Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Jaan Otter

Puuduvate andmete imputeerimine depressiooni hindavas küsimustikus

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: PhD Kelli Lehto,
MSc Anastassia Kolde

Tartu 2024

PUUDUVATE ANDMETE IMPUTEERIMINE DEPRESSIOONI HINDAVAS KÜSIMUSTIKUS

Bakalaureusetöö

Jaan Otter

Lühikokkuvõte

Andmete puudumine on oluline probleem andmestike analüüsil. Statistilise analüüsi käigus on sellest võimalik üle saada kasutades puuduvate andmete asendamist ehk imputeerimist. Imputeerimise võimalikuks puuduseks on andmeanalüüsi tulemuste korrektsus. Käesolev uurimistöö annab ülevaate erinevatest imputeerimismeetoditest ning nende rakendamisest puuduvaid andmeid sisaldavate depressiooniküsimustike analüüsil. Uurimistöö andmestiku moodustavad 87 042 TÜ Eesti geenivaramu geenidoonori vastused emotsionaalse enesetunde küsimustiku (EEK2) depressiooni alaskaala kaheksale küsimusele. Keskmiselt puudub 1,432% andmetest. Analüüsi eesmärgiks on hinnata, kas imputeerimismeetodi valik mõjutab depressiooniskoori seoseid depressioonidiagnoosiga. Koostatud ennustusmudelite põhjal võrreldakse kolme imputeerimismeetodit: listiviisiline kustutamine, keskmisega imputeerimine ning mitmene imputeerimine. Erinevaid imputeerimismeetodeid kasutades arvutatakse depressiooniskoor, mis kaasatakse kovariaadina ennustusmudelisse. Erinevatele ennustusmudelitele on leitud depressiooniskoorile šanssside suhe ning 95% usaldusintervall. Nende statistikute võrdlemisel selgub, et nende kolme imputeerimismeetodi kasutamisel on depressiooniskoori seosed depressioonidiagnoosiga sarnased.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Puuduvad andmed, imputeerimismeetodid, listiviisiline kustutamine, keskmisega imputeerimine, mitmene imputeerimine.

IMPUTATION OF MISSING DATA IN A DEPRESSION ASSESSMENT QUESTIONNAIRE

Bachelor thesis

Jaan Otter

Abstract

Missing data is a significant issue in the analysis of datasets. It can be overcome by using imputation for missing data during statistical analysis. One potential disadvantage for imputation is that the analysis results might be biased. The following research gives an overview of few imputation methods and their application in the analysis of depression questionnaires containing missing data. The dataset used for this research consists of responses of 87 042 participants in the Estonian Biobank's genetic database to eight questions of depression subset of emotional well-being questionnaire. On average, 1,432% of the data is missing. The aim of the analysis is to assess whether the choice of imputation method affects the relationship between depression score and depression diagnosis. After creating different prediction models, three imputation methods are compared: listwise deletion, mean imputation and multiple imputation. Depression scores are calculated and included as covariates for each prediction model. Odds ratios and 95% confidence intervals for depression scores are calculated for each prediction model. Comparing these statistics confirms that the use of three imputation methods does not affect the relationship between depression score and depression diagnosis.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Missing data, imputation methods, listwise deletion, mean imputation, multiple imputation.

Sisukord

Sissejuhatus

Andmete puudumine on sage probleem andmestike analüüsil. Peng et al. uurimuses esines puuduvaid andmeid 48% uuringutest, mis olid avaldatud üheteistkümnes haridus- ja psühholoogiateemalises ajakirjas aastatel $1998-2004$ [1]. Andmete puudumise mõju kvantitatiivsetele uuringutele on tõsine. See võib põhjustada parameetrite hinnangute nihet, informatsiooni kadu, standardvea suurenemist ning vähendada tulemuste üldistatavust [2].

Statistilise analüüsi käigus on sellest võimalik üle saada kasutades puuduvate andmete asendamist ehk imputeerimist [3]. Seejuures kasutatakse olemasolevat informatsiooni ja statistilisi eeldusi, et hinnata kogumi parameetreid [2]. Statistikas kasutatakse mitmeid erinevaid imputeerimismeetodeid. Imputeerimismeetodi valik võib mõjutada järgneva andmeanalüüsi tulemusi.

Käesolev uurimistöö annab ülevaate erinevatest imputeerimismeetoditest ning nende rakendamisest puuduvaid andmeid sisaldavate depressiooniküsimustike analüüsil. Töö praktiline osa uurib, kas imputeerimismeetodi valik mõjutab depressiooniskoori seoseid depressioonidiagnoosiga.

Analüüsi aluseks on TÜ Eesti geenivaramu uuringu "Heaolu ja vaimne tervis" andmed 87 042 inimese kohta. Töös kasutatakse instrumendi Emotsionaalse enesetunde küsimustik (EEK2)[4] depressiooni alaskaalat, mis koosneb kaheksast küsimusest. Mittevastamise tõttu esineb selles andmestikus puuduvaid andmeid. Kuna depressiooniskoori arvutatakse summeerides kaheksa küsimuse vastused skaalal $1-5$ [4], siis depressiooniskoori ei saa arvutada nendel inimestel, kes on vähemalt ühele küsimusele vastamata jätnud. See mõjutab andmete kasutamisvõimalusi. Minu uurimuse eesmärgiks on rakendada kolme erinevat imputeerimismeetodit ning teha kindlaks, kuidas need mõjutavad depressiooniskoori seost depressioonidiagnoosiga.

Vaimsel tervisel on inimeste elus väga oluline koht. Selle hindamiseks kasutatakse erinevaid uuringuid ja küsimustikke. Oluline on, et uuringutest tehtavad järeldused

on korrektsed. Oma uurimusega annan panuse imputeerimismeetodite võrdlusesse vaimse tervise andmestike analüüsil.

Käesolev uurimistöö koosneb kahest osast:

1. erinevate imputeerimismeetodite rakendamine emotsionaalse enesetunde küsimustiku (EEK2) depressiooni alaskaala puuduvate andmete asendamiseks;
2. hinnangu andmine, kas ja kuidas mõjutab imputeerimismeetodi valik küsimustikupõhise depressiooniskoori seost depressioonidiagnoosiga.

Esimese osa teostamiseks kasutatakse kolme erinevat imputeerimismeetodit:

- listiviisiline kustutamine;
- keskmisega imputeerimine;
- mitmene imputeerimine.

Teise osa teostamiseks koostatakse kolme erineva imputeerimismeetodiga saadud andmestiku kohta ennustusmudelid depressioonidiagnoosi ennustamiseks ning võrreldakse depressiooniskoori statistikuid.

1 Metoodika

1.1 Puuduvad andmed ning nende tekkemehhanismid

Puuduvate andmetena kirjeldatakse olukorda, kus objektil või subjektil puudub ühe või mitme vaatluse väärtus [5]. See on probleemiks paljudes teadustöödes [6]. Sageli on puuduvate andmete tekkepõhjuseks küsimusele vastamata jätmine uuritava(te) isiku(te) poolt [5].

Donald B. Rubin sõnastas 1976. aastal puuduvate andmete mehhanismide põhikontseptsioonid. Andmete puudumise mehhanisme on statistiline kirjandus tavaliselt jaganud kolmeks [7]:

- MCAR (*missing completely at random*) tähendab, et tõenäosus omada puuduvat väärtust ei sõltu mõõdetud ega mõõtmata andmetest. Praktikas esineb seda, kui vastaja unustab kogemata vastata ühele küsimusele [6];
- MAR (*missing at random*) tähendab, et andmete puudumine on süstemaatiliselt seotud mõõdetud andmetega, kuid süstemaatiline seos mõõtmata andmetega puudub [6]. Praktikas võib seda esineda siis, kui depressiooni raskusastet hindavas küsimustikus on meeste vastamismäär kõikidele küsimustele madalam kui naistel. Sellisel juhul on tõenäosus kõigile küsimustele vastamiseks seotud sootunnusega [5];
- NMAR (*not missing at random*) tähendab, et andmete puudumine on süstemaatiliselt seotud mõõtmata andmetega, kuid mõõdetud andmetega süstemaatiline seos puudub [6]. Praktikas võib seda esineda siis, kui rasket depressiooni põdevate vastajate kõikidele küsimustele vastamise määr depressiooni tõsidust hindavas küsimustikus on väiksem, kui teistel vastajatel. Seega on tõenäosus kõigile küsimustele vastamiseks seotud depressiooni tõsidusega [5].

Puuduvate andmete asendamiseks kasutatakse statistikas imputeerimist, mille puhul eristatakse ühe või mitme imputatsiooniga meetodeid.

1.2 Lihtsad imputeerimismeetodid

Sagedamini kasutatavad imputeerimismeetodid, mis ei nõua mitut imputatsiooni on:

- listiviisiline kustutamine;
- keskmise, mediaani või moodiga imputeerimine.

Kõige lihtsam metoodika on **listiviisiline kustutamine** (*listwise deletion*), kus analüüsist jäetakse välja kõik vaatlused, millel mingi uuritava tunnuse väärtus puudub. Sellel meetodil on kaks olulist puudust:

- raiskamine - informatsioon läheb kaotsi [6];
- nihe - andmed võivad olla kallutatud [8].

Kui andmete puudumise tekkemehhanismi liigiks pole MCAR, siis listiviisilise kustutamise meetodi kasutamine võib tõsiselt kallutada keskmiste, regressioonikordajate ning korrelatsioonide hinnanguid, sest olemasolevate andmete kustutamisel suureneb erinevus tegelike andmete ja imputeeritud andmete vahel [9].

Keskmisega imputeerimisel asendatakse numbrilised puuduvad väärtused selle tunnuse olemasolevate väärtuste keskmise väärtusega. Keskmisega imputeerimine on kiire meetod, sest seal on vaja teha vaid üks imputatsioon. Keskmisega tasub imputeerida vaid siis, kui puuduvaid andmeid on vähe [9].

Kuigi keskmisega imputeerimise meetod on lihtne ning kiire [6], on sellel tõsiseid puudusi:

- alahindamine - see meetod alahindab andmete hajuvust;
- tunnuste vaheliste seoste muutumine;

- nihe - see meetod kallutab peaaegu kõiki hinnanguid peale keskmise hinnangu. Kui andmete puudumise tekkemehhanismi liigiks pole MCAR, siis kallutab see meetod ka keskmise hinnangut [9].

Alternatiivselt saab **moodiga imputeerida**, kui puuduvad andmed esinevad järjestustunnustel [9] või kui andmete puudumise tekkemehhanismi liigiks on NMAR [10], sest tõenäosus omada puuduvat väärtust sõltub vaid mõõtmata andmetest.

Mediaaniga imputeerimisel asendatakse puuduvad väärtused selle tunnuse väärtuste mediaaniga. See võib olla hea valik, kui andmestikus esineb ekstreemseid väärtuseid või kui andmete puudumise tekkemehhanismi liigiks on NMAR [10]. Kuigi mediaaniga imputeerimise meetodi näol on tegemist lihtsa meetodiga, annab see kallutatuid hinnanguid [11].

Lisaks eelnimetatud lihtsamatele puuduvate andmetega toimetuleku viisidele on üha rohkem kasutusel ja laialt levinud mitmene imputeerimine.

1.3 Mitmene imputeerimine

Erinevalt eelnevalt kirjeldatud meetoditest kasutatakse mitmesel imputeerimisel mitut imputatsiooni [9].

Mitmese imputeerimise eelised:

- täpsus - kuna imputatsioonide arv on vähemalt kaks ning igal imputatsioonil asendatakse puuduvaid väärtusi erinevalt, siis tekib mitu erinevat hinnangut ning see aitab tulemuste täpsuse suurenemisele kaasa [12];
- eristatavus - kõigepealt lahendatakse puuduvate andmete probleem ning seejärel nende andmete probleem, kus ei esine puuduvaid väärtusi. See omadus lihtsustab statistilist modelleerimist [9].

Mitmese imputeerimise puudused:

- algoritm ei pruugi koonduda. Koondumine näitab imputatsioonide stabiilsust [13] ning kui seda ei juhtu, siis võib algoritm anda valesid tulemusi;
- ajamahukus.

Mitmese imputeerimise metoodika lähenemine sõltub muutujate liigist. Kui muutujad on pidevad, siis puuduvad andmed imputeeritakse kasutades tinglikku jaotust, mis pärineb lineaarse regressiooni mudelist. Kui muutujad on kategoorilised, siis kasutatakse logistilise regressiooni mudelit [6].

Käesoleva uurimistöö raames kasutatakse mitmese imputeerimise meetodit MICE (*multiple imputation by chained equations*).

1.3.1 MICE meetod

MICE meetod avaldati 2011. aastal Stef van Buureni ning Karin Groothuis-Oudshoorni poolt [14].

MICE algoritmis kasutatakse ahelat ühel muutujal põhineval imputatsioonimeetoditest, mille valik sõltub imputeeritavate tunnuste tüübist:

- arvuliste tunnuste korral kasutatavad:
  - pmm - ennustav keskmiste väärtustega sobitamine (*predictive mean matching*);
  - norm - Bayesi lineaarne regressioon *(Bayesian linear regression)*;
  - norm.nob - lineaarne regressioon (*linear regression*);
  - mean - tingimuseta keskmisega imputeerimine (*unconditional mean imputation*);
  - 2L.norm - kahetasemeline lineaarne mudel (*two-level linear model*);

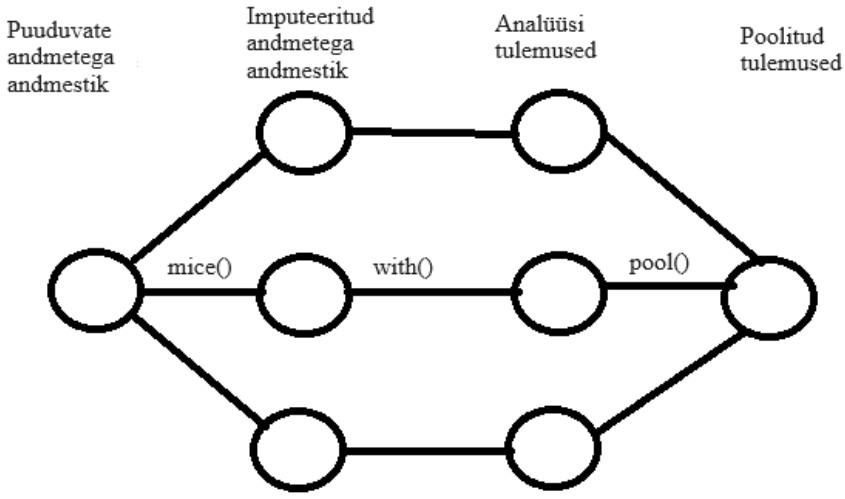
  - logreg - logistiline regressioon (*logistic regression*);

- mittearvuliste tunnuste korral kasutatavad:
  - polyreg - multinomiaalne logit mudel (*multinomial logit model*);
  - polr - järjestatud logit mudel (*ordered logit model*);
  - lda - lineaarne diskriminandianalüüs (*linear discriminant analysis*) [15].

MICE algoritmi puhul tuleb täpsustada, millist ühel muutujal põhinevat imputatsioonimeetodit kasutatakse [9]. Selles uurimistöös on mitmesel imputeerimisel ühel muutujal põhineva imputatsioonimeetodina kasutusel ennustav keskmiste väärtustega sobitamine (*pmm*).

MICE algoritm koosneb järgmistest etappidest, nagu seda on illustreeritud joonisel 1:

- puuduvate andmete imputatsioon (*mice()*). Kui m on imputatsioonide arv, siis tekib m erinevat imputeeritud andmestikku, kus pole puuduvaid andmeid;
- imputeerimisel saadud andmestike analüüs, mille käigus leitakse hinnangud kõigis imputeeritud andmestikes (*with()*);
- parameetrite hinnangute poolimine üheks lõplikuks hinnanguks (*pool()*) [9].



Joonis 1: MICE põhilised etapid [15] põhjal.

Mitmest imputeerimist kirjeldades on autorid erinevatel seisukohtadel vajaliku imputatsioonide arvu osas. Kui enamasti soovitatakse teha viis imputatsiooni [14], siis White et al. soovitab reprodutseeritava tulemuse saamiseks kasutada imputatsioonide arvu, mis on võrdne või suurem, kui puuduvate andmete määr protsentides [16]. Suurema hulga puuduvate andmete korral on soovitatav teha rohkem imputatsioone suurema standardvea kompenseerimiseks [2].

Ühe iteratsiooni moodustavad m imputatsiooni. Iteratsioone tuleb korrata kuni keskmine ja standardhälve koonduvad. Enamasti ei ole vaja teha rohkem kui viis iteratsiooni. Kui koonduvust ei saavutata, siis tuleb iteratsioonide arvu suurendada [17].

1.3.2 Ennustav keskmiste väärtustega sobitamine

Selles uurimistöös on mitmesel imputeerimisel ühel muutujal põhineva imputatsioonimeetodina kasutusel ennustav keskmiste väärtustega sobitamine (*pmm*). Ennustava keskmiste väärtustega sobitamist saab kasutada ainult pidevate tunnuste puhul [15].

Ennustav keskmiste väärtustega sobitamine toimub järgnevalt:

- muutujale x, mis sisaldab puuduvaid andmeid, sobitatakse imputatsioonimudel parameetritega α ning kovariaatidega z;
- kõikide indeksite h korral arvutatakse lineaarne ennustaja $\alpha^{obs}z_h$;
- kõikide indeksite j korral arvutatakse lineaarne ennustaja $\alpha^{mis}z_j$;
- olemasolevad andmed, mis on kõige lähemal lineaarse ennustaja väärtusele valitakse doonorite kandidaatide hulka D, kandidaatide arv on k;
- juhuslikult valitakse üks kandidaat välja ning keskmiste väärtustega sobitamise algoritm imputeerib x_h väärtuse selle kandidaadi väärtusega [18].

Seejuures on muutujad defineeritud:

- x - tunnus, kus esineb puuduvaid väärtusi;
- α - parameeter;
- z - kovariaat;
- h - indeks, kus tunnusel x väärtused on olemas;
- j - indeks, kus tunnusel x väärtused puuduvad;
- α^{obs} - parameetri väärtus, mis põhineb olemasolevatel andmetel;
- α^{mis} - parameetri väärtus, mis põhineb järeljaotusel;
- D - doonorite kandidaatide hulk;
- k - doonorite hulga võimsus (doonorite arv), enamasti on see MICE paketis viis [18].

Ennustava keskmiste väärtustega sobitamisel alustab algoritm ennustava kauguse arvutamisega iga j kohta:

$$\delta_{hj} = \alpha^{mis} z_j - \alpha^{obs} z_h$$

Seejärel valitakse k kõige väiksemat ennustavat kaugust ning lõpuks valitakse k kõige väiksema ennustava kauguse seast juhuslikult üks ennustav kaugus ning sellele ennustavale kaugusele vastav x_j imputeeritakse x_h väärtuseks [18].

Usaldusväärsuse jaoks on oluline, et mitmesel imputeerimisel kasutusel olev MICE algoritm koonduks [19] ning selleks kasutatakse imputeeritud andmestike kontrolli.

1.3.3 Imputeeritud andmestike kontroll

Enamasti saab MICE algoritmi koonduvust kontrollida tehes graafikud (*plot()*) iga imputatsiooni keskmise ning standardhälbe kohta igal iteratsioonil. Kui algoritm koondub, siis peaksid erinevad vood üksteisega vabalt segunema ning erinevate jadade varieeruvus ei ole suurem kui varieeruvus igas üksikus jadas [15].

Üldjuhul on oluline, et imputeeritud ning olemasolevate andmete vahel poleks suuri erinevusi [9].

Täiendavalt saab imputatsioonide korrektsuse kontrolliks teha joonised olemasolevate ning imputeeritud väärtuste tiheduste kohta (*densityplot()*) ning neid omavahel võrrelda. Tiheduste erinevuse puhul ei pruugi imputatsioon olla korrektne [15].

Kui andmeid on vähe, siis on kasulik visualiseerida graafik (*stripplot()*), mis näitab erinevate vaatluste väärtusi (kindlaks määratud tunnuse puhul) erineval imputatsioonil [9].

Kui vaatlusi on rohkem, siis on mõttekam teha karpdiagramm (*bwplot()*) ning sealt uurida seda, kas olemasolevate ning imputeeritud andmete vahel on erinevusi [9].

Andmestiku kontrollile järgneb mudeli koostamine.

1.3.4 Logistiline regressioonimudel

Käesolevas töös uuritakse imputeerimismeetodi valiku mõju depressiooniskoori seostele depressioonidiagnoosiga. Selleks on vaja hinnata sündmuse toimumise tõenäosust, mida saab teha logistilist regressioonimudelit kasutades.

Logistilise mudeliga hinnatakse šansi logaritmi [20]:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{1}$$

Seejuurel on tähised defineeritud järgnevalt:

- k - tunnuste arv;
- x_i - tunnus, kus $i \in \{1, 2, \ldots, k\}$;
- β_i - koeffitsiendid, kus $i \in \{0, 1, 2, \ldots, k\}$;
- Y - sündmus;
- $\pi = P(Y = 1)$ - tõenäosus, et sündmus Y esineb;
- $1 - \pi = P(Y = 0)$ - tõenäosus, et sündmus Y ei esine;
- $\Pi = \frac{\pi}{1-\pi}$ - sündmuse šanss;
- $OR = \frac{\Pi_i}{\Pi_j} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}} = \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)}$ - šansside suhe i-nda ning j-nda isiku kohta;
- $\eta = \text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$ - logit seosefunktsioon [20].

Avaldades võrdusest (1) π saame prognoosida sündmuse esinemist. Kuna $\eta = \ln\left(\frac{\pi}{1-\pi}\right)$, siis saame ka sellest võrdusest avaldada π.

Matemaatiliselt, kui $\ln x = y$, siis $x = e^y$. Seega:

$$\frac{\pi}{1-\pi} = e^{\eta},$$

$$\pi = (1-\pi)e^{\eta},$$

$$\pi = e^{\eta} - \pi e^{\eta},$$

$$\pi + \pi e^{\eta} = e^{\eta},$$

$$\pi(1+e^{\eta}) = e^{\eta},$$

$$\pi = \frac{e^{\eta}}{1+e^{\eta}}.$$

Sündmuse esinemise prognoosimiseks: $\frac{e^{\eta}}{1+e^{\eta}}$.

Kui argument muutub ühe ühiku võrra, siis muutub ka šansside suhe [20].

Oletame, et tunnuse x_h (kus $h < k$) korral suureneb argument ühe ühiku võrra.

Sellisel juhul

$\eta = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_h(x_h + 1) + \ldots + \beta_k x_k = \beta_0 + \beta_1 x_1 + \ldots + \beta_h x_h + \beta_h + \ldots + \beta_k x_k = \eta + \beta_h$.

Sündmuse esinemist saab prognoosida seosega: $\frac{e^{\eta+\beta_h}}{1+e^{\eta+\beta_h}}$.

Kasutades sündmuste šansi valemit saame $\Pi = \frac{\frac{e^{\eta+\beta_h}}{1+e^{\eta+\beta_h}}}{1-\frac{e^{\eta+\beta_h}}{1+e^{\eta+\beta_h}}} = e^{\eta+\beta_h}$.

Kui argument oleks jäänud samaks, siis oleks $\Pi = e^{\eta}$.

Kasutades šansside suhte valemit saame, et šansside suhe muutub argumendi ühe ühiku võrra suurenemisel $\frac{e^{\eta+\beta_h}}{e^{\eta}} = e^{\beta_h}$ korda.

Kui argument muutub c ühiku võrra, siis muutub šansside suhe $e^{c\beta_h}$ korda. Kui mudelis on mitu argumenti, siis kehtivad seosed šansside suhte muutuse kohta vaid siis, kui ülejäänud argumendid jäävad konstantseks [20].

Logistilist regressioonimudelit kasutatakse imputeeritud andmestike analüüsi etapis ning saadakse hinnangud iga imputeeritud andmestiku kohta. Sellele järgneb hinnangute poolimine.

1.3.5 Hinnangute poolimine

Mitmesel imputeerimisel saadakse analüüsi käigus iga erineva imputeeritud andmestiku hinnangud. Tulemuse korrektsuse seisukohast on oluline need poolimise abil ühendada üheks hinnanguks [21].

Donald Rubin avaldas 1987. aastal poolimise reeglid [9]. Rubini reeglite puhul on eelduseks, et parameetrite hinnangud on normaaljaotusega ning kui see eeldus pole täidetud, siis need hinnangud transformeeritakse [22].

Käesoleva uurimistöö tulemuste korrektsuse seisukohalt on oluline, et m imputeeritud andmestiku andmete hinnangute standardvead on õigesti kokku võetud. Seda saab teha kasutades Rubini standardvea poolimise reeglit [22]:

$$SE_{Pooled} = \sqrt{V_{Total}},$$

kus $V_{Total} = V_W + V_B + \frac{V_B}{m}$.

Seejuures imputeeritud andmete seesmine varieeruvus (*within imputation variance*) avaldub seosest [22]:

$$V_W = \frac{1}{m} \sum_{i=1}^{m} SE_i^2,$$

kus m on imputeeritud andmestike arv ning SE_i^2 on i-nda andmestiku hinnangu standardvea ruut.

Imputeeritud andmete vaheline varieeruvus (*between imputation variance*) avaldub seosest [22]:

$$V_B = \frac{\sum_{i=1}^{m} (\theta_i - \bar{\theta})^2}{m-1},$$

kus m on imputeeritud andmestike arv, $\bar{\theta}$ on poolitud hinnang ning θ_i on parameetri hinnang andmestikus i.

Et leida 95% usaldusintervall poolitud parameetrile, saame kasutada seost [22]:

$$\bar{\theta} \pm t_{df,1-\frac{\alpha}{2}} \cdot SE_{Pooled},$$

kus $\bar{\theta}$ on poolitud hinnang, t on t-jaotusest pärit statistik, df on vabadusastmete arv, α on olulisusnivoo (0,05) ning SE_{Pooled} on poolitud standardviga.

Kuna uurimistöös on statistikuks šansside suhe ning šansside suhte hinnangud pole normaaljaotusega, siis on soovituslik kasutada logaritmilist transformatsiooni [9] ning pärast seda leida poolitud standardviga ning usaldusintervall.

Õigesti kokku võetud hinnangud võimaldavad mudeli põhjal järeldusi teha.

1.4 Statistiline ennustusmudel ja imputeerimismeetodite võrdlus

Käesolevas uurimistöös kasutatakse logistilist regressioonimudelit, mille abil hinnatakse depressiooniskoori seost depressioonidiagnoosiga.

Parameetri statistilist olulisust mudelis kontrollitakse hüpoteesidega:

- H_0: parameeter ei ole ennustusmudelis statistiliselt oluline;
- H_1: parameeter on ennustusmudelis statistiliselt oluline.

Olulisusnivool $\alpha = 0{,}05$ on depressiooniskoor statistiliselt oluline siis, kui $p \leq 0{,}05$.

Töös võrreldakse kolme erinevat imputeerimismeetodit: listiviisiline kustutamine, keskmisega imputeerimine ning mitmene imputeerimine. Erinevate meetoditega saadakse imputeerimisel erinevad andmestikud, mis omakorda muudavad logistilise regressioonimudeli kovariaatide koeffitsiente ning šansside suhet.

Erinevate ennustusmudelite depressiooniskoori šansside suhte abil võrreldakse imputeerimismeetodite valiku mõju analüüsi tulemustele - kui need on sarnased, siis imputeerimismeetodi valik ei mõjuta depressiooniskoori seoseid depressioonidiagnoosiga.

Analüüsi teostamiseks kasutatakse rakendustarkvara R (4.3.2).

Kasutatakse järgnevaid pakette:

- *mice*;
- *dplyr*;
- $sjPlot$.

Paketti *mice* kasutatakse mitmesel imputeerimisel. Pakett *dplyr* on kasutusel andmestike teisendamiseks. Pakett $sjPlot$ on kasutusel mudelite võrdlemisel.

2 Analüüs

2.1 Andmestiku kirjeldus

Analüüsiks kasutatavad andmed on väljastatud TÜ Eesti geenivaramu poolt. Andmete kasutamiseks andis loa Eesti Bioeetika ja Inimuuringute Nõukogu.

Andmestikus on andmed 87 042 isiku kohta, kes on täitnud Emotsionaalse enesetunde küsimustiku (EEK2). Andmestikus sisalduvad vastajaid kirjeldavad andmed ning nende vastused depressiooni alaskaala küsimustikule.

Vastajaid kirjeldavad tunnused:

- uuringus osalemise pseudonüüm;
- sugu;
- sünniaasta;
- küsimustikule vastamise kuupäev;
- depressioonidiagnoosi olemasolu.

EEK2 depressiooni alaskaala sisaldab järgmist küsimust: "Palun lugege tähelepanelikult läbi alltoodud loetelu probleemidest ja vaevustest, mis võivad inimestel mõnikord esineda. Valige vastus, mis kõige paremini kirjeldab seda, kuivõrd see probleem on teid VIIMASE KUU jooksul häirinud". Vastajal palutakse anda hinnang järgmise kaheksa parameetri osas:

1. kurvameelsus;
2. huvi kadumine;
3. alaväärsustunne;
4. enesesüüdistused;

5. korduvad surma- või enesetapumõtted;
6. üksildustunne;
7. lootusetus tuleviku suhtes;
8. võimetus rõõmu tunda.

Vastajal on võimalik valida kuue vastusevariandi vahel:

- üldse mitte;
- harva;
- mõnikord;
- sageli;
- pidevalt;
- ei soovi vastata.

Samuti on vastajal võimalik jätta üldse küsimusele vastamata.

Analüüsi teostamiseks teisendatakse Kõigi kaheksa küsimuse vastusevariandid numbriliseks, mille summeerimisel saadakse uus tunnus *Depressiooniskoor*:

- üldse mitte - 1;
- harva - 2;
- mõnikord - 3;
- sageli - 4;
- pidevalt - 5.

Kui vastaja vastas vastusevariandiga "ei soovi vastata", siis võrdsustatakse see puuduva väärtusega.

807-l vastajal puudub informatsioon selle kohta, kas neil on varem olnud depressioonidiagnoos, seega analüüsiti 86 235 vastaja andmeid.

2.2 Vastajate kirjeldus

Vastanutest 25 325 (29,4%) on mehed ning 60 910 (70,6%) on naised.

Kuna vastajate sünnipäev pole teada, siis on vanus vastamise ajal defineeritud järgnevalt:

$$\text{Vanus} = \text{vanus 2020. aasta 31. detsembril} - \text{sünniaasta}.$$

Selline käsitlus pole väga täpne, sest tunnuse $Vanus$ väärtus võib olla tegelikust vanusest vastamise ajal ühe aasta võrra noorem. Minimaalne tunnuse $Vanus$ väärtus on 18. 2020. aasta lõpus oli vastajate mediaanvanus 48 ning keskmine vanus 48,4 aastat. Kõige vanem geenidoonor, kes vastas küsimustikule, oli 2020. aasta 31. detsembril 101-aastane.

Vastajate vanuseline jaotus on tabelis 1. Kõige rohkem on vastajaid vanusevahemikus $50 - 55$ (11,7%).

Tabel 1: Vastajate vanuseline jaotus.

Vanuserühm	**Sagedus**	**Osakaal (%)**
15-20	7	0,0
20-25	3402	3,9
25-30	5512	6,4
30-35	8811	10,2
35-40	9075	10,5
40-45	9419	10,9
45-50	10 028	11,6
50-55	10 117	11,7
55-60	8875	10,3
60-65	7705	8,9
65-70	5808	6,7
70-75	4159	4,8
75-80	2011	2,3
80-85	1003	1,2
85-90	251	0,3
90-95	41	0,0
95-100	10	0,0
100-105	1	0,0

Küsimustikule vastamine toimus 2021. aastal.

Depressioon on diagnoositud 23 609 (27,4%) vastajal.

EEK2 depressiooni alaskaala küsimustik on täielikult vastatud 84 180 (96,7%) vastaja poolt. Täielikult jättis EEK2 depressiooni alaskaala küsimustiku täitmata 1060 (2,1%) vastajat.

2.3 Andmete puudumiste mustrid

Küsimustele vastanute jaotus on esitatud tabelis 2. Vähemalt ühele küsimusele on vastamata jätnud 2055 (3,3%) vastajat.

Tabel 2: Erinevale arvule küsimustele vastanute jaotus.

Vastatud küsimusi	**Vastajate arv**	**Osakaal (%)**
0	1060	2,1
1	6	0,0
2	5	0,0
3	7	0,0
4	19	0,0
5	60	0,1
6	137	0,2
7	761	0,9
8	84180	96,7

Kõigi vastajate andmeid arvesse võttes on vastamata jäetud küsimuste arvu keskmine 0,115. 95% usaldusintervalli põhjal võime 95% kindlusega öelda, et vastamata jäetud küsimuste arvu keskmine jääb 0,109 ja 0,121 vahele [23]. Kui arvestada vaid neid vastajaid, kes on vähemalt ühele küsimusele vastamata jätnud, siis keskmine vastamata jäetud küsimuste arv on 4,8.

Kõigi vastajate andmeid arvesse võttes on vastamata jäetud küsimuste arvu mediaan ning standardhälve vastavalt 0 ja 0,9.

Depressiooniskoori pole võimalik arvutada, kui vastaja on jätnud vähemalt ühele küsimusele vastamata. Käesolevas analüüsis on depressiooniskoor puudu 2055-l vastajal.

Tabelis 3 on esitatud iga küsimuse vastamismäär.

Vastamismäär arvutatakse valemiga:

$$\text{vastamismäär} = 1 - \frac{p}{n},$$

- p on sellele küsimusele vastamata jätnud vastajate arv;
- n on kõigi analüüsi jäetud vastajate arv (86 235).

Tabel 3: Küsimuste vastamismäärad teemade kaupa.

Küsimuse teema	**Vastamata jätnud**	**Vastamismäär (%)**
kurvameelsus	1140	98,68
huvi kadumine	1122	98,70
alaväärsustunne	1277	98,52
enesesüüdistused	1207	98,60
korduvad surma- või enesetapumõtted	1542	98,21
üksildustunne	1120	98,70
lootusetus tuleviku suhtes	1256	98,54
võimetus rõõmu tunda	1214	98,59

Tabelist 3 järeldub, et kõige madalam vastamismäär on surma- või enesetapumõtteid puudutaval küsimusel. Kõige kõrgem vastamismäär on üksildustunnet puudutaval küsimusel. Keskmiselt on igale küsimusele vastamata jätnud $(1140 + 1120 + 1277 + 1207 + 1542 + 1120 + 1256 + 1214)/8 = 1234{,}75$ vastajat, mis moodustab 1,4% kõigist analüüsi jäetud vastajatest.

Logistilise regressioonimudeli kasutamiseks on EEK2 depressiooni alaskaala kaheksa küsimust teisendatud binaarseks. Seejuures 1 tähendab, et andmed puuduvad ning 0 tähendab, et andmed ei puudu. Iga küsimuse kohta eraldi on loodud logistiline regressioonimudel.

Tabelis 4 on esitatud logistilise regressioonimudeli abil leitud statistikud, mis näitavad küsimuste kaupa seost andmete puudumise ning vastaja soo vahel. Kui koefitsient on negatiivne, siis on ka seos negatiivne ning vastupidi.

Andmete analüüsil selgub, et igal küsimusel esineb negatiivse suunaga seos andmete puudumise ning vastaja soo vahel. Järelikult naissoost vastajatel on iga küsimuse puhul andmete puudumise šanss väiksem kui sama vanal meessoost vastajal.

Tabel 4: Andmete puudumise seos vastaja sooga.

Küsimuse teema	**Koeffitsient**	***p*-väärtus**	**Olulisusnivoo**
kurvameelsus	-0,524	$< 2 \cdot 10^{-16}$	0,05
huvi kadumine	-0,519	$< 2 \cdot 10^{-16}$	0,05
alaväärsustunne	-0,445	$1{,}71 \cdot 10^{-14}$	0,05
enesesüüdistused	-0,495	$< 2 \cdot 10^{-16}$	0,05
korduvad surma- või enesetapumõtted	-0,304	$1{,}55 \cdot 10^{-8}$	0,05
üksildustunne	-0,515	$< 2 \cdot 10^{-16}$	0,05
lootusetus tuleviku suhtes	-0,421	$6{,}95 \cdot 10^{-13}$	0,05
võimetus rõõmu tunda	-0,453	$2{,}61 \cdot 10^{-14}$	0,05

Sarnaselt on tabelis 5 esitatud statistikud, mis näitavad küsimuste kaupa seost andmete puudumise ning vastaja vanuse vahel.

Tabel 5: Andmete puudumise seos vastaja vanusega.

Küsimuse teema	**Koeffitsient**	***p*-väärtus**	**Olulisusnivoo**
kurvameelsus	-0,037	$< 2 \cdot 10^{-16}$	0,05
huvi kadumine	-0,038	$< 2 \cdot 10^{-16}$	0,05
alaväärsustunne	-0,026	$< 2 \cdot 10^{-16}$	0,05
enesesüüdistused	-0,031	$< 2 \cdot 10^{-16}$	0,05
korduvad surma- või enesetapumõtted	-0,020	$< 2 \cdot 10^{-16}$	0,05
üksildustunne	-0,040	$< 2 \cdot 10^{-16}$	0,05
lootusetus tuleviku suhtes	-0,025	$< 2 \cdot 10^{-16}$	0,05
võimetus rõõmu tunda	-0,030	$< 2 \cdot 10^{-16}$	0,05

Igal küsimusel esineb negatiivse suunaga seos ka andmete puudumise ning vastaja vanuse vahel. See tähendab, et samast soost vanematel vastajatel on iga küsimuse puhul andmete puudumise šanss väiksem.

2.4 Listiviisiline kustutamine

Listiviisilisel kustutamisel jäetakse analüüsist välja kõik vastajad, kellel puudub küsimustikus vastus vähemalt ühele küsimusele. Kokku kaasatakse esialgsest 86 235-st

vastajast 84 180.

Listiviisilise kustutamise põhjal arvutatud depressiooniskoori kirjeldav statistika on esitatud tabelis 6.

Tabel 6: Depressiooniskoori kirjeldav statistika listiviisilisel kustutamisel.

Statistik	**Väärtus**
Vastajate arv	84 180
Mediaan	14
Miinimum	8
Maksimum	40
Keskmine	15,956
Standardhälve	6,732
Usaldusintervall keskmisele (95%)	(15,911; 16,002)

2.5 Keskmisega imputeerimine

Keskmisega imputeerimisel asendatakse puuduvad väärtused selle tunnuse olemasolevate väärtuste keskmise väärtusega. Seda meetodit soovitatakse kasutada ainult väikese hulga puuduvate andmete korral. Käesolevas töös on puuduvate väärtuste lävendiks valitud 50%. See tähendab, et analüüsi kaasatakse nende vastajate andmed, kes on vastanud vähemalt neljale küsimusele. Seega jääb imputeeritud andmestikku tabeli 2 põhjal $84\ 180 + 761 + 137 + 60 + 19 = 85\ 157$ vastajat analüüsi.

Järgnevalt arvutatakse välja olemasolevate andmete põhjal iga küsimuse keskmised väärtused ning need imputeeritakse puuduvate andmete asemele.

Keskmisega imputeerimise abil arvutatud depressiooniskoori kirjeldav statistika on esitatud tabelis 7.

Tabel 7: Depressiooniskoori kirjeldav statistika keskmistega imputeerides.

Statistik	**Väärtus**
Vastajate arv	85 157
Mediaan	14
Miinimum	8
Maksimum	40
Keskmine	15,994
Standardhälve	6,732
Usaldusintervall keskmisele (95%)	(15,949; 16,039)

2.6 Mitmene imputeerimine

Mitmest imputeerimist rakendatakse kõikide puuduvate andmete korral.

Kuna analüüsitavas 86 235 vaatlusega andmestikus ei esine sool, vanusel ning depressioonidiagnoosil puuduvaid väärtusi, siis neid ei imputeerita.

Kõikide tunnuste korral, mida on imputeeritud, on kasutatud ennustava keskmiste väärtustega sobitamist ning imputatsioonide arvuks on võetud $m = 5$. Algoritmi koonduvuse uurimiseks tehtud 20 iteratsiooni. Lähteväärtuseks (*seed*) on võetud 123.

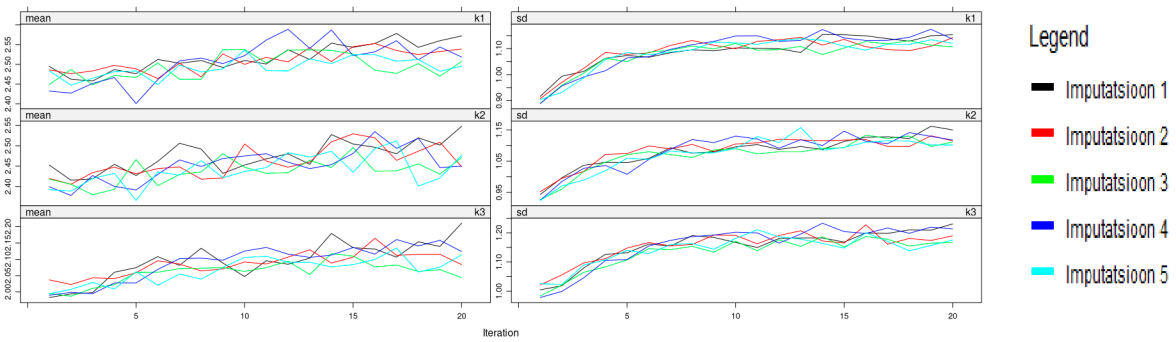
Imputeerimisel tekkinud viis andmestikku võetakse kokku üheks andmestikuks ning arvutatakse depressiooniskoor.

Seejärel leitakse hinnangud kõigis viies imputeeritud andmestikus käsuga *with()* ning poolitakse käsuga *pool()* m andmestiku parameetrite hinnangud lõplikuks hinnanguks.

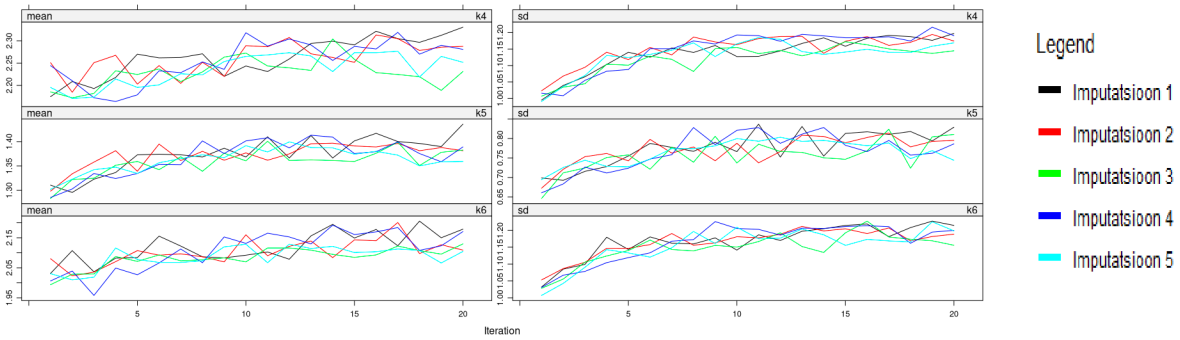
Jooniste 2, 3 ja 4 põhjal saab väita, et MICE algoritm koondub, seega on tulemused usaldusväärsed.

Lihtsuse mõttes on erinevaid teemasid käsitlevad küsimused tähistatud järgnevalt:

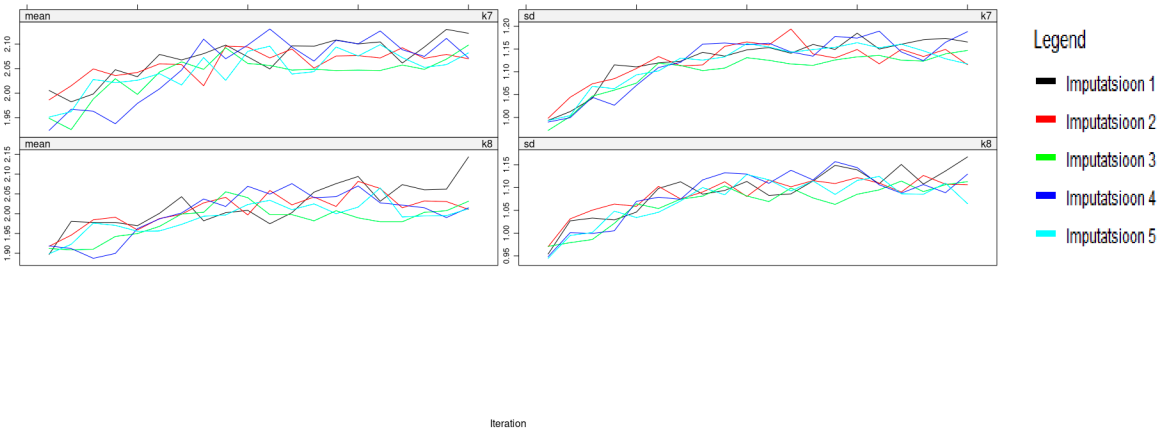
1. $k1$ - kurvameelsus;
2. $k2$ - huvi kadumine;
3. $k3$ - alaväärsustunne;
4. $k4$ - enesesüüdistused;
5. $k5$ - korduvad surma- või enesetapumõtted;
6. $k6$ - üksildustunne;
7. $k7$ - lootusetus tuleviku suhtes;
8. $k8$ - võimetus rõõmu tunda.



Joonis 2: MICE algoritmi koonduvuse kontroll tunnustel $k1$, $k2$ ning $k3$, x-telg kujutab iteratsioonide järjekorranumbrit ning y-telg keskmise (*mean*) või standardhälbe (*sd*) väärtust.

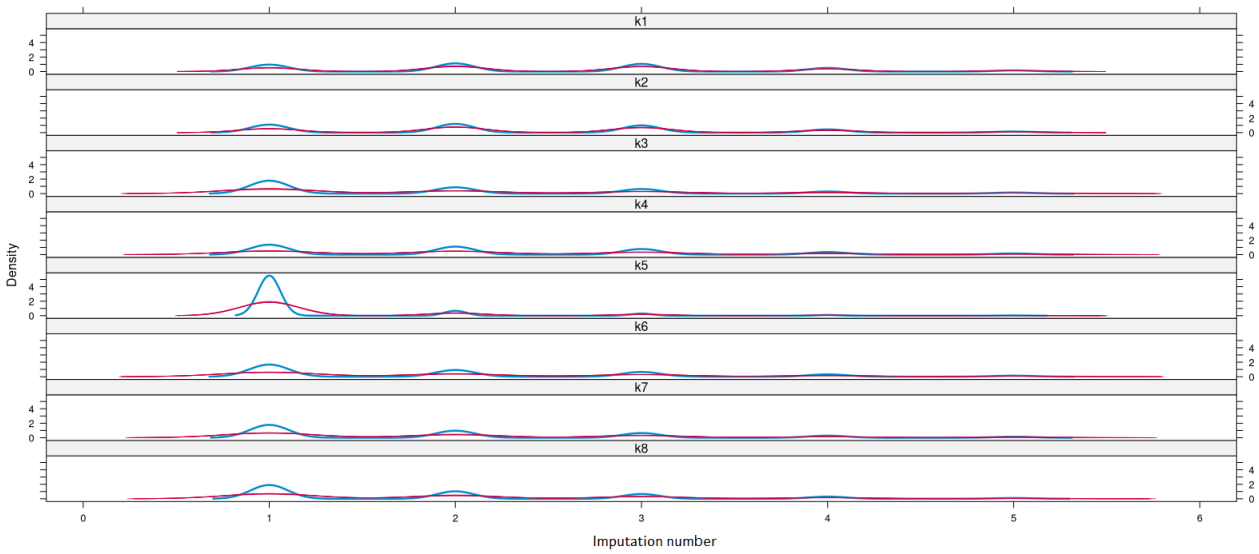


Joonis 3: MICE algoritmi koonduvuse kontroll tunnustel $k4$, $k5$ ning $k6$, x-telg kujutab iteratsioonide järjekorranumbrit ning y-telg keskmise (*mean*) või standardhälbe (*sd*) väärtust.



Joonis 4: MICE algoritmi koonduvuse kontroll tunnustel $k7$ ja $k8$, x-telg kujutab iteratsioonide järjekorranumbrit ning y-telg keskmise (*mean*) või standardhälbe (*sd*) väärtust.

Joonise 5 põhjal saab väita, et olemasolevate ning imputeeritud andmete jaotuste tihedused on sarnased. Imputeeritud väärtuste tihedusfunktsiooni graafik on joonisel punasega ning olemasolevate andmete tihedusfunktsiooni graafik on joonisel sinisega.



Joonis 5: Olemasolevate ning imputeeritud andmete tihedusfunktsioonide võrdlus tunnustel $k1$, $k2$, $k3$, $k4$, $k5$, $k6$, $k7$, $k8$. x-telg kujutab imputatsiooni järjekorranumbrit ning y-telg tihedust.

Mitmese imputeerimise abil arvutatud depressiooniskoori kirjeldav statistika viie imputeeritud andmestiku põhjal on esitatud tabelis 8.

Tabel 8: Depressiooniskoori kirjeldav statistika mitmesel imputeerimisel.

Statistik	**Väärtus**
Vastajate arv	431 175
Mediaan	14
Miinimum	8
Maksimum	40
Keskmine	16,009
Standardhälve	6,746
Usaldusintervall keskmisele (95%)	(15,989; 16,029)

2.7 Ennustusmudelid ning imputeerimismeetodite võrdlemine

Käesolevas uurimistöös kasutatakse ennustusmudelina logistilist regressioonimudelit. Logistiline regressioonimudel hindab depressioonidiagnoosi olemasolu tõenäosust ning mudeli kovariaatideks on võetud vastaja depressiooniskoor, vanus ning sugu.

Töös võrreldakse kolme erinevat imputeerimismeetodit: listiviisiline kustutamine, keskmisega imputeerimine ning mitmene imputeerimine. Selleks koostatakse ennustusmudel igale imputeerimismeetodile, kus d tähistab depressioonidiagnoosi olemasolu tõenäosust.

Olulisusnivoo α on antud töös alati 0,05.

Ennustusmudel andmestikule, mis on saadud listiviisilist kustutamist kasutades:

$$ln(\frac{d}{1-d}) = -3{,}786 + 0{,}095 \cdot \text{Depressiooniskoor} + 0{,}017 \cdot \text{Vanus} + 0{,}533 \cdot \text{Sugu}.$$

Ennustusmudel andmestikule, mis on saadud keskmisega imputeerimist kasutades:

$$ln(\frac{d}{1-d}) = -3{,}779 + 0{,}095 \cdot \text{Depressiooniskoor} + 0{,}017 \cdot \text{Vanus} + 0{,}532 \cdot \text{Sugu}.$$

Ennustusmudel andmestikule, mis on saadud mitmest imputeerimist kasutades:

$$ln(\frac{d}{1-d}) = -3{,}777 + 0{,}094 \cdot \text{Depressiooniskoor} + 0{,}017 \cdot \text{Vanus} + 0{,}537 \cdot \text{Sugu}.$$

Erinevate ennustusmudelite puhul tuleb depressiooniskoori ees olev koeffitsient sarnane (0,094 või 0,095), mis tähendab, et depressiooniskoori šansside suhe on samuti sarnane erinevate ennustusmudelite puhul. Kõikide mudelite korral on $p < 0{,}001$, seega on depressiooniskoor kõigis ennustusmudelites statistiliselt oluline.

Tabel 9: Erinevate imputeerimismeetodite depressiooniskoori statistikud.

Imputeerimismeetod	**Šansside suhe**	**Alumine usalduspiir**	**Ülemine usalduspiir**	***p*-väärtus**
Listiviisiline imputeerimine	1,099	1,097	1,102	<0,001
Mitmene imputeerimine	1,099	1,098	1,100	<0,001
Keskmised väärtused	1,099	1,097	1,102	<0,001

Tabelis 9 on esitatud erinevate imputeerimismeetodite statistikud.

Ennustusmudelite võrdlemisel võib väita, et imputeerimismeetodi valik ei mõjuta depressiooniskoori seoseid depressioonidiagnoosiga, kuna šansside suhted on võrdsed (1,099) ja usaldusintervallide vaheline erinevus on minimaalne.

Kokkuvõte

Käesolevas uurimistöös rakendatati kolme erinevat imputeerimismeetodit TÜ Eesti geenivaramu puuduvaid andmeid sisaldavate depressiooniküsimustike analüüsil ja hinnatati, kas ja kuidas mõjutab imputeerimismeetodi valik küsimustikupõhise depressiooniskoori seoseid depressioonidiagnoosiga.

Kasutati listiviisilist kustutamist, keskmisega imputeerimist ja mitmest imputeerimist. Depressioonidiagnoosi ennustusmudelite võrdlemisel võib väita, et imputeerimismeetodi valik ei mõjutanud depressiooniskoori seoseid depressioonidiagnoosiga.

Käesolevas töös saadud tulemuse põhjuseks võib olla vastanute suur arv, väike puuduvate andmete hulk ning vähene küsimuste arv. Tulemus võib olla tingitud vastajate huvist geenidoonoriks saamise vastu ja sellest, et küsimustiku täitmine oli vabatahtlik.

Täpselt sarnast imputeerimismeetodite võrdlust kirjandusest ei leidu. Erinevate autorite arvates on puuduvate väärtuste määr olulisim imputeerimise tulemust mõjutav faktor.

Kang et al. uuris puuduvate ja olemasolevate andmete suhet ning erinevate imputeerimismeetodite (MNR, kNN, CART, ANN, LLR) täpsust. Tema tulemuste põhjal osutus kNN sobivaimaks andmestikes, kus puuduvate väärtuste määr on alla 10%. LLR oli parim nendes andmestikes, kus puuduvate väärtuste määr on üle 10% [24].

Varasem uuring on näidanud, et kui puuduvate andmete määr on üle 10%, siis keskmisega imputeerimine on põhjustanud kallutatud hinnanguid [25].

Acuña ja Rodriguez üldistavad, et vähem kui 1% puuduvate andmete korral imputeerimine enamasti ei mõjuta analüüsi kvaliteeti, $1-5\%$ korral imputeerimine annab rahuldava tulemuse. Kui puuduvaid andmeid on $5-15\%$, tuleb kasutusele võtta keerukamad imputeerimismeetodid. Enam kui 15% andmete puudumine mõjutab analüüsi tulemust oluliselt [26].

Stavseth et al. võrdlesid kuut erinevat imputeerimismeetodit kategooriliste andmete korral ning leidsid, et kõik meetodid sobivad suurte andmestike (1000 vaatlust) korral. Väiksema andmehulga puhul (200 vaatlust) sõltuvad hinnangud suuresti puuduvate andmete määrast. Kui puuduvaid andmeid oli 20% või enam, siis listiviisiline kustutamine andis kallutatud tulemuse [27].

Kokkuvõtteks, see uurimistöö näitas, et suure hulga andmete ja väikse hulga puuduvate väärtuste korral ei ole listiviisilise kustutamise, keskmisega imputeerimise ning mitmese imputeerimise kasutamisel saadud tulemustes erinevusi.

Lühendite selgitus

ANN - *artificial neural network*

CART - *classification and regression tree*

EEK2 - emotsionaalse enesetunde küsimustik

kNN - *k-nearest neighbours algorithm*

LLR - *locally linear reconstruction*

MAR - *missing at random*

MCAR - *missing completely at random*

MICE - *multiple imputation by chained equations*

mids - *multiply imputed data set*

mipo - *multiply imputed pooled analysis*

mira - *multiply imputed repeated analyses*

MNR - *maximum normed residual*

NMAR - *not missing at random*

pmm - *predictive mean matching*

Kasutatud allikad

[1] CYJ Peng *et al.* "Advances in missing data methods and implications for educational research." Teoses: 2006, lk. 31–78.

[2] Y Dong ja CY Peng. "Principled missing data methods for searchers". *Springerplus* 2.222 (2013), lk. 1–17. DOI: 10.1186/2193-1801-2-222.

[3] Vikipeedia. *Imputeerimine*. 2023. URL: https://et.wikipedia.org/wiki/Imputeerimine (vaadatud 07.05.2024).

[4] A Aluoja *et al.* *Emotsionaalse enesetunde küsimustik (EEK2)*. 1999, 2002. URL: https://www.ravijuhend.ee/tervishoiuvarav/juhendid/193/taiskasvanute-unehairete-esmane-diagnostika (vaadatud 12.05.2024).

[5] Wikipedia contributors. *Missing data*. n.d. URL: https://en.wikipedia.org/wiki/Missing_data (vaadatud 29.04.2024).

[6] JR van Ginkel *et al.* "Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data". *Journal of Personality Assessment* 102.3 (2020), lk. 297–308. DOI: 10.1080/00223891.2018.1530680.

[7] A Mattei, F Mealli ja DB Rubin. "Missing Data and Imputation Methods". Teoses: *Statistical Analysis with Missing Data*. Wiley, 2011, lk. 129–154. DOI: 10.1002/9781119961154.ch8.

[8] Wikipedia contributors. *Listwise deletion*. n.d. URL: https://en.wikipedia.org/wiki/Listwise_deletion (vaadatud 12.05.2024).

[9] S van Buuren. *Flexible Imputation of Missing Data*. Second. A Chapman & Hall Book. CRC Press (Taylor & Francis Group), 2012, lk. 1–444.

[10] Chandrikasai. *Imputing Missing Values is Another Technique Used to Handle Missing Data in a Dataset*. Medium. 2023. URL: https://medium.com/@chandrikasai9997/imputing-missing-values-is-another-technique-used-to-handle-missing-data-in-a-dataset-824957ce71b4 (vaadatud 01.05.2024).

[11] Scispace. *What Are the Pros and Cons of the Median Imputation Strategy?* n.d. URL: https://typeset.io/questions/what-are-the-pros-and-cons-of-the-median-imputation-strategy-3lzrpjfjut (vaadatud 01.05.2024).

[12] P Li, EA Stuart ja DB Allison. "Multiple Imputation: A Flexible Tool for Handling Missing Data". *JAMA* 314.18 (10. november 2015), lk. 1966–1967. DOI: 10.1001/jama.2015.15281.

[13] RPubs. 2023. URL: https://rpubs.com/tamaraheijkamp/1022916 (vaadatud 01.05.2024).

[14] S van Buuren *et al. mice: Multivariate Imputation by Chained Equations in R*. R package version 3.16.0. URL: https://www.rdocumentation.org/packages/mice/versions/3.16.0/topics/mice (vaadatud 02.05.2024).

[15] S van Buuren ja K Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R". *Journal of Statistical Software* 45.3 (2011), lk. 1–67. DOI: https://doi.org/10.18637/jss.v045.i03.

[16] IR White, P Royston ja AM Wood. "Multiple imputation using chained equations: Issues and guidance for practice." *Statistics in Medicine* 30.4 (2011), lk. 377–399. DOI: 10.1002/sim.4067.

[17] S Wilson. *The MICE Algorithm*. 2021. URL: https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html (vaadatud 12.05.2024).

[18] TP Morris, IR White ja P Royston. "Tuning multiple imputation by predictive mean matching and local residual draws". *BMC medical research methodology* 14.75 (2014), lk. 1–13. URL: http://www.biomedcentral.com/1471-2288/14/75.

[19] N Erler. *Checks after Multiple Imputation?* URL: https://www.nerler.com/teaching/fgme2019/micheck (vaadatud 05.05.2024).

[20] E Käärik. *Andmeanalüüs II (MTMS.01.007) Loengukonspekt*. Loengukonspekt. Õppejõud: Ene Käärik, lk. 115–119.

[21] SN Deming. *Statistics in Laboratory: Pooling*. URL: https://www.americanlaboratory.com/353521-Statistics-in-the-Laboratory-Pooling/ (vaadatud 12.05.2024).

[22] MW Heymans ja I Eekhout. *Applied missing data analysis with SPSS and (R) Studio*. Heymans ja Eekhout: Amsterdam, The Netherlands, 2019. URL: https://bookdown.org/mwheymans/bookmi/.

[23] Wikipedia contributors. *Confidence interval*. n.d. URL: https://en.wikipedia.org/wiki/Confidence_interval (vaadatud 25.04.2024).

[24] P Kang. "Locally linear reconstruction based missing value imputation for supervised learning". *Neurocomputing* 118 (2013), lk. 65–78. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2013.02.016. URL: https://www.sciencedirect.com/science/article/pii/S0925231213002026.

[25] I Eekhout *et al.* "Missing data in a multi-item instrument were best handled by multiple imputation at the item score level". *Journal of Clinical Epidemiology* 67.3 (2014), lk. 335–342. DOI: 10.1016/j.jclinepi.2013.09.009.

[26] E Acuña ja C Rodriguez. *The Treatment of Missing Values and its Effect on Classifier Accuracy*. Springer-Verlag Berlin Heidelberg, 2004, lk. 639–647.

[27] MR Stavseth, T Clausen ja J Røislien. "How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data". *SAGE Open Medicine* 7 (2019), lk. 1–12.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Jaan Otter,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Puuduvate andmete imputeerimine depressiooni hindavas küsimustikus, mille juhendajad on Kelli Lehto ja Anastassia Kolde, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaaloman-di ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Jaan Otter

15.05.2024