Tartu Ülikool

Matemaatika-informaatikateaduskond

Matemaatilise statistika instituut

Paavo Binsol

Hindamine osakogumites abiinformatsiooni olemasolul

Bakalaureusetöö

Juhendaja:
Natalja Lepik

Tartu
2013

Sisukord

Sissejuhatus

Üldkogumi gruppide ehk osakogumite efektiivne hindamine on oluline ülesanne paljudes tänapäeva statistika uuringutes ja firmades. Traditsiooniline lähenemine sääraste hinnangute leidmiseks on otsesed hinnangud (*direct estimates*). Võib aga juhtuda, et valimimahud osakogumites on väga väikesed, mille tõttu otseste hinnangute varieeruvus muutub väga suureks. Esineb olukordi, kus valimisse ei sattu mõne osakogumi korral ühtegi vaatlust, siis pole otsest hinnangut isegi võimalik leida. Väikeste osakogumite hinnangute teooria (*Small area estimation methods theory)* tegeleb selliste probleemide uurimisega. (Saei & Chambers, 2003)

Lahenduseks kasutatakse mudelipõhiseid ehk mitteotseseid hinnanguid. Sageli on üldkogumi kohta teada abiinformatsioon (*auxiliary information*), mida on võimalik kasutada väikeste osakogumite hinnangute täpsuse parandamiseks. Selliste meetodite kasutamist on statistilises kirjanduses tõlgendatud, kui "jõu laenamisena" uuritava tunnuse ja abitunnuste vahelisest seosest (Saei & Chambers, 2003, lk 2). Siin töös on mudelipõhiste hinnangutena kasutatud GREG-i (*Generalized regression estimator*) ja segamudelit (*Mixed models*).

Käesoleva töö uuritavaks parameetriks on kindla tunnuse kogusumma osakogumis ning eesmärgiks ongi kirjeldada ja uurida, millised meetodid annavad kõige täpsema hinnangu. Samuti on eeldatud, et valimi võtmisel on kasutatud kahte valikudisaini, lihtsat juhuslikku valikut tagasipanekuta ja tagasipanekuga. Huvipakkuv on see, kas erinevate valikudisainide puhul võivad tulemused märgatavalt erineda? Mudelipõhiste hinnangute tõhususe uurimiseks on võrdlusena leitud Horwitz-Thompsoni ja Hansen-Hurwitzi hinnanguid, mille omavaheline erinevus seisneb ainult valimi võtmise meetodis. Kahe viimase nimetatud hinnangu puhul ei kasutata abi informatsiooni.

Töö on üle ehitatud järgmiselt. Esimeses peatükis esitatakse vajalikud esialgsed terminid ja teoreemid, millele hiljem toetuda. Teine kuni neljas peatükk iseloomustavad töös kasutatavaid hinnanguid (Horvitz-Thompson, Hansen-Hurwitz, GREG, segamudelil põhinev hinnang). Samuti on kirjeldatud, kuidas neid hinnanguid saab rakendada osakogumite hindamisel.

Viiendas peatükis on võrreldud nelja hinnangut simulatsioonülesandes, kus kasutatud andmestik on moodustatud ühe Kanada hüpoteetilise küla andmete põhjal. Hinnangute headuse võrdlemiseks on defineeritud täpsusnäitajad, mis arvutatakse simulatsioonis kasutatud andmestiku põhjal. Simulatsiooni läbiviimiseks ja tulemuste illustreerimiseks kasutati statistikapaketti R ning Microsoft Excelit. Lisas on esitatud hinnangute ning täpsusnäitajate leidmise ja andmestiku moodustamise R-i kood. Samuti on esitatud tööga kaasasoleval CD-l algne andmestik ning programm.

1. Valikudisainid ja hindamisteoreemid

Käesoleva bakalaureusetöö kasutatavateks valikudisainideks on lihtne juhuslik valik (LJV) tagasipanekuta (TTA) kui ka tagasipanekuga (TGA). Tagasipanekuta valiku korral ei panda juba valimisse valituks osutunud ja mõõdetud objekti üldkogumisse tagasi. Sellest tulenevalt ei saa järgnevatel valikusammudel seda objekti valida. Tagasipanekuga valiku korral valitud objekti andmed lisatakse valimisse, aga objekt jääb üldkogumisse. Seega, tagasipanekuga valiku korral võib sama objekti valida mitu korda ning ühed ja samad andmed võivad valimisse sattuda korduvalt. Rakendades kahte valikudisaini saab võrrelda saadud tulemuste erinevusi või sarnasusi. Hinnangute ning nende varieeruvuse leidmisel tuleb kasutada vastavatele disainidele kohaseid valikukarakteristikuid hindamisteoreemides, mis on defineeritud järgmises peatükis.

1.2. Valikudisaini karakteristikud

Järgmised põhilised mõisted ja valikudisainid on esitatud Traadi ja Inno põhjal (1997). Valikudisain on fundamentaalse tähtsusega mõiste valikuteoorias ja sellega on määratud kõigi hinnangute statistilised omadused. Disainide optimaalsete hinnangute konstrueerimiseks ja statistiliste omaduste esitamiseks ei kasutata otseselt disaini ennast, vaid selle karakteristikuid: kaasamis ja valikutõenäosusi.

Definitsioon 1.

Üldkogumi objekti i $(i = 1, 2, \ldots, N)$ kaasamistõenäosuseks π_i nimetatakse tõenäosust, millega see objekt kaasatakse valimisse antud disaini $p(s)$ korral.

Definitsioon 2.

Üldkogumi objekti i $(i = 1, 2, \ldots, N)$ valikutõenäosuseks p_i nimetatakse tõenäosust, millega seda objekt võidakse valida antud disaini ühel valikusammul.

Definitsioon 3.

Kaasamisindikaator I_i on iga üldkogumi objekti i $(i = 1, 2, \ldots, N)$ jaoks määratud binaarne juhuslik suurus, mis iseloomustab objekti kaasamist valimisse. TTA disainide korral, I_i on 1, kui i kaasatakse valimisse ja 0 muidu.

1.3. Lihtne juhuslik valik tagasipanekuta

Olgu antud üldkogum $U = (1,\ldots,N)$. Kõigi n mahuliste hulkade arv, mida U-st saame moodustada on C_N^n. Nende hulkade hulk on lihtsa juhusliku valiku kõigi valimite hulk $S = (s_1,\ldots,s_M)$, kus $M = C_N^n$. Vastava valikudisaini valimiteks on hulkvalimid, milles objektide järjestusel pole tähtsust. Kõikidel valimitel s_i on võrdne võimalus realiseerida. TTA disainide korral on I_i Bernoulli jaotusega ehk $I_i \sim Be(\pi_i)$.

Vajaminevad karakteristikud avalduvad kujul

$$EI_i = \frac{n}{N} \ (= f, \text{ mida nimetatakse valikusuhteks}),$$

$$\Delta_{ii} = VI_i = \frac{n}{N}(1-\frac{n}{N}) \ \forall i \neq j,$$

$$\Delta_{ij} = Cov(I_i, I_j) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -f\,\frac{1-f}{N-1}.$$

1.4. Lihtne juhuslik valik tagasipanekuga

Tagasipanekuga lihtsa juhusliku valiku korral saadakse järjestatud valim $s = (i_1,\ldots,i_k,\ldots,i_n)$, kus i_k on k-ndal sammul valitud objekt ja n on valimimaht. Iga objekt võib valimisse sattuda mitu korda, sest valik toimub igal sammul esialgsest üldkogumist. Igal valiku sammul on kõigil üldkogumi objektidel võrdne valikutõenäosus $p_i = \frac{1}{N}$, kus $i = 1,\ldots,N$. TGA disainide korral $I_i \sim B(n, p_i)$.

Vajaminevad karakteristikud avalduvad kujul

$$EI_i = np_i = \frac{n}{N}$$

$$\Delta_{ii} = V(I_i) = np_i(1-p_i)) = \frac{n}{N}(1-\frac{1}{N}) = \frac{n}{N}$$

$$\Delta_{ij} = Cov(I_i, I_j) = -np_i p_j = -\frac{n}{N^2}.$$

1.5. Hindamisteoreemid

Traadi ja Inno põhjal (1997) defineerin töös rakendatavad teoreemid.

Teoreem 1 (Üldine hindamisteoreem)

Üldkogumi kogusumma $t = \sum_{i=1}^{N} y_i$ nihketa hinnang on

$$\hat{t} = \sum_{i \in U} w_i y_i \,, \tag{1.1}$$

kus

$$w_i = \frac{I_i}{E(I_i)} \,. \tag{1.2}$$

Selle disainipõhine dispersioon on

$$V(\hat{t}) = \sum_{i \in U} \sum_{j \in U} \Delta_{ij} \breve{y}_i \breve{y}_j \,,$$

kus $\Delta_{ij} = Cov(I_i, I_j)$. Dispersiooni nihketa hinnanguks $E(I_i I_j) > 0$ korral on

$$\hat{V}(\hat{t}) = \sum_{i \in U} \sum_{j \in U} \breve{\Delta}_{ij} \breve{y}_i \breve{y}_j I_i I_j \,,$$

kus

$$\breve{\Delta}_{ij} = \frac{\Delta_{ij}}{E(I_i I_j)} \,.$$

Teoreem 2 (Alternatiivne hindamisteoreem)

Fikseeritud mahuga disaini $p(s)$ korral saab kogusumma hinnangu

$$\hat{t} = \sum_{i \in U} w_i y_i$$

dispersiooni esitada kujul

$$V(\hat{t}) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} \Delta_{ij} (\breve{y}_i - \breve{y}_j)^2$$

ja eeldusel, et $E(I_i I_j) > 0$, on dispersiooni $V(\hat{t})$ nihketa hinnanguks

$$\hat{V}(\hat{t}) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} I_i I_j \breve{\Delta}_{ij} (\breve{y}_i - \breve{y}_j)^2 \,.$$

2. Osakogum

Osakogumiks nimetatakse üldkogumi alamhulka U_d, $U_d \subset U$, kus U tähistab üldkogumit. Osakogumi mahtu tähistatakse N_d-ga ning osakogumis on sama tüüpi objektid nagu üldkogumis. Osakogumid on määratud mingi tunnuse järgi, mille väärtused peavad olema teada terves üldkogumis. Näiteks, osakogumi võivad moodustada lastega pered, kui vaadeldavaks üldkogumiks on pered või leibkonnad ning kõrgharidusega inimesed, kui üldkogumiks on inimesed. (Traat & Inno, 1997)

Siin töös võetakse valim tervest üldkogumist ja seetõttu võib väikse osakogumi puhul sattuda valimisse vähe objekte. Sel juhul ka hinnangud osakogumites ei ole väga täpsed. Väikeste osakogumite jaoks on välja töötatud omad spetsiaalsed hindamismeetodid (*Small area estimation methods*), mis põhinevad modelleerimisel. Need mudelid püüavad valimi väiksust kompenseerida teiste teadaolevate andmetega.

Kuna valikuuringute valdkonnas enamus huvipakkuvatest parameetritest avaldub uuritava tunnuse väärtuste summa kaudu, siis ka antud töös keskendutakse osakogumi kogusummale $t_d = \sum_{i \in U_d} y_i$ ja uuritakse selle erinevaid hinnanguid.

Osakogumi valim on üks osa terve üldkogumi valimist. Selles töös eeldatakse, et kõigepealt võetakse valim üle üldkogumi ja siis uuritakse, kas või kui palju objekte valimist kuuluvad uuritavasse osakogumisse.

2.1. Horvitz – Thompsoni hinnang osakogumi kogusummale

Horvitz – Thompsoni (HT) ehk ka π-hinnang on üks kõige lihtsamaid hinnanguid osakogumi kogusumma leidmiseks. Seda saab leida ainult siis, kui valim sisaldab uuritava osakogumi objekte. Nihketa t_d hinnang osakogumile U_d põhineb valemil (1.1) ja on järgmisel kujul:

$$\hat{t}_d = \sum_{i \in s} w_i y_i' \tag{2.1}$$

kus:

- $y_i' = z_i^d y_i$ ning z_i^d on indikaator, mis näitab kas objekt kuulub uuritavasse osakogumisse. Kui objekt kuulub osakogumisse U_d, siis z_i^d=1 ja vastasel korral $z_i^d = 0$;
- w_i on *i*-nda objekti kaal valimis (1.2);
- *s* on valim.

Teoreemis 1 on defineeritud nihketa hinnang $\hat{t}$-le, mis kehtib nii TTA kui ka TGA disainide korral. Nüüd on osakogumi tunnuseks y' ja hinnang (2.1) on täpselt sama nihketa hinnang Üldisest hindamisteoreemist. TTA disainide jaoks nimetatakse seda HT hinnanguks (olenemata sellest, kas ta on rakendatud osakogumi tunnusele või üldkogumi tunnusele).

Kuna LJV on fikseeritud mahuga disain, siis saab vastava disaini $V(\hat{t}_d)$ leidmiseks rakendada Teoreemi 2 ehk Alternatiivset hindamisteoreemi. Kasutades LJV TTA karakteristikuid on HT osakogumi kogusumma hinnangu, $\hat{t}_d$ (2.1) dispersioon avaldatav kujul

$$V(\hat{t}_d) = N^2(1-f)\frac{S_{y'}^2}{n}, \tag{2.2}$$

ja dispersiooni hinnang

$$\hat{V}(\hat{t}_d) = N^2(1-f)\frac{s_{y'}^2}{n} \tag{2.3}$$

Valemis (2.2) kasutatud tunnuse y' dispersioon on arvutatav valemiga

$$S^2_{y'} = \frac{1}{N-1}\sum_{i\in U_d}(y_i - \bar{Y}_d)^2 \tag{2.4}$$

ja valemis (2.3) kasutatud y' valimidispersioon on leitav valemiga

$$s^2_{y'} = \frac{1}{n-1}\sum_{i\in s_d}(y - \bar{y}_d)^2. \tag{2.5}$$

Vastavad keskmised, mis on valemites (2.4) ja (2.5) avalduvad kujudel

$$\bar{Y}_d = \frac{1}{N}\sum_{U_d} y_i$$

ja

$$\bar{y}_d = \frac{1}{n}\sum_{s_d} y_i .$$

2.2. Hansen – Hurwitzi hinnang osakogumi kogusummale

Tagasipanekuga disainide puhul nimetatakse Teoreemis 1 defineeritud nihketa hinnangut (1.1) Hansen-Hurwtiz (HH) hinnanguks. Osakogumi kogusumma hinnang avaldub samal kujul nagu TTA disainide puhul, mis on kujutatud valemis (2.1). Kahe hinnangu erinevus on, et TGA disainide korral võib üks objekt sattuda valimisse mitu korda. Huvipakkuv on, kas erinevate valikudisainide korral analoogiliste hinnangute tulemused erinevad märgatavalt.

Kogusumma hinnangu dispersiooni leidmiseks kasutatakse Teoreemi 1. Viimast teoreemi saab kasutada ka osakogumi korral. LJV TGA korral saab $\hat{t}_d$ dispersiooni avaldada kujul:

$$V(\hat{t}_d) = \frac{N(N-1)}{n} S_{y'}^2,$$

kus, $S_{y'}^2$ on defineeritud valemis (2.4).

HH hinnangu dispersiooni hinnangu leidmiseks on kasutatud Teoreemi 2 ehk Alternatiivset hindamisteoreemi, mida rakendatakse fikseeritud valimimahuga disainide puhul. Kasutades LJV TGA karakteristikuid on osakogumi kogusumma hinnangu dispersiooni hinnang avaldatav kujul:

$$V(\hat{t}_d) = \frac{N(N-1)}{n} s_{y'}^2$$

kus $s_{y'}^2$ on defineeritud valemis (2.5).

Viimase tulemusega saab võrrelda HH hinnangu varieeruvust teiste hinnangute varieeruvusega.

3. GREG (*Generalized regression estimator*)

3.1. Üldine kuju

Üldise regressiooni hinnangu (GREG) kirjeldamisel on kasutatud tööd Lepik (2011). Üldkogumi kogusumma hindamiseks kasutab GREG abiinformatsiooni, mis tõstab hinnangu täpsust. Abiinformatsioon võib tulla näiteks registritest või eelnevalt läbiviidud uuringutest, kas abitunnuste näol või agrigeeritud kujul (kogusummad).

Oletame, et kättesaadavad on p abitunnust. GREG hinnang kasutab lineaarset seost uuritava tunnuse ja abitunnuste vahel,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

,kus:

- $\boldsymbol{X}$ on abitunnuste maatriks mõõtmetega $N \times p$;
- $\boldsymbol{\beta}$ on tundmatute parameetrite vektor mõõtmetega $p \times 1$;
- $\boldsymbol{\varepsilon}$ on juhuslike vigade vektor.

Definitsioon 4. Ütleme, et üldkogumist kirjeldab regressioon mudel ξ, kui iga i, $i = 1, \ldots, N$ korral kehtivad järgmised tingimused (Traat & Inno, 1997, lk 159) :

- vektor $\boldsymbol{x_i}$ on fikseeritud, mittejuhuslik suurus;
- väärtus y_i on juhusliku suuruse $\boldsymbol{y}_i$ realisatsioon;
- $E_\xi \boldsymbol{y}_i = \boldsymbol{x_i}\beta = \sum_{j=1}^{p} \beta_j x_{ij}$, kus $E_\xi \boldsymbol{y}_i$ on uuritava tunnuse y_i keskväärtus;
- $D_\xi \boldsymbol{y}_i = \sigma^2{}_i$, mis on y_i hälve keskväärtusest.

GREG mudeli eeldused on Definitsioonis 4.

Hinnangu kuju kogusummale, $\hat{t}_{greg}$, on

$$\hat{t}_{greg} = (\boldsymbol{X}\hat{\boldsymbol{\beta}})^T \mathbf{1} + \hat{\boldsymbol{r}}^T \breve{\boldsymbol{I}}, \tag{3.1}$$

kus:

- $\hat{\boldsymbol{\beta}}$ on tundmatute parameetrite hinnangute vektor mõõtmetega $p \times 1$;

- $\mathbf{1}$ on ühtedest koosnev vektor mõõtmetega $N \times 1$;
- $\hat{\boldsymbol{r}}$ on hinnatud mudeli jääkide vektor mõõtmetega $N \times 1$. Jääk leitakse uuritava tunnuse tegeliku väärtuse ja prognoosi vahena: $\hat{\boldsymbol{r}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$, kus $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Valimist saab leida n jääki.
- $\breve{I} = (\breve{I}_1, \ldots, \breve{I}_N)$ on laiendatud valikuvektor, kus $\breve{I}_i = \dfrac{I_i}{\mathrm{E}(I_i)} = w_i$.

GREG hinnangu leidmiseks on vaja leida tundmatute parameetrite hinnangute vektor $\hat{\boldsymbol{\beta}}$. Selle arvutamiseks kasutatakse valimisse kuuluvate objektide uuritava tunnuse ning abitunnuste väärtusi.

Tundmatute parameetrite hinnangute vektor avaldub kujul (Littell, Stroup & Freund, 2002, lk 225)

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right) = \left(\sum_{i \in s} \frac{\boldsymbol{x_i} \boldsymbol{x_i}^T}{\sigma_i^2 EI_i}\right)^{-1} \sum_{i \in s} \frac{\boldsymbol{x_i} y_i}{\sigma_i^2 EI_i}, \tag{3.2}$$

kus:

- $\boldsymbol{x_i}$ on maatriki $\boldsymbol{X}$ i-s rida, ehk abitunnuste vektor i-nda objekti kohta ;
- σ_i^2 on i-nda objekti varieeruvus;
- EI_i on i-nda objekti oodatav valikute arv ($EI_i = \dfrac{n}{N}$ iga $i = 1, \ldots, n$ korral nii LJV TTA kui ka LJV TGA korral);
- s on valim.

3.2. GREG hinnang osakogumi jaoks

Kuna antud töö eesmärgiks on hinnata osakogumi kogusummat, siis GREG-i hinnangu rakendamiseks tuleb üldist valemit (3.1) muuta ning kohandada abitunnuste maatriksit $\boldsymbol{X}$ ning hinnatud jääkide vektorit $\hat{\boldsymbol{r}}$. Olgu GREG-i kogusumma hinnang d-ndas osakogumis tähistatud järgnevalt $\hat{t}_{greg}^{d}$ ning see avaldub kujul

$$\hat{t}_{greg}^{d} = (\boldsymbol{X_d}\hat{\boldsymbol{\beta}})^T \mathbf{1} + \hat{\boldsymbol{r}}_{\boldsymbol{d}}^T \breve{\boldsymbol{I}}. \tag{3.3}$$

Osakogumile vastava abiandmete maatriksi $\boldsymbol{X_d}$ moodustamisel korrutame esmalt maatriksi $\boldsymbol{X}$ kõik read läbi igale reale vastava indikaatortunnusega $z^d{}_i$, mis avaldub seosega

$$\boldsymbol{x_{d_i}} = z^d{}_i \boldsymbol{x_i}, \tag{3.4}$$

kus:

- $i = 1, \ldots, N$;
- $\boldsymbol{x_{d_i}}$ on maatriksi $\boldsymbol{X_d}$ i-s rida;
- $z^d{}_i$ on indikaatortunnus.

Tekkinud uues maatriksis $\boldsymbol{X_d}$ on osakogumisse U_d mittekuuluvate objektide abiandmed asendatud nullidega ehk nad osakogumi kogusumma hinnangule enam mõju ei avalda. Joonisel 1 on eeldatud, et esimene objekt ei kuulu osakogumisse ning i-s ja N-s kuuluvad. Indikaatortunnustega korrutamise tulemusena tekkib soovitud maatriks $\boldsymbol{X_d}$, mille 1. rida koosneb nullidest ning alles on jäänud andmed i-nda ja N-nda objekti kohta.

$$\begin{pmatrix} z^d{}_1 x_{11} & \cdots & z^d{}_1 x_{1p} \\ \vdots & & \\ z^d{}_i x_{i1} & \ddots & z^d{}_i x_{ip} \\ \vdots & & \\ z^d{}_N x_{N1} & \cdots & z^d{}_N x_{Np} \end{pmatrix} \Rightarrow \boldsymbol{X_d} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \\ x_{i1} & \ddots & x_{ip} \\ \vdots & & \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}$$

Joonis 1. Osakogumile vastava abiandmete maatriksi moodustamine

Osakogumile hinnatud mudeli jääkide vektori $\hat{r}_d$ read on samuti korrutatud läbi indikaatortunnustega:

$$\hat{r}_{d_i} = z^d{}_i \hat{r}_i, \tag{3.5}$$

kus:

- $i = 1, \ldots, N$;
- $\hat{r}_{d_i}$ on vektori $\hat{r}_d$ i-s element;
- $\hat{r}_i$ on vektori $\hat{r}$ i-s element.

Jäävad alles ainult osakogumi valimisse kuuluvad jäägid, ülejäänud objektide jäägid on nullid.

Tundmatute parameetrite hinnanguvektor $\hat{\boldsymbol{\beta}}$ (3.2), ühtedest koosnev vektor $\boldsymbol{1}$ ja laiendatud valikuvektor $\breve{\boldsymbol{I}}$ on samad nagu üldjuhul.

3.2.1. Dispersioon ja dispersiooni hinnang

Lepiku (2011) doktoritöö põhjal avaldub GREG-i hinnangu $\hat{t}^{d}_{greg}$ asümptootiline dispersioon kujul

$$\mathrm{AV}(\hat{t}^{d}_{greg}) = \boldsymbol{r_d}^{T} \breve{\boldsymbol{\Delta}} \boldsymbol{r_d}$$

kus:

- $\boldsymbol{r_d}$ on kõikide osakogumi objektide jääkide vektor mõõtmega $N \times 1$;
- $\breve{\Delta} = Cov(\breve{I})$ on laiendatud kovariatsioonimaatriks mõõtmetega $N \times N$, kus $\breve{\Delta}_{ii} = \frac{\Delta_{ii}}{E(I_i^{\,2})} = \frac{V(I_i)}{E(I_i^{\,2})}$ ning $\breve{\Delta}_{ij} = \frac{\Delta_{ij}}{E(I_i I_j)} = \frac{Cov(I_i, I_j)}{E(I_i I_j)}$.

Osakogumi kogusumma dispersiooni ning selle hinnangu leidmisel tuleb karakteristikud asendada vastavalt nende väärtustega, mis valikudisaini kasutati valimi moodustamisel.

Hinnangu $\hat{t}^{d}_{greg}$ dispersiooni hinnang avaldub kujul:

$$\hat{\mathrm{V}}(\hat{t}^{d}_{greg}) = \hat{\boldsymbol{r}}_{\boldsymbol{d}}^{\,T} \breve{\boldsymbol{\Delta}} \hat{\boldsymbol{r}}_{\boldsymbol{d}}$$

kus:

- $\hat{\boldsymbol{r}}_{\boldsymbol{d}}$ on osakogumis valimisse kuuluvate objektide jääkide vektor, mille elementide moodustamine on valemis (3.5).

4. Üldine lineaarne segamudel

Kui üldine lineaarne mudel sisaldab nii fikseerituid kui juhuslikke faktoreid, siis nimetatakse seda üldiseks lineaarseks segamudeliks. Järgnev peatükk iseloomustab segamudelit ning kirjeldab, kuidas segamudelit kasutades on võimalik leida hinnang osakogumi kogusummale.

4.1. Juhuslikud ja fikseeritud faktorid

Fikseeritud faktori puhul:

- on vähe faktori tasemeid;
- kõik faktori tasemed pakuvad iseseisvat huvi ja on valitud mittejuhuslikult;
- kõik faktori tasemed on esindatud andmetes.

Juhusliku faktori puhul:

- on faktori tasemete arv potentsiaalselt väga suur (lõpmatuhulk);
- on andmetes esindatud juhuslik valim faktori tasemetest;
- pakub huvi kõigi tasemete keskmine (andmetes esindamata) mõju.(Kaart, 2012)

Üks olulisi erinevusi fikseeritud ja juhuslike mõjude vahel on eesmärk, mida soovitakse teada vastavate mõjude analüüsis. Fikseeritud faktorite puhul on üldjuhul soov võrrelda ühte faktori taset teisega. Näiteks, meditsiini uuringus tahetakse võrrelda kontrollgrupi ja ravigrupi keskmiste erinevust ning sellest järeldada, kas ravimil on mõju tervisele. Juhusliku faktori puhul ei ole peamiselt huvipakkuvaks ühe faktori taseme objektide keskmise erinevus teise taseme objektide omast. Pigem on soov teada, missugust varieeruvust põhjustab juhuslik faktor uuritavale tunnusele ehk, milline on uuritava tunnuse keskmise varieeruvus juhusliku faktori tasemetel.

Segamudeli näitena võib käsitleda jällegi meditsiiniuuringu läbiviimise juhtu. Uurides isikute tulemuste erinevust kontrollgrupi ja ravigrupi vahel, võidakse mõõta iga isiku mõju ravidoosile või selle mitte saamisele mitu korda. Mõned uuringus osalevad inimesed võivad oma iseärasuse tõttu saada sageli suuremaid tulemusi kui teised, olenemata sellest, kas nad kuuluvad kontroll- või ravigruppi. Testides fikseeritud faktori ehk ravimi mõju, tuleb kontrollida juhuslikusest põhjustatud mõnede objektide erinevust teistest. Eesmärk oleks

kõrvaldada indiviidide tasemete varieeruvus, et testida ravimi mõju. (Littell, Stroup & Freund, 2002, lk 225)

4.2. Mudeli üldkuju

Üldine lineaarne segamudel avaldub kujul

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{4.1}$$

kus :

- $\boldsymbol{y}$ on uuritava tunnuse $N \times 1$ vektor;
- $\boldsymbol{X}$ on teadaolev abitunnuste maatriks fikseeritud faktori tasemete puhul, mõõtmetega $N \times p$;
- $\boldsymbol{Z}$ on teadaolev abitunnuste maatriks juhuslike faktorite tasemete puhul, mõõtmetega $N \times q$;
- $\boldsymbol{\beta}$ on tundmatute parameetrite vektor mõõtmetega $p \times 1$, kus p on fikseeritud faktorite tasemete arv;
- $\boldsymbol{u}$ on tundmatute parameetrite vektor mõõtmetega $q \times 1$, kus q on juhuslike faktorite tasemete arv;
- $\boldsymbol{\varepsilon}$ on juhuslike vigade vektor mõõtmetega $N \times 1$. (Schaeffer, lk 1)

Osakogumite korral sisaldab sageli just maatriks $\boldsymbol{Z}$ erinevate gruppide mõju uuritavale tunnusele. Osakogumite hindamise korral on kasutuses kahte tüüpi segamudeleid - mudelid objekti ja osakogumi tasemel.

Segamudeli kasutamiseks tuleb valemi (4.1) tundmatute parameetrite vektorid $\boldsymbol{\beta}$ ja $\boldsymbol{u}$ hinnata. Selles tulenevalt avaldub $\hat{\boldsymbol{\beta}}$ kujul (Schaeffer, lk 3)

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y},$$

kus:

- eeldame, et V on teada ning avaldub kujul $\boldsymbol{V} = Var(\boldsymbol{y})$.

Juhuslike faktorite tundmatute parameetrite vektori hinnang $\hat{\boldsymbol{u}}$ on leitav kujul (L. R. Schaeffer, lk 6)

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

- kus $Var(\boldsymbol{u}) = \boldsymbol{G}$ on teadaolev positiivselt defineeritud maatriks.

4.3. Mudel objekti tasemel (*Unit level model*)

Väikeste osakogumite puhul on probleemiks piisavalt hea hinnangu leidmine valimi abil ja abiandmete põhjal. Objekti tasemel saab mudelit esitada kujul (SAE package developers, 2007)

$$y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{u} + \varepsilon_i, \tag{4.2}$$

kus:

- $i = 1, \ldots, N$;
- y_i on i-nda objekti uuritav tunnus;
- x_i on valemis (4.1) kasutatud maatriksi $\boldsymbol{X}$ i-ndas rida;
- z_i on valemis (4.1) kasutatud maatriksi $\boldsymbol{Z}$ i-ndas rida;
- ε_i on i-nda objekti uuritav tunnuse juhuslik viga;
- $\boldsymbol{\beta}$ ja $\boldsymbol{u}$ on samad tundmatute parameetrite vektorid, mis valemis (4.1).

Näiteks kui uuritav tunnus on normaaljaotusega, mille keskväärtuseks on $\boldsymbol{x_i}\boldsymbol{\beta} + z^d{}_i\boldsymbol{u}$ ja dispersiooniks $\sigma_e{}^2$ ehk $y_i \sim N(\boldsymbol{x_i}\boldsymbol{\beta} + z^d{}_i\boldsymbol{u}, \sigma_e{}^2)$, siis sageli $z^d{}_i$ on indikaator osakogumile $\boldsymbol{U}_d$, mis lisab keskväärtusele osakogumi efekti. Sellest tulenevalt avaldub i-ndale objektile vastava osakogumi poolt tingitud mõju. Juhuslik efekt on normaaljaotusega $\boldsymbol{u} \sim N(0, \sigma_u{}^2)$, kus $\sigma_u{}^2$ on juhusliku efekti varieeruvus.(SAE package developers, 2007, lk 19)

Antud töös on eesmärgiks leida osakogumi kogusumma t_d hinnang. Selle hinnangu leidmiseks kasutatakse sünteetilist hinnangut (*synthetic estimator*). Vastav hinnang baseerub eeldusel, et väärtused, mis ei sisaldu valimis on võimalik hinnata (lineaarse) mudeliga

kasutades abiinformatsiooni. Sellest tulenevalt on *d*-nda osakogumi kogusumma hinnang avaldatav järgmisel kujul (SAE package developers, 2007, lk 8)

$$\hat{t}_{SYNTH}^{d} = \sum_{i=U_d} \hat{y}_i \,, \tag{4.3}$$

kus:

- $\hat{y}_i$ on prognoos väärtusele y_i, mis on arvutatud valemi (4.2) põhjal kasutades hinnatuid $\hat{\boldsymbol{\beta}}$ ja $\hat{\boldsymbol{u}}$ vektoreid.

4.4. Mudel osakogumi tasemel (*Area level model*)

Mudel osakogumi tasemetel hindamiseks on kasutatav, kui abiandmed on kätte saadavad ainult agrigeeritud kujul (kogusummade näol). See võib olla põhjustatud sellest, et registrites, kust abiandmeid saadakse, pole informatsiooni iga objekti kohta eraldi.

Olgu *d*-nda osakogumi kogusumma tähistatud Y_d-ga. Selle hinnangu leidmiseks ning abitunnuste agrigeeritud kuju kasutamise tõttu defineerime vektori $\boldsymbol{x}$ ja maatriksi $\boldsymbol{Z_D}$.

Vektor $\boldsymbol{x}$ sisaldab abiandmeid summeeritud kujul p tunnuse jaoks. Valemis (4.1) on defineeritud maatriks $\boldsymbol{X}$, mis sisaldab tunnuste väärtusi iga objekti jaoks eraldi. Praegusel juhul on teada ainult abitunnuste väärtused kõigi objektide peale summeerituna. Sellest tulenevalt leitakse mudel, millega on võimalik arvutada uuritav parameetri väärtus ja iga objekti tulemus pole esmatähtis.

Olgu osakogumite mõju iseloomustav maatriks $\boldsymbol{Z_D}$. Üldkogumis olevate osakogumite arv on tähistatud D-ga ning juhuslike faktorite (tunnuste) arv q-ga. Sellest tulenevalt on $\boldsymbol{Z_D}$ mõõtmetega $D \times q$. Maatriksi $\boldsymbol{Z_D}$ iga rida omab olemasolevate tunnuste väärtusi agrigeeritud kujul iga osakogumi kohta ehk iga rida iseloomustab ühte osakogumit.
Uuritava osakogumi kogusumma, kasutades abiandmeid kogusummade kujul, saab avaldada mudelina kujul

$$Y_d = \boldsymbol{x}\boldsymbol{\beta} + \mathbf{z}_d\boldsymbol{u} + \varepsilon_d \,, \tag{4.4}$$

kus:

- $d = 1,\ldots,D$;
- Y_d on d-nda osakogumi kogu uuritava tunnuse kogusumma;
- $\boldsymbol{x}$ on abiandmete vektor, mõõtmetega $1\times p$;
- $\boldsymbol{z_d}$ on d-s rida maatriksist $\boldsymbol{Z_D}$;
- ε_d on *d*-nda osakogumi kogusumma juhulik viga;
- $\boldsymbol{\beta}$ ja $\boldsymbol{u}$ analoogilised nagu valemis (4.1), aga nende leidmisel kasutatakse vektorit $\boldsymbol{x}$ ja maatriksit $\boldsymbol{Z_D}$.

Valemist (4.4) saame

$$\hat{t}_d = \hat{Y}_d,$$

kus

- $\hat{Y}_d$ on valemist (4.4) leitud väärtus kasutades hinnatud $\hat{\boldsymbol{\beta}}$ ja $\hat{\boldsymbol{u}}$ vektoreid.

Mudelil osakogumite tasemel on d-nda osakogumi jaotus esitatav kujul (SAE package developers, 2007, lk 19)

$$Y_d \sim N(\boldsymbol{x}\beta, \frac{\sigma_e^{\,2}}{n_d + \sigma_u^{\,2}}),$$

kus:

- $\sigma_e^{\,2}$ on juhuslike vigade varieeruvus;
- n_d on d-nda osakogumi maht.

5. Simulatsioon näidisandmestikuga

Antud simulatsiooni eesmärgiks on eelnevalt kirjeldatud hinnangute headuse ja tõhususe võrdlemine. Selleks koostatakse üldkogumit kirjeldav andmestik, mis sisaldab nelja tunnust, millest üks on uuritav ning on kolm abitunnust (rakendatakse mudelipõhisel hindamisel). Valikudisainidena on kasutatud lihtsat juhuslikku valikut (LJV) nii tagasipanekuta kui ka tagasipanekuga. Valimit genereeriti 1000 korda ning iga kord leiti osakogumite hinnangud ja täpsusnäitajad.

Simulatsioon reaalsete andmete peal viidi läbi statistikapaketiga R ning töö lõppu (Lisa 1) on lisatud programmikood, kui lugejal on täpsem huvi selle vastu.

5.1. Andmestik

Andmed põhinevad ühel Kanada hüpoteetilise külal, kus on 36 erineva tunnuse väärtused 1024 leibkonna kohta (Schwarz, 1997). Seda kasutades on moodustatud simulatsioonis kasutatud andmestik, mis sisaldab nelja tunnust 774 perekonna kohta, mis on ka üldkogumi mahuks. Üldkogumi maht on vähenenud 1024-lt 774-le, sest mõningate leibkondade kohta olid andmed puudulikud. Seetõttu mõne puuduva tunnusega leibkonnad kustutati andmestikust.

Uuritavaks tunnuseks oli:

- TOTINCH ehk terve leibkonna kogusissetulek (pidev tunnus) –perekonna kõigi liikmete, kes üle 15 aasta vana, kogu sissetulek 1990-ndal aastal.

Abitunnusteks olid:

- EMPINCH ehk terve leibkonna töötasu (pidev tunnus) – perekonna kõigi liikmete, kes üle 15 aasta, kogu töötasu 1990-ndal aastal;
- VALUEH ehk elamu väärtus (pidev tunnus) – elamu hinnanguline väärtus omaniku poolt, kui see läheks müüki;

Osakogumid moodustati järgneva tunnuse abil:

- HHSIZE ehk leibkonna suurus (diskreetne tunnus) – inimeste arv perekonnas. Võimalikud väärtused olid algses andmestikus 1-8-ni. Suurte leibkondade vähesuse tõttu on ühendatud simulatsioonis kasutatud andmestikus ühte gruppi kõik 4-st suurema suurema liikmelisemad leibkonnad.

Andmeid osakogumi kujul iseloomustab Tabel 1. Uuritava tunnuse ehk leibkondade kogusissetulek osakogumite kaupa on tähistatud t_d-ga. Abitunnused $t^E{}_d$ ja $t^T{}_d$ näitavad vastavalt leibkonna töötasude ning elamute kogusummasid osakogumites. Osakogumite mahud on tähistatud N_d-ga, kus $d = 1,\ldots,5$.

Tabel 1. Andmestiku kirjeldus osakogumite kujul.

d	t_d	N_d	$t^E{}_d$	$t^T{}_d$
1	2842760	90	17730000	1639825
2	12969135	248	57704000	8296144
3	9260723	136	32465000	7865126
4	13584061	185	46625000	11850000
5	9362786	115	31528000	8293654

5.2. Täpsusnäitajad

Hinnangute kvaliteedi ja headuse hindamiseks on leitud nende standardhälbed, suhteline nihe ja suhteline ruutkeskmine viga üle kõigi genereeritud valimite.
Standardhälve avaldub kujul

$$\mathrm{std}(\hat{t}_d) = \sqrt{\frac{\sum_{i=1}^{m}(\hat{t}_d^i - \bar{\hat{t}}_d)^2}{m}}, \tag{5.1}$$

suhteline nihe avaldub kujul

$$\mathrm{RB}(\hat{t}_d) = \frac{\frac{1}{m}\sum_{i=1}^{m}\hat{t}_d^i - t_d}{t_d} \tag{5.2}$$

ja suhtelise ruutkeskmise vea ruutjuur kujul

$$\mathrm{RRMSE}(\hat{t}_d) = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{t}_d^i - t_d)^2}}{t_d}, \tag{5.3}$$

kus:

- m erinevate valimite genereerimise arv;
- $\bar{\hat{t}}_d$ on d-nda osakogumi valimihinnangute keskmine üle m genereeritud valimi korral;
- $\hat{t}_d^i$ on d-nda osakogumi kogusumma valimihinnang arvutatud i-nda valimi pealt;
- t_d on d-nda osakogumi kogusumma üldkogumis.

Eesmärgiks on leida kõige täpsem meetod uuritava tunnuse hinnangu leidmiseks. Täpsusnäitajate kujust on näha, et mõõdetakse hinnangute ja tegelike kogusummade erinevusi arvestades ka kogusumma suurust. Sellest tulenevalt, mida nullile lähemale on väärtused valemites (5.1), (5.2) ja (5.3), seda paremini sobib vastav hinnang reaalsete andmetega. RRMSE ning std on ainult positiivsete väärtustega, aga RB võib saada ka negatiivseid väärtusi.

5.3. Simulatsiooni läbiviimine

Simulatsiooni läbiviimisel kasutatakse kahte valikudisaini ning mõlema disaini puhul on iga genereeritud valimi puhul valimi mahuks $n = 300$. Täpsusnäitajate arvutamiseks on vaja reaalseid väärtusi kasutada. Need on toodud Tabelis 1. Järgnevaks tehti põhiosa, milleks oli hinnangute ja täpsusnäitajate leidmine. Selleks võeti 1000 korda valimit, kasutades esmalt lihtsat juhuslikku valikut tagasipanekuta ja seejärel tagasipanekuga. Iga valikumeetodi korral leiti osakogumi kogusumma hinnangud kasutades HT (2.1), HH (ptk 2.2.), GREG-i (3.3) ja segamudeli (4.3) meetodeid. GREG-i ja segamudeli hinnangu arvutamise kasutati mudelit objekti tasemel (*Unit level model*), mida on kirjeldatud peatükis 4.3. Kõigi 1000 valimi korral jäeti meelde vastavad hinnangud ja valimi mahud osakogumi kaupa. Sellest tulenevalt oli võimalik leida tekkinud andmestikust huvipakkuvad väärtused (hinnangud, osakogumite keskmine valimimaht) ja täpsusnäitajad. Segamudelis kasutatakse juhusliku mõjuna osakogumist tulenevat mõju. Seepärast iga osakogumi tasemel on genereeritud mudeli vabaliige erinev.

5.4. Tulemused

Tulemused on kirja pandud mõlema simulatsioonis kasutatud valikudisaini kohta eraldi. Erinevate hinnangumeetodite hinnangud ja täpsusnäitajad on välja toodud peamiselt Tabelites 2-5. Samuti on huvipakkuv, kas erinevate valikudisainide põhjal oli hinnangutel mingi märgatav erinevus.

5.4.1. LJV TTA

Valikudisaini lihtsa juhusliku valiku tagasipanekuta korral on tulemused Tabelis 2. Välja on toodud

- tegelik osakogumi kogusumma väärtus t_d;
- valimimahtude keskmine $\bar{n}_d$ igas osakogumis;
- HT, GREG-i ning segamudeli keskmised hinnangud ja standardhälbed (5.1) üle 1000 genereeritud valimi.

Tabelis 2 võib näha, et valimimahtude keskmine suurus on osakogumiti suuresti erinev, näiteks esimese ja teise osakogumi erinevus on ligikaudu kolme kordne. See tuleneb sellest, et ka üldkogumis on osakogumite mahud suuresti erinevad, mida on näha Tabelist 1. Selline olukord on isegi hea, sest on võimalik uurida, millise hinnangumeetodiga leitud hinnangud parameetrile t_d on paremad väiksemate osakogumite korral. Lisame, et ühtegi tühja valimit osakogumis ei tekkinud.

Kui võrrelda hinnangute keskmisi, siis sarnanevad kõige paremini tegeliku kogusumma väärtusega HT ning segamudeli meetodid. Kuna HT hinnang on nihketa hinnang, siis HT hinnangute keskmise sarnasus tegeliku väärtusega on oodatud tulemus. GREG-i meetodil põhinev hinnang tundub visuaalsel vaatlusel kõige ebatäpsem.

Hinnangute väikseim varieeruvus on segamudeli hinnangul ning suurim HT meetodil. Keskmiselt kõige suurema valimi mahuga osakogumis on GREG-i hinnangu standardhälve isegi segamudeli omast väiksem, aga valimi mahu vähenedes muutub segamudeli varieeruvus paremaks võrreldes GREG-iga. Kuna selle töö eesmärgiks ongi väikeste osakogumite

hindamine, siis Tabelis 2 leitud näitajad viitavad segamudeli headusele ning täpsusele võrreldes HT ja GREG-i hinnangutega.

Tabel 2. Leibkonna kogusissetuleku hinnangud ja standardhälbed ($\times 10^3$) osakogumites LJV TTA korral.

d	t_d	$\bar{n}_d$	HT	std	GREG	std	Segamudel	std
1	2842760	35	2849642	451	2847905	187	2997697	132
2	12969135	96	12935189	1046	12959905	397	12503509	418
3	9260723	53	9275138	1057	9251514	233	9329422	149
4	13584061	71	13647438	1285	13589830	277	13729967	223
5	9362786	45	9325891	1204	9353358	213	9447826	152

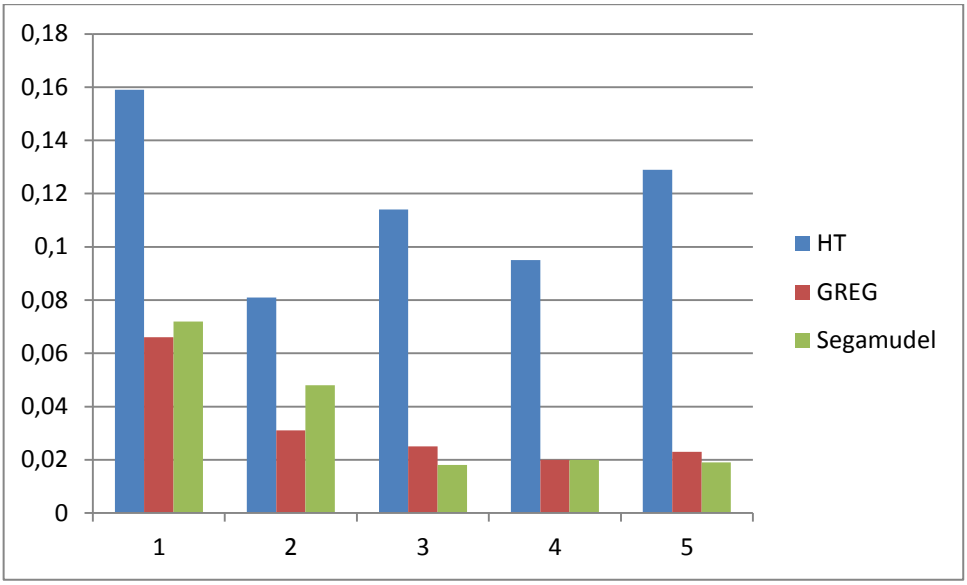
Tabelis 3 on välja toodud kõigi osakogumi kogusummade hinnangute nihked kolme erineva meetodi kaupa. HT ja GREG-i hinnangute nihked on väiksemad võrreldes segamudeliga. Suur erinevus tuleb sisse kõige väiksema mahuga osakogumis (esimeses), kus segamudeli nihe on kahe teise hinnangu nihkest mitmeid kordi suurem. Siiski kõigi kolme hinnangu nihked on nulli lähedal ning selline tulemus on hea.

Tabel 3. Suhteline nihe LJV TTA korral.

d	HT	GREG	Segamudel
1	0,002	0,002	0,055
2	-0,003	-0,001	-0,036
3	0,002	-0,001	0,007
4	0,005	0	0,011
5	-0,004	-0,001	0,009

Kolmanda headuse näitajana leiti kogusummade hinnangute suhtelise ruutkeskmise vea ruutjuur (5.3). Joonisel 2 on kujutatud RRMSE muutust tulpdiagrammil LJV TTA korral, sõltuvalt osakogumist ning kasutatud hinnangust.

Kõige suuremad väärtused kolme hinnangu puhul on mõlema esimesel osakogumil. Läbivalt kõigis osakogumites on HT hinnangu RRMSE kehvem teiste hinnangute vastavast täpsusnäitajast. Joonisel 2 on näha, et kõige suurema osakogumi (teise) korral HT hinnangu RRMSE sarnane GREG-i ja segamudeli omaga, aga osakogumi valimimahu vähenedes on GREG-i ja segamudelil põhinevad hinnangud täpsemad.



Joonis 2. Suhtelise ruutkeskmise vea ruutjuur LJV TTA korral kõikides osakogumites.

5.4.2. LJV TGA

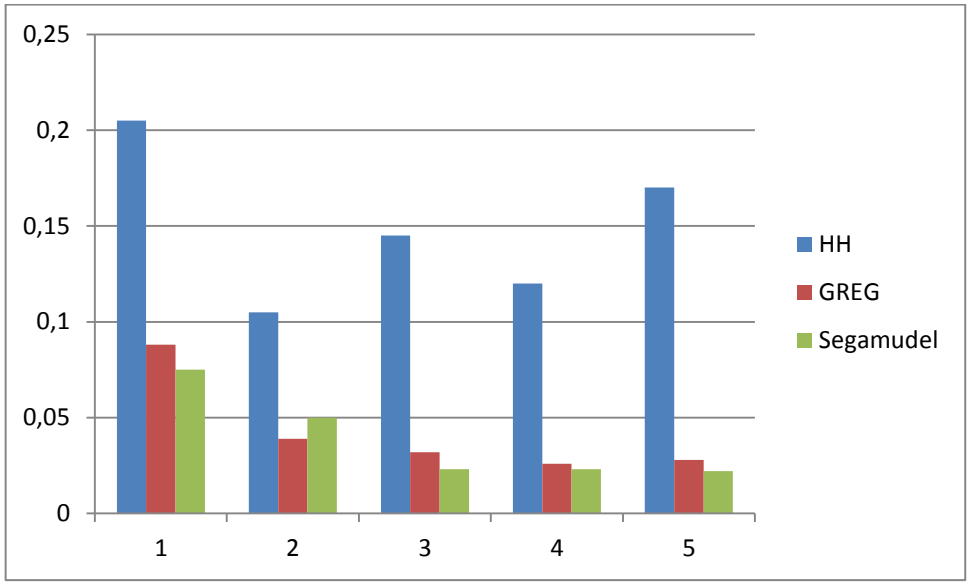
Hinnangute ja nende täpsusnäitajate leidmisel on simulatsioonis muutunud ainult valimi võtmisel kasutatud valikudisain (LJV TGA). Sellest tulenevalt on HT hinnangu asemel arvutatud huvipakkuvad väärtused HH hinnangule (1.1).

Tabelist 4 on näha, et hinnangute tulemused on väga sarnased LJV TTA juhuga, mis viitab sellele, et valimi võtmise meetod meie hinnangute headuse võrdlust nende vahel ei muuda. On märgata, et HH keskmine hinnang uuritavale parameetrile on sarnane üldkogumi omaga, aga suurem varieeruvus annab eelise segamudelil põhinevale hinnangule.

Tabel 4. Leibkonna kogusissetuleku hinnangud ja standardhälbed ($\times 10^3$) osakogumites LJV TGA korral.

d	t_d	$\bar{n}_d$	HH	std	GREG	std	Segamudel	std
1	2842760	35	2843161	582	2827388	250	2969510	173
2	12969135	96	12951041	1364	12967724	512	12574069	516
3	9260723	53	9245475	1347	9264625	295	9322653	202
4	13584061	72	13662590	1623	13584962	358	13703001	289
5	9362786	45	9356735	1590	9355443	263	9435816	196

Joonisel 3 on välja toodud uuritavate hinnangute suhtelise ruutkeskmise vea ruutjuured. On märgata, et üldine RRMSE muutuse tendents on sarnane tagasipanekuta valikudisainiga. Ainult HH hinnangu suhtelise ruutkeskmine vea ruutjuured on kõigis osakogumites mõnevõrra suuremad võrreldes LJV TTA korral leitud analoogse HT hinnangu RRMSE-ga.



Joonis 3. Suhtelise ruutkeskmise vea ruutjuur LJV TGA korral kõikides osakogumites.

Tabelis 5 on esitatud hinnangute nihked LJV TGA korral. Tulemused on jällegi sarnased TTA disainiga. Suurim nihe on segamudelil põhineval hinnangul ja kõige väiksema mahuga osakogumis on see teistes osakogumitega võrreldes märgatavalt suurem. Samuti on näha, et GREG-i nihe on kõige väiksem.

Tabel 5. Suhteline nihe LJV TGA korral.

d	HH	GREG	Segamudel
1	0	-0,005	0,045
2	-0,001	0	-0,03
3	-0,002	0	0,007
4	0,006	0	0,009
5	-0,001	-0,001	0,008

5.5. Järeldused

Simuleerimise käigus kasutatud andmestiku põhjal sai võrrelda huvipakkuvaid HT, GREG-il ja segamudelil põhinevaid hinnanguid t_d hindamisel. Kasutasime täpsusnäitajad, et välja selgitada, millise hinnangu kasutamine kolmest võiks olla eelistatud.

Standardhälbe põhjal oli kõige suurem varieeruvus HT ja HH hinnangutel. Seda oligi oodata, kuna need hinnangud ei kasuta mingisugust abiinformatsiooni. Nihketust arvestades olid kõige paremad HH, HT ja GREG hinnangud. Segamudeli eeliseks oli väiksed standardhälbed just väiksemate mahtudega osakogumites. Selliste tulemusteni jõuti kui kasutati nii lihtsat juhuslikku valikut tagasipanekuta kui ka tagasipanekuga. Sellest tulenevalt ei avaldanud valimi võtmise meetod märgatavat mõju hinnangutele.

Parima hinnangu osakogumite kogusummale annavad GREG ja segamudel, mis vihjab sellele, et abiinformatsiooni kasutamine parandab osakogumi hinnangute täpsust. Kõige suurema osakogumi korral oli GREG-i ja segamudeli varieeruvus sarnane, aga osakogumi mahu kahanedes vähendas segamudelis kasutatav osakogumi "mõju" hinnangute varieeruvust. Samuti tuli segamudel nihketa peaaegu kõikide osakogumite korral, isegi väikeste.

Edaspidi saaks kindlasti veel uurida abitunnuste valiku mõju hinnangu täpsusele. Kas erineva seosega abitunnuste ja uuritava tunnuse vahel annab segamudel paremaid tulemusi või mitte.

Domain estimations with auxiliary information

Bachelor Thesis

Paavo Binsol

Summary

In this bachelor thesis estimator based on general linear model is introduced and is compared to other small area estimators. Other estimates are Horwitz-Thompson (HT), Hansen-Hurwitz (HH) and Generalized regression estimator (GREG). The main focus is on the grand total in specific area. Also there are used different sampling methods, like simple random sampling with and without replacement are used.

Small area estimations tackles the problem of providing reliable estimates of one or several variables of interest in areas where the information available is not sufficient to provide valid estimate. The information is usually collected by conducting a survey in some or all areas.

Direct estimates such as HT and HH, provide estimates based only on the local data and the design weights for the sample. Unfortunately, when the sample sizes are small, the direct estimates are unreliable.

General linear model uses auxiliary information for estimates. Estimations using auxiliary information are called indirect or model-based. These estimates "borrow strength" from the relationship between variable of interest and auxiliary information. In this works simulation the domain levels are used as random effects and all other variables are as fixed effects.

For comparing the estimates data was composed. The sample was taken by using both simple random sampling with replacement and without. In both cases the sample was taken 1000 times. Then Mixed Model, GREG, HT and HH mean of t_d estimates and standard deviation was calculated. Also for evaluating the perfomance of different estimates following preformance criterias were found:

- standard deviation ($\mathrm{std}(\hat{t}_d)$);
- the relative bias ($\mathrm{RB}(\hat{t}_d)$);

- the relative root mean square error (RRMSE($\hat{t}_d$)).

Results from the simulation revealed that the smallest perfomance criteria measures were while using estimate that based on linear mixed models. HT and HH estimates had the biggest variance of all for estimating grand total in specific area. Using auxiliary information gave a smaller variance, which was the aim. Estimates perfomance results were not affected by different sampling designs.

Kasutatud kirjandus

1. Kaart, T. 2012. *Juhuslikud ja fikseeritud faktorid*, http://www.eau.ee/~ktanel/lineaarne_mudel/pt23.php, külastatud 03.05.2013.

2. Lepik, N. 2011. *Estimation of domains under restrictions and synthetic estimators. PhD Dissertation*, Tartu, Tartu University Press.133 lk.

3. Littell, R. C., Stroup W.W., Freund, R. J., 2002. *SAS FOR LINEAR MODELS Fourth Edition*, USA, SAS Publishing, 496 lk.

4. SAE package developers. 2007. *Introduction to Small Arema Estimation*. http://www.bias-project.org.uk/software/SAE.pdf, külastatud 05.05.2013.

5. Saei, A., Chambers, R. 2003. *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models*, http://www.unescap.org/stat/meet/disaggregated-20-23Sep2011/SAE-review.pdf, külastatud 06.05.2013.

6. Schaeffer L. R., *Prediction Theory*, http://www.aps.uoguelph.ca/~lrs/ABModels/notesx.html külastatud 06.05.2013.

7. Schwarz, C. J. 1997. StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling. Journal of Statistics Education v.5, n.2 http://www.lenato.eu/StatVillage/index.html, Maximal Village, külastatud 05.05.2013.

8. Särndal, C-E., Swensson B., Wretman J., 1992. *Model assited survey sampling*, Rootsi, Springer-Verlag New York, Inc., 694 lk.

9. Traat, I., Inno, J. 1997. *Tõenäosuslik valikuuring*, Tartu, TÜ Kirjastus, 212 lk.

Lisa 1

R-i kood

```
## AlGSE ANDMESTIKU SISSELUGEMINE
andmed=read.csv("Statvillage.csv", header=TRUE,sep=";")
head(andmed)


## UUS ANDMESTIK, TUNNUSED: TOTINCH, VALUEH, EMPINCH , HHSIZE
andmed2=data.frame(objektid=1:1024)
andmed2$hhsize=andmed$hhsize
andmed2$hhsize[andmed2$hhsize>4]=5
andmed2$totinch=as.vector(andmed$totinch)
andmed2$valueh=as.vector(andmed$valueh)
andmed2$empinch=as.vector(andmed$empinch)


## PUUDUVATE EEMALDAMINE
eemaldada1=-1*andmed2$objektid[is.na(andmed2$valueh)]
andmed2=andmed2[eemaldada1, ]

## ÜLDKOGUMI JA VALIMIMAHT
N=length(andmed2$objektid)
n=300
## KORRELATSIOONIMAATRIKS
cor(andmed2)

## ANDMESTIK, KUS KOGUSUMMAD OSAKOGUMI KAUPA
smmry=data.frame(osakogumid=1:5)
smmry$totinch=as.vector(by(andmed2$totinch,andmed2$hhsize,sum))
smmry$N_d=as.vector(by(andmed2$hhsize,andmed2$hhsize,length))
smmry$valueh=as.vector(by(andmed2$valueh,andmed2$hhsize,sum))
smmry$empinch=as.vector(by(andmed2$empinch,andmed2$hhsize,sum))




## HT ja HH HINNANGU LEIDMINE
HT_hinnang_kogusummale=function(valim){
        totinch_s=as.vector(by(andmed2$totinch[valim],andmed2$hhsize[valim],sum))
        n_d=as.vector(by(andmed2$hhsize[valim],andmed2$hhsize[valim],length))
        hinnang=(N/n)*totinch_s
        return(hinnang)
}
```

```
## GREG


hinnang_GREG=function(valim){
        d=data.frame(Y=andmed2$totinch[valim],X1=andmed2$valueh[valim],X2=andmed2$
        empinch[valim],INTERCEP=1)
        wgreg=rep(N/n,n)
        lmgreg=lm(formula=Y~-1+INTERCEP+X1+X2,data=d,weights=wgreg)
        p1=as.vector(predict(lmgreg,data.frame(INTERCEP=1,X1=andmed2$valueh,X2=and
        med2$empinch)))
        andmed2$p1=p1
        GREG_p1=as.vector(by(andmed2$p1,andmed2$hhsize,sum))
        oodatavad_s=as.vector(predict(lmgreg,data.frame(INTERCEP=1,X1=andmed2$value
        h[valim],X2=andmed2$empinch[valim])))
        GREG_jaagid=andmed2$totinch[valim]-oodatavad_s
        GREG_jaagid=(N/n)*(as.vector(by(GREG_jaagid,andmed2$hhsize[valim],sum)))
        GREG_hinnang=GREG_p1+GREG_jaagid
return(GREG_hinnang)
}


### SEGAMUDEL
library(nlme)
SM_hinnang=function(valim){
        dunit=data.frame(Y=andmed2$totinch[valim],X1=andmed2$valueh[valim],X2=andm
        ed2$empinch[valim], osakogum=andmed2$hhsize[valim])
        sm=lme(Y ~ 1 + X1 + X2 , random = ~1 | osakogum, data = dunit,method="ML")
        sm_andmed=data.frame(X1=andmed2$valueh,X2=andmed2$empinch,osakogum=and
        med2$hhsize)
        oodatud=as.vector(predict(sm,sm_andmed))
        SM_HINNANG=as.vector(by(oodatud,andmed2$hhsize,sum))
        return(SM_HINNANG)
}


## HT, GREG, SEGAMUDEL LJV TTA PUHUL
koos=function(m,valim){
valjund=data.frame(osakogumid=1:5)
HT_dispersioon=matrix(nrow=5,ncol=m)
GREG_dispersioon=matrix(nrow=5,ncol=m)
SM_dispersioon=matrix(nrow=5,ncol=m)
w_greg=rep(N/n,n)
HT_summa=0
GREG_summa=0
SM_summa=0
HT_AEMSE=0
GREG_AEMSE=0
sm_AEMSE=0
osakogumi_maht=0
```

```
for (i in 1:m){
	valim=sample(N,n)
	n_d=as.vector(by(andmed2$hhsize[valim],andmed2$hhsize[valim],length))
	osakogumi_maht=osakogumi_maht+n_d
	HT_hinnang=HT_hinnang_kogusummale(valim)
	HT_summa=HT_summa+HT_hinnang
	HT_AEMSE=HT_AEMSE+(HT_hinnang-smmry$totinch)**2
	HT_dispersioon[,i]=HT_hinnang
	##GREG
	GREG_hinnang=hinnang_GREG(valim)
	GREG_summa=GREG_summa+GREG_hinnang
	GREG_AEMSE=GREG_AEMSE+as.numeric((GREG_hinnang-smmry$totinch)**2)
	GREG_dispersioon[,i]=GREG_hinnang
	#SEGAMUDEL
	SM_HINNANG=SM_hinnang(valim)
	SM_summa=SM_summa+SM_HINNANG
	sm_AEMSE=sm_AEMSE+(SM_HINNANG-smmry$totinch)**2
	SM_dispersioon[,i]=SM_HINNANG

}
valjund$osakogumid=osakogumi_maht/m
valjund$TEGELIK=smmry$totinch
valjund$HT_KESK=HT_summa/m
valjund$HT_sd=c(sd(HT_dispersioon[1,]),sd(HT_dispersioon[2,]),sd(HT_dispersioon[3,]),sd(
HT_dispersioon[4,]),sd(HT_dispersioon[5,]))
valjund$GREG_KESK=GREG_summa/m
valjund$GREG_sd=c(sd(GREG_dispersioon[1,]),sd(GREG_dispersioon[2,]),sd(GREG_dispe
rsioon[3,]),sd(GREG_dispersioon[4,]),sd(GREG_dispersioon[5,]))
valjund$SM_KESK=SM_summa/m
valjund$SM_sd=c(sd(SM_dispersioon[1,]),sd(SM_dispersioon[2,]),sd(SM_dispersioon[3,]),s
d(SM_dispersioon[4,]),sd(SM_dispersioon[5,]))
valjund$HT_RRMSE=sqrt(HT_AEMSE/m)/smmry$totinch
valjund$GREG_RRMSE=sqrt(GREG_AEMSE/m)/smmry$totinch
valjund$SM_RRMSE=sqrt(sm_AEMSE/m)/smmry$totinch
return(valjund)
}

koos_tabel=koos(1000)

koos_tabel$HT_RB=(koos_tabel$HT_KESK-koos_tabel$TEGELIK)/koos_tabel$TEGELIK
koos_tabel$GREG_RB=(koos_tabel$GREG_KESK-
koos_tabel$TEGELIK)/koos_tabel$TEGELIK
koos_tabel$SM_RB=(koos_tabel$SM_KESK-koos_tabel$TEGELIK)/koos_tabel$TEGELIK
koos_tabel

## TULEMUSTE TRANSPORTIMINE

write.csv( koos_tabel, file="LJV_TTA.csv")
```

```
## HT, GREG, SEGAMUDEL LJV TGA DISAINI PUHUL
koos2=function(m,valim){
valjund=data.frame(osakogumid=1:5)
HH_dispersioon=matrix(nrow=5,ncol=m)
GREG_dispersioon=matrix(nrow=5,ncol=m)
SM_dispersioon=matrix(nrow=5,ncol=m)
w_greg=rep(N/n,n)
HH_summa=0
GREG_summa=0
SM_summa=0
HH_AEMSE=0
GREG_AEMSE=0
sm_AEMSE=0
osakogumi_maht=0
for (i in 1:m){
        valim=sample(N,n,replace=TRUE)
        n_d=as.vector(by(andmed2$hhsize[valim],andmed2$hhsize[valim],length))
        osakogumi_maht=osakogumi_maht+n_d
        HH_hinnang=HT_hinnang_kogusummale(valim)
        HH_summa=HH_summa+HH_hinnang
        HH_AEMSE=HH_AEMSE+(HH_hinnang-smmry$totinch)**2
        HH_dispersioon[,i]=HH_hinnang
        ##GREG
        GREG_hinnang=hinnang_GREG(valim)
        GREG_summa=GREG_summa+GREG_hinnang
        GREG_AEMSE=GREG_AEMSE+as.numeric((GREG_hinnang-smmry$totinch)**2)
        GREG_dispersioon[,i]=GREG_hinnang
        #SEGAMUDEL
        SM_HINNANG=SM_hinnang(valim)
        SM_summa=SM_summa+SM_HINNANG
        sm_AEMSE=sm_AEMSE+(SM_HINNANG-smmry$totinch)**2
        SM_dispersioon[,i]=SM_HINNANG
}
valjund$osakogumid=osakogumi_maht/m
valjund$TEGELIK=smmry$totinch
valjund$HH_KESK=HH_summa/m
valjund$HH_sd=c(sd(HH_dispersioon[1,]),sd(HH_dispersioon[2,]),sd(HH_dispersioon[3,]),s
d(HH_dispersioon[4,]),sd(HH_dispersioon[5,]))
valjund$GREG_KESK=GREG_summa/m
valjund$GREG_sd=c(sd(GREG_dispersioon[1,]),sd(GREG_dispersioon[2,]),sd(GREG_dispe
rsioon[3,]),sd(GREG_dispersioon[4,]),sd(GREG_dispersioon[5,]))
valjund$SM_KESK=SM_summa/m
valjund$SM_sd=c(sd(SM_dispersioon[1,]),sd(SM_dispersioon[2,]),sd(SM_dispersioon[3,]),s
d(SM_dispersioon[4,]),sd(SM_dispersioon[5,]))
valjund$HH_RRMSE=sqrt(HH_AEMSE/m)/smmry$totinch
```

```
valjund$GREG_RRMSE=sqrt(GREG_AEMSE/m)/smmry$totinch
valjund$SM_RRMSE=sqrt(sm_AEMSE/m)/smmry$totinch
return(valjund)
}
koos_tabel2=koos2(1000)
koos_tabel2$HH_RB=(koos_tabel2$HH_KESK-
koos_tabel2$TEGELIK)/koos_tabel2$TEGELIK
koos_tabel2$GREG_RB=(koos_tabel2$GREG_KESK-
koos_tabel2$TEGELIK)/koos_tabel2$TEGELIK
koos_tabel2$SM_RB=(koos_tabel2$SM_KESK-
koos_tabel2$TEGELIK)/koos_tabel2$TEGELIK
koos_tabel2

write.csv( koos_tabel2, file="LJV_TGA.csv")
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina: Paavo Binsol ,

(*autori nimi*)

(sünnikuupäev: 01.08.1991)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Hindamine osakogumites abiinformatsiooni olemasolul ,

(*lõputöö pealkiri*)

mille juhendaja on Natalja Lepik ,
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus **06.05.2013**