006
```

Suorat eivät selvästikään ole yhdensuuntaisia. Vakiotermien yhdensuuntaisuus voitaisiin testata seuraavasti:

```
>  anova(l1, L=c(1,-1,0,0))
F-test for linear combination(s)
   SexMale SexFemale
         1        -1
   numDF denDF   F-value p-value
1      1    25 0.5412463  0.4688
```

Estimoiduissa malleissa vakiotermin osoittautuisivat yhtäsuuriksi.

Tarkastellaan vielä testausta Sonniaineistossa (ks. esimerkki 31). Testataan ensin käyrien identtisyyttä.

```
>  f5<-update(f4, method="ML")
>  f6<-update(f5, y~t+I(t^2)+I(t^3))
```

```
>  anova(f6,f5)
    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
f6      1 12 17143.54 17212.82 -8559.770
f5      2 16 17099.82 17192.20 -8533.909 1 vs 2 51.72284  <.0001
```

Käyrät eivät selvästikään ole identtisiä. Testataan seuraavaksi kolmannen asteen termin merkitsevyys.

```
>  m<-matrix(c(0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1), nr=2)
>  m
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    0    0    0    0    0    0    1    0
[2,]    0    0    0    0    0    0    0    1
>  anova(f5, L=m)
F-test for linear combination(s)
   factor(rotu)1:I(t^3) factor(rotu)2:I(t^3)
1                    1                    0
2                    0                    1
   numDF denDF  F-value p-value
1     2  2164 408.9624  <.0001
```

Kolmannen asteen termi näyttäisi olevan merkitsevä. Testataan seuraavaksi vakiotermien yhtäsuuruus.

```
>  anova(f5, L=c(1,-1,0,0,0,0,0,0))
F-test for linear combination(s)
factor(rotu)1 factor(rotu)2
             1            -1
   numDF denDF  F-value p-value
1     1   206 1.612764  0.2055
```

Mallissa vakiotermit (syntymäpainot) näyttäisivät olevan yhtäsuuria. Testataan vielä loputkin kertoimet.

```
>  anova(f5, L=c(0,0,1,-1,0,0,0,0))
F-test for linear combination(s)
factor(rotu)1:t factor(rotu)2:t
               1              -1


   numDF denDF  F-value p-value
1      1  2164 11.19061   8e-04
>  anova(f5, L=c(0,0,0,0,1,-1,0,0))
F-test for linear combination(s)
factor(rotu)1:I(t^2) factor(rotu)2:I(t^2)
                   1                  -1
   numDF denDF   F-value p-value
1      1  2164 0.9980211  0.3179
>  anova(f5, L=c(0,0,0,0,0,0,1,-1))
F-test for linear combination(s)
factor(rotu)1:I(t^3) factor(rotu)2:I(t^3)
                   1                  -1
   numDF denDF   F-value p-value
1      1  2164 0.2810787   0.596
```

Merkitsevä ero saatiin ainoastaan lineaarisen termin kertoimeen.

14 Ennustaminen sekamallilla

Sekamallin yleinen muoto on

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

missä $\mathbf{X}$ on kuten tavallisessa lineaarisessa mallissa ja $\mathbf{Z}$ on satunnaisosan $\mathbf{u}$ suunnittelumatriisi ja $\boldsymbol{\epsilon}$ sisältää satunnaisvirheet. Yleisessä sekamallissa oletetaan, että

$$E\begin{pmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}$$

ja

$$Var\begin{pmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{pmatrix} = \begin{pmatrix}\mathbf{D} & \mathbf{O}\\ \mathbf{O} & \mathbf{R}\end{pmatrix}.$$

Sekamallin oletuksista seuraa, että

$$Var(\mathbf{y}) = Var(\mathbf{Zu}) + Var(\boldsymbol{\epsilon}) = \mathbf{ZDZ}' + \mathbf{R}.$$

Usein tehdään lisäksi normaalisuusoletus

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZDZ}' + \mathbf{R}).$$

14.1 Keskineliövirheen mielessä paras ennuste

Olkoon tuleva havainto

$$y = \mathbf{x}'\boldsymbol{\beta} + \mathbf{z}'\mathbf{u} + \epsilon,$$

Oletetaan nyt, että $(\mathbf{y}', y)'$:n jakauma on multinormaalijakauma. Tällöin keskineliövirheen mielessä paras y:n ennuste on ehdollinen odotusarvo

$$E(y \mid \mathbf{y}) = \mathbf{x}'\boldsymbol{\beta} + \mathbf{z}'E(\mathbf{u} \mid \mathbf{y}) + E(\epsilon \mid \mathbf{y}).$$

Kun nyt

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{ZDZ}' + \mathbf{R} & \mathbf{ZD}\\ \mathbf{DZ}' & \mathbf{D}\end{pmatrix}\right)$$

saadaan

$$E(\mathbf{u} \mid \mathbf{y}) = \mathbf{DZ}'\{\mathbf{ZDZ}' + \mathbf{R}\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \hat{\mathbf{u}}$$

ja kun

$$\begin{pmatrix}\mathbf{y}\\ \epsilon\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{ZDZ}' + \mathbf{R} & \mathbf{r}\\ \mathbf{r}' & r^2\end{pmatrix}\right),$$

niin

$$E(\epsilon \mid \mathbf{y}) = \mathbf{r}'\{\mathbf{ZDZ}' + \mathbf{R}\}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \hat{\epsilon}.$$

Saadaan prediktori

$$\hat{y} = \mathbf{x}'\boldsymbol{\beta} + \mathbf{z}'\hat{\mathbf{u}} + \hat{\epsilon}.$$

Voidaan lisäksi näyttää, että

$$\hat{\mathbf{u}} = \{\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\}^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

ja

$$\hat{\epsilon} = \mathbf{r}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\hat{\mathbf{u}}).$$

14.2 Paras lineaarinen ja harhaton prediktori

Edellä oletettiin, että havainnot noudattavat normaalijakaumaa. Periaatteessa hyvin samannäköiseen prediktoriin päästään ilman normaalisuusoletustakin. Tarkastellaan nyt vain yksinkertaista tilannetta $\mathbf{R} = \sigma^2\mathbf{I}$. Olkoon nyt

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon} \\ &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_*, \end{aligned}$$

missä

$$\boldsymbol{\epsilon}_* = \mathbf{Zu} + \boldsymbol{\epsilon}$$

ja

$$Var(\boldsymbol{\epsilon}) = \mathbf{ZDZ}' + \sigma^2\mathbf{I} = \mathbf{V}$$

sekä

$$Cov(\epsilon, \boldsymbol{\epsilon}_*') = \mathbf{z}'\mathbf{DZ}' = \mathbf{v}'$$

Paras lineaarinen ja harhaton prediktori (BLUP) on muotoa

$$\begin{aligned} \hat{y} &= \mathbf{x}'\tilde{\boldsymbol{\beta}} + \mathbf{v}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{x}'\tilde{\boldsymbol{\beta}} + \mathbf{z}'\mathbf{DZ}'\{\mathbf{ZDZ}' + \sigma^2\mathbf{I}\}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{x}'\tilde{\boldsymbol{\beta}} + \mathbf{z}'\hat{\mathbf{u}}, \end{aligned}$$

missä

$$\hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

ja

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Esimerkki 33. Ennustaminen toistomittausmallissa (ks. esim. Dental-aineisto). Olkoon yksilön i mittaukset vektorissa $\mathbf{y}_i$, missä $i = 1, \ldots, n$. Olkoon mallina sekamalli

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i,$$

missä $\mathbf{X}_i$ on kiinteän osan ja $\mathbf{Z}_i$ satunnaisosan suunnittelumatriisi. Oletetaan lisäksi, että

$$Var\begin{pmatrix}\mathbf{u}_i \\ \boldsymbol{\epsilon}_i\end{pmatrix} = \begin{pmatrix}\mathbf{D} & \mathbf{O} \\ \mathbf{O} & \sigma^2\mathbf{I}_i\end{pmatrix}$$

sekä

$$Cov(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{O} \text{ ja } Cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = \mathbf{O}$$

jokaisella $i \neq j$. Yhtenä sekamallina kirjoitettuna saadaan

$$\mathbf{y} = \begin{pmatrix}\mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n\end{pmatrix} = \begin{pmatrix}\mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n\end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix}\mathbf{Z}_1 & & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & & \mathbf{Z}_n\end{pmatrix}\mathbf{u} + \boldsymbol{\epsilon},$$

missä $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_n')'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \ldots, \boldsymbol{\epsilon}_n')'$ ja

$$\mathbf{V} = Var(\mathbf{y}) = \begin{pmatrix}\mathbf{Z}_1\mathbf{D}\mathbf{Z}_1' + \sigma^2\mathbf{I} & & \mathbf{O} \\ & \ddots & \\ \mathbf{O} & & \mathbf{Z}_n\mathbf{D}\mathbf{Z}_n' + \sigma^2\mathbf{I}\end{pmatrix}.$$

Tarkastellaan nyt tulevan havainnon y ennustamista yksilölle n. Paras lineaarinen ja harhaton prediktori on nyt

$$\hat{y} = \mathbf{x}'\tilde{\boldsymbol{\beta}} + \mathbf{v}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

missä

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Nyt koska

$$\mathbf{v}' = Cov(\epsilon, \boldsymbol{\epsilon}') = (\mathbf{0}', \ldots, \mathbf{0}', \mathbf{z}'\mathbf{D}\mathbf{Z}_n')$$

saadaan

$$\mathbf{v}'\mathbf{V}^{-1} = (\mathbf{0}', \ldots, \mathbf{0}, \mathbf{z}'\mathbf{D}\mathbf{Z}_n'\{\mathbf{Z}_n\mathbf{D}\mathbf{Z}_n' + \sigma^2\mathbf{I}\}^{-1})$$

ja

$$\begin{aligned} \hat{y} &= \mathbf{x}'\tilde{\boldsymbol{\beta}} + \mathbf{z}'\mathbf{D}\mathbf{Z}_n'\{\mathbf{Z}_n\mathbf{D}\mathbf{Z}_n' + \sigma^2\mathbf{I}\}^{-1}(\mathbf{y}_n - \mathbf{X}_n\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{x}'\tilde{\boldsymbol{\beta}} + \mathbf{z}'\hat{\mathbf{u}}_n, \end{aligned}$$

missä

$$\hat{\mathbf{u}}_n = \mathbf{D}\mathbf{Z}_n'\{\mathbf{Z}_n\mathbf{D}\mathbf{Z}_n' + \sigma^2\mathbf{I}\}^{-1}(\mathbf{y}_n - \mathbf{X}_n\tilde{\boldsymbol{\beta}})$$

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= (\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i, \end{aligned}$$

missä

$$\mathbf{V}_n = \mathbf{Z}_n\mathbf{D}\mathbf{Z}_n' + \sigma^2\mathbf{I}.$$

Tässä on tarkasteltu ainoastaan yhden havainnon ennustamista, mutta tulokset yleistyvät suoraan useamman havainnon ennustamiseen. Erikoistapaksena tutkitaan mallin antamia "ennusteita", koko profiilille

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}_n\tilde{\boldsymbol{\beta}} + \mathbf{Z}_n\hat{\mathbf{u}}_n \\ &= \mathbf{X}_n\tilde{\boldsymbol{\beta}} + \mathbf{Z}_n\mathbf{D}\mathbf{Z}_n'\mathbf{V}_n^{-1}(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}) \\ &= \mathbf{X}_n'\tilde{\boldsymbol{\beta}} + (\mathbf{V}_n - \sigma^2\mathbf{I})\mathbf{V}_n^{-1}(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}) \\ &= \mathbf{X}_n\tilde{\boldsymbol{\beta}} + (\mathbf{I} - \sigma^2\mathbf{V}_n^{-1})(\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}) \\ &= \sigma^2\mathbf{V}_n^{-1}\mathbf{X}_n\tilde{\boldsymbol{\beta}} + (\mathbf{I} - \sigma^2\mathbf{V}_n^{-1})\mathbf{y}_n. \end{aligned}$$

Siten koko profiilin ennuste määräytyy keskikäyrän ja ennustettavan havainnon painotettuna lineaarikombinaationa painomatriiseina $\sigma^2\mathbf{V}_n^{-1}$ ja $(\mathbf{I} - \sigma^2\mathbf{V}_n^{-1})$. Jos nyt $V_n = \sigma^2\mathbf{I}$ (ei satunnaisosaa), niin ennuste määräytyy pelkän keskikäyrän perusteella, eikä kyseisen yksilön sisältämää informaatiota voida käytää ennustettaessa.

Esimerkki 34. Ennusteiden laskenta R-ohjelmistolla

```
>  library(nlme)
Loading required package: nls
>  data(Orthodont)
>  l<-lme(Orthodont)
>  predict(l, level=1)
      M01      M01      M01      M01      M02      M02
24.81966 26.57140 28.32313 30.07487 21.43116 22.78054
>  predict(l, level=0)
      M01      M01      M01      M01      M02      M02
22.04259 23.36296 24.68333 26.00370 22.04259 23.36296
>
newOrth<-data.frame(Sex=c("Male","Male", "Female", "Female",
"Male","Male"), age = c(15, 20, 10, 12, 2, 4), + Subject =
c("M01","M01","F30","F30","M04","M04"))

>  predict(l, newOrth, level=0:1)
   Subject predict.fixed predict.Subject
1      M01      26.66389        30.95074
2      M01      29.96481        35.33008
3      F30      23.36296              NA
4      F30      24.68333              NA
5      M04      18.08148        20.95007
```

```
6     M04      19.40185        22.13869

>  intervals(predict(l))
Error in intervals(predict(l)) : no applicable method for "intervals"
```

Esimerkki 35. Approksimatiivinen ennusteväli toistomittausmallille.
Eräs approksimatiivinen ennusteväli ennustettaessa $n.$ havainnon tulevaa mittausta y voidaan laskea seuraavasti. Olkoon havainnon n mittaukset vektorissa $\mathbf{y}$ ja näihin mittauksiin liittyvä suunnittelumatriisi on $\mathbf{X}$. Tarkastellaan ensin ennustevirhettä

$$\begin{aligned}\hat{y} - y &= \mathbf{x}'\tilde{\boldsymbol{\beta}} + \mathbf{v}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) - y \\ &= (\mathbf{v}'\mathbf{V}^{-1}, -1)\begin{pmatrix}\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} \\ y - \mathbf{X}'\tilde{\boldsymbol{\beta}}\end{pmatrix} \\ &= \mathbf{A}(\mathbf{y}_* - \mathbf{X}_*\tilde{\boldsymbol{\beta}}),\end{aligned}$$

missä $\mathbf{y}_* = (\mathbf{y}', y)'$, $\mathbf{X} = (\mathbf{X}', \mathbf{x})'$ ja $\mathbf{A} = (\mathbf{v}'\mathbf{V}^{-1}, -1)$. Nyt saadaan

$$Var(\hat{y} - y \mid \mathbf{y}) = \mathbf{A}Var(\mathbf{y}_* - \mathbf{X}_*\tilde{\boldsymbol{\beta}} \mid \mathbf{y})\mathbf{A}'.$$

Oletetaan nyt, että $\mathbf{y}$ on riippumaton $\hat{\boldsymbol{\beta}}$:sta, t.s. $\boldsymbol{\beta}$ on estimoitu aikaisemmasta datasta, jossa havainto $\mathbf{y}$ ei ole mukana. Saadaan

$$\mathbf{A}Var(\mathbf{y}_* - \mathbf{X}_*\tilde{\boldsymbol{\beta}} \mid \mathbf{y})\mathbf{A}' = \mathbf{A}Var(\mathbf{y}_* \mid \mathbf{y})\mathbf{A}' + \mathbf{A}Var(\mathbf{X}_*\tilde{\boldsymbol{\beta}} \mid \mathbf{y})\mathbf{A}'.$$

Oikean puolen ensimmäinen termi on nyt

$$\mathbf{A}Var(\mathbf{y}_* \mid \mathbf{y})\mathbf{A}' = \mathbf{A}\begin{pmatrix}\mathbf{O} & \mathbf{O} \\ \mathbf{O} & \sigma_*^2\end{pmatrix}\mathbf{A}' = \sigma_*^2,$$

missä $\sigma_*^2 = Var(y \mid \mathbf{y}) = \sigma^2 - \mathbf{v}'\mathbf{V}^{-1}\mathbf{v}$, ja vasen puoli on

$$\begin{aligned}\mathbf{A}Var(\mathbf{X}_*\tilde{\boldsymbol{\beta}} \mid \mathbf{y})\mathbf{A}' &= (\mathbf{v}'\mathbf{V}^{-1}, -1)\begin{pmatrix}\mathbf{X} \\ \mathbf{x}'\end{pmatrix}Var(\tilde{\boldsymbol{\beta}})(\mathbf{X}', \mathbf{x})(\mathbf{V}^{-1}\mathbf{v}, -1) \\ &= (\mathbf{v}'\mathbf{V}^{-1}\mathbf{X} - \mathbf{x}')Var(\tilde{\boldsymbol{\beta}})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{v} - \mathbf{x}).\end{aligned}$$

Saadaan siis

$$Var(\hat{y} - y \mid \mathbf{y}) = \sigma_*^2 + \mathbf{m}'\{Var(\tilde{\boldsymbol{\beta}})\}\mathbf{m} = \sigma_\omega^2,$$

missä $\mathbf{m} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{v} - \mathbf{x}$. Jos nyt

$$\hat{y} - y \sim N(0, \sigma_\omega^2),$$

niin $100(1-\alpha)$ %:n ennusteväliksi saadaan

$$\hat{y} \pm z_{\alpha/2}\sigma_\omega,$$

missä $z_{\alpha/2}$ on standartoidun normaalijakauman vastaava fraktiili. Jos nyt σ_ω korvataan estimaatilla $\hat{\sigma}_\omega$, niin saadaan approksimatiivinen ennusteväli. Ennustevälissä ei kuitenkaan ole huomioitu σ_ω:n estimointiin liittyvää vaihtelua, jolloin siis pienellä otoskoolla saadaan liian kapeita ennustevälejä.