TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Julia Aru

INFORMATIIVSE VALIKU MÕJU TUNNUSTEVAHELISELE KOVARIATSIOONILE

Magistritöö

Juhendaja: Imbi Traat

Tallinn 2008

Sisukord

1 Sissejuhatus

Antud magistritöö eesmärgiks on uurida, kuidas mõjutab informatiivne valik tunnustevahelist kovariatsiooni. Informatiivse valiku korral sõltub valikumehhanism kas otseselt või kaudselt uuritavatest tunnustest. Selle tulemusena ei peegelda uuritavate tunnuste ühisjaotus ega marginaaljaotused valimis enam üldkogumijaotust ja ta ka ei lähene sellele valimimahu kasvades. Selle fakti arvestamata jätmine võib põhjustada suuri nihkeid hinnangutes ja valesid järeldusi.

Valiku mõju uurimiseks ja selle arvesse võtmiseks kasutatakse töös tunnuste valimijaotust, mis avaldub üldkogumijaotuse ja objektide kaasamistõenäosuste abil. Sama meetodit rakendati ka autori bakalaureusetöös, kuid seal uuriti regressiooniparameetrite hindamist. Bakalaureusetöös oli pakutud ka võimalik valiku informatiivsuse testimise meetod struktuurimudelite tehnika abil. Osutus, et selle rakendamiseks on vaja hinnata tunnuste kovariatsioonimaatriksit informatiivse valiku tingimustes, mida aga ei ole seni kirjanduses uuritud. Samas on tunnuste kovariatsioonimaatriks ka mitmete klassikaliste andmeanalüüsimeetodite (fakroranalüüs, kanooniline analüüs) aluseks, mis suurendab tema hindamise tähtsust informatiivse valikutingimustes veelgi. Sellest ongi inspireeritud antud magistritöö.

Esmalt tuuakse töös sisse vajalikud mõisted ja seosed antud valdkonnast ja selgitatakse probleemi olemust. Edasi vaadeldakse tunnuste sõltumatuse juhtu ja esitatakse tingimusi, millal sõltumatus säilib ka valimis. Järgmises peatükis pakutakse kovariatsiooni hindamise parandamise meetod suvalise üldkogumijaotuse puhul ning esitatakse analüütiline valimijaotuse avaldis mitmemõõtmelisesse eksponentsiaalsesse perre kuuluvata üldkogumijaotuse puhul. Näitena tuletatakse kahemõõtmelise normaaljaotuse ja eksponentsiaalsete kaasamistõenäosuste juhule vastav valimijaotus koos korrelatsiooni täpse avaldisega. Teise näitena vaadeldakse multinomiaaljaotust.

Erijuhuna vaadeldakse mitmemõõtmelist normaaljaotust, seejuures kasutatakse maatrikskuju, mis oluliselt lihtsustab avaldisi. Lõpuks rakendatakse eelnevalt esitatud üldist meetodit kovariatsioonimaatriksi hindamiseks simulatsiooniuuringus, mis põhineb reaalsel Eesti Sotsiaaluuringu andmestikul.

Kõik näited selles töös on teostatud statistikapaketi R abil (R Develpment Core Team, 2008), simulatsiooniuuring aga paketi SAS abil.

2 Seoseid üldkogumi- ja valimijaotuse vahel

2.1 Tähistused ja põhiseosed

Üldkogum. Vaatleme lõplikku üldkogumit U, mis sisaldab N objekti, $U = (1,2,...,i,...,N)$. Uuritavate tunnuste väärtused objektil i olgu $\mathbf{y}_i = (y_i^{\,1}, y_i^{\,2},..., y_i^{\,k})'$, kus k on uuritavate tunnuste arv. Eeldame, et uuritavad tunnused on arvulised, kas diskreetsed või pidevad. Abitunnuste väärtused tähistame $\mathbf{x}_i$ ja teiste valimi võtmiseks kasutatavate tunnuste (disainitunnuste) väärtused[1] $\mathbf{d}_i$. Üldjuhul võib vektor $\mathbf{x}_i$ samuti sisaldada disainitunnuseid, kui need on olulised $\mathbf{y}_i$ varieeruvuse seletamisel. Sellisel juhul sisaldab $\mathbf{d}_i$ ainult neid disainitunnuseid, mis ei ole $\mathbf{x}_i$-s. Vektortunnuste $\mathbf{x}_i$ ja $\mathbf{d}_i$ dimensionaalsust me praegu ei kitsenda, vajadusel tuuakse vajalikud tähistused hiljem sisse.

Abitunnuste vektor $\mathbf{x}_i$ eeldatakse olevat teada $\forall\ i \in U$ korral, suurust[2] $\mathbf{y}_i$ vaadeldakse aga juhuslikuna. Tinglikud juhuslikud suurused $\mathbf{y}_i|\mathbf{x}_i$ eeldatakse olevat sõltumatud $\forall\ i \in U$ korral. Tähistame juhusliku suuruse $\mathbf{y}_i$ tinglikku tihedusfunktsiooni $f_p(\mathbf{y}_i \mid \mathbf{x}_i)$, kus indeks p viitab üldkogumile (population). Parameetrite vektorit, mis indekseerib f_p, tähistame $\boldsymbol{\theta} = (\theta_1, \theta_2,..., \theta_m)$, kus m on parameetrite arv. Tihedusfunktsioon $f_p(\mathbf{y}_i \mid \mathbf{x}_i)$ on see, mida lõppkokkuvõttes soovitakse hinnata.

Valim. Valim s on üldkogumi U alamhulk, mis sisaldab n erinevat juhuslikult valitud objekti U-st. Üldjuhul eeldame, et objektid satuvad valimisse üksteiselt sõltumatult, st valimi väärtused $\mathbf{y}_i, i = 1,2,...,n$, on sõltumatud sama jaotusega juhuslikud suurused. Saab näidata, et keerulisemate valikuskeemide puhul kehtib see nõue vähemalt asümptootiliselt (Pfeffermann, Krieger, Rinott, 1998). Olgu π_i objekti i kaasamistõenäosus – tõenäosus, et objekt i kaasatakse valimisse. Üldjuhul võib π_i sõltuda nii uuritavate tunnuste, abitunnuste, kui ka teiste disainitunnuste üldkogumi väärtustest:

$$\pi_i = P(i \in s) = g(\mathbf{X}, \mathbf{Y}, \mathbf{D}),$$

kus $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_N), \mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2,..., \mathbf{y}_N), \mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2,..., \mathbf{d}_N)$ ja g on suvaline funktsioon. Paneme tähele, et kui $\mathbf{Y}$ on juhuslik, siis on ka π_i juhuslik suurus.

---

[1] Siin ja edasi on mitmemõõtmelised tunnused tähistatud paksus kirjas, ühemõõtmelised aga normaalses kirjas kursiiviga.

[2] Suuruse y_i tähendus sõltub tegelikult kontekstist. Näiteks tihendusfunktsiooni avaldises tähendab ta funktsiooni argumenti.

Kui kaasamistõenäosused sõltuvad vaid abitunnuste ja disainitunnuste väärtustest **X** ja **D**, saame valiku mõju uuritavate tunnuste jaotusele valimis elimineerida hoides **X** ja **D** tingimuses. Kui aga kaasamistõenäosused sõltuvad uuritavate tunnuste väärtustest, siis ei saa me valiku mõjust nii lihtsalt lahti. Edaspidi loobume disainitunnuste eraldi välja toomisest.

Olgu I_i objekti i kaasamisindikaator (I_i = 1, kui objekt i on valimis). Siis $\pi_i = P(I_i = 1)$. Valimitihedusfunktsiooni esmaseks erinevuseks üldkogumi tihedusfunktsioonist on see, et lisandub tingimus vaadeldava objekti valimisse kaasamise kohta, s.t. I_i = 1. Kasutades Bayesi teoreemi ja hoides suurust $\mathbf{x}_i$ kogu aeg tingimuses, saame, et tinglik valimitihedusfunktsioon on

$$f_s(\mathbf{y}_i \mid \mathbf{x}_i) = f_p(\mathbf{y}_i \mid \mathbf{x}_i, I_i = 1) = \frac{P(I_i = 1 \mid \mathbf{y}_i, \mathbf{x}_i) f_p(\mathbf{y}_i \mid \mathbf{x}_i)}{P(I_i = 1 \mid \mathbf{x}_i)}. \qquad (2.1)$$

Valemist (2.1) näeme, et valimi ja üldkogumi tihedusfunktsioonid ühtivad parajasti siis, kui $P(I_i = 1 \mid \mathbf{y}_i, \mathbf{x}_i) = P(I_i = 1 \mid \mathbf{x}_i)$ iga $\mathbf{y}_i$ korral. Kui see tingimus ei ole täidetud nimetatakse valikut *informatiivseks*. Kui see tingimus on täidetud, siis fikseeritud $\mathbf{x}_i$ korral, saab valikut ignoreerida.

Märkus 1. Kaasamistõenäosus π_i erineb tõenäosusest $P(I_i = 1 \mid \mathbf{y}_i, \mathbf{x}_i)$, mis määrab valimijaotuse avaldises (2.1), kuna viimases on $(\mathbf{y}_i, \mathbf{x}_i)$ fikseeritud. Vaatamata sellele saab näidata, et nende kahe tõenäosuse vahel kehtib seos (Pfeffermann, Krieger, Rinott ,1998):

$$P(I_i = 1 \mid \mathbf{y}_i, \mathbf{x}_i) = E_p(\pi_i \mid \mathbf{y}_i, \mathbf{x}_i).$$

Märkus 2. Edaspidi vaatleme tihti ka lihtsamat juhtu, kus suuruse $\mathbf{y}_i$ üldkogumitihedusfunktsioon ei sõltu abitunnustest $\mathbf{x}_i$. Seos (2.1) taandub siis järgmisele kujule:

$$f_s(\mathbf{y}_i) = f_p(\mathbf{y}_i \mid I_i = 1) = \frac{P(I_i = 1 \mid \mathbf{y}_i) f_p(\mathbf{y}_i)}{P(I_i = 1)}. \qquad (2.2)$$

Sellisel juhul loobume valemite lihtsustamiseks tihti indeksist i, kuna objektidel on sel juhul sama tihedus.

Järgnevas toome ära üldised kasulikud seosed valimi ja üldkogumi jaotuste vahl. Seosed on tõestatud autori bakalaureusetöös (Aru, 2004).

Olgu $(\mathbf{u}_i, \mathbf{v}_i)$, $i \in U$, juhuslikud vektorid, $f_p(\mathbf{u}_i \mid \mathbf{v}_i)$ ja $f_s(\mathbf{u}_i \mid \mathbf{v}_i)$ vektori $\mathbf{u}_i$ tinglikud tihedusfunktsioonid vastavalt üldkogumis ja valimis, π_i objekti i kaasamistõenäosus,

$w_i = 1/\pi_i$ objekti i valikukaal ja tähistagu E_p ja E_s keskväärtusi vastavalt üldkogumi ja valimi jaotuse suhtes. Siis kehtivad seosed:

$$f_s(\mathbf{u}_i \mid \mathbf{v}_i) = \frac{E_p(\pi_i \mid \mathbf{u}_i, \mathbf{v}_i) f_p(\mathbf{u}_i \mid \mathbf{v}_i)}{E_p(\pi_i \mid \mathbf{v}_i)}, \tag{2.3}$$

$$f_p(\mathbf{u}_i \mid \mathbf{v}_i) = \frac{E_s(w_i \mid \mathbf{u}_i, \mathbf{v}_i) f_s(\mathbf{u}_i \mid \mathbf{v}_i)}{E_s(w_i \mid \mathbf{v}_i)}, \tag{2.4}$$

$$E_p(\mathbf{u}_i \mid \mathbf{v}_i) = \frac{E_s(w_i \mathbf{u}_i \mid \mathbf{v}_i)}{E_s(w_i \mid \mathbf{v}_i)}, \tag{2.5}$$

$$E_p(\pi_i \mid \mathbf{v}_i) = E_p(E_p[\pi_i \mid \mathbf{u}_i, \mathbf{v}_i] \mid \mathbf{v}_i). \tag{2.6}$$

Avaldised (2.3) ja (2.4) võimaldavad parameetriliselt hinnata üldkogumijaotust valimijaotuse abil, avaldis (2.5) aga näitab seost üldkogumi- ja valimijaotuse momentide vahel.

2.2 Informatiivne valik

Informatiivse valiku probleemi hakati arutama statistikaalases kirjanduses alles 1970-tes aastates. Üks esimestest artiklitest oli Rubin (1976), kust pärinevad tänapäevases kirjanduses andmete puudumise ignoreerimisest rääkides laiaulatuslikult kasutatavad mõisted MAR, MCAR, NMAR jne. Järgmised väga tähtsad ja palju viidatud artiklid olid Little (1982), mis keskendus mittevastamise probleemile informatiivse valiku tekkimise põhjusena, ja Sugden ja Smith (1984), milles esitati hiljem palju kasutust leidnud valiku ignoreeritavuse definitsioonid. Käesolevas magistritöös kasutatud informatiivse valiku käsitlus põhineb Jeruusalema Ülikooli professori Danny Pfeffermanni poolt arendatud meetodile, mille järgi kasutatakse hindamises valimijaotuse analüütilist avaldist üldkogumi jaotuse ja objektide kaasamistõenäosuste abil. Tema esimene valiku informatiivsust puudutav artikkel oli Pfeffermann (1988). Mainitud meetodit rakendati aga alles artiklis Krieger ja Pfeffermann (1992), milles viidatakse töödele Patil ja Rao (1978) ja Rao (1985) kui kaalutud jaotuse idee allikatele. Esimene just sellele meetodile keskendunud artikkel oli Pfeffermann, Krieger, Rinott (1998), kus käsitleti regressiooni parameetrite hindamist informatiivse valiku tingimustes ja põhjendati suurima tõepära meetodi kasutamist valimijaotuse baasil. Järgmises artiklis, Pfeffermann ja Sverchkov (1999), esitati üldisemad seosed üldkogumi- ja valimijaotuste vahel ja tutvustati valiku informatiivsuse testimise võimalusi. Mõned aastad hiljem ilmus raamatu peatükina ka üldistatud lineaarsete mudelite hindamise kirjeldus, Pfeffermann ja Sverchkov (2003). Lisaks regressioonimudelitele

rakendati seda meetodit ka mitmetasandiliste mudelitele (Pfeffermann et al, 1998; Grilli ja Pratesi, 2004; Pfeffermann, Moura ja Silva, 2006), väikeste piirkondade hindamisele (Pfeffermann ja Sverchkov, 2005), lõpliku üldkogumi kogusumma hindamisele (Sverchkov ja Pfeffermann, 2004), longituud-andmetele ja autoregressioonimudelitele (Eideh ja Nathan, 2006).

Kõige rohkem mõjutab informatiivne valik mudelite parameetrite hindamise õigsust, kuna enamik klassikalisi meetodeid eeldab lihtsat juhuslikku valikut. Kaalutud meetodid (nt kaalutud vähimruutude meetod või pseudo suurima tõepära meetod) on aga piiratud võimalustega ja vajavad normaaljaotuse eeldusi. Mitmes ülalpool mainutud artiklis on täheldatud, et kaasamistõenäosuste ja uuritavate tunnuste vahelise seose arvesse võtmine võimaldab hinnata üldkogumi parameetreid täpsemalt, st väiksema nihke ja dispersiooniga.

Paneme tähele, et valiku informatiivsusest saab rääkida pidades meeles vaid konkreetset tunnuste komplekti, st valik võib olla informatiivne või ignoreeritav ainult mingite tunnuste suhtes. Reaalses uuringus, kus on sadu uuritavaid tunnuseid, võib valik mõjutada vaid mõningate tunnuste hindamist ja jääda ignoreeritavaks teiste suhtes. See nähtus on tihedalt seotud ühe informatiivse valiku tekkimise põhjusega – mittevastamisega. Enamik uuringuid (eriti riigistatistika juhul) on disainitud selliselt, et valikuprotseduur oleks ignoreeritav, vähemalt fikseeritud teadaolevate disaintunnuste korral. Just valimisse sattunud objektide mittevastamine on see, mis teeb valikut informatiivseks. Mittevastamise probleemiga on seotud ka üks tihti esinev informatiivse valiku meetodeid kritiseeriv tähelepanek. Nimelt osutatakse tihti tähelepanu sellele, et mittevastamise puhul ei saa uurija kunagi teada tõelisi vastamistõenäosusi, kuna nad on praktiliselt alati seotud uuringuspetsiifiliste küsimustega, mis aga ei ole mittevastajate jaoks teada. Need kaasamistõenäosused, mida kasutatakse hindamisprotseduuris, on modelleeritud nii vastajatele kui mittevastajatele teada olevate tunnuste abil. Seega hoides need tunnused ja disaintunnused fikseerituna, saame me täpse mudeli kaasamistõenäosuste jaoks, kus aga ei saagi kohta olla uuritavale tunnusele. Vastuväiteks paneme tähele, et uurijale ei ole alati huvipakkuv kasutada näiteks regressioonimudelis kõiki disaintunnuseid ja mittevastamise modelleerimisel kasutatud tunnuseid, või ei ole need andmestikust kättesaadavad. Kui aga need tunnused mudelist kõrvale jätta, siis kaasamistõenäosused võivad olla korreleeritud uuritava tunnusega, ja siis on eriti tähtis seda sõltuvust mudeli hindamise juures arvesse võtta.

Kaasamistõenäosuste seos uuritava tunnusega võib olla sisseprogrammeeritud juba valiku disaini, näiteks siis, kui uuritav tunnus on teada üldkogumis enne uuringut ja kaasamistõenäosused sõltuvad sellest otseselt, nagu retrospektiivsete uuringute puhul. Selge, et sellisel juhul ei ole uuritava tunnuse kogusumma, keskmine vms uuringu eesmärgiks, vaid

pigem selle seos teiste tunnustega, mida kogutakse uuringu käigus. Ühe sellise uuringu kirjeldavad Pfeffermann ja Sverchkov (1999). Tegemist on 1988. aastal USA-s korraldatud emade ja imikute terviseuuringuga, milles kasutati kihtvalikut, kusjuures kihid olid defineeritud ema rassi ja lapse sünnikaalu järgi. Lapse sünnikaal oli seejuures ka uuritav tunnus – uuriti selle seost tausttunnustega.

Kuna valiku informatiivsust arvestavad meetodid on palju keerulisemad kui standardsed mudelite ja üldkogumiparameetrite hindamise meetodid, siis on alati otstarbekas veenduda, et valik on informatiivne. Selleks on välja töötatud mitmed informatiivsuse testid. Üks lihtsamaid võimalusi on testida järgmist hüpoteeside seeriat (vt Pfeffermann ja Sverchkov, 1999, ning Aru, 2004):

$$E_s(\varepsilon_i^{\,k}) = E_p(\varepsilon_i^{\,k}),\ k = 1,2,\ldots,\ \text{kus}\ \varepsilon_i = y_i - E_p(y_i \mid x_i).$$

Seose (2.5) tõttu on ülaltoodud hüpoteesid samaväärsed hüpoteesidega korrelatsiooni kohta:

$$Corr_s(\varepsilon_i^{\,k}, w_i) = 0,\ \mathrm{k} = 1,2,\ldots,$$

kus $Corr_s$ tähendab korrelatsiooni valimijaotuse suhtes. Praktikas peaks piisama 2-3 korrelatsiooni kontrollimisest.

Pfeffermann ja Sverchkov (2003) pakkusid ka teist, keerulisemat testi, kus kasutatakse hindamisvõrrandite erinevust informatiivse ja ignoreeritava valiku puhul (vt ka Aru, 2004).

3 Tunnuste sõltumatus üldkogumis

Vaatame kahemõõtmelist uuritavat tunnust $\mathbf{y}_i = (y_i, z_i)'$. Uurime, kas tunnuste vaheline sõltumatus üldkogumis säilib ka valimis. Lihtsuse mõttes loobume siin ja edaspidi seletavate tunnuste vektorist $\mathbf{x}_i$. Kõik tulemused kehtivad ka fikseeritud $\mathbf{x}_i$ korral, st tingliku jaotuse suhtes.

***Näide* 1**. Uuritavad tunnused y_i ja z_i olgu sõltumatud ja standardse normaaljaotusega N(0,1), see on tunnuste üldkogumijaotus. Genereerime lõpliku üldkogumi (st tunnuste realisatsioonid) mahuga $N = 300$. Sellest üldkogumist võtame valimi mahuga $n = 40$. Selleks arvutame kõigepealt iga objekti jaoks tema kaasamistõenäosuse valemiga, mis teeb valiku informatiivseks,

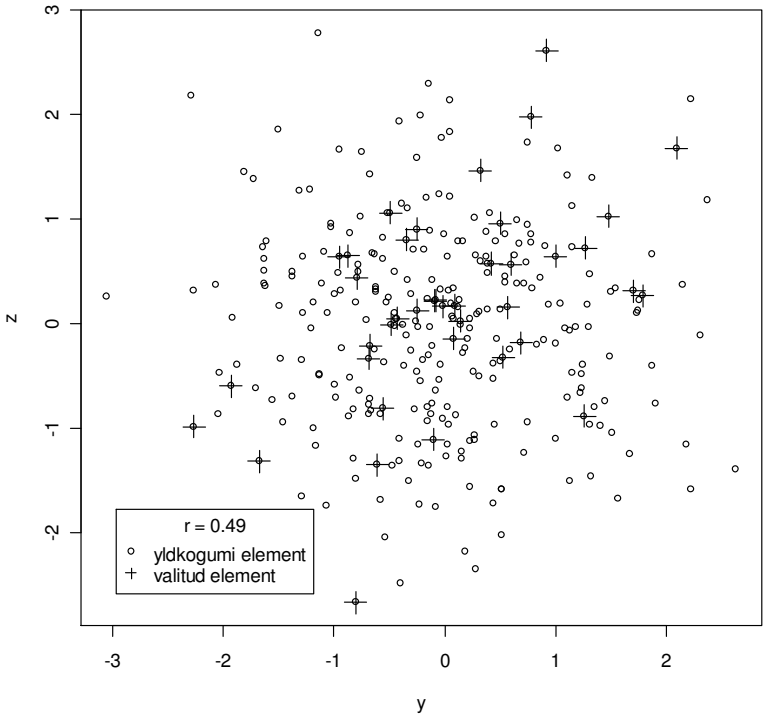
$$\pi_i = c_0 (5 + y_i \cdot z_i + 0.3\varepsilon_i),$$

kus $\varepsilon_i \sim U(-0.5, 0.5)$ ja normeeriv konstant c_0 on selline, et $\sum_{i=1}^{N} \pi_i = n$. Kaasamistõenäosuste keskmine on sel juhul

$$E_p(\pi_i \mid y_i, z_i) = c_0 \cdot (5 + y_i \cdot z_i).$$

Vastavalt saadud kaasamistõenäosustele võtame üldkogumist valimi kasutades Pareto valikut (Traat, Bondesson, Meister, 2000).

Joonis 1. Üldkogum ja valimisse sattunud elemendid



Näeme, et valimisse sattusid valdavalt elemendid samasuunalise y ja z-ga, st sellised, kus y ja z on mõlemad kas negatiivsed või positiivsed. See oli oodatav tulemus, kuna just sellisel juhul on kaasamistõenäosus π_i suur. Sellise valiku tulemusena ei ole uuritavad tunnused valimis enam sõltumatud. Pearsoni korrelatsioonikordaja suurus on $r = 0.49$.

Osutub, et kui tunnused on üldkogumis sõltumatud, siis nende vahekord valimis sõltub kaasamistõenäosustest. Järgmine teoreem annab tingimuse tunnuste sõltumatuseks valimis, kui nad on sõltumatud üldkogumis. □

***Teoreem* 1.** Olgu tunnused $y^1, y^2, ..., y^k$ sõltumatud üldkogumis. Kui kaasamistõenäosuste tinglik keskväärtus avaldub faktoriseeritud kujul $y^1, y^2, ..., y^k$ suhtes, ehk

$$E_p(\pi_i \mid \mathbf{y}_i) = E_p(\pi_i \mid y_i^1) \cdot ... \cdot E_p(\pi_i \mid y_i^k),$$

siis $y^1, y^2, ..., y^k$ on sõltumatud ka valimis.

Tõestus. Kuna $y^1, y^2, ..., y^k$ on sõltumatud üldkogumis, siis vektori $\mathbf{y}_i$ üldkogumi tihedusfunktsioon esitub marginaaltiheduste korrutisena:

$$f_p(\mathbf{y}_i) = f_p(y_i^1) \cdot f_p(y_i^2) \cdot ... \cdot f_p(y_i^k).$$

Võttes valemis (2.3) $\mathbf{u}_i = \mathbf{y}_i$ ja $\mathbf{v}_i = const$ ja kasutades teoreemi eeldusi ning edasi valemit (2.6) nimetaja jaoks, saame, et

$$f_s(\mathbf{y}_i) = \frac{E_p(\pi_i \mid \mathbf{y}_i) f_p(\mathbf{y}_i)}{E_p(\pi_i)} = \frac{[E_p(\pi_i \mid y_i^1) \cdot ... \cdot E_p(\pi_i \mid y_i^k)] \cdot [f_p(y_i^1) \cdot ... \cdot f_p(y_i^k)]}{E_p(E_p(\pi_i \mid \mathbf{y}_i))} =$$

$$= \frac{E_p(\pi_i \mid y_i^1) f_p(y_i^1)}{E_p(E_p(\pi_i \mid y_i^1))} \cdot ... \cdot \frac{E_p(\pi_i \mid y_i^k) f_p(y_i^k)}{E_p(E_p(\pi_i \mid y_i^k))} = f_s(y_i^1) \cdot ... \cdot f_s(y_i^k).$$

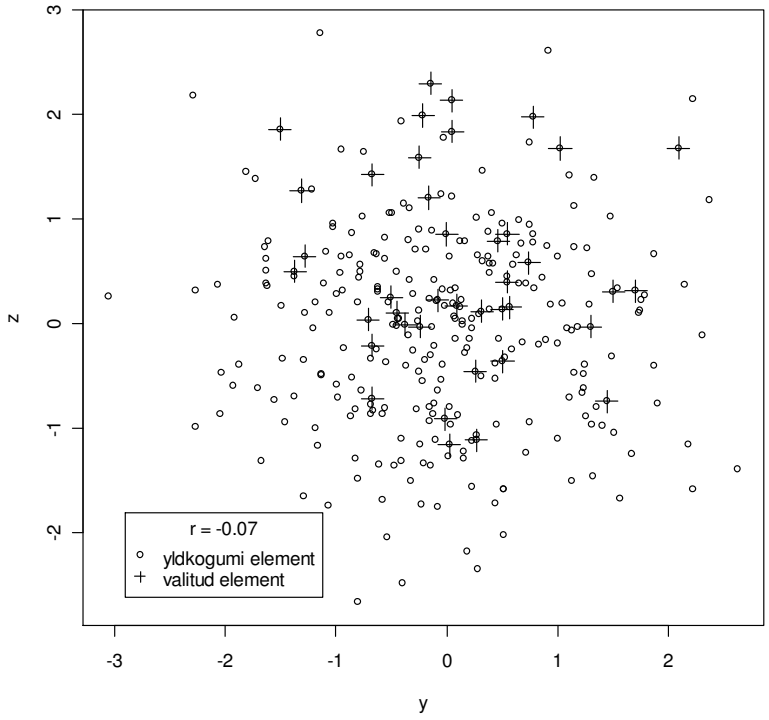
Seega tunnused on sõltumatud ka valimis. □

***Näide* 2.** Võtame sama üldkogumi, mida kasutasime näites 1. Kaasamistõenäosustel olgu nüüd kuju

$$\pi_i = c_0 \cdot [(3 + y_i) \cdot (3 + z_i) + 0.3\varepsilon_i].$$

Võttes antud kaasamistõenäosustega valimi sellest üldkogumist, näeme, et sõltumatus säilib ka valimis, $r = -0.07$.

Joonis 2. Üldkogum ja valimisse sattunud elemendid



Nii näites 1 kui näites 2 on tegu informatiivse valikuga ning valimijaotused erinevad üldkogumijaotustest. Joonistel viitab sellele erinevus punktipilvede paiknemises. Teisel juhul säilis aga sõltumatuse omadus.

4 Üldkogumi kovariatsiooni hindamine

Eesmärgiks on üldkogumi kovariatsiooni hindamine informatiivse valiku korral kasutades seoseid üldkogumi ja valimi jaotuste vahel.

4.1 Kahe suvalise jaotusega tunnuse kovariatsioon

Üks võimalus kovariatsiooni hindamiseks on lähtuda otseselt definitsioonist.

Kovariatsioon tunnuste *y* ja z vahel defineeritakse järgmiselt:

$$\mathrm{cov}_p(y,z) = E_p[y - E_p(y)]\cdot[z - E_p(z)] = E_p(yz) - E_p(y)E_p(z). \qquad (4.1)$$

Valemi (2.5) abil saame keskväärtused üldkogumijaotuse suhtes esitada valimikeskväärtuste kaudu. Seega teiseneb (4.1) kujule:

$$\mathrm{cov}_p(y,z) = \frac{E_s(wyz)}{E_s(w)} - \frac{E_s(wy)}{E_s(w)}\frac{E_s(wz)}{E_s(w)}. \qquad (4.2)$$

Keskväärtused valemi (4.2) paremal poolel on valimijaotuse suhtes. Klassikalise valimidefinitsiooni kohaselt on valimi elemendid sõltumatud ja sama valimijaotusega.

Seega saame iga keskväärtust avaldises (4.2) hinnata valimikeskmise abil. Paneme tähele, et asendades keskväärtused avaldises (4.2) valimikeskmistega, saame tegelikult valemi tavalise kaalutud valimikovariatsioonikordaja jaoks, mis on teadaolevalt mõjus hinnang üldkogumi kovariatsioonile:

$$\hat{\text{cov}}_p(y,z;w) = \frac{\sum_s w_i y_i z_i}{\sum_s w_i} - \frac{\sum_s w_i y_i}{\sum_s w_i} \cdot \frac{\sum_s w_i z_i}{\sum_s w_i}. \tag{4.3}$$

Märkus 3. Kaalutud kovariatsioonikordaja (4.3) on mõjus hinnang üldkogumi kovariatsioonile, kuid ta ei ole nihketa hinnang. Jagades hinnangu kordajaga

$$c^{korr} = 1 - \sum_s \left(\frac{w_i}{\sum_s w_i} \right)^2 \tag{4.4}$$

saame nihketa hinnangu (R Development Core Team, 2008).

Paneme tähele, et avaldises (4.2) võetakse keskväärtus kaalu ja uuritavate tunnuste korrutisest. Kui kaalud (ja järelikult ka kaasamistõenäosused) on sõltumatud uuritavatest tunnustest, siis kaalude keskväärtus $E_s(w)$ taandub avaldise (4.2) liikmete lugejast ja nimetajast ja üldkogumikovariatsioon ühtib kovariatsiooniga valimijaotuse suhtes. Arvestades seda, tundus töö autorile, et kaalude ja uuritavate tunnuste vahelise seose ulatuslikum kasutamine peaks parandama kovariatsiooni hindamise täpsust. Selle seose võimaldaks paremini arvesse võtta näiteks tinglike keskväärtuste $E_s(w \mid y,z)$ kasutamine kaalude w asemel valemis (4.3). Teadaolevalt on tinglike keskväärtuste $E_s(w \mid \cdot)$ varieeruvus väiksem, kui esialgsete kaalude w oma, mis peaks vähendama ka kaalutud hinnangute varieeruvust. See asendus on õigustatud ka seetõttu, et avaldises (4.2) $E_s(wyz) = E_s(E_s[w \mid y,z]yz)$. Lähemalt illustreerib seda võtet näide 3.

***Näide* 3.** Genereerime üldkogumi mahuga 10 000 objekti. Kuna reaalsetes uuringutes on tunnused tihti asümmeetrilise jaotusega, siis uuritavad tunnused y ja z genereerime kahemõõtmelisest kald-normaalsest jaotusest. Selle jaotuse tihedusfunktsioon on mitmemõõtmelisel juhul järgmine (Azzalini, Capitanio, 1999):

$$f_p(\mathbf{y}) = 2g(\mathbf{y}-\boldsymbol{\mu})\Phi(\boldsymbol{\alpha}'(\mathbf{y}-\boldsymbol{\mu})),$$

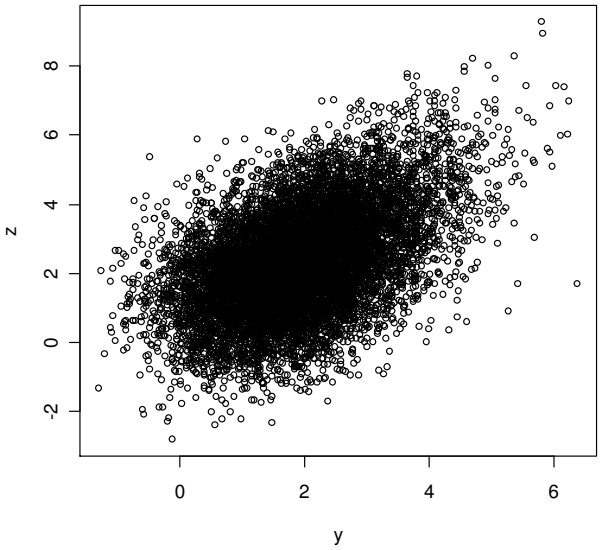
kus $g(\mathbf{y})$ on mitmemõõtmelise normaaljaotuse $N_k(0,\boldsymbol{\Sigma})$ tihedusfunktsioon, $\Phi(\cdot)$ on standardse ühemõõtmelise normaaljaotuse jaotusfunktsioon ja $\boldsymbol{\alpha}$ ja $\boldsymbol{\mu}$ on k-mõõtmelised

vektorid. Jaotuse asümmeetrilisuse suuna ja tugevuse määrab vektor $\boldsymbol{\alpha}$, vektor $\boldsymbol{\mu}$ on paiknemisparameeter. Käesoleva näite tarvis valime

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix},\ \boldsymbol{\alpha} = (1,1)\ \text{ja}\ \boldsymbol{\mu} = (1,1).$$

Uuritavate tunnuste ühisjaotust üldkogumis illustreerib joonis 3.

Joonis 3. Tunnuste jaotus üldkogumis



Uuritavate tunnuste keskmised üldkogumis on vastavalt 1.9 ja 2.3, tunnuste vaheline korrelatsioon on 0.53.

Objektide kaasamistõenäosused arvutame järgmise valemi abil:

$$\pi_i = c_0 \cdot (8 - (y_i - 1.9) \cdot (z_i - 2.3) + \varepsilon_i),$$

kus $\varepsilon_i \sim U(-0.5,0.5)$ ja c_0 on normeeriv konstant. Seega valik on selgelt informatiivne: valimisse satuvad suurema tõenäosusega objektid, mille y ja z on üheaegselt kas suuremad või väiksemad keskmisest ehk valik suurendab antud juhul korrelatsiooni.

Üldkogumist võtame valimi mahuga n = 1000 vastavalt arvutatud kaasamistõenäosustele ja püüame hinnata üldkogumi kovariatsiooni tunnuste y ja z vahel. Kasutame kaalutud kovariatsioonihinnangut, täpsemalt selle nihketa versiooni:

$$\hat{\text{cov}}_p^{\,korr}(y,z,w) = \frac{1}{c^{korr}} \hat{\text{cov}}_p(y,z,w),$$

kus c^{korr} on parandustegur (4.4) ja $\hat{\text{cov}}_p(y,z,w)$ on antud valemiga (4.3). Kaaludena vaatleme kahte võimalust:

- valikukaalud $w_i^1 = 1/\pi_i$,
- kaalude tinglikud keskmised $w_i^2 = E_s(w_i^1 \mid y_i, z_i)$.

Kaalude w_i^2 leidmiseks peame sobitama mudeli, kus funktsioontunnuseks on kaalud w^1 ja seletavateks tunnusteks on *y* ja *z*. Erinevaid mudeleid läbi proovides (kasutame funktsiooni *glm* paketist *R*) leiame, et kõige paremini sobib (mitmese korrelatsioonikordaja alusel) üldistatud lineaarne mudel, kus funktsioontunnuse jaotuseks on eeldatud normaaljaotus ja seosefunktsiooniks pöördteisendus. Sobitatud mudeli abil prognoositav väärtus konkreetsete *y* ja *z* väärtuste puhul ongi keskväärtus w_i^2, mida tahtsime hinnata.

Kordasime seda protseduuri (alates valimi võtmisest) 1000 korda, arvutasime $\hat{\text{cov}}_p^{korr}(\cdot)$ nii kaaludega w^1 kui w^2, ja leidsime ruutkeskmise vea, st hälvete ruutude keskmise kovariatsiooni tõelisest väärtusest. Kaalude w^2 abil leitud hinnangu ruutkeskmine viga oli $2.43 \cdot 10^{-3}$, esialgsete kaalude w^1 puhul aga $2.52 \cdot 10^{-3}$. Seega antud juhul saab kaalude ja tunnuste vahelise seose modelleerimise abil saada mõnevõrra täpsema kovariatsiooni hinnangu kui tavalise kaalumisega. Tuleb aga märkida, et hea hinnangu leidmiseks on vaja head mudelit kaalude jaoks.

Peatükis 6 rakendame seda meetodit ka reaalsetele andmetele.

4.2 Kovariatsioon mitmemõõtmelise eksponentsiaalse pere korral

Praktikas avaldub tihti tunnustevaheline kovariatsioon teiste jaotuseparameetrite (näiteks keskväärtuse ja dispersiooni) kaudu. Sellepärast piisab tihti, kui oskame määrata üldkogumijaotuse parameetrid valimijaotuse parameetritest ja kaasamistõenäosustest. Osutub, et kui üldkogumijaotus kuulub eksponentsiaalsesse jaotuste perre ja kaasamistõenäosused omavad teatud kuju, siis on ka valimijaotuse kuju täpselt teada. Seega on võimalik analüütiliselt arvutada üldkogumiparameetrid valimiparameetritest ja järelikult ka kovariatsiooni. Eksponentsiaalse pere juhtu on ühemõõtmelisel juhul uuritud ka artiklis (Pfeffermann, Krieger, Rinott, 1998).

Vaatleme *k*-mõõtmelist uuritavate tunnuste vektorit $\mathbf{y} = (y_1, ..., y_k)$. Eeldame, et vektori $\mathbf{y}$ jaotus kuulub mitmemõõtmelisse eksponentsiaalsesse jaotuste perre (Lehman, Casella, 1998), st

$$f_p(\mathbf{y} \mid \boldsymbol{\theta}) = h(\mathbf{y}) \exp\left\{ \sum_{j=1}^{m} g_j(\boldsymbol{\theta}) T_j(\mathbf{y}) - B(\boldsymbol{\theta}) \right\},$$

kus $\boldsymbol{\theta} = (\theta_1,...,\theta_m)$ on parameetrite vektor, $h(\cdot): \Re^k \to \Re$ ja $T_j(\cdot): \Re^k \to \Re$ on suvalised argumendist **y** sõltuvad funktsioonid, ning $g_j(\cdot): \Re^m \to \Re$ ja $B(\cdot): \Re^m \to \Re$ on suvalised parameetrist $\boldsymbol{\theta}$ sõltuvad funktsioonid.

Edasises on mugavam kasutada mitmemõõtmelise eksponentsiaalse pere kanoonilist kuju:

$$f_p(\mathbf{y} \mid \boldsymbol{\eta}) = h^*(\mathbf{y}) \exp\left\{\sum_{j=1}^{m} \eta_j T_j(\mathbf{y}) - B^*(\boldsymbol{\eta})\right\}, \tag{4.5}$$

kus $\boldsymbol{\eta} = (\eta_1,...,\eta_m)$ on kanooniliste parameetrite vektor. Vektori $\boldsymbol{\eta}$ väärtuste ruumi tähistame $\mathbf{H} \subset \Re^m$.

Vaatleme juhtu, kus kaasamistõenäosused on samuti eksponentsiaalsel kujul:

$$E_p(\pi \mid \mathbf{y}) = c_0 \exp\left\{\sum_{j=1}^{m} p_j T_j(\mathbf{y})\right\},$$

kus c_0 on normeeriv konstant, $T_j(\mathbf{y})$ on sama **y**-funktsioon, mis avaldises (4.5), ja p_j on konstandid.

Valemi (2.3) abil saame valimijaotuse avaldise:

$$f_s(\mathbf{y} \mid \boldsymbol{\eta}) = \frac{f_p(\mathbf{y} \mid \boldsymbol{\eta}) E_p(\pi \mid \mathbf{y})}{E_p(\pi)} = h^*(\mathbf{y}) \exp\left\{\sum_{j=1}^{m} \eta_j T_j(\mathbf{y}) - B^*(\boldsymbol{\eta})\right\} \cdot c_0 \exp\left\{\sum_{j=1}^{m} p_j T_j(\mathbf{y})\right\} \cdot \frac{1}{E_p(\pi)} =$$

$$= \frac{c_0 h^*(\mathbf{y})}{E_p(\pi)} \exp\left\{\sum_{j=1}^{m} (\eta_j + p_j) T_j(\mathbf{y}) - B^*(\boldsymbol{\eta})\right\}.$$

Seega valimijaotus kuulub samasse jaotuste perre, mis üldkogumi jaotus, kuid teiste kanooniliste parameetritega, $\eta_j^* = \eta_j + p_j$, tingimusel et $\boldsymbol{\eta}^* \in \mathbf{H}$. Samasugune seos kehtib ka ühemõõtmelise eksponentsiaalse pere korral (Pfeffermann, Krieger, Rinott, 1998).

Teades valimijaotuse kanooniliste parameetrite vektorit $\boldsymbol{\eta}^* = (\eta_1^*,...,\eta_m^*)$ ja kaasamistõenäosusi indekseerivate parameetrite vektorit $\mathbf{p} = (p_1,...,p_m)$, saame nüüd analüütiliselt avaldada üldkogumi kanoonilised parameetrid $\boldsymbol{\eta}$, ja järelikult ka teised parameetrid nagu jaotuse keskväärtus, dispersioon, korrelatsioon või kovariatsioon. Probleemiks osutub see, et seos teiste parameetrite ja kanooniliste parameetrite vahel võib olla üsna keeruline ja meid huvitava kovariatsiooni avaldamine valimijaotuse kanooniliste parameetrite vektorist ei ole sugugi lihtne ülesanne. Järgmine näide illustreerib seda probleemi.

***Näide* 4** (normaaljaotus). Uurime kahemõõtmelist normaaljaotust, uuritavad tunnused tähistame $\mathbf{y} = (y, z)$. Kui mitte kasutada maatrikskuju, siis sellel jaotusel on viis parameetrit: kaks keskväärtust (μ_y ja μ_z), kaks hälvet (σ_y ja σ_z) ja korrelatsioon (*r*). Parameetrite vektorit tähistame $\boldsymbol{\theta} = (\mu_y, \mu_z, \sigma_y, \sigma_z, r)$. Jaotuse tihedusfunktsioon avaldub kujul:

$$f_p(y, z \mid \boldsymbol{\theta}) = \frac{\exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(y-\mu_y)^2}{\sigma_y{}^2} - \frac{2r(y-\mu_y)(z-\mu_z)}{\sigma_y\sigma_z} + \frac{(z-\mu_z)^2}{\sigma_z{}^2}\right]\right\}}{2\pi\sigma_y\sigma_z\sqrt{1-r^2}}. \qquad (4.6)$$

Avades sulud, näeme, et tihedusfunktsioon (4.6) avaldub kanoonilisel kujul (4.5), kusjuures:

$$\begin{aligned}
&h^*(\mathbf{y}) = \frac{1}{2\pi};\\
&\eta_1 = -\frac{1}{2(1-r^2)\sigma_y{}^2},\ T_1(\mathbf{y}) = y^2;\\
&\eta_2 = -\frac{1}{2(1-r^2)}\left(\frac{-2\mu_y}{\sigma_y{}^2} + \frac{2r\mu_z}{\sigma_y\sigma_z}\right),\ T_2(\mathbf{y}) = y;\\
&\eta_3 = \frac{r}{(1-r^2)\sigma_y\sigma_z},\ T_3(\mathbf{y}) = yz; \qquad (4.7)\\
&\eta_4 = -\frac{1}{2(1-r^2)}\left(\frac{-2\mu_z}{\sigma_z{}^2} + \frac{2r\mu_y}{\sigma_y\sigma_z}\right),\ T_4(\mathbf{y}) = z;\\
&\eta_5 = -\frac{1}{2(1-r^2)\sigma_z{}^2},\ T_5(\mathbf{y}) = z^2;\\
&B^*(\boldsymbol{\eta}) = \ln(\sigma_y\sigma_z\sqrt{1-r^2}) + \frac{1}{2(1-r^2)}\left[\frac{\mu_y{}^2}{\sigma_y{}^2} - \frac{2r\mu_y\mu_z}{\sigma_y\sigma_z} + \frac{\mu_z{}^2}{\sigma_z{}^2}\right].
\end{aligned}$$

Kaasamistõenäosuste keskmine olgu eksponentsiaalsel kujul:

$$E_p(\pi \mid \mathbf{y}) = c_0 \exp\left\{a_1 y - a_2 y^2 + b_1 z - b_2 z^2 + cyz\right\}. \qquad (4.8)$$

Eelnevast teame, et valimijaotuseks on nüüd samuti normaaljaotus kanooniliste parameetritega:

$$\begin{aligned}
\eta_1^* &= \eta_1 - a_2, \\
\eta_2^* &= \eta_2 + a_1, \\
\eta_3^* &= \eta_3 + c, \\
\eta_4^* &= \eta_4 + b_1, \\
\eta_5^* &= \eta_5 - b_2.
\end{aligned} \tag{4.9}$$

Teiselt poolt, avalduvad valimijaotuse kanoonilised parameetrid analoogselt üldkogumi kanooniliste parameetritele valimi keskväärtuste $\tilde{\mu}_y$ ja $\tilde{\mu}_z$, standardhälvete $\tilde{\sigma}_y$, $\tilde{\sigma}_z$ ja korrelatsiooni $\tilde{r}$ kaudu järgmiselt:

$$\begin{aligned}
\eta_1^* &= -\frac{1}{2(1-\tilde{r}^2)\tilde{\sigma}_y^{\,2}}, \\
\eta_2^* &= -\frac{1}{2(1-\tilde{r}^2)}\left(\frac{-2\tilde{\mu}_y}{\tilde{\sigma}_y^{\,2}} + \frac{2\tilde{r}\tilde{\mu}_z}{\tilde{\sigma}_y\tilde{\sigma}_z}\right), \\
\eta_3^* &= \frac{\tilde{r}}{(1-\tilde{r}^2)\tilde{\sigma}_y\tilde{\sigma}_z}, \\
\eta_4^* &= -\frac{1}{2(1-\tilde{r}^2)}\left(\frac{-2\tilde{\mu}_z}{\tilde{\sigma}_z^{\,2}} + \frac{2\tilde{r}\tilde{\mu}_y}{\tilde{\sigma}_y\tilde{\sigma}_z}\right), \\
\eta_5^* &= -\frac{1}{2(1-\tilde{r}^2)\tilde{\sigma}_z^{\,2}}.
\end{aligned} \tag{4.10}$$

Asendades võrdustes (4.9) kanoonilised parameetrid nende avaldistega üldkogumis (4.7) ja valemis (4.10), saame järgmise võrrandite süsteemi:

$$
\begin{aligned}
&-\frac{1}{2(1-r^2)\sigma_y{}^2}-a_2=-\frac{1}{2(1-\tilde{r}^2)\tilde{\sigma}_y{}^2},\\
&-\frac{1}{2(1-r^2)}\left(\frac{-2\mu_y}{\sigma_y{}^2}+\frac{2r\mu_z}{\sigma_y\sigma_z}\right)+a_1=-\frac{1}{2(1-\tilde{r}^2)}\left(\frac{-2\tilde{\mu}_y}{\tilde{\sigma}_y{}^2}+\frac{2\tilde{r}\tilde{\mu}_z}{\tilde{\sigma}_y\tilde{\sigma}_z}\right),\\
&\frac{r}{(1-r^2)\sigma_y\sigma_z}+c=\frac{\tilde{r}}{(1-\tilde{r}^2)\tilde{\sigma}_y\tilde{\sigma}_z},\\
&-\frac{1}{2(1-r^2)}\left(\frac{-2\mu_z}{\sigma_z{}^2}+\frac{2r\mu_y}{\sigma_y\sigma_z}\right)+b_1=-\frac{1}{2(1-\tilde{r}^2)}\left(\frac{-2\tilde{\mu}_z}{\tilde{\sigma}_z{}^2}+\frac{2\tilde{r}\tilde{\mu}_y}{\tilde{\sigma}_y\tilde{\sigma}_z}\right),\\
&-\frac{1}{2(1-r^2)\sigma_z{}^2}-b_2=-\frac{1}{2(1-\tilde{r}^2)\tilde{\sigma}_z{}^2}.
\end{aligned}
\tag{4.11}
$$

Sõltuvalt sellest, kas soovime hinnata üldkogumi või valimi parameetreid, saame võrrandite süsteemi (4.11) lahendada kas (μ_y, μ_z, σ_y, σ_z, r) või ($\tilde{\mu}_y$, $\tilde{\mu}_z$, $\tilde{\sigma}_y$, $\tilde{\sigma}_z$, $\tilde{r}$) suhtes.

Lihtsuse mõttes eeldame edasises, et üldkogumi jaotus on standardne, st $\mu_y=\mu_z=0$ ja $\sigma_y=\sigma_z=1$. Süsteemi (4.11) lahend valimiparameetrite jaoks on siis järgmine:

$$
\begin{aligned}
\tilde{r}&=\frac{r+cR}{\sqrt{(1+2a_2R)(1+2b_2R)}}\\
\tilde{\mu}_y&=\frac{R[b_1(r+cR)+a_1(1+2b_2R)]}{(1+2a_2R)(1+2b_2R)-(r+cR)^2}\\
\tilde{\mu}_z&=\frac{R[a_1(r+cR)+b_1(1+2a_2R)]}{(1+2a_2R)(1+2b_2R)-(r+cR)^2}\\
\tilde{\sigma}_y{}^2&=\frac{R(1+2b_2R)}{(1+2a_2R)(1+2b_2R)-(r+cR)^2}\\
\tilde{\sigma}_z{}^2&=\frac{R(1+2a_2R)}{(1+2a_2R)(1+2b_2R)-(r+cR)^2}\\
R&=1-r^2
\end{aligned}
\tag{4.12}
$$

Analoogsed avaldised valimiparameetrite jaoks on võimalik saada ka üldjuhul, kitsendamata üldkogumi parameetrite väärtusi. Mugavam on neid tulemusi esitada maatrikskujul, mida näeme järgmises peatükis. Samas teeme mõned huvitavad järeldused elemendiviisilistest seostest (4.12). □

***Näide* 5** (multinomiaaljaotus). Olgu tunnused $\mathbf{y} = (t, y, z)$ multinomiaaljaotusega parameetritega n ja $\mathbf{p} = (p_t, p_y, p_z)$, st

$$f_p(y \mid n, \mathbf{p}) = \frac{n!}{t!\,y!\,z!}\, p_t^{\,t}\, p_y^{\,y}\, p_z^{\,z}, \quad p_t + p_y + p_z = 1, \quad t + y + z = n.$$

Eeldame, et katsete arv n on teada, tundmatud on tõenäosused p_t, p_y ja p_z. Multinomiaaljaotus kuulub eksponentsialsesse perre, kuna

$$f_p(\mathbf{y} \mid n, \mathbf{p}) = \frac{n!}{t!\,y!\,z!} \exp\{t \log p_t + y \log p_y + z \log p_z\}.$$

Kanooniliste parameetrite vektor on seega $\boldsymbol{\eta} = (\log p_t, \log p_y, \log p_z)$, $T_i(\mathbf{y}) = y_i$, kus $y_1 = t$, $y_2 = y$, $y_3 = z$.

Kui kaasamistõenäosuste keskmine on kujul

$$E_p(\pi \mid \mathbf{y}) = c_0 \exp\{a \cdot t + b \cdot y + c \cdot z\},$$

siis valimis on $\mathbf{y}$ samuti multinomiaaljaotusega kanooniliste parameetritega $\boldsymbol{\eta}^* = (\log p_t + a,\ \log p_y + b,\ \log p_z + c)$. Valimijaotuse traditsioonilised parameetrid ehk tõenäosuste vektor on seega $\mathbf{p}^* = (p_t e^a, p_y e^b, p_z e^c)$, $c = \log(1 - p_t e^a - p_y e^b) - \log p_z$. Tunnuste t ja y vaheline korrelatsioon on üldkogumis ja valimis vastavalt $\rho_p(t, y) = -\sqrt{\dfrac{p_t p_y}{(1-p_t)(1-p_y)}}$ ja $\rho_s(t, y) = -\sqrt{\dfrac{p_t e^a p_y e^b}{(1-p_t e^a)(1-p_y e^b)}}$. Seega näiteks kui a ja b on positiivsed, mis tähendab, et valimisse satuvad objektid suuremate t ja y väärtustega, siis valimis on negatiivne korrelatsioon t ja y vahel tugevam kui üldkogumis. □

5 Kovariatsioonimaatriksi hindamine normaaljaotuse puhul

Eelmises peatükis nägime, et kui üldkogumis on tunnused kahemõõtmelise normaaljaotusega ja kaasamistõenäosused on eksponentsiaalsel kujul, siis on ka valimis tunnused normaaljaotusega, kuid teiste parameetritega. Seda tulemust saab üldistada mitmemõõtmelisele normaaljaotusele kitsendamata tunnuste arvu. Kasutame maatriksesitust, mis lihtsustab avaldisi.

Mitmemõõtmelise normaaljaotuse tiheduse maatrikskuju on:

$$f_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}, \tag{5.1}$$

kus $\boldsymbol{\mu}$: ($k\times1$) on keskväärtuste vektor ja $\boldsymbol{\Sigma}$: ($k\times k$) on kovariatsioonimaatriks. Avades sulud ja kasutades üksnes argumendist $\mathbf{y}$ sõltuvaid liikmeid, saame tiheduse tuuma avaldise:

$$f_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left\{-\frac{1}{2}(\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - 2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{y})\right\}.$$

Objektide kaasamistõenäosused on nagu ka eelmises peatükis eksponentsiaalsel kujul, kuid nüüd esitame need maatriksite abil:

$$E_p(\pi \mid \mathbf{y}) = c_0 \exp(\mathbf{y}'\mathbf{A}\mathbf{y} + \mathbf{b}'\mathbf{y}). \tag{5.2}$$

Siin $\mathbf{b}$ on ($k\times1$) vektor ja $\mathbf{A}$ on ($k\times k$) sümmeetriline maatriks, selline et maatriks $(\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})^{-1}$ on positiivselt määratud, ja c_0 on normeeriv konstant.

Teoreem 2. Kui vektori $\mathbf{y}$ jaotus üldkogumis on mitmemõõtmeline normaaljaotus (5.1) ja objektide kaasamistõenäosused on kujul (5.2), siis $\mathbf{y}$ jaotus valimis on jälle mitmemõõtmeline normaaljaotus keskväärtusega $\boldsymbol{\lambda}$ ja kovariatsioonimaatriksiga $\boldsymbol{\Omega}$:

$$\boldsymbol{\lambda} = (\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{b}) \tag{5.3}$$

$$\boldsymbol{\Omega} = (\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})^{-1} \tag{5.4}$$

Tõestus. Kasutame seost (2.3) ja jätame välja argumendist $\mathbf{y}$ mittesõltuvad konstandid.

$$f_s(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{E_p(\pi \mid \mathbf{y}) f_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{E_p(\pi)} \propto \exp(\mathbf{y}'\mathbf{A}\mathbf{y} + \mathbf{b}'\mathbf{y}) \exp\left\{-\frac{1}{2}[\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - 2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{y}]\right\} =$$

$$= \exp\left\{-\frac{1}{2}[-2\mathbf{y}'\mathbf{A}\mathbf{y} - 2\mathbf{b}'\mathbf{y} + \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - 2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{y}]\right\} =$$

$$= \exp\left\{-\frac{1}{2}[\mathbf{y}'(\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})\mathbf{y} - 2(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1} + \mathbf{b}')\mathbf{y}]\right\}.$$

Seega valimis on $\mathbf{y}$ jälle normaaljaotusega, mille keskväärtuste vektori $\boldsymbol{\lambda}$ ja kovariatsioonimaatriksi $\boldsymbol{\Omega}$ saame kätte võrdustest $\boldsymbol{\Sigma}^{-1} - 2\mathbf{A} = \boldsymbol{\Omega}^{-1}$ ja $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1} + \mathbf{b}' = \boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1}$, ehk $\boldsymbol{\Omega} = (\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})^{-1}$ ja $\boldsymbol{\lambda}' = (\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1} + \mathbf{b}')(\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})^{-1}$. □

Valimi keskväärtuste vektori ja kovariatsioonimaatriksi kujud on kooskõlas näite 4 tulemustega ja võimaldavad teha järgmisi järeldusi üldkogumi ja valimi jaotuste vahekorrast:

Järeldus 1. Valimi kovariatsioonimaatriks (ja järelkult ka tunnuste vahelised korrelatsioonid) sõltub kaasamistõenäosustest vaid läbi maatriksi **A** ehk tunnuste ruutude ja korrutiste kordajatest. Valimi keskväärtuste vektor sõltub nii **A**-st kui **b**-st. Seega muudab informatiivne valik **A** = 0 korral küll keskväärtusvektorit aga mitte sõltuvusstruktuuri.

Järeldus 2. Kui tunnused on üldkogumis sõltumatud, ehk maatriks $\boldsymbol{\Sigma}$ on diagonaalne, siis sõltumatus valimis säilib vaid juhul kui lisaks on ka maatriks **A** diagonaalne ehk kaasamistõenäosuste valemis on tunnuste korrutistele vastavad liikmed võrdsed nulliga. Vastasel juhul muutuvad valimis nii tunnuste keskmised, dispersioonid kui ka kovariatsioonid. See tulemus ühtib kolmanda peatüki tulemusega, et sõltumatuseks peavad kaasamistõenäosused olema faktoriseeritud kujul.

Järeldus 3. Maatriksi **A** sobiva valikuga saab $\boldsymbol{\Omega}$ teha diagonaalseks, mis ütleb, et teatava valikuga saame üldkogumis sõltuvad tunnused muuta valimis sõltumatuteks.

Nüüd esitame mõned näited, mis illustreerivad selle peatüki tulemusi.

***Näide* 6.** Vaatame üldkogumit kolme uuritava tunnusega, $\mathbf{y} = (t, y, z)'$, mis on kolmemõõtmelise normaaljaotusega keskväärtuste vektoriga $\boldsymbol{\mu}$ ja kovariatsioonimaatriksiga $\boldsymbol{\Sigma}$,

$$\boldsymbol{\mu} = \begin{pmatrix} 2 & 4 & 6 \end{pmatrix}', \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 2 & 0 \\ 0.3 & 0 & 3 \end{pmatrix}.$$

Kaasamistõenäosused olgu eksponentsiaalsel kujul (5.2) parameetritega **A** ja **b**,

$$\mathbf{A} = \begin{pmatrix} -0.5 & 0.2 & 0 \\ 0.2 & -0.3 & 0.04 \\ 0 & 0.04 & -0.1 \end{pmatrix}, \ \mathbf{b} = (0.1 \quad 0.1 \quad 0.1)'.$$

Teoreemi 1 põhjal on $\mathbf{y}$ valimis samuti normaaljaotusega keskväärtusega $\boldsymbol{\lambda}$ ja kovariatsioonimaatriksiga $\boldsymbol{\Omega}$,

$$\boldsymbol{\lambda} = \begin{pmatrix} 1.35 & 2.67 & 4.30 \end{pmatrix}', \ \boldsymbol{\Omega} = \begin{pmatrix} 0.58 & 0.35 & 0.16 \\ 0.35 & 1.07 & 0.17 \\ 0.16 & 0.17 & 1.88 \end{pmatrix}.$$

Maatriks $\boldsymbol{\Omega}$ on positiivselt määratud, $|\boldsymbol{\Omega}| = 0.911$. Tulemuste kontrollimiseks viisime läbi väikese simulatsiooni. Genereerisime üldkogumi suurusega N = 10 000 kolmemõõtmelisest normaaljaotusest keskväärtusega $\boldsymbol{\mu}$ ja kovariatioonimaatriksiga $\boldsymbol{\Sigma}$. Antud üldkogumist võtsime 500 valimit suurusega n = 1000 vastavalt kaasamistõenäosustele (5.2) parameetritega **A** ja **b**. Igas valimis arvutasime tunnuste valimikeskmised ja valimikovariatsioonimaatriksid. Keskväärtuste keskmine üle 500 valimi oli $\hat{\boldsymbol{\lambda}}$ ja kovariatsioonimaatriksite keskmine $\hat{\boldsymbol{\Omega}}$,

$$\hat{\boldsymbol{\lambda}} = \begin{pmatrix} 1.39 & 2.76 & 4.40 \end{pmatrix}', \ \hat{\boldsymbol{\Omega}} = \begin{pmatrix} 0.57 & 0.33 & 0.11 \\ 0.33 & 1.05 & 0.09 \\ 0.11 & 0.09 & 1.77 \end{pmatrix}.$$

Hinnangud $\hat{\boldsymbol{\lambda}}$ ja $\hat{\boldsymbol{\Omega}}$ on kooskõlas teoreetiliste tulemustega $\boldsymbol{\lambda}$ ja $\boldsymbol{\Omega}$. □

***Näide* 7.** Vaatame sama üldkogumit, mis näites 6. Kui informatiivne valik oleks selline, et

$$\mathbf{A} = \frac{1}{2}\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 0.592 & -0.148 & -0.059 \\ -0.148 & 0.287 & 0.015 \\ -0.059 & 0.015 & 0.173 \end{pmatrix}, \ \mathbf{b} = (0.1 \quad 0.1 \quad 0.1)',$$

siis tunnuste valimijaotus on 3-dimensionaalne sõltumatu normaaljaotus, kus $\boldsymbol{\lambda} = (0.573 \quad 1.982 \quad 2.053)'$, mis on kooskõlas järeldusega 3.

***Näide* 8.** Huvipakkuv on uurida, kui palju võib kahe tunnuse korrelatsioon valimis erineda üldkogumikorrelatsioonist. Selleks viime läbi veel ühe väikese simuleerimisuuringu. Genereerime üldkogumi suurusega N = 10 000. Tunnused y ja z olgu üldkogumis standardse normaaljaotusega korrelatsiooniga r. Kasutades näite 4 tulemusi vaatame, kuidas muutub valimi korrelatsioonikordaja sõltuvalt r väärtusest. Kasutame eksponentsiaalseid kaasamistõenäosusi (4.7) ja võtame $a_1 = a_2 = b_1 = b_2 = c = 1$, ehk

$$\mathbf{A} = \begin{pmatrix} -1 & 1/2 \\ 1/2 & -1 \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} 1 & 1 \end{pmatrix}'.$$

Tabelis 1 on r üldkogumi korrelatsioonikordaja, $\tilde{r}$ valemite (4.9) abil arvutatud valimi korrelatsioonikordaja ja $\hat{r}$ empiiriline valimi korrelatsioonikordaja (valimimaht 1000 objekti, keskmine üle 1000 korduse).

Tabel 1. Üldkogumi ja valimi korrelatsioonikordaja võrdlus

r	$\tilde{r}$	$\hat{r}$
-1	-1	-1
-0.8	-0.256	-0.252
-0.6	0.018	0.009
-0.4	0.164	0.166
-0.2	0.26	0.27
0	0.333	0.335
0.2	0.397	0.392
0.4	0.463	0.462
0.6	0.544	0.532
0.8	0.674	0.676
1	1	1

Näeme, et nii sõltuvuse suund kui tugevus võib valimis võrreldes üldkogumiga muutuda, negatiivne korrelatsioon võib muutuda positiivseks, sõltumatud tunnused üldkogumis võivad olla sõltuvad valimis. Empiiriline korrelatsioonikordaja $\hat{r}$ kinnitab tuletatud valemite õigsust. □

Lahendades avaldised (5.3) ja (5.4) $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ suhtes saame valemid üldkogumiparameetrite hindamiseks valimiparameetrite abil:

$$\boldsymbol{\Sigma} = (\boldsymbol{\Omega}^{-1} + 2\mathbf{A})^{-1} \tag{5.5}$$

$$\boldsymbol{\mu}' = (\boldsymbol{\lambda}'\boldsymbol{\Omega}^{-1} - \mathbf{b}')(\boldsymbol{\Omega}^{-1} + 2\mathbf{A})^{-1} \tag{5.6}$$

Valemite (5.5) ja (5.6) kasutamist illustreerib järgmine näide.

***Näide* 9.** Võtame ühe näites 6 kasutatud valimitest ja proovime hinnata üldkogumiparameetrid $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$, mis on samuti antud näites 6. Valimimaht on n = 1000. Kõigepealt peame hindama kaasamistõenäosuste parameetrid $\mathbf{A}$ ja $\mathbf{b}$. Oletame, et meil on alust eeldada, et seos kaasamistõenäosuste ja uuritavate tunnuste vahel on eksponentsiaalkujul (5.2). Kuna tegemist on üldistatud lineaarse mudeliga, saame kasutada tuntud meetodeid parameetrite hindamiseks. Seletavateks tunnusteks võtame t, y, z nende

ruudud ja korrutised kahe kaupa. Valimihinnangud suurustele **A** ja **b** saame *R* mooduliga *glm* ja need on antud valimi puhul järgmised:

$$\hat{\mathbf{A}} = \begin{pmatrix} -0.49 & 0.20 & 0.00 \\ 0.20 & -0.30 & 0.04 \\ 0.00 & 0.04 & -0.10 \end{pmatrix},\ \hat{\mathbf{b}} = \begin{pmatrix} 0.09 & 0.09 & 0.13 \end{pmatrix}'$$

Valimi kovariatsioonimaatriks ja keskväärtuste vektor on antud juhul:

$$\hat{\mathbf{\Omega}} = \begin{pmatrix} 0.56 & 0.29 & 0.10 \\ 0.29 & 1.04 & 0.13 \\ 0.10 & 0.13 & 1.83 \end{pmatrix},\ \hat{\boldsymbol{\lambda}} = \begin{pmatrix} 1.40 & 2.79 & 4.44 \end{pmatrix}'.$$

Need on hinnanguteks valimijaotuse kovariatsioonimaatriksile $\mathbf{\Omega}$ ja keskväärtusele $\boldsymbol{\lambda}$.

Asendades valemites (5.5) ja (5.6) **A**, **b**, $\mathbf{\Omega}$ ja $\boldsymbol{\lambda}$ ülal mainitud hinnangutega saame järgmised hinnangud üldkogumi parameetritele:

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 0.99 & 0.31 & 0.19 \\ 0.31 & 1.98 & -0.09 \\ 0.19 & -0.09 & 2.89 \end{pmatrix},\ \hat{\boldsymbol{\mu}} = \begin{pmatrix} 1.90 & 4.10 & 6.02 \end{pmatrix}.$$

Need on lähedased parameetrite $\boldsymbol{\mu}$ ja $\mathbf{\Sigma}$ tõelistele väärtustele. □

6 Rakendus reaalsele andmestikule

Katsetame kovariatsioonimaatriksi hindamist informatiivse valiku tingimustes reaalsetel andmetel. Võrdleme selles töös väljapakutud uusi hinnanguid tavalise disain-kaalutud hinnanguga. Andmestikuna kasutame Eesti Sotsiaaluuringu 2006 andmeid. Sotsiaaluuring on Eesti Statistikaameti poolt korraldatud igaaastane paneeluuring, mille abil kogutakse andmed inimeste sissetulekute ja elamistingimuste kohta. Võrdlemaks tulemusi tõeliste kovaritsioonimaatriksi väärtustega käsitleme olemasolevat valimit (täpsemalt selle alamvalimit) üldkogumina. Täpsemalt kirjeldavad simulatsiooniuuringu ülesehitust järgmised paragrahvid. Uuring on läbi viidud statistikapaketi SAS abil, kasutatud kood on toodud Lisas 1.

6.1 Üldkogumi moodustamine

Eesti Sotsiaaluuring on paneeluuring, mille valim koosneb kuni neljast mittekattuvast erineval ajal uuringusse tulnud alamvalimist. Aastal 2006 kuulus valimisse 6993 leibkonda, millest 3850 sattus esmakordselt valimisse 2004. aastal, 648 2005. aastal ja 2495 2006. aastal. Välitööde käigus õnnestus saada 5680 leibkonna täidetud ankeet. Lihtsuse mõttes käsitleme aga antud uuringus eri aastatel valimisse sattunud leibkondi ühtemoodi, nagu oleksid nad kõik 2006. aastal valitud.

Sotsiaaluuringus, nagu ka paljudes teistes Statistikaameti uuringutes, kasutatakse valikuskeemina suurusega võrdelist valikut. Leibkonnad valitakse isikute kaudu, st esialgu valitakse Rahvastikuregistrist juhuslikult isikud ja siis võetakse valimisse isikute leibkonnad. Tulemusena on leibkonna kaasamistõenäosus võrdeline isikute arvuga leibkonnas. Simulatsioonuuringus kasutame sama valikuskeemi.

Üldkogumi moodustamiseks kasutame 5680 vastanud leibkonna andmeid. Selleks, et simulatsioonis kasutatav üldkogum oleks struktuuri poolest sarnane rahvastikuga, peame osast suurematest leibkondadest lahti saama, kuna algses valimis on suuremad leibkonnad üleesindatud. Kuna leibkonna kaasamistõenäosus on võrdeline leibkonna suurusega (14-aastate ja vanemate isikute arv leibkonnas, tähistame t), siis võtame algsest valimist alamvalimi järgmise skeemi järgi: t-liikmelistest leibkondadest võtame alamvalimisse juhuslikult $1/t$ leibkonda. Seega

- üheliikmelised leibkonnad jätame kõik alles,
- kaheliikmelistest leibkondadest jätame pool jne.

Uuritavateks tunnusteks valime järgmised:

u – hinnanguline minimaalne leibkonna sissetulek, et leibkond ots-otsaga kokku tuleks;

x – leibkonna summaarne netosissetulek 2005. aastal (summeeritud nii kõigi liikmete individuaalsed kui ka leibkonna sissetulekud);

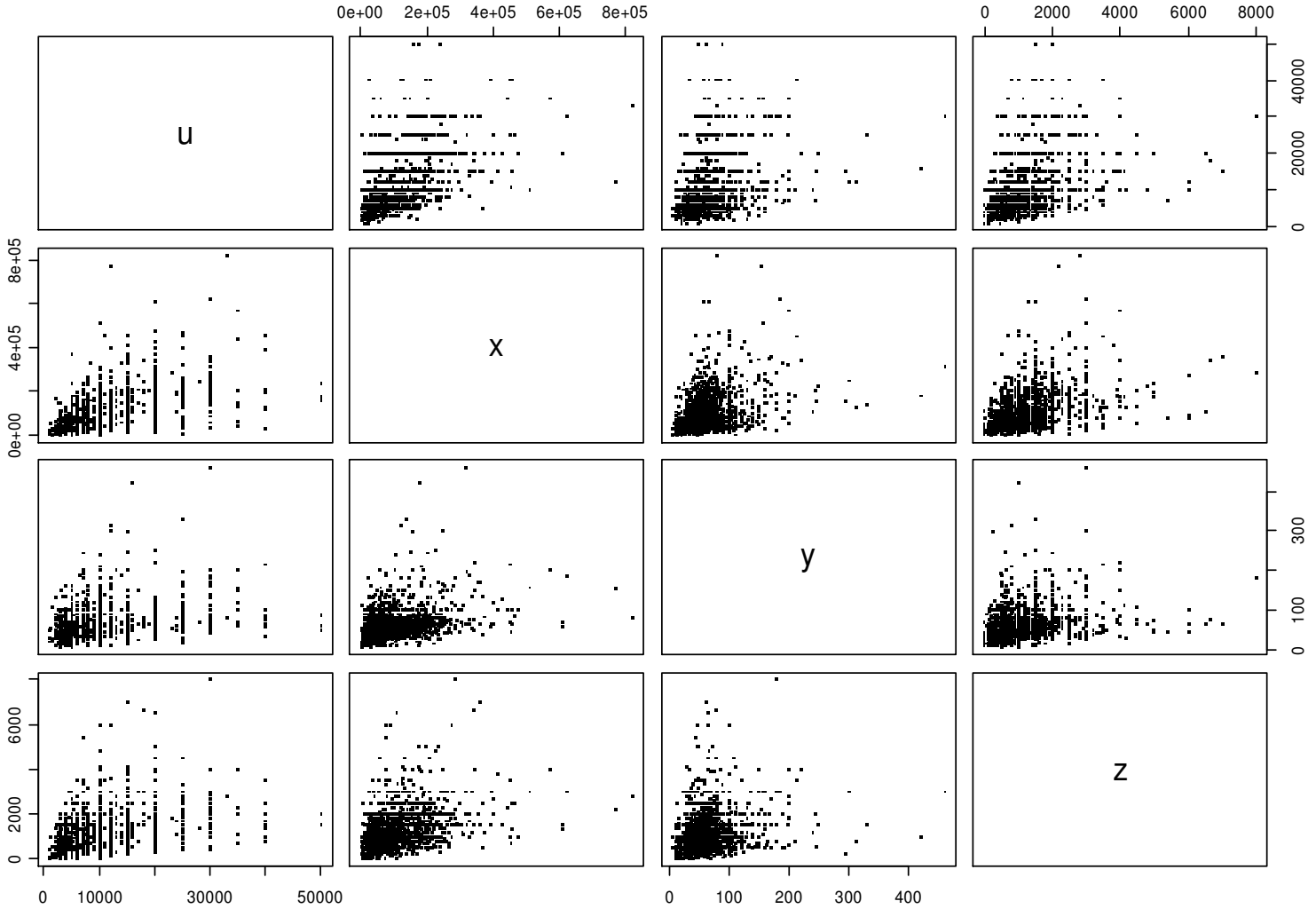
y – leibkonna käsutuses olev pind ruutmeetrites;

z –keskmised kogukulud leibkonna eluruumile kuus.

Kõigi liikmete arvu leibkonnas tähistame q. Paneme tähele, et t ja q võivad peres erineda. Eeldame, et t ei ole uurijale kättesaadav.

Tekitatud alamvalimi võtame üldkogumiks. Pärast mõningate puuduvate väärtustega leibkondade eelmaldamist, jääb meil üldkogumisse 2102 leibkonda. Tunnuste hajuvusdiagrammid kahe kaupa on näidatud joonisel 4.

Joonis 4. Tunnuste hajuvusdiagrammid üldkogumis



6.2 Valimi võtmine ja hindamine

Valimi võtmisel kasutame sarnaselt reaalse uuringuga leibkonnasuurusega t võrdelisi kaasamistõenäosusi. Selle valikuskeemi abil võtame üldkogumist valimi suurusega 500 leibkonda, kasutame protseduuri *surveyselect* paketist SAS ja selle Hanurav-Vijayan algoritmi suurusega võrdeliste tõenäosustega valiku jaoks. Valikuprotseduuri tulemusena saame valimi koos valiku- ehk disainikaaludega. Paneme tähele, et leibkonna kaasamistõenäosus sõltub sissetulekust vaid kaudselt, läbi leibkonna suuruse.

Suvaliste tunnuste t_1 ja t_2 vahelise üldkogumikovariatsiooni hindamiseks valimist kasutame kaalutud valimikovariatsiooni:

$$\mathrm{cov}(t_1, t_2, w) = \frac{1}{c^{korr}} \left(\frac{\sum_s w_i t_{1i} t_{2i}}{\sum_s w_i} - \frac{\sum_s w_i t_{1i}}{\sum_s w_i} \cdot \frac{\sum_s w_i t_{2i}}{\sum_s w_i} \right), \qquad (6.1)$$

kus $w_i, i \in s,$ on kaalud, c^{korr} on nihke parandustegur (4.4). Dispersiooni hindamiseks kasutame samuti avaldist (6.1) võttes $t_1 = t_2$.

Antud simulatsiooniuuringu tarbeks kasutame viit tüüpi kaalusid, milledest kolm viimast on antud töös väljapakutud uued kaalud,:

- võrdsed kaalud $w = 2102/500 = 4.204$, need vastavad juhule kui valikut ignoreeritakse;
- disainkaalud w^d;
- disainkaalude tinglikud keskmised kõikide uuritavate tunnuste ja liikmete arvu q suhtes: $w_i^e(q_i, u_i, x_i, y_i, z_i) = E_s(w_i^d \mid q_i, u_i, x_i, y_i, z_i)$;
- disainkaalude tinglikud keskmised kõikide uuritavate tunnuste kuid mitte liikmete arvu suhtes: $w_i^e(u_i, x_i, y_i, z_i) = E_s(w_i^d \mid u_i, x_i, y_i, z_i)$;
- disainkaalude tinglikud keskmised t_1 ja t_2 suhtes: $w_i^e(t_{1i}, t_{2i}) = E_s(w_i^d \mid t_{1i}, t_{2i})$, kus $t_j = u, x, y$ või z, $j = 1,2$, sõltuvalt sellest, milliste tunnuste kovariatsiooni hetkel arvutatakse.

Kaalude $w^e(t_1, t_2)$ näol on tegemist sisuliselt kuue erineva kaaluga. Näiteks tunnuste u ja x vahelise kovariatsiooni hindamisel kasutame kaalu $w^e(u, x)$ jne. Dispersiooni hindamisel võetakse tinglik keskväärtus vaid ühe tunnuse suhtes.

Kaalude $w^e(s,u,x,y,z)$, $w^e(u,x,y,z)$ ja $w^e(t_1,t_2)$ leidmiseks peame koostama mudeli kaalu w^d prognoosimiseks uuritavate tunnuste abil. Kaal w^e on selle mudeli prognoositav väärtus. Üldiselt võiksime siin koostada kaheksa erinevat mudelit, kuid lihtsuse mõttes kasutame ühte mudeli ülesehitust, muudame vaid seletavaid tunnuseid.

Kuna kaalude jaotus andmestikus ei ole pidev, st esineb vaid seitse erinevat väärtust, siis kõige loogilisem valik tundub olevat multinomiaalne logistiline mudel. Mudelisse kaasame nii peaefektid (tunnused ühe kaupa) kui ka kõik koosmõjud (tunnuste korrutised kahe, kolme, nelja ja viie kaupa). Selle mudeli abil saame arvutada igale objektile seitse tõenäosust, mis vastavad igale kaalu väärtusele. Kaalu väärtus, mis saab kõige suurema tõenäosuse ongi mudeli prognoositav väärtus. Asendades mudelisse vastavad seletavad tunnused, saame kaalud $w^e(s,u,x,y,z)$, $w^e(u,x,y,z)$ ja $w^e(t_1,t_2)$. Valemi (6.1) abil saame nüüd arvutada kovariatsioonide hinnangud.

Hinnangute täpsuse iseloomustamiseks vaatame nende keskmist suhtelist nihet ja ruutkeskmist viga. Esimene on hinnangu keskmine erinevus tõelisest väärtusest väljendatud suhtena tõelisse väärtusesse, teine on erinevuste ruutude keskmine. Tähistades tõelist väärtust θ-ga, ja selle hinnangut i-ndal iteratsioonil $\hat{\theta}_i$-ga, saab need karakteristikud kirja panna järgmiselt (g on iteratsioonide arv):

keskmine suhteline nihe: $\frac{1}{g}\sum_{i=1}^{g}\frac{\hat{\theta}_i-\theta}{\theta}$,

ruutkeskmine viga: $\frac{1}{g}\sum_{i=1}^{g}(\hat{\theta}_i-\theta)^2$.

6.3 Tulemused

Kirjeldatud skeemi järgi võtsime üldkogumist 1000 valimit ja igaühest arvutasime tunnuste kovariatsioonid viit tüüpi kaalude abil. Tulemusi võtavad kokku tabelid 2 ja 3.

Kõigepealt märgime, et valikut ignoreerivate kaalude *w* puhul on hinnangud nihkega (välja arvatud kovariatsioon *u* ja *x* ning *x* ja *z* vahel). Erinevate valikut arvestavate kaalude abil saadud hinnangud käituvad aga üsna sarnaselt. Kõikide hinnangute nihked on väikesed ja ühesuunalised, mis on oodatav tulemus, kuna hinnangud on omavahel hästi korreleeritud. Mõnel juhul õnnestus uute kaaludega nihet veelgi vähendada. Nihke suund on teine vaid kaalude $w^e(t_1,t_2)$ puhul, kuid ka siin on nihe absoluutväärtuselt väike. Seevastu õnnestus hinnangute hajuvust uute kaaludega märkimisväärselt vähendada võrreldes disainkaalude

juhuga, eriti kehtib see kaalude $w^e(u, x, y, z)$ ja mõnel juhul ka $w^e(t_1, t_2)$ kohta. Kaalud $w^e(u, x, y, z)$ tunduvad antud simulatsiooniuuringus olevat optimaalne valik, kuna nad parandavad hinnangut kõigi kuue kovariatsiooni ja nelja dispersiooni puhul. Seega sobivad need kaalud kovariatsioonimaariksi hindamiseks tervikuna. Võit kaalude $w^e(t_1, t_2)$ puhul on mõnel juhul isegi suurem, kuid mitte alati: *x*-iga seotud kovariatsioonide cov(*u*,*x*), cov(*x*,*y*) ja cov(*x*,*z*) hinnangud ei paranenud.

Reaalses uuringus ei ole selline võrdlemine muidugi võimalik, kuid, nagu käesolev simulatsiooniuuring näitab, on kaalude ja uuritavate tunnuste seose modelleerimisega võimalik tõsta kovariatsioonihinnangute täpsust. Isegi kui ei õnnestu tõsta kõigi hinnangute täpsust, ei tee see halba.

Tabel 2. Kovariatsioonihinnangu keskmine suhteline nihe (1000 simulatsiooni)

Kaalud	Cov(u, x)	Cov(u, y)	Cov(u, z)	Cov(x, y)	Cov(x, z)	Cov(y, z)	Var(u)	Var(x)	Var(y)	Var(z)
w^d	-0.05	-0.06	-0.04	-0.04	-0.04	-0.04	-0.01	-0.01	0.01	0.01
$w^e(s,u,x,y,z)$	-0.05	-0.06	-0.03	-0.05	-0.04	-0.05	-0.02	-0.02	-0.01	0.00
$w^e(u,x,y,z)$	-0.05	-0.06	-0.03	-0.04	-0.04	-0.04	-0.05	-0.05	-0.03	-0.02
$w^e(t_1,t_2)$	-0.04	-0.03	0.00	0.01	0.02	0.06	-0.10	-0.18	-0.07	0.09
w	0.05	0.16	0.11	0.15	0.04	0.24	0.22	0.19	0.34	0.13

Tabel 3. Kovariatsioonihinnangu ruutkeskmine viga (1000 simulatsiooni)

Kaalud	Cov(u, x)	Cov(u, y)	Cov(u, z)	Cov(x, y)	Cov(x, z)	Cov(y, z)	Var(u)	Var(x)	Var(y)	Var(z)
w^d	$1.76 \cdot 10^{15}$	$1.78 \cdot 10^{8}$	$8.44 \cdot 10^{10}$	$3.16 \cdot 10^{10}$	$1.45 \cdot 10^{13}$	$1.98 \cdot 10^{6}$	$7.29 \cdot 10^{13}$	$1.09 \cdot 10^{18}$	$8.10 \cdot 10^{4}$	$8.25 \cdot 10^{9}$
$w^e(s,u,x,y,z)$	$1.74 \cdot 10^{15}$	$1.79 \cdot 10^{8}$	$8.24 \cdot 10^{10}$	$3.20 \cdot 10^{10}$	$1.48 \cdot 10^{13}$	$1.94 \cdot 10^{6}$	$7.36 \cdot 10^{13}$	$1.06 \cdot 10^{18}$	$7.53 \cdot 10^{4}$	$8.24 \cdot 10^{9}$
$w^e(u,x,y,z)$	$1.53 \cdot 10^{15}$	$1.66 \cdot 10^{8}$	$7.43 \cdot 10^{10}$	$2.78 \cdot 10^{10}$	$1.39 \cdot 10^{13}$	$1.76 \cdot 10^{6}$	$7.27 \cdot 10^{13}$	$1.07 \cdot 10^{18}$	$7.91 \cdot 10^{4}$	$7.80 \cdot 10^{9}$
$w^e(t_1,t_2)$	$1.78 \cdot 10^{15}$	$1.57 \cdot 10^{8}$	$5.21 \cdot 10^{10}$	$3.38 \cdot 10^{10}$	$1.60 \cdot 10^{13}$	$1.47 \cdot 10^{6}$	$7.35 \cdot 10^{13}$	$1.77 \cdot 10^{18}$	$7.28 \cdot 10^{4}$	$9.53 \cdot 10^{9}$
w	$2.35 \cdot 10^{15}$	$4.11 \cdot 10^{8}$	$16.50 \cdot 10^{10}$	$5.05 \cdot 10^{10}$	$1.78 \cdot 10^{13}$	$4.18 \cdot 10^{6}$	$15.70 \cdot 10^{13}$	$2.21 \cdot 10^{18}$	$25.15 \cdot 10^{4}$	$13.86 \cdot 10^{9}$

7 Kokkuvõte

Antud töö põhiteemaks oli tunnustevaheline kovariatsioon informatiivse valiku tingimustes. Nägime, et valiku informatiivsus ei tähenda alati, et tunnustevaheline kovariatsioon valimis erineb üldkogumi kovariatsioonist. Töös on esmakordselt näidatud, et sõltumatus üldkogumis säilib sellise informatiivse valiku puhul ka valimis, kus tunnuste mõjud kaasamistõenäosustele saab üksteisest eraldada, st kus kaasamistõenäosused on esitatavad faktoriseeritud kujul uuritavate tunnuste suhtes. Samas nägime, et informatiivse valiku puhul võib tunnuste sõltuvusstruktuur valimis drastiliselt erineda sõltuvusstruktuurist üldkogumis; valikuga saab tunnused muuta koguni omavahel sõltumatuteks.

Mõnel juhul saab valimikovariatsiooni arvutada üldkogumi jaotust ja kaasamistõenäosusi indekseerivate parameetrite abil analüütiliselt. Antud töös on käsitletud eksponentsiaalse pere juhtu, mis on võrreldes varasemate artiklitega selles vallas laiendatud mitmemõõtmelisele juhule. Samuti on esmakordselt tuletatud mitmemõõtmelise normaaljaotuse ja eksponentsiaalsete kaasamistõenäosuste juhule vastav valimijaotus maatrikskujul.

Praktilise võttena valimikovariatsiooni hindamiseks on töös välja arendatud idee kasutada tavaliste disainkaalude asemel kaalude tinglikke keskmisi uuritavate tunnuste suhtes, mida saab arvutada modelleerides kaasamistõenäosusi uuritavate tunnuste abil. Meetodi testimiseks läbiviidud simulatsiooniuuring reaalsete andmete baasil, näitas selle meetodi potentsiaalset kasulikkust.

Kirjandus

Aru, J. (2004). Regressiooniparameetrite hindamine informatiivse valiku tingimustes. Bakalaureusetöö. Käsikiri Tartu Ülikooli matemaatilise statistika instituudis.

Azzalini, A., Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of Royal Statistical Society* B 61, 579-602.

Eideh, A. A. H., Nathan, G. (2006). The analysis of data from sample surveys under infromative sampling. *Acta et commentationes universitatis tartuensis de mathematica* 10, 41-51.

Grilli, L., Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology* 30, 93-103.

Krieger, A. M., Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology* 18, 225-239.

Lehman, E. L., Casella, G. (1998). Theory of point estimation. Second edition. New Jork, Springer, p.23.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association* 77, 237-250.

Patil, G. P., Rao, C. R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics* 34, 179-189.

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association* 83, 824-833.

Pfeffermann, D., Krieger, A. M., Rinott, Y. (1998) Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* 8, 1087-1114.

Pfeffermann, D., Moura, F. A. D. S., Silva, P. L. D. N. (2006). Multi-level modelling under informative sampling. *Biometrika* 94, 943-959.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society* B 60, 23-40.

Pfeffermann, D., Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *The Indian Journal of Statistics*. Special Issue on Sample Surveys, Volume 61, Series B, 166-186.

Pfeffermann, D., Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In *Analysis of Survey Data*, C. J. Skinner and R. L. Chambers, eds., New York: Wiley, 175-195.

Pfeffermann, D., Sverchkov, M. (2005). Small area estimation under informative sampling. *Statistics in transition* 7, 675-684.

Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: what population does a sample represent? In *A celebration in statistics*, ISI Centenary Volume, A.C. Atkinson and S.E. Fienberg, eds., New York: Springer-Verlag, 543-569.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581-592.

Sugden, R. A. , Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* 71, 495-506.

Sverchkov, M., Pfeffermann, D. (2004). Prediction of final population totals based on sample distribution. *Survey Methodology* 30, 79-92.

Traat, I., Bondesson, L., Meister, K. (2004). Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference.* Volume 132, 395-413.

Influence of informative sampling on dependence between variables

Summary

A purpose of the present paper is to investigate the influence that informative sampling has on the covariance matrix between variables. In case of informative sampling the sampling scheme explicitly or implicitly depends on the response variables. As a result, neither sample distribution of response variables, nor covariance matrix reflects corresponding population counterparts.

Fist, the case of independence between variables in population is considered. It is shown that when inclusion probabilities can be presented in a factorised form with respect to study variables, the independence between variables is preserved in the sample.

In general case covariance matrix can be estimated with common methods modified to take sampling weights into account. As illustrated in the simulation study based on real data, using conditional expectations of weights with respect to study variables instead of ordinary design weights can potentially decrease the variability of estimates.

Possibilities of estimating sample covariance analytically are illustrated on the basis of multivariate exponential family of distributions. It is shown that with a special form of inclusion probabilities multivariate exponential family is invariant under sampling, i.e. sample distribution has the same form as population distribution but with different parameters. Multivariate normal distribution is examined closer and the parameters of sample distribution are derived explicitly in matrix form.

Lisa 1. Simuleerimisuuringu programm

Programm arvutab välja kovariatsioonide ja dispersioonide hinnangud peatükis 6 toodud viit tüüpi kaalude abil. Kasutatakse parandusteguriga kovariatsiooni arvutamise valemit (6.1). Igal sammul võtab programm üldkogumist etteantud mahuga valimit, iga valimi puhul sobitab nelja viimase kaalu arvutamiseks vajalikud mudelid (protseduuri *logistic* abil), arvutab hinnangud, nende suhtelised nihked ja ruutkeskmised vead, ning salvestab tulemused. Peamised kasutatud SASi vahendid on makrokeel ja IML.

```
/*kogu log eraldi faili*/
proc printto log='C:\WINNT\Profiles\juliaa\My Documents\log.txt' new;
run;

/*makro kovariatsioonide arvutamiseks*/
%macro kovar (data, weight, out);
data temp; set &data;
ux=D24*HY020;
uy=D24*C03a;
uz=D24*D04;
xy=hy020*c03a;
xz=hy020*d04;
yz=c03a*d04;
uu=D24*D24;
xx=HY020*HY020;
yy=C03A*C03A;
zz=D04*D04;
kaal2=&weight*&weight;
run;
/*arvutame kaalutud keskmised*/
proc summary data=temp;
weight &weight;
output out=temp2 mean(D24 HY020 C03A D04 ux uy uz xy xz yz uu xx yy zz)=u x y z ux
uy uz xy xz yz uu xx yy zz;
run;
/*arvutame koefitsient nihke vähendamiseks*/
proc summary data=temp;
output out=temp3 sum(&weight kaal2)=sumkaal1 sumkaal2;
run;
data temp3; set temp3; korr=(sumkaal1*sumkaal1-sumkaal2)/sumkaal1/sumkaal1; run;

/*arvutame hinnangud*/
data &out;
merge temp2 temp3 (keep=korr);
cov_ux=(ux-u*x)/korr;
cov_uy=(uy-u*y)/korr;
cov_uz=(uz-u*z)/korr;
cov_xy=(xy-x*y)/korr;
cov_xz=(xz-x*z)/korr;
cov_yz=(yz-y*z)/korr;
cov_uu=(uu-u*u)/korr;
cov_xx=(xx-x*x)/korr;
cov_yy=(yy-y*y)/korr;
cov_zz=(zz-z*z)/korr;
keep cov_ux cov_uy cov_uz cov_xy cov_xz cov_yz cov_uu cov_xx cov_yy cov_zz;
run;
%mend;

/*tõelised väärtused*/
%kovar(mag.frame_lopp, yks, truev);
```

```
/*simulatsioon*/
data mse1; set mag.nullid; run;  /*andmestikud tavaliste kaaludega arvutatud*/
data mean1; set mag.nullid; run; /*hinnangute, nihete ja mse hoidmiseks*/
data kovar1; set mag.nullid; run;

data mse2; set mag.nullid; run;  /*andmestikud (s,u,x,y,z)-kaaludega arvutatud*/
data mean2; set mag.nullid; run; /*hinnangute, nihete ja mse hoidmiseks*/
data kovar2; set mag.nullid; run;

data mse3; set mag.nullid; run;  /*andmestikud (u,x,y,z)-kaaludega arvutatud*/
data mean3; set mag.nullid; run; /*hinnangute, nihete ja mse hoidmiseks*/
data kovar3; set mag.nullid; run;

data mse4; set mag.nullid; run;  /*andmestikud (t1,t2)-kaaludega arvutatud*/
data mean4; set mag.nullid; run; /*hinnangute, nihete ja mse hoidmiseks*/
data kovar4; set mag.nullid; run;

data mse5; set mag.nullid; run;  /*andmestikud SRS-kaaludega arvutatud*/
data mean5; set mag.nullid; run; /*hinnangute, nihete ja mse hoidmiseks*/
data kovar5; set mag.nullid; run;

%macro simu(k);
%do i=1 %to &k;
      /*võtame valimi PPS valikuga*/
      proc surveyselect noprint
            data=mag.frame_lopp
            out=sample
            method=PPS
            sampsize=500;
            size suurus;
      run;

      /***** sobitame mudelid kaaludele *****/
      /*täismudel*/
      proc logistic data=sample noprint;
            class samplingweight;
            model samplingweight=D24|HY020|C03A|D04;
            output out=pred predprobs=I;
      run;
      data pred; set pred; newweight=_into_+0; run;

      /*täismudel leibkonnasuurusega*/
      proc logistic data=sample noprint;
            class samplingweight;
            model samplingweight=BN|D24|HY020|C03A|D04 ;
            output out=pred_bn predprobs=I;
      run;
      data pred_bn; set pred_bn; newweight=_into_+0; run;


      /*mudelid kahe kaupa : u ja x*/
      proc logistic data=sample noprint;
            class samplingweight;
            model samplingweight=D24|HY020;
            output out=pred_ux predprobs=I;
      run;
      data pred_ux; set pred_ux; newweight=_into_+0; run;

      /*mudelid kahe kaupa : u ja y*/
      proc logistic data=sample noprint;
            class samplingweight;
            model samplingweight=D24|C03A;
            output out=pred_uy predprobs=I;
      run;
      data pred_uy; set pred_uy; newweight=_into_+0; run;

      /*mudelid kahe kaupa : u ja z*/
```

```
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=D24|D04;
       output out=pred_uz predprobs=I;
 run;
 data pred_uz; set pred_uz; newweight=_into_+0; run;

 /*mudelid kahe kaupa : x ja y*/
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=HY020|C03A;
       output out=pred_xy predprobs=I;
 run;
 data pred_xy; set pred_xy; newweight=_into_+0; run;

 /*mudelid kahe kaupa : x ja z*/
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=HY020|D04;
       output out=pred_xz predprobs=I;
 run;
 data pred_xz; set pred_xz; newweight=_into_+0; run;

 /*mudelid kahe kaupa : y ja z*/
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=C03A|D04;
       output out=pred_yz predprobs=I;
 run;
 data pred_yz; set pred_yz; newweight=_into_+0; run;

 /*mudelid yhe kaupa : u */
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=D24;
       output out=pred_uu predprobs=I;
 run;
 data pred_uu; set pred_uu; newweight=_into_+0; run;

 /*mudelid yhe kaupa : x */
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=HY020;
       output out=pred_xx predprobs=I;
 run;
 data pred_xx; set pred_xx; newweight=_into_+0; run;

 /*mudelid yhe kaupa : y */
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=C03A;
       output out=pred_yy predprobs=I;
 run;
 data pred_yy; set pred_yy; newweight=_into_+0; run;

 /*mudelid yhe kaupa : z */
 proc logistic data=sample noprint;
       class samplingweight;
       model samplingweight=C03A;
       output out=pred_zz predprobs=I;
 run;
 data pred_zz; set pred_zz; newweight=_into_+0; run;

 /*hinnangud uute kaaludega*/
 %kovar(sample, samplingweight, hinnang1);    * disainkaaludega;
%kovar(pred_bn, newweight, hinnang2);         *(s,u,x,y,z)-kaaludega;
%kovar(pred, newweight, hinnang3);                  *(u,x,y,z)-kaaludega;
```

```
	%kovar(pred_ux, newweight, uus_ux (keep=cov_ux));
	%kovar(pred_uy, newweight, uus_uy (keep=cov_uy));
	%kovar(pred_uz, newweight, uus_uz (keep=cov_uz));
	%kovar(pred_xy, newweight, uus_xy (keep=cov_xy));
	%kovar(pred_xz, newweight, uus_xz (keep=cov_xz));
	%kovar(pred_yz, newweight, uus_yz (keep=cov_yz));
	%kovar(pred_uu, newweight, uus_uu (keep=cov_uu));
	%kovar(pred_xx, newweight, uus_xx (keep=cov_xx));
	%kovar(pred_yy, newweight, uus_yy (keep=cov_yy));
	%kovar(pred_zz, newweight, uus_zz (keep=cov_zz));

	data hinnang4; set uus_ux uus_uy uus_uz uus_xy uus_xz uus_yz uus_uu uus_xx
uus_yy uus_zz; run; *(t1,t2)-kaaludega;

	data sample; set sample; srsweight=2102/500; run; *kaalumata hinnang;
	%kovar(sample, srsweight, hinnang5);

	/*leiame hinnangute erinevused tõelistest väärtustest*/
	proc iml;
	start;
		/*loeme sisse tõelised väärtused ja kõik hinnangud*/
		use truev; read all var _all_ into truev; close truev;
		use hinnang1; read all var _all_ into hinnang1;
			close hinnang1;
		use hinnang2; read all var _all_ into hinnang2;
			close hinnang2;
		use hinnang3; read all var _all_ into hinnang3;
			close hinnang3;
		use hinnang4; read all var _all_ into hinnang4;
			close hinnang4;
		use hinnang5; read all var _all_ into hinnang5;
			close hinnang5;
		hinnang4=t(vecdiag(hinnang4));

/*loeme sisse andmestikud kus hoitakse vastavad hinnangud, nihked ja mse-d*/
		use mean1; read all var _all_ into mean1; close mean1;
		use mse1; read all var _all_ into mse1; close mse1;
		use kovar1; read all var _all_ into kovar1; close kovar1;

		use mean2; read all var _all_ into mean2; close mean2;
		use mse2; read all var _all_ into mse2; close mse2;
		use kovar2; read all var _all_ into kovar2; close kovar2;

		use mean3; read all var _all_ into mean3; close mean3;
		use mse3; read all var _all_ into mse3; close mse3;
		use kovar3; read all var _all_ into kovar3; close kovar3;

		use mean4; read all var _all_ into mean4; close mean4;
		use mse4; read all var _all_ into mse4; close mse4;
		use kovar4; read all var _all_ into kovar4; close kovar4;

		use mean5; read all var _all_ into mean5; close mean5;
		use mse5; read all var _all_ into mse5; close mse5;
		use kovar5; read all var _all_ into kovar5; close kovar5;

		/*arvutame suhtelised nihked*/
		mean1 = mean1 + (hinnang1-truev)/truev;
		mean2 = mean2 + (hinnang2-truev)/truev;
		mean3 = mean3 + (hinnang3-truev)/truev;
		mean4 = mean4 + (hinnang4-truev)/truev;
		mean5 = mean5 + (hinnang5-truev)/truev;

		/*arvutame ruutkeskmised vead*/
		mse1 = mse1 + ((hinnang1-truev)#(hinnang1-truev));
		mse2 = mse2 + ((hinnang2-truev)#(hinnang2-truev));
		mse3 = mse3 + ((hinnang3-truev)#(hinnang3-truev));
		mse4 = mse4 + ((hinnang4-truev)#(hinnang4-truev));
		mse5 = mse5 + ((hinnang5-truev)#(hinnang5-truev));
```

```
            /*salvestame hinnangud*/
            kovar1=kovar1//hinnang1;
            kovar2=kovar2//hinnang2;
            kovar3=kovar3//hinnang3;
            kovar4=kovar4//hinnang4;
            kovar5=kovar5//hinnang5;

/*kirjutame  üle andmestikud, kus hoitakse hinnangud, nihked ja mse-d */
            create mean1 from mean1; append from mean1;
            create mean2 from mean2; append from mean2;
            create mean3 from mean3; append from mean3;
            create mean4 from mean4; append from mean4;
            create mean5 from mean5; append from mean5;

            create mse1 from mse1; append from mse1;
            create mse2 from mse2; append from mse2;
            create mse3 from mse3; append from mse3;
            create mse4 from mse4; append from mse4;
            create mse5 from mse5; append from mse5;

            create kovar1 from kovar1; append from kovar1;
            create kovar2 from kovar2; append from kovar2;
            create kovar3 from kovar3; append from kovar3;
            create kovar4 from kovar4; append from kovar4;
            create kovar5 from kovar5; append from kovar5;

      finish;
      run; quit;
%end;
%mend;

%let k=1000; *simulatsioonide arv;
%simu(&k);

/*jagame simulatsioonide arvuga, et saada keskmised*/
data keskmised;
set mean1 mean2 mean3 mean4 mean5;
col1=col1/&k; col2=col2/&k; col3=col3/&k; col4=col4/&k; col5=col5/&k;
col6=col6/&k; col7=col7/&k; col8=col8/&k; col9=col9/&k; col10=col10/&k;
run;

data vead;
set mse1 mse2 mse3 mse4 mse5;
col1=col1/&k; col2=col2/&k; col3=col3/&k; col4=col4/&k; col5=col5/&k;
col6=col6/&k; col7=col7/&k; col8=col8/&k; col9=col9/&k; col10=col10/&k;
run;

/*eksportime Exceli tabeliteks*/
PROC EXPORT DATA= WORK.keskmised OUTFILE= "U:\Diplom\ESU06\keskmised.xls"
DBMS=EXCEL REPLACE; SHEET="tabel"; RUN;
PROC EXPORT DATA= WORK.vead OUTFILE= "U:\Diplom\ESU06\vead.xls" DBMS=EXCEL
REPLACE; SHEET="tabel"; RUN;
PROC EXPORT DATA= WORK.kovar1 OUTFILE= "U:\Diplom\ESU06\kovar1.xls" DBMS=EXCEL
REPLACE; SHEET="tabel"; RUN;
PROC EXPORT DATA= WORK.kovar2 OUTFILE= "U:\Diplom\ESU06\kovar2.xls" DBMS=EXCEL
REPLACE; SHEET="tabel"; RUN;
PROC EXPORT DATA= WORK.kovar3 OUTFILE= "U:\Diplom\ESU06\kovar3.xls" DBMS=EXCEL
REPLACE; SHEET="tabel"; RUN;
PROC EXPORT DATA= WORK.kovar4 OUTFILE= "U:\Diplom\ESU06\kovar4.xls" DBMS=EXCEL
REPLACE; SHEET="tabel"; RUN;
PROC EXPORT DATA= WORK.kovar5 OUTFILE= "U:\Diplom\ESU06\kovar5.xls" DBMS=EXCEL
REPLACE; SHEET="tabel"; RUN;
PROC EXPORT DATA= WORK.truev OUTFILE= "U:\Diplom\ESU06\truev.xls" DBMS=EXCEL
REPLACE; SHEET="tabel"; RUN;

/*kogu log tagasi logi aknasse*/
proc printto; run;
```