Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Grete Ojala

Kao kompenseerimine mittejuhusliku vastamise korral latentse tunnuse abil

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendaja Natalja Lepik

Tartu 2016

Kao kompenseerimine mittejuhusliku vastamise korral latentse tunnuse abil

Mittevastamist esineb peaaegu igas uuringus ning see võib põhjustada nihkeid hinnangutes. Tavaliselt on mittevastamine juhuslik ning leidub lisainformatsiooni valimi objektide kohta. Sel juhul saab andmete kadu kompenseerida omitus- või kaalumismeetoditega. Käesoleva bakalaureusetöö eesmärk on aga välja selgitada, kuidas kompenseerida mittejuhuslikku mittevastamist jälgides vaid inimeste üldist vastamise mustrit olukorras, kus puudub lisainformatsioon. Selleks kasutatakse latenseid tunnuseid, mis iseloomustavad objektide tahet vastata uuringu küsimustele. Simuleerimisülesandes demonstreeritakse latentsete tunnuste abil üldkogumi kogusumma hindamist ja võrreldakse erinevaid hinnanguid omavahel.

Märksõnad: valikuuringud, puuduvad andmed

P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Compensation of missing data in the case of nonignorable response using latent variable

Nonresponse is present in almost all surveys and may produce a bias in estimates. Usually nonresponse is ignorable and auxiliary information is available for sample units. In this case, it is possible to compensate missing data by imputation and reweighting methods. The purpose of this bachelor thesis is to study how to deal with nonignorable nonresponse by following only people's pattern of response in the situation where auxiliary information is not available. To deal with it, latent variables that describe units' willing to answer survey questionnaire are used. The estimation of population total with latent variables is described and different estimators are compared in simulation study.

Key words: sample surveys, missing data

P160 Statistics, operation research, programming, actuarial mathematics

Sisukord

Sissejuhatus

Mittevastamist esineb peaaegu igas uuringus ning see võib põhjustada nihkeid hinnangutes. Mittevastamise kompenseerimiseks kasutatakse peamiselt kahte meetodit: omistus- ja kaalumismeetodeid. Nende kasutamise eelduseks on juhuslik mittevastamine ning lisainformatsiooni olemasolu üldkogumi tasemel.

Käesoleva bakalaureusetöö eesmärk on aga välja selgitada, kuidas kompenseerida mittejuhuslikku mittevastamist jälgides vaid inimeste üldist vastamise mustrit olukorras, kus puudub lisainformatsioon. Selleks kasutatakse latenseid ehk varjatud tunnuseid, mis iseloomustavad objektide tahet vastata uuringu küsimustele ja on leitavad kõigi valimisse sattunud objektide jaoks.

Latentse tunnusega lähenemise korral on esmatähtis eeldus see, et mehhanism, mis innustab inimesi osalema küsitluses üldiselt, on sama, mis innustab vastama ka konkreetsetele küsimustele. Latentse tunnuse mudeli abil leitakse varjatud tunnused, mida omakorda kasutatakse vastamistõenäosuste hindamisel ning nende abil leitakse üldkogumi hinnangud.

Bakalaureusetöö esimeses osas kirjeldatakse mittevastamist, selle liike ja mõningaid kao kompenseerimise meetodeid. Teises peatükis antakse ülevaade hinnangute leidmisest kasutades latentseid tunnuseid. Kolmandas peatükis tehakse läbi artiklis (Matei ja Ranalli, 2015) kirjeldatud simuleerimisülesanne ning võrreldakse täpsusnäitajate põhjal erinevaid üldkogumi kogusumma hinnanguid.

Töö kirjutamiseks on kasutatud tekstitöötlusprogrammi LaTeX. Simulatsiooniülesanne viidi läbi tarkvarapaketiga *R*.

1 Mittevastamine ja selle kompenseerimise meetodid

Mittevastamine on sageli esinev probleem valikuuringutes ja seetõttu püütakse leida efektiivseid meetodeid selle kompenseerimiseks. Valikuuringutes tehakse otsused üldkogumi kohta valimi baasil. Mittevastamine tähendab seda, et kõikide või mõnede tunnuste väärtusi ei ole võimalik saada uuringu teostamiseks koostatud valimi kõigilt objektidelt. Sellega kaasneb andmete puudumine ning lüngad andmestikus põhjustavad nihkeid hinnangutes. Kao määr ehk osakaal andmestikus ulatub sageli $30 - 40\%$-ni (Traat ja Inno, 1997, lk 191).

Mittevastamise põhjuseid võib olla mitmeid, näiteks inimene on küsitluse ajaks kolinud mujale ja temaga ei ole võimalik kontakti saada või pole inimest küsitluse ajal kodus. Mõnikord keeldutakse mingil põhjusel uuringust või jäetakse vastamata tundlikele küsimustele. Posti teel läbiviidavate uuringute korral on peamiseks mittevastamise põhjuseks tagastamata küsimustikud.

Andmestikus võib esineda kahte liiki kadu (Traat ja Inno, 1997, lk 191):

1. Tunnuse väärtuse kadu (ehk kadu väärtuse tasemel) – sel juhul puudub vaadeldaval objektil mõne tunnuse väärtus, esinevad lüngad andmestikus.

2. Objekti kadu (ehk kadu objekti tasemel) – andmestikust puudub terve objekt, puuduvad objekti kõigi tunnuste väärtused.

Tabelis 1 tähistab „X” vastamist ja „ .” mittevastamist. Objektidel 2, 3 ja 4 esineb väärtuse kadu ning objektil 6 kadu objekti tasemel (st tema kohta on teada vaid registritunnuste väärtused).

Tabel 1: Näide andmestikust, kus esineb mittevastamist.

	Registritunnused		Uuritavad tunnused		
Objekti nr	1	2	1	2	3
1	×	×	×	×	×
2	×	×	×	·	×
3	×	×	·	·	×
4	×	×	·	×	×
5	×	×	×	×	×
6	×	×	·	·	·

Mittevastamine ja suur kadu andmetes mõjuvad halvasti hinnangute kvaliteedile. Olukord, kus vastanute ja mittevastanute karakteristikud on üksteisest oluliselt erinevad, põhjustab nihkega hinnanguid (Traat ja Inno, 1997, lk 191-192). Mittevastamise tagajärjeks võib osutuda suurem hinnangute dispersioon, mis on lünkadeta andmestiku saamiseks tegeliku valimimahu vähendamise tulemus. Samas esineb oht dispersiooni ka alahinnata, sest puuduvad väärtused on sageli erandlikud väärtused, mida küsitletavad ei taha avaldada. Nende asendamine olemasolevate vähem erandlike väärtustega muudab tunnuse dispersiooni väiksemaks ja ka selle tunnuse põhjal arvutatud hinnangute dispersiooni väiksemaks.

Traat ja Inno (1997, lk 192-204) esitavad kaks peamiselt kasutatavat meetodit mittevastamise kompenseerimiseks: tunnuse väärtuse kaoga andmestikes kasutatakse kao kompenseerimiseks omistusmeetodeid, objekti kaoga andmestikes aga kaalumismeetodeid. Omistusmeetodite üldine eesmärk on lünkadeta andmestiku tagamine, mis on vajalik paljude andmetöötlusprogrammide kasutamiseks. Sel meetodil asendatakse puuduvad väärtused hinnanguliste väärtustega (hindamiseks kasutatakse vastanute andmeid). Kaalumismeetodite puhul omistatakse vastanud objektidele kaalud, mis näitavad kui suurt osa üldkogumist nad esindavad. Sel moel saadakse uuritava tunnuse hinnangud üldkogumi jaoks. Erinevaid mittevas-

tamise kompenseerimise meetodeid on varem käsitlenud oma bakalaureusetöös Prostakova (2007).

Omistusmeetodid töötavad hästi, kui mittevastamine on juhuslik ja ei sõltu uuritavatest tunnustest. Kaalumismeetodeid saab kasutada siis, kui registris leidub sobivaid tunnuseid, mille abil saab objekti kohta vajalikku lisainformatsiooni. Käesoleva töö eesmärk on aga välja selgitada, kuidas kompenseerida mittejuhuslikku mittevastamist valikuuringutes jälgides inimeste üldist vastamise mustrit olukorras, kus puudub lisainformatsioon.

1.1 Mittejuhuslik mittevastamine ja latentsed tunnused

Matei ja Ranalli (2015) järgi esineb andmestikus mittejuhuslik mittevastamine juhul, kui puuduvate andmete mehhanism on mittejuhuslik, st inimene jätab küsimusele vastamata teatud põhjusel. Põhjuseks võib olla näiteks soov mitte avaldada oma tavatult madalat või kõrget sissetulekut. Mittejuhuslik mittevastamine sõltub uuritavatest tunnustest, mille väärtused on saadud vaid vastanud objektidelt või on täielikult puudu. Selline andmete puudumine on tüüpiline tundlike küsimustega uuringutes. Mittejuhuslik mittevastamine on tavaline näiteks uuringutes, mis puudutavad seksuaalkäitumist või narkootikumide kuritarvitamist.

Mittejuhusliku mittevastamise korral kasutatakse üldkogumi parameetrite hindamiseks varjatud ehk latentseid tunnuseid, mida saab hinnata kasutades latentse tunnuse modelleerimise meetodeid. Põhjalikumalt saab tutvuda vastava metoodikaga näiteks raamatute (Beaujean, 2014) ja (Skrondal ja Rabe-Hesketh, 2004) kaudu. Siin töös toome välja vaid vajalikud mõisted. Latentseid tunnuseid saab kasutada vastamistõenäosuste leidmiseks ja neid omakorda on vaja kaalumismeetodite jaoks hinnangute leidmiseks. Latentsed tunnused iseloomustavad vastamiskäitumist uuringus, väljendavad objekti tahet vastata uuringu küsimustele. Latentset tunnust on võimalik leida kõigi valimi objektide jaoks ning selleks pole vaja lisainformatsiooni. Täpsemalt kirjeldatakse latentsete tunnuste kasutamist hindamisprotsessis järgnevas peatükis.

2 Kompenseerimine latensete tunnuste abil

Edasine teooria põhineb artiklil (Matei ja Ranalli, 2015).

2.1 Tähistused ja uuritava tunnuse üldkogumi kogusumma hinnangud

Olgu U lõplik üldkogum objektide arvuga N, objekti tähistatakse $k = 1, ..., N$. Olgu s mingi tõenäosusliku valikudisainiga $p(s)$ üldkogumist U võetud valim mahuga n. Tõenäosusliku valikuuringu korral on iga üldkogumi objekti kohta teada tema valimisse sattumise tõenäosus ehk kaasamistõenäosus. Objekti k kaasamistõenäosust tähistatakse järgmiselt $\pi_k = \sum_{s;k\in s} p(s)$. Eeldame, et $\pi_k > 0 \quad \forall\, k = 1, \ldots, N$ korral. Kõik valimisse sattunud objektid ei osale uuringus ja seega ei vasta uuringu küsimustele. Vastanute hulka tähistatakse $r \subseteq s$ ja mittevastanute hulka $\overline{r} = s \setminus r$.

Eeldame, et vastamise mehhanism ei ole juhuslik ja on antud jaotusega

$$q(r \,|\, s) = P(\text{vastanute hulk on } r \mid \text{saadakse valim } s),$$

kusjuures iga fikseeritud valimi s korral vastanute hulga r saamise tõenäosus

$$q(r \,|\, s) \geqslant 0 \quad \forall\, r \in R_s \text{ korral ja } \sum_{r\in R_s} q(r \,|\, s) = 1,$$

kus $R_s = \{r \mid r \subseteq s\}$ on kõikvõimalike vastanute hulkade hulk fikseeritud valimi s korral.

Objekti tasemel kao korral defineeritakse vastamisindikaator R_k:

$$R_k = \begin{cases} 1, \text{kui } k \in r, \\ 0, \text{kui } k \in \overline{r}. \end{cases}$$

Siis saab kirjutada, et vastanute hulk on $r = \{k \in s \mid R_k = 1\}$. Eeldame, et juhuslikud suurused R_k on üksteisest sõltumatud, st inimesed vastavad üksteisest sõltumatult, ning on

sõltumatud ka valimi võtmise mehhanismist. Kui on olemas ainult vastanute (hulka r kuuluvate objektide) tunnuste väärtused, saame vastamistõenäosuse iga objekti $k \in U$ jaoks leida vastamismudeli põhjal, seega $p_k = P(k \in r \mid k \in s) = P(R_k = 1 \mid k \in s)$.

Oletame, et uuringus on m huvipakkuvat tunnust ja olgu eesmärk hinnata uuritava tunnuse y_j ($j = 1, ..., m$) kogusummat üldkogumis, st hinnata juhuslikku suurust

$$Y_j = \sum_{k=1}^{N} y_{kj}, \tag{1}$$

kus y_{kj} on objekti k tunnuse y_j väärtus. Tagasipanekuta valikudisainide korral saab kogusumma (1) hindamiseks kasutada Horvitz-Thompsoni hinnangut:

$$\hat{Y}_{j,s} = \sum_{s} y_{kj}\omega_k, \tag{2}$$

kus $\omega_k = \frac{1}{\pi_k}$ on disaini kaal.

Oletame veel, et lisaks üldisele mittevastamisele (objekti tasemel) esineb mittevastamist ka tunnuste tasemel. Olgu r_j tunnusele y_j vastanud objektide hulk. Nagu terve objekti kao korral, eeldame ka siin, et objektid hulgas r_j vastavad üksteisest sõltumatult. Kui j. tunnusel on mõned väärtused puudu, siis viib üle vastanute hulga r_j arvutatud hinnang

$$\hat{Y}_{j,r_j} = \sum_{r_j} y_{kj}\omega_k \tag{3}$$

alahinnanguni, sest ei arvestata seda, kuidas hinnang laieneb vastanute hulgalt r_j valimile s. Selle asemel kasutatakse sageli alternatiivset kogusumma hindamise valemit:

$$\hat{Y}_{j,alt} = \hat{\overline{Y}}_{j,alt} \cdot N = \frac{\hat{Y}_j}{\hat{N}} \cdot N = \frac{\sum_{r_j} y_{kj}\omega_k}{\sum_{r_j} 1 \cdot \omega_k} \cdot N, \tag{4}$$

kus $\hat{\overline{Y}}_{j,alt}$ on tunnuse y_j üldkogumi kogusumma keskmise hinnang ja $\hat{N}$ üldkogumimahu hinnang. Siin kompenseeritakse alahindamine sellega, et tunnuse y_j üldkogumi kogusumma hinnang $\hat{Y}_j$ jagatakse üldkogumimahu hinnanguga $\hat{N}$, mis on samuti alahinnang, ja korrutatakse üldkogumi tegeliku mahuga N.

Kui vastamise jaotus $q(r \mid s)$ oleks teada, siis oleks teada ka vastamistõenäosus p_k ja saaksime kogusumma Y_j valemis (1) hinnata kasutades kaalumismeetodit, kus objektidele uue kaalu omistamisel kasutaksime ka vastamistõenäosuse pöördväärtust $\frac{1}{p_k}$ ja valem (3) saaks kuju:

$$\hat{Y}_{j,r_j} = \sum_{r_j} \frac{y_{kj}\omega_k}{p_k} = \sum_{r_j} \frac{y_{kj}}{\pi_k p_k}.$$

Vastamistõenäosuse pöördväärtus $\frac{1}{p_k}$ laiendab kogusumma hinnagut vastanute hulgalt valimile ning disaini kaal $\frac{1}{\pi_k}$ laiendab kogusumma hinnangut valimilt üldkogumile.

Tähistagu $q_{kj} = P(\text{objekt } k \text{ vastas küsimusele } y_j \mid k \in r)$.

Lõplik kaalude hulk, mida kasutatakse kaalumismeetodiga lähenemise korral tegelemaks objekti ja tunnuse tasemel kaoga, on $\frac{1}{\pi_k p_k q_{kj}}$, iga $k \in r_j$ korral, eeldusel $q_{kj} > 0$. Siin arvestatakse objektile uue kaalu omistamisel peale vastamistõenäosuse ja valimisse sattumise tõenäosuse ka objekti tõenäosust vastata küsimusele y_j. Nende kaalude abil saab parandada Horvitz-Thompsoni hinnagut:

$$\hat{Y}_{j,lin,tegelik} = \sum_{r_j} \frac{y_{kj}}{\pi_k p_k q_{kj}}. \tag{5}$$

Kuna praktikas on vastamismehhanism tundmatu, siis on ka p_k ja q_{kj} tundmatud ning neid tuleb hinnata. Mittevastamisele kohandatud hinnang on konstrueeritud nii, et p_k ja q_{kj} on valemis asendatud hinnangutega $\hat{p}_k$ ja $\hat{q}_{kj}$. Lineaarne hinnang (Horvitz-Thompsoni parandatud hinnang) saab kuju:

$$\hat{Y}_{j,lin} = \sum_{r_j} \frac{y_{kj}}{\pi_k \hat{p}_k \hat{q}_{kj}}. \tag{6}$$

Hinnagute $\hat{p}_k$ ja $\hat{q}_{kj}$ leidmiseks kasutatakse mitmeid erinevaid meetodeid. Järgnevas peatükis vaatleme ühte nendest - modelleerimine logistilise regressiooni ja latentse tunnuse abil.

2.2 Vastamistõenäosuste hindamine

2.2.1 Vastamistõenäosuse hindamine kasutades logistilist regressiooni

Mittejuhusliku mittevastamise korral on uuritav tunnus ise mingi kindla vastamismustri tekkimise põhjus või üks selle tekkimise põhjustest. Seega esineb otsene seos vastamiskäitumise ja vastamistõenäosuse vahel. Sel juhul vastamistõenäosus p_k iga $k \in s$ jaoks leitakse kasutades üht järgmistes logistilise regressiooni mudelitest:

$$p_k = P(R_k = 1 \mid y_{kj}) = \frac{1}{1 + e^{-(a_0 + a_1 y_{kj})}} \tag{7}$$

või

$$p_k = P(R_k = 1 \mid y_{kj}, z_k) = \frac{1}{1 + e^{-(a_0 + a_1 y_{kj} + z_k' \alpha)}}, \tag{8}$$

kus $z_k = (z_{k1}, \ldots, z_{kt})'$ on vektor, mis koosneb $t \geqslant 1$ objekti k tunnusest ning a_0, a_1 ja α on parameetrid.

Mittevastamisest põhjustatud nihe kogusumma hinnangus, mis on saadud vastanud objektide tunnuse y_j väärtuste põhjal, sõltub väärtuse y_{kj} ja tõenäosuse p_k vahelisest seosest. Suhtumine uuringu teemasse on üks näide tunnusest, mis võib ära kirjeldada seost y_{kj} ja p_k vahel. Seletavate tunnuste hulk z_k võib samuti olla seotud uuritava tunnusega y_j ja seega vähendada mittevastamisest tingitud nihet.

Kui uuritava tunnuse väärtused y_{kj} on teada ainult vastanud objektidel, ei saa mudeleid (7) ja (8) hinnata. Järgnevas mudelis kasutatakse väärtuseid z_k, mis on tavaliselt teada nii vastajate kui ka mittevastajate korral ja on väärtustega y_{kj} tugevalt seotud:

$$p_k = P(R_k = 1 \mid z_k) = \frac{1}{1 + e^{-(a_0 + z_k' \alpha)}}. \tag{9}$$

Leides hinnangud parameetritele a_0 ja α, saame hinnata vastamistõenäosused järgmiselt:

$$\hat{p}_k = P(R_k = 1 \mid z_k) = \frac{1}{1 + e^{-(\hat{a}_0 + z_k' \hat{\alpha})}}. \tag{10}$$

Kui z_k ennustab hästi vastamistõenäosuse ja/või uuritava tunnuse väärtuse, siis selline protseduur vähendab mittevastamisest põhjustatud hinnangute nihet.

2.2.2 Latentne tunnus kui lisainformatsioon

Vastamistendentsi näitaja ehk latentse tunnuse uurimiseks vaatleme olukorda, kus esineb ka huvipakkuvate tunnuste väärtuste puudumist (kadu väärtuse tasemel). Eeldame, et faktorid, mis innustavad inimesi vastama kõikidele küsimustele, on samad, mis innustavad vastama ka huvipakkuvate tunnuste küsimustele. Selliste faktorite hindamiseks saab kasutada latentse tunnuse mudeleid ja neid faktoreid on võimalik kasutada logistilises vastamismudelis seletavate tunnustena, näiteks mudelis (10).

Nagu varem märgitud, eeldame, et mittevastamine mõjutab m huvipakkuvat tunnust, tähistatakse $l = 1, ..., m$. Defineerime vastamisindikaatori iga tunnuse l ja iga objekti k jaoks. Binaarne tunnus x_{kl} saab väärtuse 1, kui objekt k vastab tunnuse l küsimusele, ja 0 vastasel juhul:

$$x_{kl} = \begin{cases} 1, \text{objekt } k \text{ vastab küsimusele } l, \\ 0, \text{vastasel juhul.} \end{cases}$$

Olgu $x_k = (x_{k1}, .., x_{kl}, .., x_{km})'$ vastamisindikaatorite vektor iga objekti k ja iga tunnuse l jaoks ning $y_k = (y_{k1}, \ldots, y_{kl}, \ldots, y_{km})'$ huvipakkuvate tunnuste väärtuste vektor objekti k jaoks. Seega y_{kl} on objekti k tunnuse l väärtus ja x_{kl} on selle vastamisindikaator.

Oletame, et väärtused x_{kl} on seotud varjatud ehk latentse tunnusega θ. Varjatud tunnuse väärtus θ_k väljendab objekti k soovi osaleda uuringus. Eeldame esialgu, et väärtus θ_k on teada iga valimi objekti jaoks ja seda saab kasutada argumenttunnusena nagu tavalist lisainformat-

siooni. Teiste seletavate tunnuste puudumise korral saab mudeli (9) kirjutada kujul:

$$p_k = P(R_k = 1 \mid \theta_k) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 \theta_k)}}. \tag{11}$$

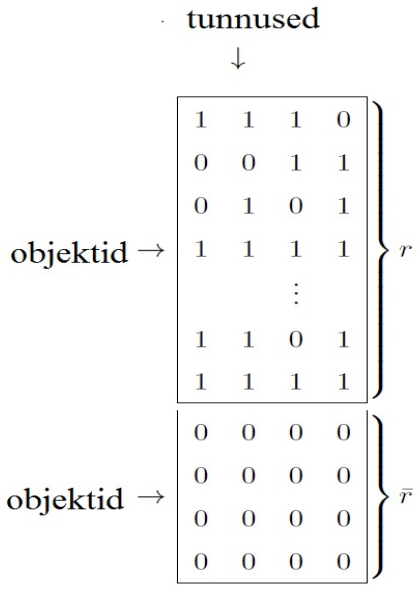
Argumenttunnust θ iseloomustab uuringu teemaga seotud käitumist. Seega on sel head võimalused ära kirjeldada seost uuritava tunnuse y_{kj} ja vastamistõenäosuse p_k vahel ning vähendada mittevastamisest tingitud nihet.

Kui on saadaval veel sobivat lisainformatsiooni (nt registrist), siis saab selle lisada mudelisse täiendava tunnusena. Selleks, et hinnata mudeli (11) parameetreid, peab tunnuse θ väärtus θ_k olema teada kõikide valimi objektide kohta. Peatükk 2.3.2 annab ülevaate, kuidas saada θ_k hinnangud nii vastanute kui ka mittevastanute jaoks.

2.3 Vastamistõenäosuste arvutamine latentse tunnuse mudeli abil

Tunnuse θ väärtusi θ_k leitakse latentse tunnuse mudeli abil. Tavaliselt on latentse tunnuse mudelid mitme argumendiga regressioonmudelid, mis seovad vaadeldud tunnuseid nende tunnustega, mis antud küsitluse raames ei olnud mõõdetud. Latentse tunnuse mudel on põhiliselt faktoranalüüsi mudel binaarsete andmete jaoks.

Moodustame maatriksi elementidega $\{x_{kl}\}_{k \in s; l=1,\ldots,m}$, kus k. reas on k. objekti m huvipakkuva tunnuse indikaatorid ja l. veerus on kõigi valimisse sattunud objektide tunnuse l indikaatorid. Maatriksit $\{x_{kl}\}_{k \in s; l=1,\ldots,m}$ illustreerib Joonis 1.



Joonis 1: Skemaatiline näide maatriksist $\{x_{kl}\}_{k \in s; l=1,\ldots,m}$ (Matei ja Ranalli, 2015).

Eeldame, et kadu objekti tasemel põhjustavad samad faktorid, mis viivad uuritava tunnuse väärtuse kaoni, st objekti ja väärtuse kaod on mittejuhuslikud. See on latentse tunnuse abil lähenemise puhul fundamentaalne eeldus, mis peab kindlasti täidetud olema.

Olgu q_{kl} objekti k küsimusele l vastamise tõenäosus iga $l = 1, \ldots, m$ jaoks, kus $k \in r$. Nagu ka objekti kao puhul, esitatakse q_{kl} siin uuritava tunnuse funktsioonina kasutades logistilist regressiooni:

$$q_{kl} = P(x_{kl} = 1 \mid y_{kl}, \theta_k, R_k = 1) = \frac{1}{1 + e^{-(\beta_{l0} + \beta_{l1}\theta_k + \beta_{l2} y_{kl})}} \tag{12}$$

iga $l = 1, \ldots, m$ ja $k \in r$ korral, kus β_{l0}, β_{l1} ja β_{l2} on mudeli paramteerid. Kui väärtused y_{kl} on teada ainult objektidel, kelle puhul $x_{kl} = 1$ ja $k \in r$, siis mudelit (12) ei saa hinnata, sest igal tunnusel esineb ka väärtuse kadu ja osad y_{kl} on tundmatud. Objekti kao korral saab kasutada q_{kl} hindamiseks vaid latentset tunnust θ. Mudel (12) saab kuju:

$$q_{kl} = P(x_{kl} = 1 \mid \theta_k, R_k = 1) = \frac{1}{1 + e^{-(\beta_{l0} + \beta_{l1}\theta_k)}} \tag{13}$$

iga $l = 1, \ldots, m$ ja $k \in r$ korral. Mudel (13) ei ole tavapärane logistilise regressiooni mudel, sest väärtused θ_k on tundmatud. Sellises olukorras hinnatakse q_{kl}, θ_k ja mudeli parameetrid latentse tunnuse mudeli abil.

Üks lihtsamaid latentse tunnuse modelleerimise variante on nn Raschi mudel. Selle korral on mudelis (13) kordaja β_{l1} sama kõikide tunnuste jaoks. Kirjeldatud Raschi mudeli kuju on järgmine:

$$q_{kl} = \frac{1}{1 + e^{-(\beta_{l0} + \beta_1 \theta_k)}} \tag{14}$$

iga $l = 1, \ldots, m$ ja $k \in r$ korral. Parameeter β_{l0} hinnatakse iga tunnuse l jaoks ja see peegeldab tunnuse l vastamisosakaalu: suurem väärtus vastab suuremale vastamise protsendile. Parameeter β_1 on ühine kõikide tunnuste korral. Siiski see nõue võib osutuda liiga piiravaks ning mudel (13) töötab paremini kui (14).

2.3.1 Latentse tunnuse mudeli eeldused

Latentse tunnuse mudeli kasutamiseks peavad olema täidetud konkreetsed eeldused. Esimeseks eelduseks on nö tingliku sõltumatuse eeldus, mis nõuab tunnuse väärtuste omavahelist sõltumatust etteantud latentse tunnuse väärtuse korral. See tähendab, et latentne tunnus kajastab kogu sõltuvust vaadeldud x_{kl} korral. Tingliku sõltumatuse eeldust saab kontrollida vaid testides mudeli kooskõla andmetega. Latentse tunnuse mudel töötab hästi, kui latentsed tunnused kirjeldavad enamuse huvipakkuvate tunnuste seosest.

Teiseks eeldatakse monotoonsust: kui latentse tunnuse θ väärtus kasvab, siis ka tunnusele vastamise tõenäosus kasvab või on tunnusega θ samas intervallis. Mida suurem on väärtus θ_k, seda suurem on objekti k vastamise tahe.

Viimane ja ilmselt tähtsaim eeldus on ühemõõtmelisus, mis tähendab, et vaid üks latentne tunnus suudab ära seletada täielikult objekti k tahet vastata kõikidele huvipakkuvatele küsimustele.

Kõik need põhilised eeldused viitavad sellele, et tunnuste x_{kl} omavahelised sõltuvused on ära kirjeldatud latentse tunnuse θ poolt. Tunnuse θ väärtus θ_k peegeldab objekti k soovi vastata küsimustele ning objekti k tõenäosus vastata etteantud küsimusele kasvab koos latentse tunnuse väärtuse θ_k kasvuga.

2.3.2 Vastamistõenäosuse p_k hindamine

Latentse tunnuse mudelist saadud informatsiooni abil vastamistõenäosuse p_k hindamiseks pakutakse artiklis (Matei ja Ranalli, 2015) järgmine lahendus:

Esimene samm: Esiteks, leitakse θ_k hinnang $\hat{\theta}_k$, kus $k \in r$, mudeli (13) abil. Antud töö simuleerimise osas kasutatakse selle jaoks *R* paketti *ltm* (Rizopoulos, 2006). Selle abil saadakse ka hinnanguid vastamistõenäosustele q_{kl}, kus $k \in r$.

Teine samm: Hinnagu $\hat{\theta}_k$ leidmiseks iga $k \in \overline{r}$ jaoks eeldame, et objekti tasemel kadu on tunnuse väärtuse kao erijuht. See tähendab, et mittevastaja ei vasta ühelegi tunnusele l ja iga $l = 1, \ldots, m$ korral $x_{kl} = 0$. Hinnag $\hat{\theta}_k$ iga $k \in \overline{r}$ jaoks leitakse järgmiselt: vastanute hulka r lisatakse kujutletav vastaja $\widetilde{k}$, kelle vastamisindikaator $x_{\widetilde{k}l} = 0$ iga $l = 1, \ldots, m$ korral. See uus hulk tähistatakse $\widetilde{r} = r \cup \widetilde{k}$. Taas kasutatakse mudelit (13), kuid seekord objektide $k \in \widetilde{r}$ jaoks, ja leitakse hinnangute $\hat{\theta}_k$ uued väärtused, kus $k \in \widetilde{r}$. Lisatud objekti $\widetilde{k}$ latentse tunnuse hinnang on $\hat{\theta}_{\widetilde{k}}$. Iga $k \in \overline{r}$ korral määratakse $\hat{\theta}_k = \hat{\theta}_{\widetilde{k}}$. Seega, iga objekti $k \in \overline{r}$ korral on väärtus $\hat{\theta}_k$ sama. Selle meetodi korral on iga objekt $k \in s$ seotud hinnanguga $\hat{\theta}_k$, kus objekti $k \in r$ hinnang $\hat{\theta}_k$ on hinnatud esimesel sammul ning objekti $k \in \overline{r}$ hinnang $\hat{\theta}_k$ teisel sammul.

Kolmas samm: Esimestel sammudel leitud hinnanguid $\hat{\theta}_k$ (iga $k \in s$ jaoks) kasutatakse argumenttunnustena mudelis (11) tundmatute väärtuste θ_k asemel:

$$p_k = P(R_k = 1 \mid \hat{\theta}_k) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 \hat{\theta}_k)}}. \tag{15}$$

Mudel (15) on tavaline logistilise regressiooni mudel, mille abil saab leida vastamistõenäo-

suste p_k hinnangud $\hat{p}_k$ iga $k \in s$ jaoks.

2.4 Hinnang kogusummale latentse tunnuse meetodil

Tuletame meelde, et meie uuritavaks tunnuseks on y_j, mille puhul esines tunnuse väärtuse kadu. Uuritava tunnuse kogusumma hindamiseks kasutame lineaarset hinnangut (6), kus vastamistõenäosuse hinnang $\hat{p}_k$ on saadud mudeli (15) abil ja tunnusele y_j vastamistõenäosuse hinnag $\hat{q}_{kj}$ on saadud mudeli (13) abil. Hinnangu (6) omadused sõltuvad objekti ja tunnuse väärtuse tasemel mittevastamise mehhanismist.

3 Simuleerimisülesanne

3.1 Andmestiku kirjeldus

Antud töö praktilises osas on läbi tehtud artiklis (Matei ja Ranalli, 2015) esitatud esimese simulatsiooni näide. Üldkogumist on võetud korduvalt valimeid ja võrreldud erinevate hinnangute omadusi omavahel. Simuleerimisülesande teostamiseks on kasutatud rakendustarkvara *R* paketti *ltm*, mis on vajalik latentsete tunnuste modelleerimiseks ja vastamistõenäosuste hinnangute leidmiseks.

Vastamise modelleerimiseks on kasutatud samu valemeid, mida artikliski kasutatud on, kuid kõik tarkvaraprogrammi *R* koodid on koostatud antud töö autori poolt.

Andmestik koosneb neljast binaarsest tunnusest, mis on saadud 1986. aasta Suurbritannia sotsiaalseid hoiakuid käsitlevast uuringust (Social and Community Planning Research, 1988). Andmed kirjeldavad inimeste suhtumist aborti. Kasutatav andmestik on kättesaadav tarkvaraprogrammi *R* paketis *ltm*.

Neljale uuringuküsimusele vastas 379 inimest, seega on üldkogumi suurus $N = 379$. Inimestelt küsiti, kas seadused peaksid lubama teha aborti järgmistes olukordades:

1. Naine otsustab ise, et ta ei taha last endale jätta.
2. Paar otsustab ühiselt, et ei soovi last.
3. Naine ei ole abielus ja ei taha mehega abielluda.
4. Paar ei saa endale rohkem lapsi lubada.

Tunnuse väärtus on 1, kui inimene arvab, et antud olukorras võiks abordi tegemine olla seadusega lubatud, ning 0, kui ta on sellele vastu. Uuritavaks tunnuseks on valitud teise tunnuse ehk y_2, mille üldkogumi kogusumma tegelik väärtus on $Y_2 = 255$.

3.2 Hinnangud ja täpsusnäitajad

Valimi võtmiseks on kasutatud lihtsat juhuslikku valikut tagasipanekuta. Lihtsa juhusliku valiku korral on kõikidel objektidel võrdne valimisse sattumise tõenäosus ehk igal üldkogumi objektil on π_k väärtus sama. Kuna $\omega_k = \frac{1}{\pi_k}$, siis ka ω_k on igal objektil sama. Tähistame $\omega_k = \omega \quad \forall\, k \in U$ korral.

Töös on käsitletud järgmisi hinnanguid:

1. **Horvitz-Thompsoni hinnang** (2). Selle korral on kõik valimisse sattunud objektid vastanud küsimustikule, st hinnang leitakse kogu valimi objektide andmete põhjal. See hinnang tuuakse välja võrdluseks hinnangutega, mis on leitud ainult vastanute andmete põhjal. Kuna kõikidel objektidel on valimisse sattumise tõenäosus võrdne, siis saame valemit (2) lihtsustada:
$$\hat{Y}_{2,s} = \omega \sum_{s} y_{k2}$$

2. **Alternatiivne kogusumma hinnang** (4). Lihtsa juhusliku valiku korral lihtsustub see kujule:
$$\hat{Y}_{2,alt} = \frac{\sum_{r_2} y_{k2}}{n_{r_2}} \cdot N,$$
kus n_{r_2} on 2. küsimusele vastanute arv.

3. **Lineaarne hinnang** (5), kus kasutame vastamistõenäosuste p_k ja q_{k2} tegelikke väärtusi. Lihtsa juhusliku valiku korral:
$$\hat{Y}_{2,lin,tegelik} = \omega \sum_{r_2} \frac{y_{k2}}{p_k q_{k2}}$$

4. **Lineaarne hinnang** (6), kus kasutame vastamistõenäosuste p_k ja q_{k2} hinnanguid $\hat{p}_k$ ja $\hat{q_{k2}}$. Lihtsa juhusliku valiku korral:
$$\hat{Y}_{2,lin} = \omega \sum_{r_2} \frac{y_{k2}}{\hat{p}_k \hat{q_{k2}}}$$

Hinnangute iseloomustamiseks on kasutatud erinevaid täpsusnäitajaid. Lihtsuse mõttes on tähistatud edaspidi hinnangut $\hat{Y}_2$ lihtsalt $\hat{Y}$. Hinnangute analüüsimiseks kasutatud täpsusnäitajad on järgmised:

1. **Monte-Carlo nihe**:
$$B = E_{sim}(\hat{Y}) - Y,$$
kus $E_{sim}(\hat{Y}) = \frac{\sum_{i=1}^{R} \hat{Y}_i}{R}$, $\hat{Y}_i$ on hinnangu $\hat{Y}$ väärtus i. simulatsioonis ja R on kõikide simulatsioonide arv;

2. **suhteline nihe**:
$$RB = \frac{B}{Y};$$

3. **Monte-Carlo standardhälve**:
$$\sqrt{VAR} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(\hat{Y}_i - E_{sim}(\hat{Y}))^2};$$

4. **Monte-Carlo ruutkeskmine viga**:
$$MSE = B^2 + VAR.$$

3.3 Mittevastamise modelleerimine

Vastamine peab sõltuma huvipakkuvatest tunnustest, mis omakorda peavad olema kirjeldatud latentse tunnuse θ poolt. Seetõttu arvutatakse latentse tunnuse θ väärtused θ_k^a terves üldkogumis kõigi nelja tunnuse alusel kasutades selleks mudelit (13). Vaadeldavate tunnuste ja latentse tunnuse vahelise seose hindamiseks on vaadeldud korrelatsioone y_{kl} ja θ_k^a vahel iga $l = 1, ..., 4$ korral. Saadud korrelatsioonide väärtused on toodud Tabelis 2.

Tabel 2: Tunnuste y_{kl} ja θ_k^a vahelised korrelatsioonid.

	θ_k^a
y_{k1}	0.85
y_{k2}	0.85
y_{k3}	0.87
y_{k4}	0.81

Kõik saadud korrelastioonid on tugevad, seega võib arvata, et latentse tunnuse abil lähenemine töötab nende andmete puhul hästi.

Mittejuhusliku mittevastamise simuleerimiseks genereeritakse kõigepealt vastamistõenäosused p_k üldkogumis kasutades järgnevat vastamise mudelit:

$$p_k = \frac{1}{1 + e^{-(0.7+y_{k2}+\theta_k+0.2\varepsilon_k)}}, \tag{16}$$

kus $\varepsilon_k \sim U(0, 1)$. Vastamistõenäosuste p_k keskmine üldkogumis on ligikaudu 0.75.

Igale konkreetsele tunnusele vastamist genereeritakse üldkogumis kasutades mudelit:

$$q_{kl} = \frac{1}{1 + e^{-(3\theta_k+a_l+y_{kl})}} \; \forall \; l = 1, ..., 4 \text{ jaoks}, \tag{17}$$

kus a_l võtab erinevaid väärtusi vastavalt l väärtusele: $a_1 = 1$, $a_2 = 0$, $a_3 = -0.5$ ja $a_4 = 1$. Vastavad tunnusele vastamise keskmised osakaalud üldkogumis on ligikaudu 35%, 43%, 47% ja 32%. Vastavad *R*-i käsud on Lisas 1.

Seejärel võetakse üldkogumist 1000 valimit mahuga $n = 50$ ning hiljem 1000 valimit mahuga $n = 100$. Igas valimis s saadakse vastanud objektide hulk r Poissioni valikuga kasutades vastamistõenäosusi p_k mudelist (16). Saadud vastanute hulga r korral konstrueeritakse maat-

riksi $\{x_{kl}\}_{k\in r;l=1,\ldots,4}$, kus väärtused x_{kl} on saadud samuti Poissioni valikuga mudeli (17) abil leitud vastamistõenäosuste q_{kl} järgi.

Iga koostatud valimi korral leitakse pärast mittevastamise genereerimist latentse tunnuse hinnangud $\hat{\theta}_k$ vastanute hulgal r. Kõikide vastanute hulka kuuluvate objektide korral $k \in r$ leitakse hinnangud $\hat{q}_{kl}$ (vt Lisa 2).

Vastamistõenäosuste p_k hindamiseks kasutatakse kahte hulka: vastanute hulka r ja mittevastanute hulka $\overline{r}$, kusjuures $s = r \cup \overline{r}$. Hulgal s leitakse latentse tunnuse hinnangud $\hat{\theta}_{\tilde{k}}$. Hulgal $\overline{r}$ hinnangute leidmiseks kasutatakse metoodikat, mis on kirjeldatud peatükis 2.3.2. Lõpuks kujunevad latentse tunnuse hinnangud $\hat{\theta}_k$ valimis nii, et vastanute hulka kuuluvate objektide $k \in r$ latentse tunnuse hinnangud on $\hat{\theta}_k$, mille leidsime algselt ainult vastanute hulgal r, ja mittevastanute hulka kuuluvate objektide $k \in \overline{r}$ latentse tunnuse hinnangud on $\hat{\theta}_{\tilde{k}}$, mille leidsime hulgal s. Saadud hinnanguid $\hat{\theta}_k$ kasutatakse vastamistõenäosuste p_k hindamiseks tavalise logistilise regressiooni mudelis (15). Hinnanguid $\hat{p_k}$ ja $\hat{q_{kl}}$ kasutatakse lineaarse hinnangu (6) leidmisel (vt Lisa 2).

Keskmised mittevastamise osakaalud vaadeldava nelja tunnuse puhul üle simulatsioonide on ligikaudu 26%, 33%, 38% ja 23%.

3.4 Tulemused

Järgnevalt esitame peale simulatsioonide teostamist saadud hinnangute täpsusnäitajad valimimahtude $n = 50$ ja $n = 100$ korral.

Tabel 3: Saadud hinnangute näpsusnäitajad.

n=50				
Hinnang	B	$\sqrt{VAR}$	MSE	RB
$\hat{Y}_{2,s}$	0.7	26.1	679.9	0.003
$\hat{Y}_{2,alt}$	126.7	19.6	16451.1	0.563
$\hat{Y}_{2,lin,tegelik}$	1.7	36.6	1341.7	0.008
$\hat{Y}_{2,lin}$	-18.7	33.8	1495.5	0.563
n=100				
Hinnang	B	$\sqrt{VAR}$	MSE	RB
$\hat{Y}_{2,s}$	-0.07	16.2	263.7	-0.0003
$\hat{Y}_{2,alt}$	126.4	13.5	16165.8	0.562
$\hat{Y}_{2,lin,tegelik}$	-0.6	24.0	574.1	-0.003
$\hat{Y}_{2,lin}$	-17.2	22.5	803.7	-0.077

Hinnangute $\hat{Y}_{2,s}$ ja $\hat{Y}_{2,lin,tegelik}$ korral tulid nihked väiksed. See on oodatav, sest esimesel juhul leitakse hinnang valimi tegelike väärtuste abil ja teisel juhul kasutatakse tegelikke vastamistõenäosusi p_k ja q_{kl}. Alternatiivsel hinnangul on suur nihe, kuna objektidel, kellel on uuritava tunnuse väärtus 0, on väiksem vastamistõenäosus. Seda hinnangu leidmisel aga ei arvestata.

Võrreldes omavahel hinnanguid $\hat{Y}_{2,lin,tegelik}$ ja $\hat{Y}_{2,lin}$, selgub, et hinnang $\hat{Y}_{2,lin}$ põhjustab suuremat nihet. Põhjus on selles, et vastamistõenäosuste p_k ja q_{kl} hindamisel vastamismudeli

abil pole võimalik leida täpseid p_k ja q_{kl} väärtusi ning seega esineb ebatäpsusi juba varem. Hinnangu $\hat{Y}_{2,lin}$ nihe on küll suurem kui hinnangul $\hat{Y}_{2,s}$, kuid palju väiksem kui hinnangul $\hat{Y}_{2,alt}$. Samuti on hinnangu $\hat{Y}_{2,lin}$ ruutkeskmine viga (MSE) mitu korda väiksem kui alternatiivsel hinnangul.

Võrreldes omavahel erinevate valimimahtudega ($n = 50$ ja $n = 100$) teostatud simulatsioonidest saadud tulemusi, siis täpsusnäitajate põhjal on näha, et hinnangute nihked väga palju ei muutunud. Küll aga vähenesid hinnangute standardhälbed ja MSE-d. Kõige vähem paranes alternatiivne hinnang, ikka jääb selle MSE väga suureks.

Kuigi hinnangu $\hat{Y}_{2,lin}$ nihe on suurem kui hinnangutel $\hat{Y}_{2,lin,tegelik}$ ja $\hat{Y}_{2,lin}$, on see hinnang siiski hea. Praktikas pole üldjuhul teada uuritava tunnuse väärtused kogu valimi jaoks või vastamistõenäosuste p_k ja q_{kl} tegelikud väärtused. Lisaks sellele on vaadeldav hinnang parem alternatiivsest hinnangust.

Kokkuvõte

Mittejuhusliku mittevastamise korral on vastamise mehhanism mittejuhuslik. See tähendab, et inimesed jätavad küsimustele vastamata teatud põhjustel. Antud töö eesmärk on välja selgitada, kuidas kompenseerida mittejuhuslikku mittevastamist valikuuringutes jälgides inimeste üldist vastamise mustrit. Selleks kasutatakse latentseid ehk varjatud tunnuseid. Need hinnatakse logistilise regressioonmudeli abil, mida nimetatakse latentse tunnuse mudeliks. Varjatud tunnuste abil leitakse vastamistõenäosuste hinnangud, mida kasutatakse üldkogumi hinnangute leidmiseks.

Latentse tunnuse abil lähenemise korral ei ole mittevastamise kompenseerimiseks vaja teada lisainformatsiooni, vaid vastamise mustrit. Kui sobiv lisainformatsiooni siiski leidub, saab seda vastamistõenäosuste hindamisel kasutada .

Simuleerimisülesande tulemustest järeldub, et latentsete tunnuste abil leitud hinnang põhjustab suuremat nihet kui tegelike vastamistõenäosuste abil leitud hinnang. Selle põhjuseks on vastamistõenäosuste p_k ja q_{kl} hindamisel tekkivad mõningased ebatäpsused. Latentsete tunnuste abil leitud hinnang annab veidi kehvema tulemuse kui valimi objektide tegelike väärtuste järgi leitud hinnang ja tegelike vastamistõenäosuste abil leitud hinnang, kuid annab kindlalt parema hinnangu kui alternatiivne hinnang, mille leidmisel ei kasutata vastamistõenäosusi p_k ja q_{kl}. Praktikas on latense tunnuse abil leitud hinnang siiski hea, sest üldjuhul pole teada nii uuritava tunnuse väärtused kogu valimi jaoks kui ka vastamistõenäosuste p_k ja q_{kl} tegelikud väärtused.

Kasutatud kirjandus

Beaujean, A. A. (2014). Latent Variable Modeling Using R: A Step-by-Step Guide. New York: Routledge.

Matei, A., Ranalli, M. G. (2015). Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach. Survey Methodology, Vol. 41, No.1, pp. 145-164.

Prostakova, J. (2007). Mittevastamine ja selle kompenseerimine. Tartu.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. Journal of Statistical Software, Volume 17, Issue 5. Catholic University of Leuven.

Skrondal, A., Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Boca Raton [etc.] : Chapman & Hall/CRC.

Social and Community Planning Research. (1988). British Social Attitudes Survey, 1986. UK Data Service. SN: 2315, http://dx.doi.org/10.5255/UKDA-SN-2315-1.

Traat, I., Inno, J. (1997). Tõenäosuslik valikuuring. Tartu: Tartu Ülikool.

Traat, I. (2006). Matemaatilise statistika põhikursus. Tartu: Tartu Ülikool.

Lisad

Lisa 1. Programmi *R* koodi esimene osa.

```
library(ltm)
#Leiame latentsed tunnused, mis väljendab
#naise kalduvust teha aborti
fit<-ltm(Abortion~z1)
deeta<-factor.scores(fit, resp.patterns = Abortion)$score.dat$z1

#Leiame korrelatsioonid algse deeta ja uuritavate
#tunnuste vahel:
cor(Abortion,deeta)

#Genereerime p_k-d ja q_kl-d nagu töös kirjeldatud.
p=1/(1+exp(-(0.7+Abortion$`Item 2`+
               deeta+0.2*runif(nrow(Abortion)))))
q1=1/(1+exp(-(3*deeta+1+Abortion$`Item 1`)))
q2=1/(1+exp(-(3*deeta+0+Abortion$`Item 2`)))
q3=1/(1+exp(-(3*deeta-0.5+Abortion$`Item 3`)))
q4=1/(1+exp(-(3*deeta+1+Abortion$`Item 4`)))

#Tunnusele vastamise keskmised osakaalud üldkogumis
1-mean(q1)
1-mean(q2)
1-mean(q3)
1-mean(q4)
```

Lisa 2. Programmi *R* koodi teine osa.

Koodi teist osa ei saa kasutada ilma esimese osata.

Selle koodiga on näidatud, kuidas on genereeritud valimeid mahuga $n = 50$. Suurusega $n = 100$ valimite genereerimiseks tuleb vaid muuta muutuja Rep väärtust ($Rep = 100$).

```
# Võtame 1000 valimit lihtsa juhuvaliku abil.
# Igal sammul valimis:
# 1. Genereeritakse 1-0 tunnus p_k alusel Poissoni
# valiku abil vastanute hulga r genereerimiseks.
# Leitakse ka objektide arv vastanute valimis.
# 2. Seejärel vastanute hulgal r genereeritakse
# maatriks {x_{kl}} jällegi Poissoni valiku alusel
# Leitakse x_{kl} keskmised vastamise osakaalu saamiseks.
# 3. Saadud andmestiku põhjal leitakse prognoosid
# vastamistõenäosustele p_k ja q_{kl} töös kirjeldatud
# mudelite järgi (latentse mudeli abil).
# 4. Leitakse töös kirjeldatud neli hinnangut

N=nrow(Abortion)
Jrk=1:N
Rep=1000
n=50
n_r=rep(0,Rep) #vastanute arvud üle simulatsioonide
n1_r=rep(0,Rep)
n2_r=rep(0,Rep)
n3_r=rep(0,Rep)
n4_r=rep(0,Rep)
```

```
w=N/n #kaal iga objekti jaoks

HT=rep(0,Rep)
Alt=rep(0,Rep)
True=rep(0,Rep)
Lat=rep(0,Rep)

# Fun. vastanute hulga leidmiseks
vastamine <- function(tn){
  u=runif(n)
  r_n=rep(0,n)
  r_n[u<tn]=1
  return(r_n)}

for(i in 1:Rep){
  #valimi võtmine LJV TTA
  s=sample(Jrk,n)
  valimiandmed=Abortion[s,]
  p_v=p[s]
  q1_v=q1[s]
  q2_v=q2[s]
  q3_v=q3[s]
  q4_v=q4[s]

  #vastanute valimi r genereerimine
  u=runif(n)
  r=rep(0,n)
  r[u<p_v]=1
```

```
n_r[i]=sum(r)

#maatriksi {x_{kl}} moodustamine
r1=vastamine(q1_v)*r #need, kes pole valimis, ei saa vastata
n1_r[i]=sum(r1)
r2=vastamine(q2_v)*r
n2_r[i]=sum(r2)
r3=vastamine(q3_v)*r
n3_r[i]=sum(r3)
r4=vastamine(q4_v)*r
n4_r[i]=sum(r4)

# Hinnangud: HT, alternatiivne, Y_j,lin,tegelik
HT[i]=sum(valimiandmed$`Item 2`)*w
True[i]=w*sum(valimiandmed$`Item 2`[r2==1]/
                  (p_v[r2==1]*q2_v[r2==1]))
Alt[i]=N*sum(valimiandmed$`Item 2`[r2==1])/n2_r[i]

# Latentse tunnuse hinnangud leiame leiame eraldi
# vastanute hulgas ja kogu valimis
# Vastamistõenäosused q_{kl} hindame vaid vastanute põhjal
# Moodustame maatriksi x (mille maht on n_r rida),
# ehk ainult vastanute jaoks:
x=cbind(r1,r2,r3,r4)
rownames(x)=rownames(valimiandmed)
x=x[r==1,]

# Leiame latentse tunnuse hinnangu ja q_{k2}-d:
```

```
  fit2=ltm(x~z1)
  deeta_hat<-factor.scores(fit2, resp.patterns = x)$score.dat$z1

  q_k6ik=fitted(fit2, type="conditional-probabilities",
                resp.patterns = x)
  q2_hinnangud=rep(0,n)
  q2_hinnangud[r==1]=data.frame(q_k6ik)$r2

  # p_k hinnaguid leiame terve valimi põhjal:
  # selleks tuleb leia uued latentse tunnuse hinnangud
  # (väljaspool vastanute hulka) terve valimi s põhjal:
  x1=cbind(r1,r2,r3,r4)
  rownames(x1)=rownames(valimiandmed)
  fit3=ltm(x~z1)
  deeta_hat1<-factor.scores(fit3, resp.patterns = x1)$score.dat$z1
  deeta_hat1[r==1]=deeta_hat
  glm.fit1=glm(r~deeta_hat1,  family=binomial)

  # Saame p_k hinnangud
  pk_hat=predict(glm.fit1, type="response")
  koos=cbind(valimiandmed,q2_hinnangud,pk_hat,r2)

  # lin. hinnang, kus kasutame p_k ja q_{kl} hinnanguid
  Lat[i]=w*sum(koos$`Item 2`[koos$r2==1]/
                 (koos$q2_hinnangud[koos$r2==1]*
                    koos$pk_hat[koos$r2==1]))
}
```

```
# Mittevastanute osakaalud üle vastanute
1-mean(n1_r/n_r)
1-mean(n2_r/n_r)
1-mean(n3_r/n_r)
1-mean(n4_r/n_r)

# Monte-Carlo nihe
B1=mean(HT)-sum(Abortion$`Item 2`)
B2=mean(Alt)-sum(Abortion$`Item 2`)
B3=mean(True)-sum(Abortion$`Item 2`)
B4=mean(Lat[Lat<400])-sum(Abortion$`Item 2`)
# viimase hinnangu korral kasutame vaid neid saadud hinnangute
# väärtusi, mis on väiksema kui 400, sest vastasel juhul on
# ilmselt mingid p_k või q_{kl} hinnatakse 0-ga ja see
# annab meile vigased tulemused
c(B1,B2,B3,B4)

# Suhteline nihe
RB1=B1/sum(Abortion$`Item 2`)
RB2=B2/sum(Abortion$`Item 2`)
RB3=B3/sum(Abortion$`Item 2`)
RB4=B4/sum(Abortion$`Item 2`)

c(RB1,RB2,RB3,RB4)

# Monte-Carlo standardhälve
sqrt_var1=sqrt(1/(Rep-1)*sum((HT-mean(HT))^2))
sqrt_var2=sqrt(1/(Rep-1)*sum((Alt-mean(Alt))^2))
```

```
sqrt_var3=sqrt(1/(Rep-1)*sum((True-mean(True))^2))
sqrt_var4=sqrt(1/(Rep-1)*sum((Lat[Lat<400]-
                                mean(Lat[Lat<400]))^2))
c(sqrt_var1,sqrt_var2,sqrt_var3,sqrt_var4)

# Monte-Carlo MSE
MSE1=B1^2+sqrt_var1^2
MSE2=B2^2+sqrt_var2^2
MSE3=B3^2+sqrt_var3^2
MSE4=B4^2+sqrt_var4^2

c(MSE1,MSE2,MSE3,MSE4)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Grete Ojala,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
   „Kao kompenseerimine mittejuhusliku vastamise korral latentse tunnuse abil”,
   mille juhendaja on Natalja Lepik,

   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, 29.04.2016