ÕPIOBJEKT
„Sissejuhatus üldiste lineaarsete mudelite teooriasse“

Tanel Kaart
http://www.eau.ee/~ktanel/lineaarne_mudel/

Eesti Maaülikool
Estonian University of Life Sciences
www.emu.ee

Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

Õpiobjektid -> Sissejuhatus üldiste lineaarsete mudelite teooriasse

SISSEJUHATUS ÜLDISTE LINEAARSETE MUDELITE TEOORIASSE

Õpiobjekti kirjeldus
Õpijuhis

Sissejuhatus
1. Põhimõisted

2. Üldiste lineaarsete mudelite ja faktorite liigitus
- Diskreetsed ja pidevad faktorid
- Faktorite vahekord mudelis
- Juhuslikud ja fikseeritud faktorid
- Tasakaalus ja mittetasakaalus andmed (mudelid)

3. Üldise lineaarse mudeli esitused
- Mudeli esitus objektiviisi
- Üldine lineaarne mudel maatrikskujul
- Üldine lineaarne segamudel maatrikskujul

4. Keskväärtused ja dispersioonid
- Vaatluste keskväärtused ja kovariatsioonistruktuur fikseeritud mudeli korral
- Vaatluste keskväärtused ja kovariatsioonistruktuur segamudeli korral

5. Eeldused ja kitsendused
- Jaotused
- Kovariatsioonistruktuur
- Täisastakuga ja mitte täisastakuga mudelid, reparametriseerimis-tingimused

6. Fikseeritud efektide hindamine (BLUE)
- Vähimruutude hinnangud
- Hinnatavad funktsioonid
- Vähimruutkeskmised

7. Juhuslike faktorite realiseerunud väärtuste prognoosimine (BLUP)

Õpiobjekti kirjeldus

Õppekava: Loomakasvatus (449)

Õppeaine: VL.0192 Loomade aretusväärtuse hindamine ja aretusprogrammid

Maht: 10 tundi

Sihtrühm: Loomakasvatuse õppekava magistrandid

Eesmärk: Õpiobjekti eesmärk on toetada õppeaine omandamist

Õpiobjekti läbinu:

- mõistab üldiste lineaarsete mudelite esitust nii objektiviisi kui ka maatrikskujul;
- oskab hinnata mudelite parameetrite väärtuseid lihtsamate mudelite korral ja mõistab nende tähendust;
- omab võimalust enesekontrolliks.

Sisu ja tehniline teostus: Tanel Kaart

Eesti Maaülikool
kevadsemester 2012

Järgmine >

ÜLDINE LINEAARNE MUDEL

Eelmises materjalides kirjeldatud selektsiooniindeksite rakendamine suurtes reaalsetes populatsioonides on sageli pärsitud mitmel põhjusel.

- Esiteks on keerulisemate põlvnemisskeemide puhul tülikas ja vägagi töömahukas kõikvõimalikke sugulussidemeid arvestavate indeksite välja töötamine.
- Teiseks eeldab selektsiooniindeksite teooria kõigi mitte aditiivgeneetiliste faktorite mõjude täpset teadmist ja mõõtmistulemuste nende suhtes korrigeerimist, mida on aga suuremamahuliste, mitmeid aastaid ja loomakasvatusettevõtteid hõlmavate uuringute korral pea võimatu teostada. Seetõttu võivad selektsiooniindeksite abil saadud aretusväärtuste hinnangud küll rahuldada parima lineaarse prognoosi (BLP) tingimusi, aga olla kokkuvõttes ikkagi positiivselt kallutatud paremais ettevõtteis või aastatel mõõdetud loomade suhtes (tulemuseks on nihkega hinnangud).
- Kolmandaks ei saa siiski alati lugeda selektsiooniindeksites kasutatavaid populatsiooni iseloomustavaid geneetilisi parameetreid (päritavuskoefitsient, korduvus, geneetilised korrelatsioonikordajad jne) teadaolevaiks, nende väärtuste hindamiseks selektsiooniindeksid aga enamasti ei sobi.

Kõigi kolme probleemi lahendus on kasutada **üldisi lineaarseid segamudeleid** (inglise keeles *general linear mixed models*, GLMM). Nende mudelite olemuses on nii tundmatute keskkonnamõjude kui ka aretusväärtuste samaaegne üksteise mõju arvesse võttev hindamine, mistõttu on tulemuseks nihketa hinnangud – nii keskkonnamõjude **parimad lineaarsed nihketa hinnangud** (*best linear unbiased estimate*, BLUE) kui ka aretusväärtuste **parimad lineaarsed nihketa prognoosid** (*best linear unbiased predictor*, BLUP). Samuti võimaldavad need mudelid sobivalt valitud kovariatsioonistruktuuri abil arvesse võtta kõikvõimalikke sugulussidemeid ning hinnata polügeensete tunnuste geneetilist determineeritust ja omavahelist seotust populatsioonis iseloomustavaid parameetrid.

Mitmetel populatsioonigeneetikas ja aretusteoorias kasutatavatel üldiste lineaarsete segamudelite erijuhtudel peatutakse pikemalt järgmistes materjalides. Käesolev materjal esitab üldiste lineaarsete mudelite ja üldiste lineaarsete segamudelite korrektseks rakendamiseks ja tulemuste tõlgendamiseks vajalikud põhiteadmised.

1 PÕHIMÕISTED

Üldise lineaarse mudeli rakendamiseks jagatakse registreeritud tunnused kahte ossa – uuritavateks e sõltuvateks tunnusteks (need, mille käitumine huvi pakub) ja argument- e sõltumatuteks tunnusteks e faktoriteks (need, mille mõju uuritavatele tunnustele soovitakse selgitada).

Definitsiooni kohaselt on **üldine lineaarne mudel** (*general linear model*, GLM) uurija poolt eeldatav faktorite ja uuritavate tunnuste vahekorra üldskeem (mudel), mille detailid tuleb hinnata algandmetest (valimist).

Üldise lineaarse mudeli alusel tehtud järeldused on õiged kui

a) uuritav mudel vastab tegelikkusele – on õigesti paika pandud faktorite olemus ja vahekord;
b) kehtivad vajalikud matemaatilised eeldused (normaaljaotus, vaatluste sõltumatus jm).

Tänu oma lihtsusele ja tõsiasjale, et faktorite mõjude lineaarne kombinatsioon vastab ligilähedaselt tegelikkusele paljudes reaalsetes analüüsides ning sobib lähendama lineaarsest märksa keerulisemaid funktsioone, on üldine lineaarne mudel tänapäeva eluteadustes enim rakendatav matemaatilise statistika meetod.

Iga üldise lineaarse mudeli püstitamine eeldab kolme asja:

a) mudeli esitust faktorite mõjude summana,
b) uuritava tunnuse ning faktorite keskväärtuste ja dispersioonimaatriksite struktuuri fikseerimist vastavalt andmestiku ülesehitusele ja faktorite olemusele, ning
c) tõenäosusjaotuslike eelduste tegemist mudeli liikmete kohta võimaldamaks määrata hinnangute täpsust ja kontrollida hüpoteese.

2 ÜLDISTE LINEAARSETE MUDELITE JA FAKTORITE LIIGITUS

2.1 Diskreetsed ja pidevad faktorid

Faktortunnuse erinevaid väärtusi nimetatakse **tasemeteks** e **nivoodeks** (inglise keeles *levels*). Iga faktor jaguneb vastavalt oma tasemete iseloomule **diskreetseks** või **pidevaks**, arvuliseks (kvantitatiivne) või klassifitseerivaks (kvalitatiivne).

> **N**äiteks on lehma sünniaasta, laktatsioon jne diskreetsed arvulised faktorid; farm tasemetega (väärtustega) 'Vorbuse', 'Ülenurme' jne on diskreetne klassifitseeriv faktor; laktatsiooni pikkus, piimatoodang, pekipaksus jne (mõõdetud tunnused) on aga pidevad faktorid.

Üldine lineaarne mudel on **dispersioonanalüüsi**, **regressioonanalüüsi** või **kovariatsioonanalüüsi** mudel vastavalt sellele, kas faktorid on diskreetsed, pidevad või esineb mõlemaid.

2.2 Faktorite vahekord mudelis

Faktorid on lihtsad ja tuletatud. Lihtsate faktorite väärtused on vahetult mõõdetud või registreeritud, tuletatud faktorid moodustatakse lihtsatest. Tüüpilised tuletatud faktorid on **interaktsioonid** e. **koosmõjud** ning arvuliste faktorite korrutised.

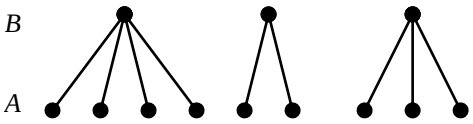
> **N**äiteks on farm, isa ja laktatsiooni pikkus lihtsad faktorid; farm*isa (koosmõju) ja laktatsioon*laktatsioon (arvulise faktori kõrgem järk) aga tuletatud faktorid.

Praktikas on faktorite vahel sageli ka alluvusseosed. Faktor *A* **allub** faktorile *B* – tähistatakse enamasti *A*(*B*) –, kui *A* iga nivoo (tase) esineb koos vaid ühe *B* nivooga (joonis 1).

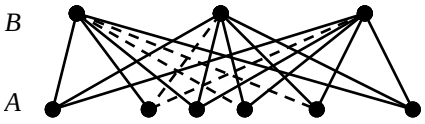
> **N**äiteks võime me tavaliselt lugeda farmi allutatuks maakonnale; kui iga ema on ristatud kindla isaga, on ema allutatud isale.

Faktorid *A* ja *B* on **ristseoses**, kui *A* iga nivoo kombineerub (saab põhimõtteliselt kombineeruda) *B* kõigi nivoodega (joonis 2).

> **N**äiteks kui viiel aastal on uuritud pullide tütarde jõudlusandmeid ja igal pullil on igal aastal tütreid, on pull ja aasta ristseoses; kui aga igal aastal on valitud uued pullid, allub pull aastale.



Joonis 1. Alluvusseoses faktorid



Joonis 2. Ristseoses faktorid

2.3 Juhuslikud ja fikseeritud faktorid

Sõltuvalt uurija eesmärkidest ja analüüsitavatest andmetest omistatakse igale faktorile kindel tüüp – **juhuslik** või **fikseeritud**.

Fikseeritud faktoril on:

a) vähe nivoosid,
b) iga nivoo pakub iseseisvat huvi ja on valitud mittejuhuslikult,
c) andmetes on või saavad põhimõtteliselt olla esindatud kõik nivood.

Seetõttu käsitletakse fikseeritud faktorite tasemetele vastavaid efekte kui konstante.

Juhuslikul faktoril on

a) potentsiaalselt väga palju (lõpmatu hulk) nivoosid,
b) andmetes on neist esindatud juhuslik valim,
c) huvi pakub kõigi (ka andmetes esindamata) nivoode keskmine mõju e see, kui suur osa uuritava tunnuse koguvarieeruvusest on kirjeldatud antud faktori poolt.

Juhuslike faktorite tasemetele vastavaid efekte käsitletakse kui mingi teoreetilise jaotusega juhuslike suuruste realiseerunud väärtusi, kus selle teoreetilise jaotuse näol mõistetakse üldjuhul normaaljaotust.

Näiteks, kui andmestik sisaldab kümne spetsiaalselt valitud isaslooma järglaste andmeid, loetakse isa mõju (faktor 'isa') fikseerituks, kuna me soovime võrrelda just antud isasid ega soovi teha järeldusi teiste isade kohta. Kui aga huvi pakub isa kui geneetilise faktori üldine mõju (isade poolt järglastele pärandatavate juhuslike geenidekomplektide mõju) ja järglased on valitud juhuslikult, siis loetakse isa mõju juhuslikuks.

See, kas faktor on juhuslik või fikseeritud, ei kajastu mudeli võrrandis, küll aga sõltuvad sellest mudeli liikmete keskväärtused ja dispersioonid ning seeläbi ka mõjude hinnangud ja nende statistiline olulisus.

Vastavalt sellele, kas üldine lineaarne mudel sisaldab üksnes fikseeritud faktoreid, üksnes juhuslikke faktoreid või siis mõlemaid, nimetatakse kasutatavat mudelit kas **fikseeritud mudeliks** (*fixed model*), **juhuslikuks mudeliks** (*random model*) või **segamudeliks** (*mixed model*). Aretusväärtuste ja teiste geneetiliste parameetrite hindamisel rakendatakse just viimast.

2.4 Tasakaalus ja mittetasakaalus andmed (mudelid)

Tasakaalus (*balanced*) andmed on sellised, kus kõigil faktorite ja nende kombinatsioonide tasemetel on sooritatud võrdne arv mõõtmisi/vaatlusi/vmt.

Mittetasakaalus (*unbalanced*) andmed on sellised, kus vastupidiselt tasakaalulisele juhule ei vasta kõigile faktorite tasemetele võrdne arv objekte. Andmete mittetasakaalulisus eeldab vähegi keerulisema struktuuriga mudelite puhul ligikaudsete arvutusmeetodite kasutamist, mistõttu tulemused ei pruugi olla enam ühesed.

Näide 1. Järgnevalt toodud andmetabeleist esimene kujutab tasakaalus ja teine mittetasakaalus andmestikku. Esimeses on mõlemast majandist ühepalju eesti holsteini ja eesti punast tõugu lehmi, lisaks on ühepalju kõikvõimalikke majandite, tõugude ja laktatsioonide kombinatsioone. Teise andmestiku korral pole enamus nimetatud tingimustest täidetud.

Tasakaalus andmed

Farm	Tõug	Laktatsioon
Põlula	EHF	1
Põlula	EHF	2
Põlula	EHF	3
Põlula	EPK	1
Põlula	EPK	2
Põlula	EPK	3
Põlva	EHF	1
Põlva	EHF	2
Põlva	EHF	3
Põlva	EPK	1
Põlva	EPK	2
Põlva	EPK	3

Mittetasakaalus andmed

Farm	Tõug	Laktatsioon
Põlula	EHF	1
Põlula	EHF	2
Põlula	EPK	2
Põlula	EPK	1
Põlula	EPK	2
Põlva	EHF	1
Põlva	EHF	2
Põlva	EHF	2
Põlva	EHF	3
Põlva	EPK	1
Põlva	EPK	3

3 ÜLDISE LINEAARSE MUDELI ESITUSED

3.1 Mudeli esitus objektiviisi

Mudeli objektiviisi esitamisel avaldatakse igal objektil mõõdetud väärtus üldkeskmise (nn vabaliikme, inglise keeles *intercept*), objektile vastavate mõjude ja konkreetse objekti omapära e faktoritega kirjeldamata jääva osa (mudeli vea) summana.

Erinevate objektide ja faktorite tasemete üheseks identifitseerimiseks kasutatakse indeksite abi: igale diskreetsele faktorile omistatakse põhiindeks (harilikult *i*, *j*, *k*, *l*, …), mis näitab iga objekti kohta just temale omase faktori taseme mõju, peale selle tähistab üks lisaindeks objekti (mõõtmist) ennast. Faktorite koosmõjude korral kirjutatakse indeksiks mõlema faktori indeksid, allutatud faktorite puhul tuuakse tavaliselt faktori enese põhiindeksi järel sulgudes ära kõigi nende faktorite indeksid, millele antud faktor allub. Pideva faktori tähis omab kogu indeksite komplekti, sest märgib konkreetsel objektil ja diskreetsete faktorite kombinatsioonil sooritatud mõõtmise tulemust, mille mõju samal objektil ja diskreetsete faktorite kombinatsioonil mõõdetud uuritava tunnuse väärtusega väljendab vastav regressioonikordaja.

Näide 2. Uuritavaks tunnuseks on tallede võõrutusmass ja analüüsi eesmärgiks on võrrelda ühes lambafarmis kasutatud kahte isa. Objektiivsemate tulemuste huvides soovitakse arvestada ka talle soo ja pesakonna suuruse kui diskreetsete faktorite ning võõrutusvanuse kui pideva faktori mõjudega. Andmed on kirjas järgnevas tabelis.

Sugu	Pesakonna suurus	Võõrutus-mass	Võõrutus-vanus	Isa kood
1	3	22	91	1025
1	1	39	90	1025
1	2	24	86	1025
2	2	27	101	1025
2	2	24	86	1025
1	2	31	117	1027
1	2	30	152	1027
2	1	31	110	1027
2	2	26	107	1027

Üldise lineaarse mudeli võib toodud ülesande püstituse korral esitada järgmiselt:

$$y_{ijkl} = \mu + I_i + S_j + P_k + b \times v_{ijkl} + e_{ijkl}, \quad (1)$$

kus parameetrit μ võib mõista kui nö keskmist tallede 100 päeva kehamassi (mis siiski ei pruugi võrduda aritmeetilise keskmisega, sest on tegelikult faktorite keskmine mõju);

y_{ijkl} tähistab objektil *l* sooritatud mõõtmist e antud juhul *l*. talle 100 päeva kehamassi (ülejäänud indeksid fikseerivad täpselt ära vaatlusaluse talle isa, sünniaasta jne), $l = 1,...,9$;

I_i tähistab isa *i* mõju, $i = 1025, 1027$;

S_j märgib soo *j* mõju, $j = 1, 2$;

P_k on pesakonna suuruse *k* mõju, $k = 1, 2, 3$;

v_{ijkl} tähistab konkreetse talle võõrutusvanust ja *b* on võõrutusmassi lineaarset sõltuvust võõrutusvanusest iseloomustav regressioonikordaja;

e_{ijkl} on talle *l* omapära (e juhuslik viga).

3.2 Üldine lineaarne mudel maatrikskujul

Üldiste lineaarsete mudelite maatrikskujul esitamiseks kasutatakse eriliste maatriksite abi, mis seostavad iga objekti just temal sooritatud mõõtmistega või temale vastava faktori tasemega. Selliseid maatrikseid, mille ridade arv võrdub uuritud objektide arvuga ja veergude arv mudelist hinnatavate efektide arvuga, nimetatakse **plaani**- e **disainimaatriksiteks**.

Regressioonivõrrand, kirja panduna objekti *i* kohta,

$$y_i = \mu + bx_i + e_i ,$$

näeb maatrikskujul esitatuna välja järgmiselt:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e} , \qquad (2)$$

kus $\mathbf{y} = (y_1\ y_2\ \dots\ y_n)^{\mathrm{T}}$ ja $\mathbf{e} = (e_1\ e_2\ \dots\ e_n)^{\mathrm{T}}$ on vastavalt funktsioontunnuse väärtuste ja prognoosivigade vektorid, $\boldsymbol{\beta} = (\mu\ b)^{\mathrm{T}}$ on hinnatavate parameetrite vektor ja plaanimaatriks $\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}^{\mathrm{T}}$ sisaldab ühte veergu kummagi hinnatava parameetri tarvis – esimene, ühtedest koosnev veerg tähendab, et kõigil objektidel mõõdetud uuritava tunnuse väärtusi püütakse prognoosida sama keskmise μ abil, teine, argumenttunnuse väärtusi sisaldav veerg seob igal objektil mõõdetud uuritava tunnuse väärtuse läbi kordaja *b* just temale vastava argumenttunnuse väärtusega.

Dispersioonanalüüsi korral koosneb plaanimaatriks vaid nullidest ja ühtedest – mingile objektile vastavas reas ja faktori tasemele vastavas veerus on 1, kui mõõtmine antud objektil on sooritatud just sellel tasemel, ja 0 vastupidisel juhul. Näiteks ühefaktorilise dispersioonanalüüsi mudel

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

on maatriksite abil esitatav kujul

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e} \quad \text{ehk} \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ 1 & \mathbf{0} & \mathbf{1}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{n_q} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_q \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} .$$

Siin n_i tähistab faktori *i*-ndale tasemele vastavate objektide arvu, *q* on faktori tasemete arv ($i = 1,...,q$), $\mathbf{1}_{n_i}$ on ühtedest koosnev $n_i \times 1$-vektor: $\mathbf{1}_{n_i} = (\underbrace{1\ 1\ \dots\ 1}_{n_i \text{ tk}})^{\mathrm{T}}$, ja $\mathbf{0}$ on sobiva dimensiooniga nullidest koosnev vektor.

Kovariatsioonanalüüsi mudel, milles on nii mittearvulisi kui ka arvulisi faktortunnuseid, sisaldab plaanimaatriksis nii nulle ja ühtesid sisaldavaid veerge kui ka pidevate argumenttunnuste väärtuseid.

Kokkuvõttes kujutab üldise lineaarse mudeli maatriksesitus enesest tavalise lineaarvõrranditesüsteemi maatriksesitust, kus kordajate maatriksi rollis on plaanimaatriks, tundmatud parameetrid on koondatud vektorisse $\boldsymbol{\beta}$ ja vabaliikmete veerg moodustub vahest $\mathbf{y} - \mathbf{e}$.

Näide 3. Paneme näite 2 mudeli (1) kirja ka maatrikskujul (2).

Vektor **y** on tallede võõrutusmasside vektor kujul

$$\mathbf{y} = (22\ 39\ 24\ 27\ 24\ 31\ 30\ 31\ 26)^{\mathrm{T}} .$$

Plaanimaatriks **X** sisaldab ühte rida iga vaatluse ja ühte veergu iga vektoris $\boldsymbol{\beta}$ kirjas oleva mudeli parameetri kohta.

Hinnatavate parameetrite vektor $\boldsymbol{\beta}$ sisaldab 9 elementi (parameetrite tähistused on üle võetud sama mudeli objektiviisi esitusest):

$$\boldsymbol{\beta} = (\mu\ I_{1025}\ I_{1027}\ S_1\ S_2\ P_1\ P_2\ P_3\ b)^{\mathrm{T}}$$

Plaanimaatriks **X**, mis seob iga vaatluse just temale vastava faktori tasemega või väärtusega, näeb praeguse tallede järjestuse korral välja järgmine:

$$\mathbf{X} = \left(\begin{array}{c|cc|cc|ccc|c} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 91 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 90 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 86 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 101 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 86 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 117 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 152 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 110 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 107 \end{array}\right) \tag{3}$$

(parema jälgitavuse huvides on erinevatele faktoritele vastavad blokid eraldatud punktiirjoontega).

Juhuslike vigade vektor on kujul

$$\mathbf{e} = (e_1 \;\; e_2 \;\; e_3 \;\; e_4 \;\; e_5 \;\; e_6 \;\; e_7 \;\; e_8 \;\; e_9)^{\mathrm{T}} .$$

Näitamaks, et eelnevalt defineeritud plaanimaatriks tõepoolest iga talle võõrutusmassi just temale vastavate faktorite tasemetega seob, asendame kõik väljakirjutatud vektorid ja maatriksid mudelisse (2) ning sooritame korrutustehte **Xβ**. Tulemuseks saame maatriksvõrduse

$$\begin{pmatrix} 22 \\ 39 \\ 24 \\ 27 \\ 24 \\ 31 \\ 30 \\ 31 \\ 26 \end{pmatrix} = \overbrace{\begin{pmatrix} \mu + I_{1025} + S_1 + P_3 + b \times 91 \\ \mu + I_{1025} + S_1 + P_1 + b \times 90 \\ \mu + I_{1025} + S_1 + P_2 + b \times 86 \\ \mu + I_{1025} + S_2 + P_2 + b \times 101 \\ \mu + I_{1025} + S_2 + P_2 + b \times 86 \\ \mu + I_{1027} + S_1 + P_2 + b \times 117 \\ \mu + I_{1027} + S_1 + P_2 + b \times 152 \\ \mu + I_{1027} + S_2 + P_1 + b \times 110 \\ \mu + I_{1027} + S_2 + P_2 + b \times 107 \end{pmatrix}}^{\mathbf{X\beta}} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \end{pmatrix} .$$

3.3 Üldine lineaarne segamudel maatrikskujul

Juhul, kui üldine lineaarne mudel sisaldab nii fikseeritud kui ka juhuslikke faktoreid (tegu on segamudeliga), esitatakse need eraldi liidetavatena, jagades kaheks nii hinnatavate parameetrite vektori kui ka plaanimaatriksi:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e} , \tag{4}$$

siin **β** ja **u** tähistavad vastavalt fikseeritud ja juhuslike efektide vektoreid ning **X** ja **Z** vastavaid plaanimaatrikseid.

Näide 4. Oletame nüüd, et näites 2 vaadeldud jäärade alusel soovitakse iseloomustada isa kui geneetilise faktori üldist mõju ning jäärad on valimisse sattunud juhuslikult. Sellisel juhul tuleb isa mõjusid käsitleda juhuslike efektidena ning mudeli võib välja kirjutada valemi (4) alusel:

$$\underbrace{\begin{pmatrix} 22 \\ 39 \\ 24 \\ 27 \\ 24 \\ 31 \\ 30 \\ 31 \\ 26 \end{pmatrix}}_{\mathbf{y}} = \underbrace{\left(\begin{array}{c|cc|cc|ccc|c} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 91 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 90 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 86 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 101 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 86 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 117 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 152 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 110 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 107 \end{array}\right)}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ S_1 \\ S_2 \\ P_1 \\ P_2 \\ P_3 \\ b \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}}_{\mathbf{Z}} \underbrace{\begin{pmatrix} I_{1025} \\ I_{1027} \end{pmatrix}}_{\mathbf{u}} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \end{pmatrix}}_{\mathbf{e}} . \tag{5}$$

4 KESKVÄÄRTUSED JA DISPERSIOONID

Üldise lineaarse mudeli esituse teine osa seisneb mudeli juhuslike liikmete keskväärtuste ja dispersioonimaatriksite defineerimises (et fikseeritud mõjusid käsitletakse konstantsetena, ei ole nende keskväärtuste ja dispersioonide leidmisel mõtet).

4.1 Vaatluste keskväärtused ja kovariatsioonistruktuur fikseeritud mudeli korral

Igas lineaarses mudelis on vähemalt kaks juhuslikku liiget (mis eeldatakse traditsiooniliselt käituvat vastavalt normaaljaotuse seaduspäradele) – uuritav tunnus, kui populatsioonist juhuslikult valitud indiviididel sooritatud mõõtmiste tulemusi koondav suurus, ja juhuslik viga. Juhul, kui mudel sisaldab faktoritena üksnes fikseeritud mõjusid, on need kaks ka mudeli ainsad mittekonstantsed ja seega mittenullilist dispersiooni omavad suurused.

On loomulik eeldada, et juhuslike vigade keskväärtus võrdub nulliga (mudeli keskmine viga on null): $\mathrm{E}(\mathbf{e}) = \mathbf{0}$. Uuritava tunnuse keskväärtus avaldub fikseeritud mudeli (2) korral kujul

$$\mathrm{E}(\mathbf{y}) = \mathrm{E}(\mathbf{X\beta} + \mathbf{e}) = \mathrm{E}(\mathbf{X\beta}) + \mathrm{E}(\mathbf{e}) = \mathbf{X\beta}$$

ja dispersioonimaatriks võrdusena

$$\mathrm{var}(\mathbf{y}) = \mathrm{var}(\mathbf{X\beta} + \mathbf{e}) = \underbrace{\mathrm{var}(\mathbf{X\beta})}_{0} + \mathrm{var}(\mathbf{e}) = \mathrm{var}(\mathbf{e}) \tag{6}$$

4.2 Vaatluste keskväärtused ja kovariatsioonistruktuur segamudeli korral

Segamudeli (4) korral tuleb arvesse võtta ka juhuslike efektide keskväärtust ja varieeruvust. Analoogselt juhuslike vigadega eeldatakse ka juhuslike faktorite mõjude keskväärtuste nulliga võrdumist: $\mathrm{E}(\mathbf{u}) = \mathbf{0}$. Seetõttu jääb uuritava tunnuse keskväärtus endiselt määratuks üksnes fikseeritud mõjude poolt:

$$\mathrm{E}(\mathbf{y}) = \mathrm{E}(\mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}) = \mathrm{E}(\mathbf{X\beta}) + \mathrm{E}(\mathbf{Zu}) + \mathrm{E}(\mathbf{e}) = \mathbf{X\beta}\ .$$

Arvestades dispersioonide defineerimisel ka juhuslike mõjude varieeruvusega, saame, et

$$\mathrm{var}(\mathbf{y}) = \mathrm{var}(\mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}) = 0 + \mathrm{var}(\mathbf{Zu}) + \mathrm{var}(\mathbf{e}) - 2\underbrace{\mathrm{cov}(\mathbf{Zu}, \mathbf{e}^{\mathrm{T}})}_{0} = \mathbf{Z}\,\mathrm{var}(\mathbf{u})\mathbf{Z}^{\mathrm{T}} + \mathrm{var}(\mathbf{e})\ .$$

Eelnevas võrduste reas on eeldatud, nagu üldiste lineaarsete mudelite teoorias tavaks, et faktorite mõjud ja juhuslikud vead on sõltumatud: $\mathrm{cov}(\mathbf{u}, \mathbf{e}^{\mathrm{T}}) = \mathbf{0}$.

Tuues sisse tähistused $\mathbf{V} = \mathrm{var}(\mathbf{y})$, $\mathbf{G} = \mathrm{var}(\mathbf{u})$ ja $\mathbf{R} = \mathrm{var}(\mathbf{e})$, saame segamudeli (4) keskväärtused ja kovariatsioonistruktuuri määrata maatriksitena

$$\mathrm{E}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{X\beta}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} \tag{7}$$

ja

$$\mathrm{var}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{ZGZ}^{\mathrm{T}} + \mathbf{R} & \mathbf{GZ}^{\mathrm{T}} & \mathbf{R}\\ \mathbf{ZG} & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix}. \tag{8}$$

Traditsioonilises segamudelis eeldatakse, et nii kõik uuritavad objektid objektid kui ka kõik juhuslikud efektid on omavahel sõltumatud, mistõttu on nii jääkdispersioonimaatriks **R** kui ka juhuslike efektide dispersioonimaatriks **G** diagonaalmaatriksid:

$$\mathbf{R} = \mathbf{I}_n\sigma_e^2\,, \tag{9}$$

$$\mathbf{G} = \mathbf{I}_a\sigma_u^2\,, \tag{10}$$

siin *n* tähistab uuritavate objektide arvu ja *a* juhusliku faktori tasemete arvu (enam kui ühe juhusliku faktori korral on maatriks **G** blokkdiagonaalne – igale faktorile vastab tema tasemete arvule vastava dimensiooniga diagonaalmaatriks).

Dispersioone σ_e^2 ja σ_u^2 valemeis (9) ja (10) nimetatakse **dispersioonikomponentideks** (*variance components*), sest nad kujutavad enesest vaatluste dispersiooni komponente: $\mathrm{var}(y_i) = \sigma_u^2 + \sigma_e^2$.

Näide 5. Paneme kirja mudeli (5) kovariatsioonimaatriksid (isade mõjusid käsitleme juhuslike efektidena).

Vastavalt valemitele (9) ja (10) saame, et

$$\operatorname{var}(\mathbf{e}) = \mathbf{R} = \begin{pmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 \end{pmatrix}, \quad \operatorname{var}(\mathbf{u}) = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}. \tag{11}$$

Valemist (8) järeldub, et $\operatorname{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}^{\mathrm{T}} + \mathbf{R}$, seega

$$\mathbf{V} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_u^2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_e^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \sigma_u^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ 0 & 0 & 0 & 0 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ 0 & 0 & 0 & 0 & \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 \\ 0 & 0 & 0 & 0 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \sigma_u^2 \\ 0 & 0 & 0 & 0 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_e^2 + \sigma_u^2 \end{pmatrix}.$$

Viimasest maatriksist ilmneb ka, et vaatluste dispersioonid esituvad summana $\operatorname{var}(y_i) = \sigma_u^2 + \sigma_e^2$ ja kovariatsioonid kujul

$$\operatorname{cov}(y_i, y_j) = \begin{cases} \sigma_u^2, \text{ kui } i \neq j \text{ ja loomad } i \text{ ja } j \text{ on sama isa järglased,} \\ 0, \text{ kui } i \neq j \text{ ja loomad } i \text{ ja } j \text{ on erinevate isade järglased.} \end{cases}$$

Märkus: käsitledes isade mõjusid fikseeritutena, avaldub vaatluste dispersioonimaatriks seosena (6):

$$\operatorname{var}(\mathbf{y}) = \mathbf{V} = \begin{pmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 \end{pmatrix} = \mathbf{R} = \operatorname{var}(\mathbf{e}),$$

millest omakorda järeldub, et $\operatorname{var}(y_i) = \sigma_e^2$ ja $\operatorname{cov}(y_i, y_j) = 0$, kui $i \neq j$.

5 EELDUSED JA KITSENDUSED

5.1 Jaotused

Üldiste lineaarsete mudelite analüüsil eeldatakse uuritava(te) tunnus(t)e jaotumist vastavalt normaaljaotuse seaduspäradele (pt 2.3.1). Matemaatiliselt väljendudes tähendab see, et vektor **y** on normaaljaotusega keskväärtusega **Xβ** ja dispersioonimaatriksiga **V**:

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

Taolise eelduse puhul on garanteeritud, et vähimruutude printsiibil (pt 2.4.2) leitud efektide hinnangud on tõepoolest väikseima varieeruvusega, osutub võimalikuks faktorite mõjude täpsuse ja statistilise olulisuse kontrollimine (*F*-test dispersioon- ja regressioonanalüüsil) ning suurima tõepära meetodi rakendamine dispersioonikomponentide ja seeläbi ka päritavuse ja geneetiliste korrelatsioonide hindamiseks.

5.2 Kovariatsioonistruktuur

Traditsioonilisel üldiste lineaarsete mudelite analüüsil (nagu näiteks ka tavalise regressioon- ja dispersioonanalüüsi korral) eeldatakse dispersioonimaatriksite diagonaalset kuju. Segamudelite tarvis on selline lihtne kovariatsioonistruktuur defineeritud seostega (9) ja (10). Et aga tegelikkus (uuritava(te) tunnus(t)e varieeruvus) sageli nii triviaalselt kirjeldatav pole, leiab reaalsete andmete analüüsil kasutust suur hulk rohkem või vähem keerukaid dispersioonistruktuure, mis on faktorite mõjude ja juhuslike vigade sõltumatuse eeldusel kokkuvõtvalt esitatavad dispersioonimaatriksina (8).

Fikseeritud mudelite puhul on levinuimad kõiksugu korduvate mõõtmiste analüüsi mudelid, kus püütakse arvesse võtta ühel ja samal objektil sooritatud mõõtmistele vastavate mudeli jääkide korrelleeritust.

Näide 6. Olgu meil uuritavaiks objektideks lehmad, kellel kõigil on fikseeritud esimese kolme laktatsiooni piimatoodangud (väljavõte andmetabelist on toodud teksti paremas servas).

ID	Lakt.	Piim, kg
3396	1	4119
3396	2	5857
3396	3	6260
3990	1	3106
3990	2	3934
3990	3	5171
4390	1	2473
4390	2	3301
4390	3	2958
…		

Ükskõik, milliste faktortunnuste mõju me ka uurida tahame, on loomulik eeldada, et sama lehma piimatoodangud ei ole sõltumatud. Mudelisse saab viimase kirja panna jääkdispersioonimaatriksi **R** struktuuri abil, kus endiselt vastab kõigile vaatlustele juhuslike vigade dispersioon σ_e^2, aga lisaks võib defineerida (ja lasta arvutil ka andmetest hinnata) samale loomale vastavate mudeli jääkliikmete vahelist varieeruvust peegeldava dispersioonikomponendi σ^2, misjärel jääkdispersioonimaatriks esitub kompaundsümmeetrilisel kujul

$$\mathbf{R} = \begin{pmatrix} \sigma_e^2+\sigma^2 & \sigma^2 & \sigma^2 & 0 & 0 & 0 & \cdots \\ \sigma^2 & \sigma_e^2+\sigma^2 & \sigma^2 & 0 & 0 & 0 & \cdots \\ \sigma^2 & \sigma^2 & \sigma_e^2+\sigma^2 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \sigma_e^2+\sigma^2 & \sigma^2 & \sigma^2 & \cdots \\ 0 & 0 & 0 & \sigma^2 & \sigma_e^2+\sigma^2 & \sigma^2 & \cdots \\ 0 & 0 & 0 & \sigma^2 & \sigma^2 & \sigma_e^2+\sigma^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} .$$

Sammu võrra keerulisema dispersioonistruktuuri saame, kui eeldame, et mida suurem on mõõtmiste ajaline vahe, seda nõrgem on nende vaheline seos, ehk antud näite puhul – 1. ja 2. laktatsiooni toodangud on omavahel tugevamini seotud, kui 1. ja 3. laktatsiooni toodangud. Üks võimalus seda mudelisse kirja panna on kasutada jääkdispersioonimaatriksi defineerimisel esimest järku autoregressiivset struktuuri kujul

$$\mathbf{R} = \sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & 0 & 0 & 0 & \cdots \\ \rho & 1 & \rho & 0 & 0 & 0 & \cdots \\ \rho^2 & \rho & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \rho & \rho^2 & \cdots \\ 0 & 0 & 0 & \rho & 1 & \rho & \cdots \\ 0 & 0 & 0 & \rho^2 & \rho & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

kus parameetrit ρ võib mõista kui ühe ja sama looma järjestikuste vaatluste vahelist autokorrelatsioonikordajat, $|\rho| < 1$.

Segamudelite puhul tuleb arvestada ka juhuslike faktorite mõjude võimaliku korreleeritusega, mis tähendab dispersioonimaatriksi **G** struktuuri erinevust traditsioonilisest diagonaalsest kujust (10). Loomade aretuses väljendub see enamasti kõikvõimalike sugulussidemete matemaatilisel kujul väljendamises ja mudelisse kaasamises – taolisel kovariatsioonistruktuuri defineerimisel põhineb näiteks aretusväärtuste prognoosimine looma mudelist .

5.3 Täisastakuga ja mitte täisastakuga mudelid, reparametriseerimistingimused

Üldine lineaarne mudel on **täisastakuga**, kui täisastakuga (kõik read-veerud lineaarselt sõltumatud, pööratav) on tema plaanimaatriksi transponeeritud maatriksi ja plaanimaatriksi enese korrutismaatriks, näiteks $\mathbf{X}^T\mathbf{X}$. Nagu järgmisest peatükist nähtub, kujutab just taoline korrutismaatriks enesest mudeli parameetrite hindamiseks konstrueeritud võrrandi kordajatemaatriksit.

Mudeli parameetrid on **üheselt hinnatavad** üksnes täisastakuga mudeli korral, sest vaid siis leidub kasutataval kordajatemaatriksil ühene pöördmaatriks; mitte täisastakuga mudeli korral on erinevaid parameetrite hinnanguid palju (teoreetiliselt lõpmatu arv).

Tavalise fikseeritud mudeli (2) korral piisab mudeli parameetrite üheseks hinnanguks ka plaanimaatriksi enese veergude lineaarsest sõltumatusest. Täisastakuga on näiteks regressioonanalüüsi mudel, sest plaanimaatriksi ühtedest koosnev vabaliikmele vastav veerg pole üldjuhul mingi lineaarkombinatsiooni tulemusena teisendatav argumenttunnuse väärtuste veeruks.

Näide 7. Püstitades näite 2 alusel lihtsa lineaarse regressioonimudeli prognoosimaks tallede võõrutusmassi võõrutusvanuse abil, saame regressioonivõrrandi maatrikskujul

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

kus

$$\mathbf{X} = \begin{pmatrix} 1 & 91 \\ 1 & 90 \\ 1 & 86 \\ 1 & 101 \\ 1 & 86 \\ 1 & 117 \\ 1 & 152 \\ 1 & 110 \\ 1 & 107 \end{pmatrix} \text{ ja } \boldsymbol{\beta} = \begin{pmatrix} \mu \\ b \end{pmatrix}.$$

Plaanimaatriksi **X** veeruastak on 2, sest vabaliikmele ja regressioonikordajale vastavad veerud on lineaarselt sõltumatud, ja seega on parameetervektor **β** üheselt hinnatav.

Täisastakuga pole aga dispersioonanalüüsi mudel, sest näiteks mistahes faktori kõigile tasemetele vastavate veergude summeerimine annab tulemuseks ühtedest koosneva vabaliikmele vastava veeru. Seega pole ka dispersioonanalüüsi korral mudeli parameetrite ühene hindamine võimalik. Lahenduseks on hinnata mudeli parameetrite funktsioone, mis on üheselt hinnatavad (nn **hinnatavad funktsioonid** – *estimable functions*, vt pt 6.6.2) või kasutada mudeli parameetrite **reparametriseerimist**. Viimane tähendab parameetritele mingi lisakitsenduse rakendamist, mille all võib mõista näiteks iga faktortunnuse viimase taseme mõju nulliga võrdsustamist ja teiste tasemete mõjude hindamist selle

suhtes (kasutab näiteks statistikapakett SAS ja kasutatakse ka antud õppevahendi järgnevais näidetes) või traditsioonilisema reparametriseerimistingimusena iga faktori mõjude summa nulliga võrdsustamist. Seejuures peab selliseid lisakitsendusi olema sama palju, kui on plaanimaatriksis lineaarselt sõltuvaid veerge.

Näide 8. Mudel (1) näites 2 ei ole täisastakuga, sest plaanimaatriksi

$$\mathbf{X} = \left(\begin{array}{c:cc:cc:ccc:c} 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 91 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 90 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 86 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 101 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 86 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 117 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 152 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 110 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 107 \end{array}\right)$$

veeruastak on 6, mis on väiksem maatriksi **X** veergude arvust (9) – veergude lineaarne sõltuvus ilmneb selles, et kõigi diskreetsete faktorite korral annavad neile vastavad plaanimaatriksi veerud summeeritult esimese, vabaliikmele vastava ühtedest koosneva veeru.

Faktorite mõjude üheseks hindamiseks on vajalik teatud kitsenduste püstitamine. Näiteks võime nõuda kõigi faktorite korral nende mõjude summa nulliga võrdumist ($I_{1025} + I_{1027} = 0$, $S_1 + S_2 = 0$, $P_1 + P_2 + P_3 = 0$) või siis iga faktori viimase efekti nulliga võrdumist ($I_{1027} = 0$, $S_2 = 0$, $P_3 = 0$).

6 FIKSEERITUD EFEKTIDE HINDAMINE (BLUE)

6.1 Vähimruutude hinnangud

Faktorite mõjud mudelist (2) hinnatakse vähimruutude meetodil, st et parameetervektori $\boldsymbol{\beta}$ hinnang $\hat{\boldsymbol{\beta}}$ valitakse selliselt, et mudeli vigade ruudud oleks minimaalsed. Maatrikskujul on vähimruutude tingimus mudeli (2) tarvis väljendatav seosena

$$\min_{\boldsymbol{\beta}}(\mathbf{e}^{\mathrm{T}}\mathbf{e}) = \min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) .$$

Hinnangu $\hat{\boldsymbol{\beta}}$ avaldamiseks tuleb mudeli (2) vigade ruutude summast

$$\mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

võtta tuletis $\boldsymbol{\beta}$ järgi ja võrdsustada tulemus nulliga. Diferentseerimise tagajärjel saame

$$\partial\mathbf{e}^{\mathrm{T}}\mathbf{e}/\partial\boldsymbol{\beta} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

ning viimase avaldise nulliga võrdsustamisest järeldub, et

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} , \tag{12}$$

millest

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} . \tag{13}$$

Võrrand (12) on tuntud kui **normaalvõrrand** (*normal equation*) ja hinnang (13) kujutab enesest parameetervektori **β vähimruutude hinnangut** (OLS, *ordinary least squares*), mis diagonaalse võrdsete dispersioonikomponentidega jääkdispersioonimaatriksi (9) korral on **parim lineaarne nihketa hinnang** (BLUE, *best linear unbiased estimator*). Sõna "parim" hinnangu nimetuses tähendab, et tegu on täpseima e väikseima dispersiooniga $\mathrm{var}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$ hinnanguga (efektiivse hinnanguga, pt 2.4.3) ning sõna "nihketa" märgib hinnangu nn keskmist õigsust (süstemaatilise vea puudumist).

Kui meil on aga tegu keerulisema kovariatsioonistruktuuriga, kus $\mathbf{V} = \mathbf{R} = \mathrm{var}(\mathbf{e})$ ei ole võrdsete dispersioonikomponentidega diagonaalmaatriks, siis pole ka valemiga (13) defineeritud hinnang enam BLUE parameetervektorile $\boldsymbol{\beta}$. Hädast aitab välja kaval teisendus. Nimelt, kui teame dispersioonimaatriksit $\mathbf{V}$, mis tänu dispersiooni definitsioonile on alati positiivselt määratud ja mistõttu leiduvad ka $\mathbf{V}^{-1}$ ja selline maatriks $\mathbf{V}^{1/2}$, et $\mathbf{V} = \mathbf{V}^{1/2}\mathbf{V}^{1/2}$, võime võrrandi (2) mõlemaid pooli vasakult korrutada maatriksiga $\mathbf{V}^{-1/2}$, saades tulemuseks mudeli

$$\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\mathbf{e} .$$

Tähistades $\mathbf{y}^{*} = \mathbf{V}^{-1/2}\mathbf{y}$, $\mathbf{X}^{*} = \mathbf{V}^{-1/2}\mathbf{X}$ ja $\mathbf{e}^{*} = \mathbf{V}^{-1/2}\mathbf{e}$ saame viimase mudeli esitada võrrandina

$$\mathbf{y}^{*} = \mathbf{X}^{*}\boldsymbol{\beta} + \mathbf{e}^{*} , \tag{14}$$

kusjuures

$$\mathbf{V}^{*} = \mathrm{var}(\mathbf{e}^{*}) = \mathrm{var}(\mathbf{V}^{-1/2}\mathbf{e}) = \mathbf{V}^{-1/2}\underbrace{\mathrm{var}(\mathbf{e})}_{\mathbf{V}}\mathbf{V}^{-1/2} = (\underbrace{\mathbf{V}^{-1/2}\mathbf{V}^{1/2}}_{\mathbf{I}})(\underbrace{\mathbf{V}^{-1/2}\mathbf{V}^{1/2}}_{\mathbf{I}}) = \mathbf{I} .$$

Et nüüd on meil dispersioonimaatriks $\mathbf{V}^{*}$ diagonaalsel kujul, nagu on vaja BLUE esitamiseks valemi (13) abil, rakendame viimast mudelile (14) ja saame parameetervektori hinnangu kujul

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{y}^{*} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} . \tag{15}$$

Viimast hinnangut nimetatakse **üldistatud vähimruutude hinnanguks** (GLS, *generalized least squares*).

Juhul, kui analüüsitav mudel ei ole täisastakuga, ei ole normaalvõrrandi kordajate maatriksid $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ ja $\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}$ (vastavalt tavalise ja üldistatud vähimruutude hinnangu puhul) pööratavad ja kasutada tuleb üldistatud pöördmaatrikseid. Vähimruutude hinnangud (13) ja (15)saavad siis üldisemad kujud

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{16}$$

ja

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} . \tag{17}$$

Näide 9. Vaatleme edasi tallede võõrutusmassi andmestikku ja hindame talle isa, soo, pesakonna suuruse ning võõrutusvanuse kui fikseeritud faktorite mõjud. Garanteerimaks

hinnangute ühesust, võrdsustame nulliga iga faktori viimase efekti ($I_{1027}=0$, $S_2=0$, $P_3=0$). Ülejäänud efektide hindamiseks paneme kirja täisastakuga (ilma nulliga võrdsustatud efektidele vastavate veergudeta) plaanimaatriksi $\mathbf{X}^*$:

$$\mathbf{X}^* = \left(\begin{array}{ccccc|c} 1 & 1 & 1 & 0 & 0 & 91 \\ 1 & 1 & 1 & 1 & 0 & 90 \\ 1 & 1 & 1 & 0 & 1 & 86 \\ 1 & 1 & 0 & 0 & 1 & 101 \\ 1 & 1 & 0 & 0 & 1 & 86 \\ 1 & 0 & 1 & 0 & 1 & 117 \\ 1 & 0 & 1 & 0 & 1 & 152 \\ 1 & 0 & 0 & 1 & 0 & 110 \\ 1 & 0 & 0 & 0 & 1 & 107 \end{array}\right).$$

Et meil vaatluste dispersioonimaatriks on diagonaalsel kujul, $\mathbf{V}=\mathbf{I}_9\sigma_\varepsilon^2$, ja plaanimaatriks $\mathbf{X}^*$ on täisveeruastakuga, saame faktorite mõjude vektori

$$\boldsymbol{\beta}^* = \left(\mu \;\; I_{1025} \;\; S_1 \;\; P_1 \;\; P_2 \;\; b\right)^{\mathrm{T}}$$

hindamiseks kasutada valemit (13):

$$\underbrace{\begin{pmatrix} \hat{\mu} \\ \hat{I}_{1025} \\ \hat{S}_1 \\ \hat{P}_1 \\ \hat{P}_2 \\ \hat{b} \end{pmatrix}}_{\hat{\mathbf{b}}^*} = \left[\underbrace{\left(\begin{array}{ccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 91 & 90 & 86 & 101 & 86 & 117 & 152 & 110 & 107 \end{array}\right)}_{\mathbf{X}^{*\mathrm{T}}} \underbrace{\left(\begin{array}{ccccc|c} 1 & 1 & 1 & 0 & 0 & 91 \\ 1 & 1 & 1 & 1 & 0 & 90 \\ 1 & 1 & 1 & 0 & 1 & 86 \\ 1 & 1 & 0 & 0 & 1 & 101 \\ 1 & 1 & 0 & 0 & 1 & 86 \\ 1 & 0 & 1 & 0 & 1 & 117 \\ 1 & 0 & 1 & 0 & 1 & 152 \\ 1 & 0 & 0 & 1 & 0 & 110 \\ 1 & 0 & 0 & 0 & 1 & 107 \end{array}\right)}_{\mathbf{X}^*}\right]^{-1}$$

$$\times \underbrace{\left(\begin{array}{ccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 91 & 90 & 86 & 101 & 86 & 117 & 152 & 110 & 107 \end{array}\right)}_{\mathbf{X}^{*\mathrm{T}}} \underbrace{\begin{pmatrix} 22 \\ 39 \\ 24 \\ 27 \\ 24 \\ 31 \\ 30 \\ 31 \\ 26 \end{pmatrix}}_{\mathbf{y}} = \begin{pmatrix} 11{,}38 \\ 1{,}10 \\ 3{,}30 \\ 14{,}59 \\ 6{,}03 \\ 0{,}07 \end{pmatrix}.$$

Peale kitsenduste lisamist saame mudeli (1) parameetrite hinnangud kujul

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu} \\ \hat{I}_{1025} \\ \hat{I}_{1027} \\ \hat{S}_1 \\ \hat{S}_2 \\ \hat{P}_1 \\ \hat{P}_2 \\ \hat{P}_3 \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 11{,}4 \\ 1{,}1 \\ 0 \\ 3{,}3 \\ 0 \\ 14{,}6 \\ 6{,}0 \\ 0 \\ 0{,}07 \end{pmatrix}. \qquad (18)$$

Võrreldes faktorite tasemete kaupa leitud aritmeetiliste keskmistega (vt tabelit näite lõpus) hakkab silma jäärade paremusjärjestuse muutus – kui järglaste keskmiste väärtuste alusel võinuks eelistada jäära numbriga 1027, siis peale teiste faktorite mõjude arvesse võtmist on ilmselge hoopis jäära nr 1025 paremus. Andmetabelist (pt 6.3.1) põhjust otsides hakkab eelkõige silma, et jäära nr 1027 järglased on võõrutatud hiljem ja saanud seega kauem aega kasvada (positiivset seost võõrutusvanuse ja võõrutusmassi vahel näitab ka regressioonikordaja positiivne väärtus, $\hat{b}=0{,}07$). Jäärade mõjude hindamine mudeli (1) abil võimaldab nende seostega arvestada.

Faktorite tasemed ja neile vastavad keskmised võõrutusmassid:

Isa		Sugu		Pesakonna suurus	
1025	27,2	1	29,2	1	35,0
1027	29,5	2	27,0	2	27,0
				3	22,0

6.2 Hinnatavad funktsioonid

Olgugi, et üldjuhul parameetervektor $\boldsymbol{\beta}$ ise ei ole üheselt hinnatav, on seda hulk tema elementide lineaarkombinatsioone $\mathbf{k}^{\mathrm{T}}\boldsymbol{\beta}$. Viimaseid nimetatakse **hinnatavateks funktsioonideks** (*estimable functions*). Hinnatavuse olulisemad omadused lineaarsete mudelite korral on, et

1. vabaliige μ ei ole üheselt hinnatav (va regressioonanalüüsi korral);
2. erinevused ühe faktori tasemete vahel on üheselt hinnatavad tingimusel, et mudelis puuduvad selle ja teiste faktorite vahelised interaktsioonid ning uuritava faktori tasemete mõjud ei kattu mõne teise faktori tasemete mõjudega;
3. sisulist tähendust omavad üksnes hinnatavate funktsioonide väärtused.

Näide 10. Et tallede võõrutusmassi mõjutavate faktorite tasemete hinnangute ei ole üheselt hinnatavad, saame hinnangutele $\hat{\boldsymbol{\beta}}_1$ (18) lisaks leida suure hulga erinevaid hinnanguid. Näiteks järgmised hinnangute vektorid on saadud kasutades erinevaid reparametriseerimistingimusi või rakendades erinevaid üldistatud pöördmaatrikseid valemis (16):

$$\hat{\boldsymbol{\beta}}_2 = \begin{pmatrix} \hat{\mu} \\ \hat{I}_{1025} \\ \hat{I}_{1027} \\ \hat{S}_1 \\ \hat{S}_2 \\ \hat{P}_1 \\ \hat{P}_2 \\ \hat{P}_3 \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 30,4 \\ 0 \\ -1,1 \\ 0 \\ -3,3 \\ 0 \\ -8,6 \\ -14,6 \\ 0,07 \end{pmatrix}, \quad \hat{\boldsymbol{\beta}}_3 = \begin{pmatrix} \hat{\mu} \\ \hat{I}_{1025} \\ \hat{I}_{1027} \\ \hat{S}_1 \\ \hat{S}_2 \\ \hat{P}_1 \\ \hat{P}_2 \\ \hat{P}_3 \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 20,45 \\ 0,55 \\ -0,55 \\ 1,65 \\ -1,65 \\ 7,71 \\ -0,84 \\ -6,87 \\ 0,07 \end{pmatrix}, \quad \hat{\boldsymbol{\beta}}_4 = \begin{pmatrix} \hat{\mu} \\ \hat{I}_{1025} \\ \hat{I}_{1027} \\ \hat{S}_1 \\ \hat{S}_2 \\ \hat{P}_1 \\ \hat{P}_2 \\ \hat{P}_3 \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 20,7 \\ 1,1 \\ 0 \\ 0 \\ -3,3 \\ 8,5 \\ 0 \\ -6,0 \\ 0,07 \end{pmatrix} \quad \text{jne.}$$

Hinnatavad funktsioonid on näiteks

$$\underbrace{\begin{pmatrix} 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}}_{\mathbf{k}_1^{\mathrm{T}}} \hat{\boldsymbol{\beta}} = \hat{S}_1 - \hat{S}_2 = 0,$$

mis hindab sugude vahelist erinevust ja on sama väärtusega mistahes parameetrite hinnangute vektori $\hat{\boldsymbol{\beta}}$ korral, või

$$\underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix}}_{\mathbf{k}_2^{\mathrm{T}}} \hat{\boldsymbol{\beta}} = \hat{P}_2 - \hat{P}_3 = -2,$$

mis hindab kahe- ja kolmetallelisest pesakonnast pärit tallede võõrutusmasside erinevust ja on jällegi ühene mistahes parameetrite hinnangute vektori $\hat{\boldsymbol{\beta}}$ korral.

6.3 Vähimruutkeskmised

Vähimruutkeskmised (VRK; inglise keeles *least square means*, LSM), mis esitatakse sageli teadusartiklites ja mille hindamisvõimalust pakuvad paljud statistikaprogrammid, on oma olemuselt samuti hinnatavad parameeterfunktsioonid. Nende laialdane kasutamine on tingitud sellest, et ühelt poolt on tegu tõepoolest lihtsalt mõistetavate hinnangutega uuritava tunnuse keskmistele väärtustele faktorite tasemete alusel moodustatud gruppides, teiselt poolt on aga vähimruutkeskmiste hindamise metoodika märksa täiuslikum võrreldes tavaliste aritmeetiliste keskmiste arvutamisega – VRK-d leitakse baseeruvana mudelil ja on korrigeeritud mudelis sisalduvate segavate faktorite mõjude suhtes ega ole tundlikud vaatluste arvu erinevusele võrreldavais gruppides (seetõttu erinevad vähimruutkeskmised üldjuhul tavalistest keskmistest).

Näide 11. Illustreerime vähimruutkeskmiste leidmist tallede võõrutusmassi näite varal. Üks variant mudeli (1) parameetrite hinnanguist on esitatud vektorina (18).

Kõikvõimalikud vähimruutkeskmised koos arvutusvalemitega (ja võrdluseks ka tavalised aritmeetilised keskmised) on esitatud järgnevas tabelis.

Efekt	$\overline{x}$	VRK
I_{1025}	27,2	$LSM(\hat{I}_{1025}) = \hat{\mu} + \hat{I}_{1025} + \frac{1}{2}(\hat{S}_1 + \hat{S}_2) + \frac{1}{3}(\hat{P}_1 + \hat{P}_2 + \hat{P}_3) + \hat{b}_1 \times \overline{v}$ $= 11,4 + 1,1 + \frac{1}{2}(3,3 + 0) + \frac{1}{3}(14,6 + 6,0 + 0) + 0,07 \times 104,4 = 28,14$
I_{1027}	29,5	$LSM(\hat{I}_{1027}) = \hat{\mu} + \hat{I}_{1027} + \frac{1}{2}(\hat{S}_1 + \hat{S}_2) + \frac{1}{3}(\hat{P}_1 + \hat{P}_2 + \hat{P}_3) + \hat{b}_1 \times \overline{ev} + \hat{b}_2 \times \overline{v}$ $= 11,4 + 0 + \frac{1}{2}(3,3 + 0) + \frac{1}{3}(14,6 + 6,0 + 0) + 0,07 \times 104,4 = 27,04$
S_1	29,2	$LSM(\hat{S}_1) = \hat{\mu} + \frac{1}{2}(\hat{I}_{1025} + \hat{I}_{1027}) + \hat{S}_1 + \frac{1}{3}(\hat{P}_1 + \hat{P}_2 + \hat{P}_3) + \hat{b}_1 \times \overline{v} = 29,24$
S_2	27,0	$LSM(\hat{S}_2) = \hat{\mu} + \frac{1}{2}(\hat{I}_{1025} + \hat{I}_{1027}) + \hat{S}_2 + \frac{1}{3}(\hat{P}_1 + \hat{P}_2 + \hat{P}_3) + \hat{b}_1 \times \overline{v} = 25,94$
P_1	35,0	$LSM(\hat{P}_1) = \hat{\mu} + \frac{1}{2}(\hat{I}_{1025} + \hat{I}_{1027}) + \frac{1}{2}(\hat{S}_1 + \hat{S}_2) + \hat{P}_1 + \hat{b}_1 \times \overline{v} = 35,30$
P_2	27,0	$LSM(\hat{P}_2) = \hat{\mu} + \frac{1}{2}(\hat{I}_{1025} + \hat{I}_{1027}) + \frac{1}{2}(\hat{S}_1 + \hat{S}_2) + \hat{P}_2 + \hat{b}_1 \times \overline{v} = 26,75$
P_3	22,0	$LSM(\hat{P}_3) = \hat{\mu} + \frac{1}{2}(\hat{I}_{1025} + \hat{I}_{1027}) + \frac{1}{2}(\hat{S}_1 + \hat{S}_2) + \hat{P}_3 + \hat{b}_1 \times \overline{v} = 20,72$

7 JUHUSLIKE FAKTORITE REALISEERUNUD VÄÄRTUSTE PROGNOOSIMINE (BLUP)

Et juhuslikud efektid pole konstandid, nagu fikseeritud efektid, vaid kujutavad enesest mingist jaotusest pärinevaid juhusliku suuruse (juhusliku faktori) realiseerunud väärtusi, tuleb nende hindamisel arvestada ka selle jaotuse dispersiooniga (ehk teisiti väljendatuna, juhuslike efektide poolt määratud osaga uuritava tunnuse koguvarieeruvusest). Mudeli (4) korral tähendab see dispersioonimaatriksite (8) teadmist.

Juhul, kui dispersioonimaatriks (8) on teada, avaldub juhuslike faktorite realiseerunud väärtuste **u** **parim lineaarne nihketa prognoos** (BLUP, *best linear unbiased prediction*) kujul

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{19}$$

kus $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$ on fikseeritud efektide üldine vähimruutude hinnang. Seejuures tehakse siin ingliskeelses kirjanduses selget vahet sõnadel hinnang (*estimate*) ja prognoos (*prediction*) – hinnatakse midagi fikseeritut, konstantset, prognoositakse aga midagi, mille ilmnemine sõltub juhusest.

Fikseeritud efektide hindamine ning juhuslike efektide prognoosimine valemite (17) ja (19) alusel nõuab vaatluste dispersioonimaatriksi **V** (8) pöördmaatriksi leidmist, milline operatsioon võib vähegi suurema andmestiku korral problemaatiliseks kujuneda. Üks 20. sajandi suurimaid segamudelite ja aretusteooria arendajaid C. Henderson pakkus 1950. aastal välja normaalvõrrandile (12) sarnase maatriksvõrduse, mis võimaldab korraga leida nii BLUE(**β**) kui ka BLUP(**u**) ja seda ilma vaatluste dispersioonimaatriksit pööramata. Tänapäeval tuntaksegi järgnevat maatriksvõrdust **Hendersoni segamudeli võrrandina** (või lihtsalt segamudeli võrrandina, inglise keeles *mixed model equation*, MME):

$$\begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \tag{20}$$

Traditsioonilise segamudeli korral, kus vealiikmele ja juhuslikele efektidele vastavad dispersioonimaatriksid **R** ja **G** esituvad kujul (9) ja (10), on võrrand (20) kirjutatav ka kujul

$$\begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{y} \end{pmatrix}, \tag{21}$$

millest

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathrm{BLUE}(\boldsymbol{\beta}) \\ \mathrm{BLUP}(\mathbf{u}) \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{y} \end{pmatrix}.$$

Kuna praktikas sageli dispersioonimaatrikseid võrrandis (20) või dispersioonikomponente võrrandis (21) teada pole, tuleb need enne efektide prognoosimist andmetest hinnata, misjärel tehakse segamudeli võrrandis asendused $\mathbf{G} = \hat{\mathbf{G}}$ ja $\mathbf{R} = \hat{\mathbf{R}}$ (ehk $\sigma_u^2 = \hat{\sigma}_u^2$ ja $\sigma_e^2 = \hat{\sigma}_e^2$). On tõestatud, et juhul, kui uuritav tunnus on sümmeetrilise jaotusega (näiteks normaaljaotusega), on selliselt saadud fikseeritud efektide hinnangud ning juhuslike efektide prognoosid nihketa hinnangud BLUE(**β**)-le ja BLUP(**u**)-le.

Segamudeli võrrandis on juhuslikele efektidele vastavad read/veerud lineaarselt sõltumatud ega vaja hinnangu ühesuseks erinevalt fikseeritud efektidest lisakitsenduste rakendamist (lisakitsendused võivad olla endiselt vajalikud fikseeritud efektide hindamisel).

Andmaks pisut ettekujutust, mis vahe on faktori mõjude hinnanguil sõltuvalt sellest, kas käsitleda faktorit fikseerituna või juhuslikuna, vaatleme lihtsaimat dispersioonanalüüsi mudelit, mis sisaldab vaid vabaliiget μ ja üht diskreetset faktorit u:

$$y_{ij} = \mu + u_i + e_{ij}. \tag{22}$$

Käsitledes faktorit u fikseerituna, võime parima lineaarse nihketa hinnangu [valem (13)] faktori u tasemele u_i kirjutada kujul

$$\hat{u}_i = \bar{y}_{i\cdot} - \mu, \tag{23}$$

kus $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ ja n_i tähistab vaatluste arvu faktori i. tasemel – seega on vaid üht faktorit sisaldava dispersioonanalüüsi mudeli korral iga faktori mõju hinnanguks temale vastavate uuritava tunnuse väärtuste aritmeetilise keskmise erinevus üldkeskmisest μ.

Lugedes aga u juhuslikuks, avaldub tema i. taseme mõju u_i parim lineaarne nihketa prognoos (19) seosena

$$\tilde{u}_i = \frac{n_i\sigma_u^2}{\sigma_e^2 + n_i\sigma_u^2}(\bar{y}_{i\cdot} - \mu)\,. \tag{24}$$

Valemite (23) ja (24) võrdlemisel hakkab silma, et juhuslike efektide hinnangud on alati väiksemad, võrreldes fikseeritud efektide hinnangutega (sest $n_i\sigma_u^2/(\sigma_e^2 + n_i\sigma_u^2) \le 1$). Seejuures kehtib intuitiivselt mõistetav loogika – mida tugevam on faktori mõju, seda sarnasemad on fikseeritud ja juhusliku mudeli eeldusel leitud efektide hinnangud (seda suurem on dispersioonikomponendi σ_u^2 väärtus võrreldes jääkvarieeruvusega σ_e^2). Samuti nähtub valemeist (23) ja (24), et faktori *i*. tasemel sooritatud mõõtmiste arvu n_i suurenedes väheneb erinevus selle taseme mõju hinnangute $\hat{u}_i$ ja $\tilde{u}_i$ (leituna vastavalt kas fikseeritud või juhuslikust mudelist) vahel.

Näide 12. Käsitleme nüüd mudeliga (1) kirjeldatud tallede võõrutusmassi näiteanalüüsis jäära mõju juhuslikuna (andmestikus oleva kahe jäära ja nende kokku üheksa järglase alusel hinnatavad isa mõjud esindavad vaid juhuslikku väljavõtet kõigist isalt järglsele päranduda võivatest geenikomplektidest). Mudeli fikseeritud ja juhuslikele efektidele vastavad plaanimaatriksid on

$$\mathbf{X} = \left(\begin{array}{cc:cc:cc:c} 1 & 1 & 0 & 0 & 1 & 91 \\ 1 & 1 & 0 & 1 & 0 & 0 & 90 \\ 1 & 1 & 0 & 0 & 1 & 0 & 86 \\ 1 & 0 & 1 & 0 & 1 & 0 & 101 \\ 1 & 0 & 1 & 0 & 1 & 0 & 86 \\ 1 & 1 & 0 & 0 & 1 & 0 & 117 \\ 1 & 1 & 0 & 0 & 1 & 0 & 152 \\ 1 & 0 & 1 & 1 & 0 & 0 & 110 \\ 1 & 0 & 1 & 0 & 1 & 0 & 107 \end{array}\right) \text{ ja } \mathbf{Z} = \left(\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array}\right)$$

ning hinnatavate parameetrite vektorid esituvad kujul

$$\boldsymbol{\beta} = \left(\begin{array}{c} \mu \\ \hdashline S_1 \\ S_2 \\ \hdashline P_1 \\ P_2 \\ P_3 \\ \hdashline b_1 \end{array}\right) \text{ ja } \mathbf{u} = \begin{pmatrix} I_{1025} \\ I_{1027} \end{pmatrix}.$$

Üheste lahendite saamise huvides võrdsustame nulliga fikseeritud faktorite viimaste tasemete mõjud ($S_2 = 0$, $P_3 = 0$) ning eemaldame neile vastavad veerud plaanimaatriksist **X**, saades maatrikstehete läbiviimisel kasutatava täisastakuga plaanimaatriksi

$$\mathbf{X}^* = \left(\begin{array}{cc:cc:c} 1 & 1 & 0 & 0 & 91 \\ 1 & 1 & 1 & 0 & 90 \\ 1 & 1 & 0 & 1 & 86 \\ 1 & 0 & 0 & 1 & 101 \\ 1 & 0 & 0 & 1 & 86 \\ 1 & 1 & 0 & 1 & 117 \\ 1 & 1 & 0 & 1 & 152 \\ 1 & 0 & 1 & 0 & 110 \\ 1 & 0 & 0 & 1 & 107 \end{array}\right).$$

Et juhuslike efektide **u** ja **e** dispersioonimaatriksid on diagonaalsel kujul (11), on mudeli parameetrid hinnatavad segamudeli võrrandist (21). Viimase rakendamiseks vajaliku dispersioonikomponentide suhte σ_e^2/σ_u^2 saame avaldada tallede võõrutusmassi päritavusest h^2, lugedes viimase teadaolevaks:

$$\frac{\sigma_e^2}{\sigma_u^2} = \frac{4 - h^2}{h^2}\,.$$

Võttes $h^2 = 0{,}2$, saame $\sigma_e^2/\sigma_u^2 = 19$.

Seega avaldub segamudeli võrrandi (21) kordajate maatriksi parem alumine nurk $\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \sigma_e^2/\sigma_u^2\,\mathbf{I}$ kujul

$$\begin{pmatrix}1&1&1&1&1&0&0&0&0\\0&0&0&0&0&1&1&1&1\end{pmatrix}\begin{pmatrix}1&0\\1&0\\1&0\\1&0\\1&0\\0&1\\0&1\\0&1\\0&1\end{pmatrix}+19\begin{pmatrix}1&0\\0&1\end{pmatrix}=\begin{pmatrix}5+19&0\\0&4+19\end{pmatrix}=\begin{pmatrix}24&0\\0&23\end{pmatrix}$$

ning kogu segamudeli võrrand on peale plaanimaatriksite korrutamistehete sooritamist järgmine:

$$\left(\begin{array}{ccccc:cc}9&5&2&6&940&5&4\\5&5&1&3&536&3&2\\2&1&2&0&200&1&1\\6&3&0&6&649&3&3\\940&536&200&649&101716&454&486\\\hdashline5&3&1&3&454&24&0\\4&2&1&3&486&0&23\end{array}\right)\left(\begin{array}{c}\mu\\S_1\\P_1\\P_2\\b_1\\\hdashline I_{1025}\\I_{1027}\end{array}\right)=\left(\begin{array}{c}254\\146\\70\\162\\26746\\\hdashline136\\118\end{array}\right).$$

Selle võrrandi lahendina saame parameetrite hinnangute vektori

$$\begin{pmatrix}\hat{\mu}\\\hat{S}_1\\\hat{P}_1\\\hat{P}_2\\\hat{b}_1\\\hdashline\hat{I}_{1025}\\\hat{I}_{1027}\end{pmatrix}=\begin{pmatrix}14{,}12\\3{,}51\\14{,}37\\5{,}98\\0{,}05\\\hdashline-0{,}04\\0{,}04\end{pmatrix}.$$

Võrreldes fikseeritud mudelist saadud hinnangutega (18) on ootuspäraselt vähenenud jääradevaheline erinevus – on ju jäära mõjuga seletatav võõrutusmasside erinevus 19 korda väiksem juhusliku vea arvele jäänud varieeruvusest ($\sigma_e^2/\sigma_u^2=19$), mis, nagu nähtus valemeist (23) ja (24), toob kaasa väiksemad juhuslike isamõjude hinnangud. Loomulikult muutusid tänu mudeli muutmisele mingil määral ka fikseeritud efektide hinnangud, aga faktorite tasemete järjestus jäi endiseks – jäärtallede võõrutusmass on suurem; suurima võõrutusmassiga on üksiktalled, seejärel kaksik- ja kolmiktalled; võõrutusvanuse kasvades suureneb ka tallede võõrutusmass.

Eelnevas näites rakendatud mudelit, kus uuritava tunnuse geneetilist determineeritust väljendava juhusliku faktorina käsitletakse isa ning ülejäänud faktorid on fikseeritud ja kirjeldavad mitte aditiiv-geneetilisi mõjusid, nimetatakse põllumajandusloomade aretuses **isa mudeliks** (inglise keeles *sire model*). Lähemalt isa mudelist ja ka teistest segamudeli rakendustest geneetiliste parameetrite hindamisel järgmises peatükis.

8 DISPERSIOONIKOMPONENTIDE HINDAMINE *

Nii fikseeritud efektide hindamine üldistatud vähimruutude meetodil (17), juhuslike efektide parim lineaarne nihketa prognoos (19) kui ka Hendersoni segamudeli võrrandi (20) või (21) lahendamine eeldab dispersioonikomponentide kaudu defineeritud dispersioonimaatriksite teadmist. Nagu nähtus eelneva peatüki näitest, piisab sageli ka dispersioonikomponentide suhte või mõne populatsiooni geneetilise parameetri väärtuse teadmisest. Samas aga ei ole meil kuidagi võimalik teada dispersioonikomponentide suhet kõigi ette tulla võivate tunnuste või mudelite korral ning ka meie poolt analüüsitava populatsiooni geneetiline struktuur ei pruugi vastata teadaolevate geneetiliste parameetritega kirjeldatud populatsioonile. Valede dispersiooniparameetrite kasutamisega kaasnevad aga ka valed mõjuhinnangud, aretusväärtused ja seeläbi ka valed otsused tuleviku strateegiate osas.

Dispersioonikomponentide hindamiseks on välja töötatud hulk erinevaid meetodeid. Nende mitmesus on tingitud sellest, et vähegi keerulisema mudeli puhul ei ole võimalik välja kirjutada dispersioonikomponentide parimate omadustega analüütilisi lahendeid, kasutada tuleb kas teatud lihtsustavaid eeldusi või siis ligilähedasi matemaatilisi arvutusmeetodeid. Meetodite matemaatiline keerukus on ka põhjuseks, miks üldiste lineaarsete segamudelite rakenduste tudeerimisel dispersioonikomponentide hindamine kõrvale jäetakse. Samas baseerub peatükis 3 tutvustatud peamiste populatsioonigeneetiliste parameetrite hindamine enamasti just dispersioonikomponentide hindamisel.

Andmete tasakaalulisuse (vt pt 6.2.4) ja vaatluste sõltumatuse eeldusel on dispersioonikomponentide hinnangud leitavad tavalise dispersioonanalüüsi abil. Mittetasakaalulise andmestiku ja keerulisema kovariatsioonistruktuuri korral on kõige universaalsemaks osutunud REML-meetod.

8.1 Dispersioonanalüüsi meetod (ANOVA-meetod)

Võtame vaatluse alla ühefaktorilise dispersioonanalüüsi mudeli (22). Eeldame et faktor *u* on juhuslik ning et efektid u_i ja e_{ij} on sõltumatud ja normaaljaotusega. Samuti eeldame, et faktoril *u* on *a* taset, igal tasemel on sooritatud täpselt *n* mõõtmist ning kokku on mõõtmisi $N = an$. Dispersioonanalüüsi arvutused koondatakse nn dispersioonanalüüsi tabelisse, mis vaatlusaluse mudeli tarvis on toodud tabelis 1.

Tabel 1. Dispersioonanalüüs tabel mudeli (22) korral

Varieeruvuse allikas	Ruutude summa *	Vabadus-astmete arv	Keskruut	Keskruudu ooteväärtus
Faktor *u*	$SS(u) = \sum_{i=1}^{a} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N}$	$a-1$	$MS(u) = SS(u)/a-1$	$n\sigma_u^2 + \sigma_e^2$
Jääk	$SS(e) = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{a} \frac{y_{i.}^2}{n}$	$N-a$	$MS(e) = SS(e)/N-a$	σ_e^2

* siin $y_{i.} = \sum_{j=1}^{n} y_{ij}$ ja $y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$

ANOVA-hinnangud dispersioonikomponentidele σ_u^2 ja σ_e^2 saadakse, võrdsustades keskruudud nende ooteväärtustega:

$$\hat{\sigma}_u^2 = \frac{1}{n}\left[MS(u) - MS(e)\right]$$

ja

$$\hat{\sigma}_e^2 = MS(e).$$

Mittetasakaaluliste andmete ja ühefaktorilise dispersioonanalüüsi mudeli (22) korral on juhusliku faktori poolt määratud osa uuritava tunnuse varieeruvusest, σ_u^2, ANOVA-meetodil hinnatav valemist

$$\hat{\sigma}_u^2 = \frac{1}{d}\left[MS(u) - MS(e)\right],$$

kus $d = \frac{1}{a-1}\left(N - \frac{1}{N}\sum_{i=1}^{a} n_i^2\right)$ ja n_i tähistab mõõtmiste arvu faktori *u* tasemel *i*.

8.2 REML-meetod

REML-meetod sobib dispersioonikomponentide hindamiseks ükskõik millise dispersioonistruktuuriga mudelite korral, kusjuures tasakaalulise traditsioonilise segamudeli korral on dispersioonikomponentide REML-hinnangud võrdsed ANOVA-hinnangutega. Lühend REML tähendab **kitsendatud** e **jääkide** e **vähendatud suurima tõepära meetodit** (*REstricted* e *REsidual* e *REduced Maximum Likelihood method*) ja baseerub nagu suurima tõepära meetodid ikka mingi teoreetilise jaotuse tõepärafunktsioonil (vt ka pt 2.4.1). Üldise lineaarse mudeli (4) korral on selleks teoreetiliseks jaotuseks normaaljaotus – $\mathbf{y} \sim N(\mathbf{X\beta}, \mathbf{V})$. Kitsendatud suurima tõepära meetodi idee seisneb selles, et kuigi üldine lineaarne mudel sisaldab kaht tüüpi hinnatavaid parameetreid – fikseeritud faktorite efekte ja juhuslike faktorite dispersiooniparameetreid – püütakse viimaste hindamiseks teisendada esialgset mudelit kujule, kus juhuslike faktorite mõjudest tingitud varieeruvus oleks sama, aga fikseeritud faktorite mõju võrduks nulliga. Sellisel juhul on kogu arvutusprotsess suunatud dispersioonikomponentide hindamisele, mis peaks garanteerima täpsemad hinnangud võrreldes tavalise suurima tõepära meetodiga (ML, *Maximum Likelihood method*), mis hindab samaaegselt nii fikseeritud faktorite mõjusid kui ka dispersioonikomponente. Mudeli (4) teisendust võib esitada korrutisena

$$\mathbf{K}^{\mathrm{T}}\mathbf{y} = \mathbf{K}^{\mathrm{T}}\mathbf{X\beta} + \mathbf{K}^{\mathrm{T}}\mathbf{Zu}$$

kus maatriks **K** on defineeritud selliselt, et $\mathbf{K}^{\mathrm{T}}\mathbf{X} = \mathbf{0}$, mistõttu

$$\mathbf{K}^{\mathrm{T}}\mathbf{y} \sim N(\mathbf{0}, \mathbf{K}^{\mathrm{T}}\mathbf{VK}) .$$

Teisendatud mudeli vasakut poolt $\mathbf{K}^{\mathrm{T}}\mathbf{y}$ võib mõista ka kui fikseeritud efektide poolt kirjeldamata jäänud osa uuritavast tunnusest **y** e erinevust (jääki) fikseeritud efektide poolt kirjeldatu ja uuritava tunnuse **y** tegelike väärtuste vahel (siit ka nimetus *REsidual ML method*).

REML-meetodi korral maksimeeritakse teisendatud uuritava tunnuse logaritmiline tõepärafunktsioon

$$l(\mathbf{V};\mathbf{y}) = -\tfrac{1}{2}\ln\left|2\pi\mathbf{K}^{\mathrm{T}}\mathbf{VK}\right| - \tfrac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}(\mathbf{K}^{\mathrm{T}}\mathbf{VK})^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{y} \qquad (25)$$

dispersiooniparameetrite suhtes, võttes selleks funktsioonist (25) esimest järku osatuletise hinnatava parameetri suhtes, võrdsustades tulemuse nulliga ning lahendades – tulemuseks ongi dispersiooniparameetri REML-hinnang. Üldjuhul ei ole dispersioonikomponentide hinnangute analüütiline leidmine võimalik, logaritmilise tõepärafunktsiooni maksimeerimise tulemusena saadakse keeruline maatriksvõrdus

$$\mathrm{tr}\left[(\mathbf{K}^{\mathrm{T}}\mathbf{VK})^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\mathbf{K}\right] = \mathbf{y}^{\mathrm{T}}\mathbf{K}(\mathbf{K}^{\mathrm{T}}\mathbf{VK})^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}\mathbf{K}(\mathbf{K}^{\mathrm{T}}\mathbf{VK})^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{y} ,$$

kus mudeli dispersiooniparameetrid sõltuvad üksteisest ja ka juhuslike faktorite realiseerunud väärtustest (maatriks $\mathbf{Z}_i$ koosneb *i*.-le juhuslikule faktorile vastavatest plaanimaatriksi **Z** veergudest). Üksikute parameetrite hinnangute leidmiseks kasutatakse ligikaudseid järkjärgulise lähendamise e iteratsioonimeetodeid: algselt (nn nullsammul) antakse mudeli parameetritele mingid algväärtused, neist lähtuvalt leitakse 1. sammul kõigi parameetrite uued hinnangud, seejärel lähtutakse 2. sammul parameetrite hinnangute arvutamisel juba 1. sammul leitutest jne – igal järgneval sammul kasutatakse eelmisel sammul leitud hinnanguid; protsess koondub parameetrite lõplikeks hinnanguteks, kui mingil sammul leitud hinnangute erinevus eelneval sammul leitutest on piisavalt väike. Algoritme, mida sellisel ligikaudsel arvutamisel kasutatakse, on jälle rohkem kui üks – mõni neist koondub väiksema, mõni suurema arvu sammude järel; teisalt ei pruugi enamasti kiiremini koonduvad algoritmid mõnel juhul üleüldse koonduda – kõik see teeb dispersioonikomponentide hindamise veel keerulisemaks.

9 ÜLESANDED

Järgnev tabel sisaldab andmeid seitsme lehma piimatoodangu, isa, vanuse ja karja kohta.

Loom	Isa	Vanus	Kari	Toodang
2	1	7	1	7000
4	3	7	1	6600
6	5	6	1	6800
7	5	4	1	7200
8	3	3	2	8000
9	3	2	2	8300
10	5	2	2	8900

1. Lugedes kõik faktorid fikseerituteks, pange kirja üldine lineaarne mudel hindamaks isa, vanuse ja karja mõju piimatoodangule, seejuures käsitlege isa ja karja diskreetsete faktoritena ning vanust pidava faktorina
    a. objektiviisi;
    b. maatrikskujul, pannes kirja ka plaanimaatriksi **X**.
2. Pange kirja mudeli dispersioonimaatriksid **G**, **R** ja **V**, lugedes isa mõju juhuslikuks.
3. Lugedes kõik faktorid fikseerituteks, hinnake nende mõjud vähimruutude meetodil (13), kasutades SAS-i reparametriseerimistingimusi.
4. Leidke isadele ja karjadele vastavad vähimruutkeskmised.
5. Lugedes isa mõjud juhuslikuks ja võttes piimatoodangu päritavuseks $h^2 = 0,25$, leidke Hendersoni segamudeli võrrandi (21) abil fikseeritud efektide hinnangud ja juhusliku faktori 'Isa' realiseerunud väärtuste prognoosid.