Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Kristiina Uusna

Statistilise analüüsi rakendamine Eesti Haigekassa raviarvetele

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Laur, PhD

Tartu 2018

Statistilise analüüsi rakendamine Eesti Haigekassa raviarvetele

Bakalaureusetöö

Kristiina Uusna

Lühikokkuvõte. Käesolevas bakalaureusetöös analüüsitakse Eesti Haigekassale saadetud raviarveid. Eesmärk on anda ülevaade regressioonanalüüsist diskreetsete argumentidega. Täpsemalt vaadatakse võtteid, kuidas suurte andmekoguste korral tagada arvutuslik efektiivsus ning kuidas parameetrite rohke mudeli korral anda sellele lihtne visuaalne interpretatsioon. Uuritakse mudeli headuse ja täpsuse näitajaid. Eelkõige pööratakse tähelepanu efekti suurusele ja selle tähtsusele. Olulisuse määramiseks tegeliku keskmise ja mudeli poolt ennustatud keskmise vahel võetakse kasutusele permutatsioonitest ja bootstrap-meetod.

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

Märksõnad: üldine lineaarne mudel, regressioonanalüüs diskreetsete argumentidega, vähimruutude meetod, permutatsioonitest, bootstrap-meetod

Application of Statistical Analysis to the Treatment Bills of Estonian Health Insurance Fund

Bachelor's thesis

Kristiina Uusna

Abstract. The given Bachelor's thesis analyses the treatment bills of the Estonian Health Insurance Fund. The purpose of this thesis is to give an overview of the regression analysis with dummy variables. More specifically, the aim is to look at methods for how to make effective calculations when we work with big data sets and how to give a simple visual interpretation when the model has a lot of parameters. For each model goodness-of-fit and accuracy was determined. Additionally, the size of the effect and its importance was computed and highlighted. Permutation test and bootstrap method are used to attach significance scores for difference between the actual mean and mean, which is predicted by the model.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics
Keywords: general linear model, regression with dummy variables, least squares method, permutation test, bootstrap method

Sisukord

Sissejuhatus

Eestis organiseerib riiklikku ravikindlustust Eesti Haigekassa. Haigekassa peamiseks ülesandeks on tervishoiuteenuste korraldamine ja nende eest tasumine. Sellega seoses peab haigekassa tagama eelarve efektiivse ja otstarbeka kasutamise. Eesti Haigekassa hüvitab iga haigla kulusid kuluarvestussüsteemi alusel. Kuluarvestussüsteem võimaldab haigekassal eelarve piires meditsiiniasutustele õiglaselt raha jagada ja meditsiiniasutused omakorda saavad selle alusel oma kulusid kontrolli all hoida. Teenuse eest tasumiseks peab tervishoiuteenuse osutaja haigekassale esitama nõuetekohaselt vormistatud raviarve. Ravijuhud klassifitseeritakse erialade alusel. Igale ravijuhule koostatakse eraldi raviarve, mille vormistamisel võetakse aluseks teenuse osutamise kuupäeval kehtinud hinnakirja ning ravijuhtumi ja haigla eripäradest lähtuvaid korrektsioone.

Selleks, et igale haiglale koostada õiglane eelarve, tuleb teha põhjalik analüüs eelnevalt saadetud raviarvete põhjal. Eesmärk on teada saada, kas erinevate haiglate ravijuhtude keskmine maksumus (RJKM) erineb oluliselt, kui võtta arvesse patsiendi eristuskategooriaid (sugu, vanus) ja ravi iseloomu. Selleks tuleb leida tunnused, mille mõju raviarvele on piisavalt suur ja samas ka statistiliselt oluline. Lisaks on oluline teada, kui suurel määral statistiliselt oluliseks osutunud tunnused raviarvet mõjutavad. Samal ajal tuleb uurida ka nende tunnuste lõikes ravijuhu maksumuse jaotust, et RJKM-st ei tõlgendataks valesti olukorras, kus kõrge keskmise põhjuseks on vaid üksikud raviarved.

Selles bakalaureusetöös tegeletakse analüüsiks vajaliku metoodika väljatöötamise ja valideerimisega läbi juhtumiuuringu. Tunnusetüüpe arvesse võttes on statistilise mudeli leidmisel läbi viidud regressioonanalüüs diskreetsete argumentidega. Kuna uuritav andmestik on äärmiselt mahukas, siis vaadatakse lähemalt võtteid, kuidas arvutuslikult efektiivsemalt soovitud mudel kätte saada. Uuritakse mudeli headust ja täpsust. Lisaks kasutatakse permutatsioonitesti olulisuse määramiseks tegeliku ja ennustatud RJKM-te erinevuste vahel.

Töö on koostatud tekstikujundustarkvaraga LaTeX. Praktilise osa läbiviimiseks on kasutatud statistikatarkvara R ja permutatsioonitest on realiseeritud programmeerimiskeeles Python.

Autor soovib tänada töö juhendajat Sven Lauri kasulike nõuannete ja suunamise eest sedavõrd mahukate andmete analüüsimise osas.

1 Andmestiku ülevaade

Töös uuritav valim on moodustatud 2014.-2016. aasta haigekassa raviarvetest. Täpsemalt uuritakse kahte haiglat, millest üks on Tallinna Lastehaigla ja teine Tartu Ülikooli Kliinikum. Valiku tegemisel lähtuti sellest, et omavahel võrreldavad haiglad oleksid võimalikult erinevad aga samal ajal oma suuruselt ja teenindusvõimelt sarnased. Antud olukorras on mõlemad haiglad piirkondlikud ja asuvad erinevates maakondades. Eeldatavasti on Tallinna Lastehaigla patsientide hulgas rohkem nooremaid inimesi, samal ajal kui Tartu Ülikooli Kliinikumi patsiendid jagunevad ühtlasemalt vanusegruppide vahel. Sedasi võib huvitavaid tulemis anda ka patsiendi eristuskategooriate mõju uurimine ravijuhu keskmisele maksumusele. Vaatluse alla on võetud patsiendid, kes on läbinud ravi, mis on seotud meditsiiniharudega nagu ortopeedia, pediaatria, lastekirurgia või üldkirurgia.

Töös uuritava andmestiku maht on 263 422 arvet. Üldkogumi moodustavad kõigi haiglate raviarved aastatel 2014-2016, mis on koostatud eelnimetatud meditsiiniharudesse kuuluvatele teenustele. Üldkogumis on ligikaudu 9.9 miljonit arvet, mis on umbes 37 korda rohkem kui vaatluse alla võetud valimis. Kui aga vaadata sama iseloomuga raviarveid aastatel 2010-2018, siis tõuseks raviarvete arv koguni 22.5 miljonini. Seega on äärmiselt oluline pöörata tähelepanu arvutuste viimisele võimalikult efektiivseks.

1.1 Uuritav tunnus ja argumendid

Antud töös on uuritavaks tunnuseks ravijuhu maksumus. Argumenttunnuseid, mille mõju ravijuhu maksumusele uuritakse, on kokku 11 (Tabel 1).

Patsiendi põhidiagnoos on välja toodud Rahvusvahelise Haiguste Klassifikatsiooni RHK-10 abil (inglise keeles ICD-10 *International Statistical Classification of Diseases and Related Health Problems*)[7]. Igal haigusel, häirel, vigastusel või seisundil on oma kood, mis koosneb tähest ja kahest numbrist, millele võib veel järgneda punktiga eraldatud numbreid. Mida pikem on kood, seda täpsemalt see diagnoosi kirjeldab. RHK-10 kood on spetsifikatsioon eraldi tunnusena välja toodud meditsiinierialale.

Tabel 1: Uuritav tunnus ja argumenttunnused

Uuritav tunnus		
reimbursed_sum	pidev tunnus	haigekassa ravikindlustuse fondist välja makstud summa
Argumenttunnused		
fiscal_year	järjestustunnus	arve esitamise aasta
tto_code	nominaaltunnus	haigla, kus patsient viibis
official_gender	binaarne tunnus	patsiendi sugu
age_group	järjestustunnus	vanusegrupp, kuhu patsient kuulus, kui ta viibis haiglas
patient_residence_code	nominaaltunnus	maakond, kuhu on registreeritud patsiendi elukoht
tto_location	nominaaltunnus	maakond, kus asub haigla (kohati linna täpsusega)
main_icd10_diagnosis	nominaaltunnus	patsiendi põhidiagnoosi RHK-10 kood
main_specialty	nominaaltunnus	meditsiinierialade klassifikatsioon (spetsiaalne kood kirjeldamaks kogu ravi, mis patsiendile haiglas tehti)
treatment_duration	diskreetne tunnus	ravi pikkus päevades
ttl_count	diskreetne tunnus	teenuste arv, mis patsiendile tehti
treatment_type	binaarne tunnus	ravi tüüp : A-ambulatoorne, S-statsionaarne ravi.

2 Kasutatud metoodika

Bakalaureusetöös on soov leida mudel, kus on nii diskreetsed argumendid, pidevad argumendid kui ka mitmesugused koosmõjud. Selles peatükis tutvustatakse lähemalt, milliseid meetodeid soovitud tulemuse saamiseks kasutatakse.

Toome esmalt välja statistilise mudeli üldkuju:

$$Y = f(X, \beta_0, \beta_1, \ldots, \beta_p) + \epsilon \ ,$$

kus Y on uuritav tunnus, $f(\cdot)$ on mingi deterministlik funktsioon, mis on uuritava tunnuse arvutamise eeskirjaks, X on argumenttunnuste vektor ehk need tunnused, millest sõltub uuritav tunnus, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ on mudeli parameetrid ja ϵ on mudeli juhuslik viga. Selles bakalaureusetöös olgu fikseeritud valmimaht n, mudelis esinevate argumentide arv k ja mudeli parameetrite arv p.

Uuritava tunnuse oodatava väärtuse sõltuvust argumenttunnuse väärtusest nimetatakse regressiooniks. Kui vaadatakse uuritava tunnuse sõltuvust mitmest argumenttunnusest, siis seda nimetatakse mitmeseks regressioonanalüüsiks. Selleks, et jõuda sobiva mudelini, moodustatakse mingi hulk mudeleid

$$\mathcal{F} = \{f(X, \beta_0, \ldots, \beta_p) : \boldsymbol{\beta} \in \mathcal{B}\} \ ,$$

kust regressioonülesande lahendamise tulemusel valitakse meelepärase meetodiga välja sobivaim. Selles bakalaureusetöös kasutatakse selleks vähimruutude meetodit.

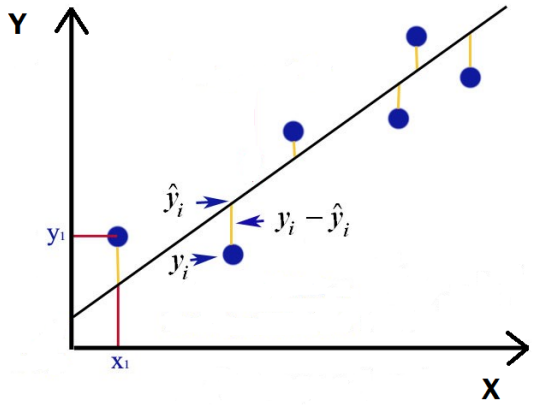
2.1 Vähimruutude meetod

Vaatame lähemalt vähimruutude meetodi rakendamist lineaarregressiooni korral, kus vaatluse all on üks argumenttunnus ja uuritav tunnus.

Uuritava tunnuse väärtuste arvutamise eeskirja nimetatakse funktsiooniks, mida tähistatakse $y = f(x)$ (tunnus y on tunnuse x funktsioon).

Lineaarset regressiooniseost saab kirjeldada ka graafilisel kujul, kus y-i ja x-i vahelist seost iseloomustab regressioonisirge. Nimelt tõmmatakse regressioonisirge läbi

ristteljestikus oleva punktide parve nii, et punktide y-koordinaatide summaarne kõrvalekalle regressioonisirgest oleks võimalikult väike (Joonis 1).



Joonis 1: Vähimruutude meetod [1]

Moodustatud sirge ja i-nda punkti vahelist kaugust nimetatakse juhusliku vea hinnanguks:

$$e_i = y_i - \hat{y}_i \ ,$$

kus y_i on i-nda punkti tegelik Y-i väärtus ja $\hat{y}_i$ on väärtus sirgel, mida nimetatakse ka y_i prognoosiks. Oluline on selle juures tähele panna, et juhusliku vea hinnang e_i langeb kokku mudeli veaga ϵ_i ainult siis, kui andmetest leitud parameetrid $\hat{\beta}_0, \ldots, \hat{\beta}_p$ langevad kokku tegelikku seost määravate parameetritega $\beta_0, \ldots, \beta_p$. See ei juhtu peaaegu mitte kunagi. Küll aga on enamasti piisava andmemahu korral $e_i \approx \epsilon_i$.

Järgnevalt toome teieni valemid, mis kehtivad igat tüüpi mudelite kohta, mitte ainult lineaarse regressioonimudeli korral.

Vigade ruutude summa võime kirjutada järgmise võrdusena:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \ ,$$

kus n on punktide arv. Vähimruutude meetodi rakendamise idee seisnebki vigade ruutude summade minimiseerimises. Seega lahendatakse järgmine ekstreemum-

ülesanne mudeli parameetrite suhtes:

$$\sum_{i=1}^{n}(y_i - f_{\boldsymbol{\beta}}(x_i))^2 \longrightarrow min \ ,$$

kus $f_{\boldsymbol{\beta}}(\cdot)$ on parameetrite $\boldsymbol{\beta}$ poolt määratud funktsioon ja $\hat{y}_i = f_{\boldsymbol{\beta}}(x_i)$ on selle põhjal arvutatud i-nda vaatluse prognoos. Hinnang juhuslike vigade hajuvusele ehk keskmisele ruutveale avaldub aga kujul:

$$MSE = \frac{SSE}{n-k-1} = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 \ . \tag{1}$$

Mudeli standardviga ehk mudeli täpsus on ruutjuur sellest $\sqrt{MSE}$.
Seose headust näitab aga determinatsioonikordaja R^2. See näitab, kui suure osa uuritava tunnuse koguhajuvusest moodustab regressioonhajuvus. Determinatsioonikordaja avaldub järgmiselt:

$$R^2 = 1 - \frac{SSE}{SST} \ ,$$

kus $SST = SSR + SSE$ on uuritava tunnuse koguhajuvus ja SSR on regressioonhajuvus.

Selleks, et vähimruutude meetod annaks võimalikult täpse ennustuse peaksid olema täidetud järgnevalt toodud eeldused.

(a) Juhuslikud vead on erinevate vaatluste korral sõltumatud, millest järeldub $cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$.

(b) Juhuslikud vead on normaaljaotusega: $\epsilon_i \sim N(0, \sigma^2)$, kus σ on juhusliku vea standardhälve.

(c) Juhuslikud vead on konstantse hajuvusega: $D(\epsilon_i) = \sigma^2$.

2.2 Üldine lineaarne mudel

Selles peatükis on toetutud Ene Kääriku loengukonspektile „Andmeanalüüs II" [2].

Üldise lineaarne mudeli moodustamise üks võimalikke viise on vähimruutude meetodi rakendamine. Üldist lineaarset mudelit võib nimetada ka regressioonanalüüsi mudeliks, kui argumentideks on pidevad arvtunnused või dispersioonanalüüsiks, kui argumendid on diskreetsed. Maatrikskujul avaldub üldine lineaarne mudel järgmiselt:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \dots & x_1 \\ 1 & \dots & x_2 \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \ .$$

Eelnevat võime esitada ka lühemalt:

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

kus $\mathbf{y}$ on $n \times 1$ uuritava tunnuse vektor, $\mathbf{X}$ on $n \times p$ plaanimaatriks, $\boldsymbol{\beta}$ on $p \times 1$ tundmatute parameetrite vektor ja $\boldsymbol{\epsilon}$ on $n \times 1$ juhuslike vigade vektor.

Fikseeritud i-nda vaatluse korral y_i avaldub järgmisel kujul:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \ . \tag{2}$$

Mudeli parameetrid hinnatakse vähimruutude meetodil. Seega tuleb minimiseerida vigade ruutude summad. Vigade ruutude summa maatriksesituses on järgmine:

$$SSE(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \ .$$

Minimiseerimisülesannet lahendades jõutakse normaalvõrrandisüsteemini

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \ ,$$

mille ühene lahend avaldub kujul:

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

mida nimetatakse parameetervektori $\boldsymbol{\beta}$ vähimruutude hinnanguks. Seega mudeli põhjal arvutatud uuritava tunnuse väärtus avaldub järgmiselt:

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

Viimast nimetatakse ka prognoosiks ehk ennustuseks. Fikseeritud i-nda vaatluse korral avaldub $\hat{y}_i$ järgmisel kujul:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik} \ .$$

2.2.1 Regressioonanalüüs diskreetsete argumentide korral

Vaatame täpsemalt üldise lineaarse mudeli kasutamist juhul, kui kõik ennustuseks kasutatavad argumendid omavad vaid diskreetseid väärtusi.

Kui me kaasame diskreetse argumendi regressioonimudelisse, siis võetakse kasutusele indikaatortunnued. Diskreetne argumenttunnuse erinevaid väärtusi nimetatakse tasemeteks. Olgu $i = 1, ..., k$ argumenttunnuse tasemete arv. Kui vaadata x_i-d indikaatortunnusena, siis ta määrab argumenttunnuse taseme järgmiselt:

$$x_i = \begin{cases} 1, & \text{kui argumenttunnusel on tase i} \\ 0, & \text{teistel juhtudel} \end{cases} \ .$$

Näiteks, kui meil on vaatluse all argumenttunnus, millel on kolm taset, siis saame mudeli kujul:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \ ,$$

kus x_1, x_2 ja x_3 on kasutusele võetud kolm indikaatortunnust ja $\beta_0, \beta_1, \beta_2, \beta_3$ on mudeli parameetrid. Kuna indikaatortunnused on lineaarselt sõltuvad $x_1+x_2+x_3 = 1$, siis mitu parameetrikomplekti määravad ära sama funktsiooni. Antud probleemi lahendamiseks on kaks standardset võimalust. Esiteks võib vaadelda vabaliikmeta mudelit

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \ .$$

Teiseks võib jätta välja viimase indikaatortunnuse ja vaadelda mudelit

$$y = \tau + \gamma_1 x_1 + \gamma_2 x_2 + \epsilon \ . \tag{3}$$

Mõlemad regresioonülesanded on ekvivalentsed, sest

$$\begin{aligned} \beta_1 &= \gamma_1 + \beta_3 \\ \beta_2 &= \gamma_2 + \beta_3 \\ \beta_3 &= \tau - \beta_0 \end{aligned} \ .$$

Kasutades sedasi indikaatortunnuseid, saame fikseeritud i-nda vaatluse korral kirjutada eelneva mudeli (3) üldise lineaarse mudeli kujul (2). Kui mudel sisaldab kahe diskreetse tunnuse koosmõju, siis saame käituda analoogselt eelnevale. Ainult nüüd on indikaatortunnus x_{ij} kujul:

$$x_{ij} = \begin{cases} 1, & \text{kui esimesel argumenttunnusel on tase i ja} \\ & \text{teisel argumenttunnusel on tase j} \\ 0, & \text{teistel juhtudel.} \end{cases} \ .$$

Analoogselt on võimalik indikaatortunnus määrata ka olukorras, kus vaatame koosmõju rohkem kui kahe diskreetse tunnuse vahel.

Illustreeriva näitena vaatame olukorda, kus me soovime ennustada patsiendi raviarvet vanuse ja soo põhjal. Tabelist 2 näeme, et vaatluse all on kuus erinevat kombinatsiooni.

Tabel 2: Keskmine raviarve patsiendi soo ja vanusegrupi järgi

sugu / vanusegrupp	M	N
[0,1)	95.43498	77.15644
[1,4)	105.92692	96.83235
[5,10)	107.58841	97.76498

Seega saame välja kirjutada kuus võimalikku indikaatortunnust, millest üks avaldub teiste kaudu. Näiteks indikaatormuutuja x_{12} avaldub järgmiselt:

$$x_{12} = \begin{cases} 1, & \text{naissoost patsient kuulub vanusegruppi } [0,1) \\ 0, & \text{muu} \end{cases} .$$

Mudeli parameetreid on võimalik hinnata vähimruutde meetodil. Juba eelneva põhjal teame, et selleks tuleb minimiseerida vigade ruutude summasid. Kui aga diskreetseid tunnuseid on palju ja lisaks nende tasemete arv on suur, muutub sobiva mudeli leidmine väga töömahukaks. Siinkohal tasuks tähele panna, et minimiseerimisülesannet lahendades on võimalik jõuda sellisele normaalvõrrandisüsteemini, kus parameetri hinnnag on võimalik leida teades ainult uuritava tunnuse keskmist vaadeldavas tasemes ja vastavat kaalu, mis on võrdne selle taseme alla kuuluvate objektide arvuga. Näitame seda pikemalt järgmise aruteluga.

Defineerime esmalt, mida mõistetakse antud töös sõna lahtri all. Kahe uuritava argumenttunnuse korral on lahter veeru ja rea ristumiskoht tabelis. Veeru ja rea alguses asuvad tabeli päised, kus on ära toodud vastavate tunnuste tasemete nimetused. Lähtudes veeru ja rea päistest jagatakse väärtused lahtritesse, mille tulemusel tekib latrisse $\mathcal{C}_\kappa$ alamhulk, kus rea ja veeru poolt määratud tunnuste taseme väärtused $x_1, ..., x_{n_k}$ on konstantsed. Siinkohal n_k tähistab lahtrisse $\mathcal{C}_\kappa$ kuuluvate väärtuste koguarvu. Rohkemate argumenttunnuste korral pole andmeid võimalik esitada kahemõõtmelise tabelina, kuid põhimõte on sama.

Kõik lahtri $\mathcal{C}_\kappa$ põhjal tehtud arvutused on saadud sellesse lahtrisse kuuluvate väärtuste põhjal. Tabelis 2 näeme olukorda, kus lahtritesse on jagatud ravijuhu maksumused ja selle põhjal arvutatud välja iga lahtri RJKM. Selles näites on tabeli päisteks vastavalt soo ja vanusegrupi tasemed.

Järgnev arutelu kehtib igat tüüpi mudeli korral, mitte ainult lineaarse mudeli korral. Olgu lahtreid kokku ℓ tükki. Sealjuures kehtib loomulikult

$$n = n_1 + n_2 + \ldots + n_\ell \ .$$

Olgu $\mathcal{I}_\kappa$ nendele objektidele vastavate indeksite hulk, mis kõik kuuluvad κ-ndasse lahtrisse ja tähistagu $\boldsymbol{x}_\kappa$ lahtrit defineerivate tunnuste väärtusi. Siis on lihtne aru saada, et ükskõik millise mudeli korral on selle ennustus kahe lahtris oleva andmepunkti $i, j \in \mathcal{I}_\kappa$ korral sama

$$\hat{y}_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) = f(\boldsymbol{x}_\kappa, \boldsymbol{\beta}) = f(\boldsymbol{x}_j, \boldsymbol{\beta}) = \hat{y}_j$$

eeldusel, et mudel võtab arvesse vaid lahtrit defineerivaid tunnuseid. Meie näite korral peab mudel ennustama ravijuhu maksumust vaid soo ja vanusegrupi põhjal. Samas ei pea mudel kõiki lahtrit defineerivaid tunnuseid arvesse võtma ning võib teha ennustuse näiteks vaid vanuse järgi. Seega eelneva arutelu põhjal võime lahtri $\mathcal{C}_\kappa$ kohta käivat ennustust tähistada $\hat{y}_\kappa = f(\boldsymbol{x}_\kappa, \boldsymbol{\beta})$.

Olgu $\hat{\boldsymbol{y}}$ mudeli poolt prognoosistud väärtuste vektor. Antud eeldusel saame vigade ruutude summa avaldada kujul:

$$SSE(\hat{\boldsymbol{y}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{\kappa=1}^{\ell} \sum_{i \in \mathcal{I}_\kappa} (y_i - \hat{y}_\kappa)^2 \ .$$

Defineerides vigade ruutude summa igas lahtris

$$SSE_\kappa(\hat{y}_\kappa) = \sum_{i \in \mathcal{I}_\kappa} (y_i - \hat{y}_\kappa)^2 \ ,$$

saame esitada vigade ruutude summa järgneva summana

$$SSE(\hat{\boldsymbol{y}}) = \sum_{\kappa=1}^{\ell} SSE_\kappa(\hat{y}_\kappa) \ .$$

Olgu $\bar{y}_\kappa$ ühe lahtri keskmine, mis avaldub kujul:

$$\bar{y}_\kappa = \frac{1}{n_\kappa} \sum_{i \in \mathcal{I}_\kappa} y_i \ . \tag{4}$$

Vigade ruutude summa igas lahtris on lahtri keskmise abil võimalik omakorda välja kirjutada kujul:

$$\begin{aligned} SSE_\kappa(\hat{y}_\kappa) &= \sum_{i\in\mathcal{I}_\kappa}(y_i - \bar{y}_\kappa + \bar{y}_\kappa - \hat{y}_\kappa)^2 \\ &= \sum_{i\in\mathcal{I}_\kappa}(y_i - \bar{y}_\kappa)^2 + \sum_{i\in\mathcal{I}_\kappa}(\bar{y}_\kappa - \hat{y}_\kappa)^2 + \sum_{i\in\mathcal{I}_\kappa} 2(y_i - \bar{y}_\kappa)(\bar{y}_\kappa - \hat{y}_\kappa) \ . \end{aligned} \tag{5}$$

Vaatame lähemalt viimast liidetavat:

$$2\sum_{i\in\mathcal{I}_\kappa}(y_i - \bar{y}_\kappa)(\bar{y}_\kappa - \hat{y}_\kappa) = 2(\bar{y}_\kappa - \hat{y}_\kappa)\cdot\left[\sum_{i\in\mathcal{I}_\kappa} y_i - n_\kappa\bar{y}_\kappa\right] \ . \tag{6}$$

Paneme tähele, et lahtri keskmise (4) võime lahti kirjutada kujul:

$$\bar{y}_\kappa = \frac{1}{n_\kappa}\sum_{i\in\mathcal{I}_\kappa} y_i \quad \Leftrightarrow \quad \bar{y}_\kappa n_\kappa = \sum_{i\in\mathcal{I}_\kappa} y_i \quad \Leftrightarrow \quad \sum_{i\in\mathcal{I}_\kappa} y_i - n_\kappa\bar{y}_\kappa = 0 \ .$$

Lähtudes viimasest võrdusest, saame et $SSE_\kappa(\hat{y}_\kappa)$ viimane liidetav (6) on võrdne nulliga ja seega on meil võimalik võrdus (5) viia järgmisele kujule, kus

$$SST_\kappa = \sum_{i\in\mathcal{I}_\kappa}(y_i - \bar{y}_\kappa)^2$$

iseloomustab vaatluste varieeruvus κ-nda lahtri keskmise ümber:

$$\begin{aligned} SSE_\kappa(\hat{y}_\kappa) &= \sum_{i\in\mathcal{I}_\kappa}(y_i - \bar{y}_\kappa)^2 + \sum_{i\in\mathcal{I}_\kappa}(\bar{y}_\kappa - \hat{y}_\kappa)^2 \\ &= \sum_{i\in\mathcal{I}_\kappa}(y_i - \bar{y}_\kappa)^2 + n_\kappa(\bar{y}_\kappa - \hat{y}_\kappa)^2 = SST_\kappa + n_\kappa(\bar{y}_\kappa - \hat{y}_\kappa)^2 \ . \end{aligned}$$

Seega kogu vigade ruutude summa on võrdne:

$$SSE(\hat{\boldsymbol{y}}) = \sum_{\kappa=1}^{\ell} SST_\kappa + \sum_{\kappa=1}^{\ell} n_\kappa(\bar{y}_\kappa - \hat{y}_\kappa)^2 \ .$$

Kuna SST_κ ei sisalda mudeli poolt prognoositud väärtusi, siis saame minimiseerimisülesannet rakendada vaid teise liidetava suhtes. Teame, et antud juhul $f(\mathbf{x}_\kappa, \boldsymbol{\beta}) = \hat{y}_\kappa$ on funktsioon, mille abil arvutatakse uuritava tunnuse väärtus. Sel-

lisel juhul lahendatakse ekstreemumülesanne mudeli parameetrite suhtes:

$$\sum_{\kappa=1}^{\ell} n_\kappa(\bar{y}_\kappa - f(\mathbf{x}_\kappa, \boldsymbol{\beta}))^2 \longrightarrow \min \quad .$$

Sedasi jõutakse normaalvõrranditeni ja saadakse parameetritele hinnangud. Näeme, et selleks on vaja teada ainult iga vaadeldava lahtri keskmist ja lahtri moodustanud objektide arvu. Saadud tulemus kehtib iga mudeli kuju korral, mitte ainult lineaarmudeli puhul.

Sellist lahtritel baseeruvat funktsiooni võib nimetada piisavaks statistikuks. Piisav statistik on funktsioon, mis kasutab kogu andmestiku asemel väärtusi, mis on saadud eelnevalt nende andmetega tehtud arvutuste tulemusel, säilitades sedasi informatiivsuse [6].

2.3 Statistilise olulisuse määramine

Regeresioonimudelite kasutamine võimaldab igale lahtrile leida optimaalse ennustuse $\hat{y}_\kappa$ ning ennustada selle abil ravijuhu keskmist maksumust konkreetses haiglas. Seda suurust saab omakorda kõrvutada ravijuhu tegeliku keskmise maksumusega. Nende suuruste erinevus näitab kas haigla kulutab ravile keskmiselt rohkem või vähem raha kui peaks. Probleem on aga selles, et see erinevus võib olla tingitud juhusest, kus näiteks kõrge keskmise põhjuseks on vaid üksikud ülisuured ravijuhu maksumused. Selleks et seda probleemi lahendada tuleb anda erinevusele statistiline olulisus. Selleks tuleb fikseerida nullhüpotees andmete tekkimise kohta. Vaatame lähemalt, kuidas on sõnastatud nullhüpotees kahe erineva olulisust määrava testi korral.

Permutatsioonitest. Sõnastame nullhüpoteesi andmete tekkimise kohta. Haiglate järgi moodustatud gruppide vahel jaotatakse raviarved juhuslikult arvestades selle juures ainult lahtreid defineerivate tunnuste marginaaljaotusi.

Bootstrap. Uuritavatele haiglatele antud raviarvete jaotus on tundmatu, kuid seda saab simuleerida võttes iga kord juhuslikult tagasipanekuga arve kogutud arvete hulgast. Raviarvete valimi genereerimisel arvestakse lahtreid defineerivate

tunnuste marginaaljaotusi.

See, millist meetodit uuritava tunnuse statistilise olulisuse määramisel kasutada, oleneb ülesande püstitusest. Täpsemalt, milline on huvipakkuv suurus ja kuidas seda suurust arvutatakse. Vaatame lähemalt kolme huvipakkuvat ülesannet.

1. **Kahe haigla ravijuhtude keskmiste võrdlemine.**
   Soovime teada, kas kahe haigla ravijuhtude keskmised on statistiliselt olulised. Uuritav tunnus on jaotatud haiglate vahel ära kahte gruppi. Eesmärk on välja selgitada, kas gruppide vaheline jaotus on sama ehk kas raviarved on gruppide vahel jaotatud juhuslikult.

2. **Ühe haigla ravijuhtude keskmise võrdlemine kõigega.**
   Soovime teada, kas ühe haigla ravijuhtude keskmine maksumuse erinevus üldkeskmisest on statistiliselt oluline. Eesmärk on välja selgitada, kas uuritav tunnus käitub sarnaselt üldkogumile või mitte.

3. **Mingi alamgrupi ravijuhtude keskmiste võrdlemine.**
   Soovime teada, kas kahe alamgrupi ravijuhtude keskmised on statistiliselt olulised. Uuritav tunnus on jaotatud haiglate vahel ära kahte gruppi mingi tunnus(t)e taseme(te) kitsendusega. Näiteks vaadatakse mehi vanuses $[30, 35)$ kahe haigla lõikes. Eesmärk on välja selgitada, kas gruppide vaheline jaotus on sama ehk kas raviarved gruppide vahel on jaotatud juhuslikult.

Defineeritud ülesannetest esimeses ja kolmandas oleks mõistlik kasutada statistilise olulisuse määramisel permutatsioonitesti. Permutatsioonitest seisneb esialgse analüüsi paljukordses kordamises juhuslikult ümberpaigutatud andmetega. Kui me soovime määrata kahe haigla poolt moodustatud gruppide keskmiste erinevuse olulisust fikseerides selle juures gruppide suurused n_1 ja n_2, siis esmalt tuleb meil leida teststatistik:

$$D_t = \bar{X}_1 - \bar{X}_2 \ ,$$

kus $\bar{X}_1$ on esimese grupi (esimese haigla) uuritava tunnuse keskmine ja $\bar{X}_2$ on teise grupi (teise haigla) sama tunnuse keskmine. Seejärel moodustatakse ühine andmestik suurusega $n_1 + n_2$ ja paigutatakse andmed andmestikus juhuslikult ümber.

Juhuslikustamist korratakse m korda ja igal juhul leiame ka teststatistiku väärtuse:

$$D_i = \bar{X}_1' - \bar{X}_2' \qquad i = 1, .., m \ ,$$

kus $\bar{X}_1'$ on juhuslikustatud andmestiku esimese n_1 väärtuse põhjal arvutatud uuritava tunnuse keskmine ja $\bar{X}_2'$ ülejäänud n_2 väärtuse põhjal arvutatud keskmine. Tulemuseks saadud teststatistiku jaotuse alusel leitakse, kui suure tõenäosusega tulid teststatistiku väärtused võrdsed või suuremad reaalsete andmete põhjal leitud väärtusest. Sedasi saame teada, kui ekstreemne on reaalsete andmete põhjal arvutatud väärtus võrreldes saadud teststatistiku väärtustega. Saame välja kirjutada p-väärtuse kahepoolse hüpoteesi jaoks:

$$p = \frac{\sum_{i=1}^{m} C_i}{m}, \text{ kus} \qquad C_i = \begin{cases} 1 & \text{,kui } |D_i| \geqslant |D_t| \\ 0 & \text{,kui } |D_i| < |D_t| \end{cases} .$$

Seega, mida väiksem on p-väärtus seda ekstreemsem on reaalsete andmete pealt saadud väärtus, mis viitab sellele, et kahe haigla keskmiste erinevus on oluline. Nullhüpotees väidab aga, et kaks andmestikku on sama jaotusega ja keskmiste erinevus ei ole oluline [5]

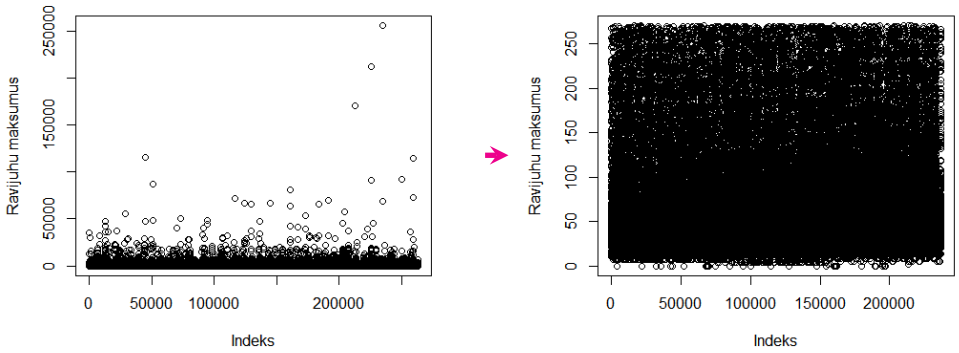
Kui meil on aga vaja lahendada eelnevalt toodud ülesanne 2, siis on mõistlikum kasutada bootstrap-meetodit. Kui eelnevalt kasutati juhuslikku taasvalikut tagasipanekuta, siis bootstrapi korral kasutatakse sama meetodit tagasipanekuga ühe grupi siseselt.

3 Kirjeldav analüüs

Kirjeldava analüüsi idee seisneb seaduspärasuste otsimises andmetes ilma eelnevalt eeldusi või hüpoteese kasutamata. Kirjeldav analüüs on vajalik ülevaate saamiseks andmetest, vajalike teisenduste tegemiseks ja edasiste uuringute paremaks kavandamiseks.

Vaadates, milliseid väärtusi omab uuritav tunnus ravijuhu maksumus, võib näha, et teistest erinevad üksikud väga suured arved. Mõistlik oleks need andmestikust eraldada ja vaadata ülejäänud arvetest eraldi. Antud juhul loeme erandlikeks summadeks kõik ravijuhu maksumused, mis ületavad 0.9-kvantiili. Andmete hajuvuse paranemist võib näha Jooniselt 2. Saame uue andmestiku, mis sisaldab 237 077 raviarvet.

Andmestikus on kahe haigla raviarved, milleks on Tallinna Lastehaigla ja Tartu Ülikooli Kliinikum. Esimeses neist on 153 305 raviarvet ja teises on 83 772 arvet.

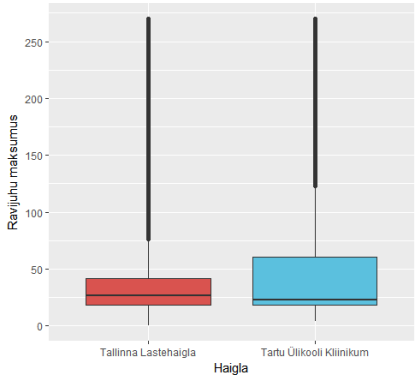


(a) Kõik andmed (b) Eemaldatud arved 0.9-kvantiiliga

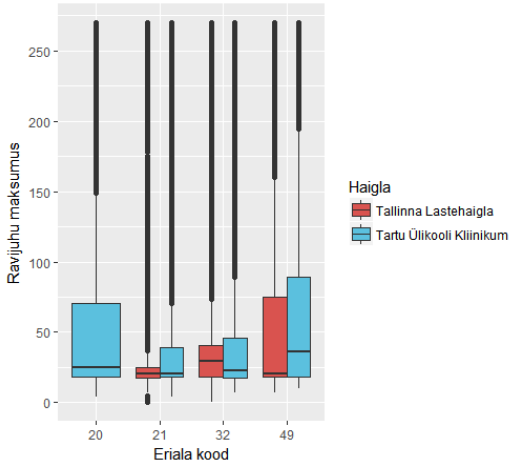
Joonis 2: Ravijuhu maksumuse ja patsiendi vaheline hajuvusdiagramm

Raviarve jaotuse hajuvust mõlemas haiglas illustreerib karpdiagramm joonisel 3. Ravijuhu maksumused haiglates on jagunenud üsna võrdselt. Mõlemas haiglas kõige suurem raviarve on 270 € ja mediaan on Tallinna Lastehaiglal 26.7 € ning Tartu Ülikooli Kliinikumil 22.7 €. Tartu Ülikooli Kliinikumis aga on suuremaid raviarveid rohkem.

Ravijuhu maksumuse hajuvust erialade kaupa mõlemas haiglas on võimalik näha jooniselt 4. Kood 20 tähendab üldkirurgiat, 21 lastekirurgiat, 32 ortopeediat ning 49 pediaatriat. Jooniselt võime näha, et üldkirurgia on esindatud ainult Tartu Ülikooli Kliinikumis. Kõige suurem ravijuhu maksumus on ka sellel joonisel kõigil sama, 270 €. Suuri mediaanide erinevusi erialade vahel pole märgata. Välja võiks tuua, et kõige kõrgem ravijuhu maksumuste mediaan on Tartu Ülikooli Kliinikumis pediaatria erialal 36 €. Võrreldes Tallinna Lastehaiglaga on mediaan sellel erialal 16 € võrra suurem.

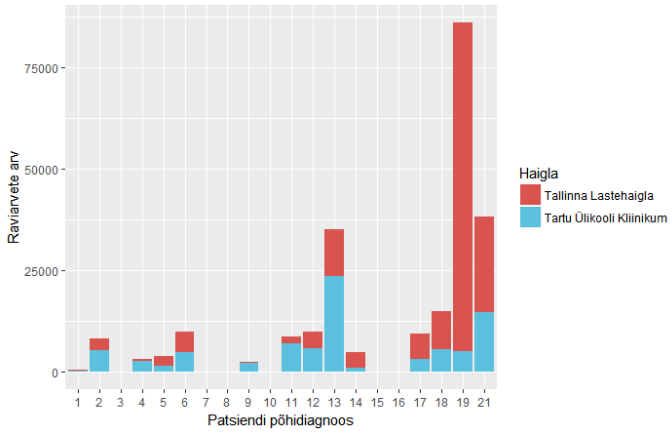


Joonis 3: Karpdiagramm ravijuhu maksumuse kohta erinevates haiglates



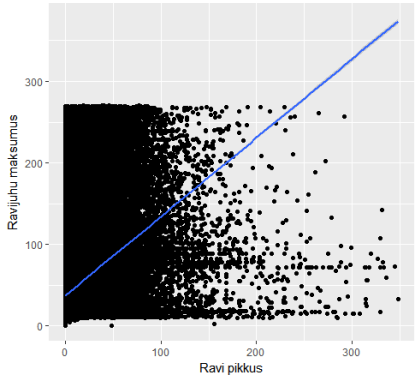
Joonis 4: Karpdiagramm ravijuhu maksumuse kohta erinevates haiglates eriala lõikes

Kuna andmestikus on erinevaid RHK-10 koode kokku 2056, siis on mõistlik nad jagada RHK-10 kõige üldisema klassifikatsiooni järgi plokkidesse. Indeksitele vastavaid koode ja nende tähendusi võib leida lisast 1. Jooniselt 5 võime näha, et andmestikus suurem osa patsientidest on haiglat külastanud mõne saadud vigastuse tõttu, eriti suur on see arv Tallinna Lastehaiglas. Tartu Ülikooli Kliinikumis on moodustatud arveid kõige rohkem patsientide pealt, kes on põdenud lihasluukonna ja sidekoehaigusi.

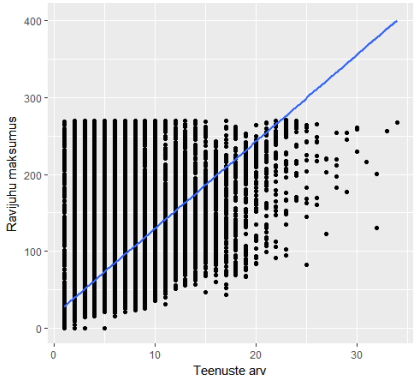


Joonis 5: Tulpdiagramm. RHK-10 koodide sagedused haiglates

Uuritavate tunnuste seas on kaks kvantitatiivset tunnust, millest üks on ravi pikkus päevades ja teine teenuste arv, mis selle raviarve ulatuses patsiendile tehti. Vastavalt joonistelt 6 ja 7 võime näha, et märkimisväärselt tugevat seost ravijuhu maksumuse ja uuritavate kvantitatiivsete tunnuste vahel pole. Mõlemal juhul, kui ravi pikkus on lühike või teenuste arv väike, võime leida ravijuhu maksumusi igas summas. Teenuste arvu kasvades on aga selgelt näha, et seda suurem on ka ravijuhu maksumus. Seega mingi seos leidub, mida võime näha ka mõlemale joonisele tõmmatud regressioonisirge abil.



Joonis 6: Ravi pikkuse ja raviarve vaheline hajuvusdiagramm



Joonis 7: Teenuste arvu ja raviarve vaheline hajuvusdiagramm

Selleks, et saaksime regressioonanalüüsi ainult diskreetsete tunnuste peal rakendada, tuleks eelmainitud kvantitatiivsed tunnused jagada gruppidesse. Moodus-

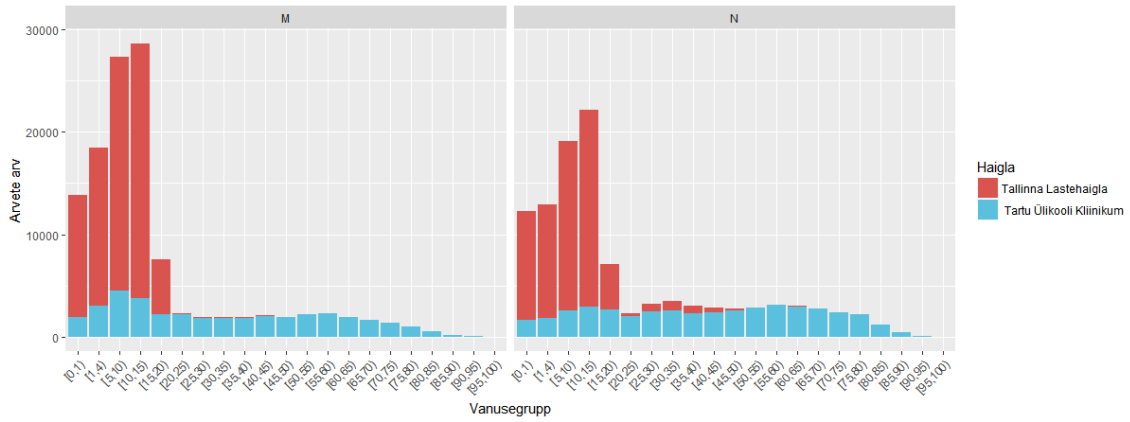
tatakse enam-vähem võrdse suurusega grupid, kus tunnuse ravi pikkus päevades väärtusteks saab

$$0, [1, 3), [3, 8), [8, 17), [17, 40), [40, 349)$$

ning tunnuse teenuste arv väärtusteks

$$1, 2, 3, 4, 5 \quad \text{ja} \quad 6 \quad \text{või rohkem} \ .$$

Vaadates lähemalt patsiendi eristuskategooriaid jooniselt 8 paistab silma, et enamus arvete taga on patsiendid vanuses $0-20$. See on ka loomulik, sest andmestikus on Tallinna Lastehaigla andmed, mis moodustavad suurema enamuse võrreldes Tartu Ülikooli Kliinikumi raviarvetega. Jooniselt 8 võime veel näha, et meessoost patsiente on rohkem kui naissoost patsiente. Kui vaadata sugu eraldi haiglates, siis Tartu Ülikooli Kliinikumi patsientide hulka kuulub siiski rohkem naisi.



Joonis 8: Tulpdiagramm. Vanusegruppide sagedused soo kaupa erinevates haiglates

4 Statistiline mudel

Selles bakalaureusetöös on moodustatud 5 mudelit iga huvipakkuva argumenttunnuse jaoks eraldi. Jooniste abil uuritakse kui täpselt suudavad saadud mudelid ennustada ravijuhtude keskmist maksumust Tallinna Lastehaigla ja Tartu Ülikooli Kliinikumi jaoks eraldi. Lisaks on joonistele toodud tärniga esile need keskmised, mille korral permutatsioonitesti või bootstrap-meetodi rakendamisel ei suudetud nullhüpoteesi kummutada. Permutatsioonitesti tulemus viitab sellele, et kahe haigla vahel tärniga märgitud väärtuste jaotused ei erine ja arved nendesse lahtritesse on saadud juhuslikult. Bootstrap-meetodit on kasutatud tunnuste tasemete peal, mis ühes haiglas esinevad aga teises mitte. Tärniga on märgitud need tunnuse väärtused, mille korral lahtri jaotus vastab üldkogumi jaotusele. Kui aga tärn puudub, siis viitab see jaotuste erinevusele vastavalt kahe haigla vahel või üldkogumi ja haigla vahel, mis võib olla tekitatud üksikutest väga kõrgetest ravijuhu maksumustest.

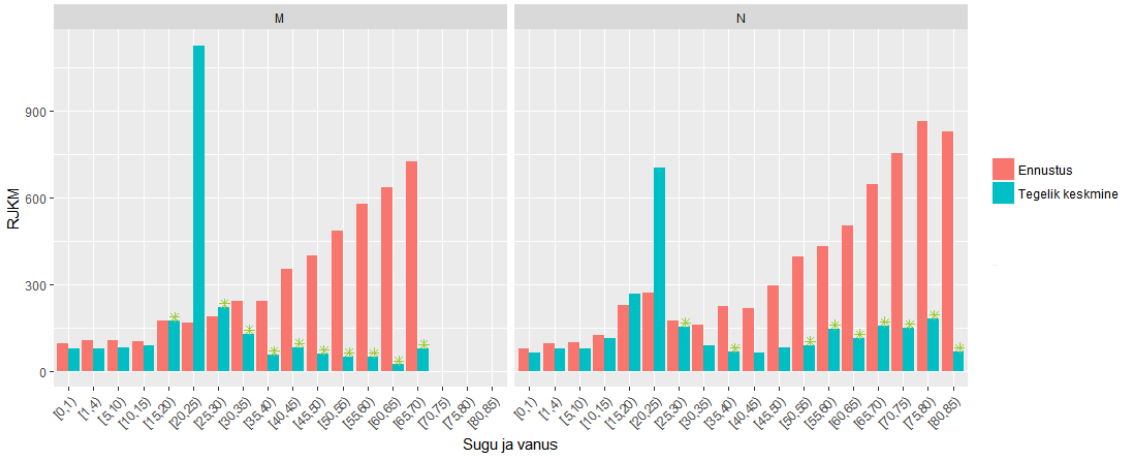
4.1 Soo ja vanusegrupi mõju ravijuhtude keskmisele maksumusele

Esmalt vaatame lähemalt, millist mõju avaldavad raviarvele patsiendi eristuskategooriad (sugu ja vanus). Sedasi saame teada, kas haigekassa peab eelarve moodustamisel arvestama sellega, millise eristuskategooriaga patsiendid haiglat kõige enam külastavad. Moodustatakse mudel üle kahe haigla, mille argumenttunnusteks on sugu, vanusegrupp ja nende koosmõju. Saadud mudel kirjeldab 1.97% uuritava tunnuse ravijuhu maksumuse hajuvusest. Kui aga moodustada mudel üle kõigi 11ne argumenttunnuse, siis saadud mudel kirjeldaks 90% uuritava tunnuse hajuvusest. Seega ei ole meie vaadeldav mudel üle soo, vanusegrupi ja nende koosmõju kuigi hea. Mudeli täpsust on võimalik arvutada juhusliku vea standardhälbe abil, milleks antud olukorras saame 1374 €.

Jooniselt 9 võime näha Tallinna Lastehaigla tegelikku RJKM ja mudeli poolt ennustatud RJKM. Paneme tähele, et Tallinna Lastehaigla korral ennustab mudel üsna täpselt ära keskmise raviarve kuni vanusegrupini $[15, 20)$ ja ka vanusegrupis $[25, 30)$. Vanemates vanusegruppides võib näha ennustatud RJKM kindlat tõusu,

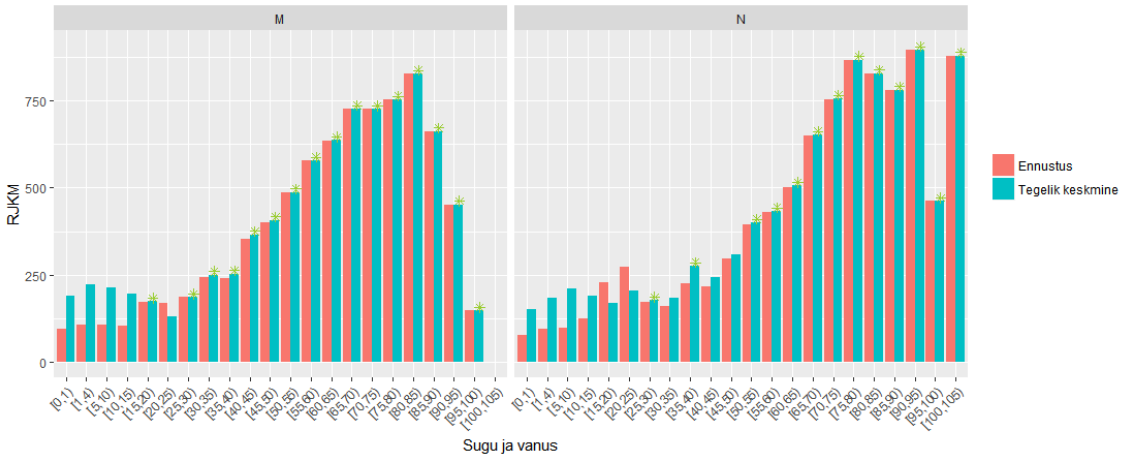
mis ei vasta Tallinna Lastehaigla tegelikule keskmisele. Vaadates andmetele peale on 35-85 aastased Tallinna Lastehaiglas kõik olnud statsionaarsel ravil. Seega võib vanemate inimeste madal ravijuhu maksumus olla tingitud sellest, et nad pole päris patsiendi rollis vaid on patsiendi saatjad ja kulud tulevad suuremas osas elamistingimuste pakkumisest haiglas.

Permutatsioonitesti tulemus näitab meile, et jaotus vanusegrupis $[20, 25)$ haiglate vahel on erinev. See võib tähendada seda, et kõrge tegelik keskmine vanusegrupis tuleneb üksikutest väga suurtest ravijuhtude maksumustest.



Joonis 9: Tallinna Lastehaigla tegelik keskmine ja ennustatud keskmine soo ja vanusegrupi järgi

Samal ajal jooniselt 10 võib näha, et Tartu Ülikooli Kliinikumi keskmist raviarvet soo ja vanusegrupi järgi ennustab mudel üsnagi täpselt. Permutatsioonitesti tulemus näitab, et Tartu Ülikooli Kliinikumis alla 15 aastaste laste ravijuhu maksumuste jaotus ei ole Tallinna Lastehaiglaga sama. Sellest võib olla tingitud ennustatud ja tegeliku keskmise erinevus.



Joonis 10: Tartu Ülikooli Kliinikumi tegelik keskmine ja ennustatud keskmine soo ja vanusegrupi järgi

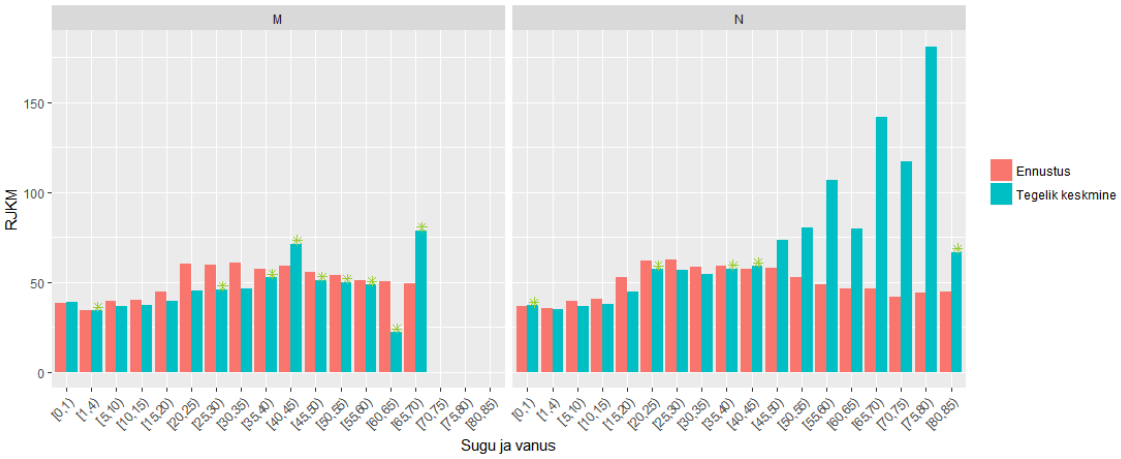
Joonisel 11 on välja toodud mudeli parameetrid. Rohelisega on märgitud vastava vanusegrupi parameetri väärtus, kollasega vabaliikme väärtus ning oranžiga vastava vanusegrupi ja vastava soo koosmõju. Jooniselt näeme, et puudub parameetri väärtus soo üksikmõjule, mis tähendab seda, et see tunnus osutus mudelis ebaoluliseks. Need parameetri väärtused, mis on RJKM-se teljel alla nulli, avaldavad RJKM-le negatiivset mõju.

Jooniselt näeme, et vanusegrupp on oluline RJKM kujunemisel. Tuleb välja, et mida vanem on patsient seda rohkem läheb maksma ka tema ravi. Joonise põhjal võib öelda, et suurem osa ennustusest on olulisusnivoo 0.05 juures osutunud oluliseks. Kõrgetes vanusegruppides, kus ennustus on ebaoluline, on liiga vähe patsiente selle andmestiku pealt järelduste tegemiseks.

Joonis 11: Mudeli parameetrid (vanusegrupp, sugu ja nende koosmõju)

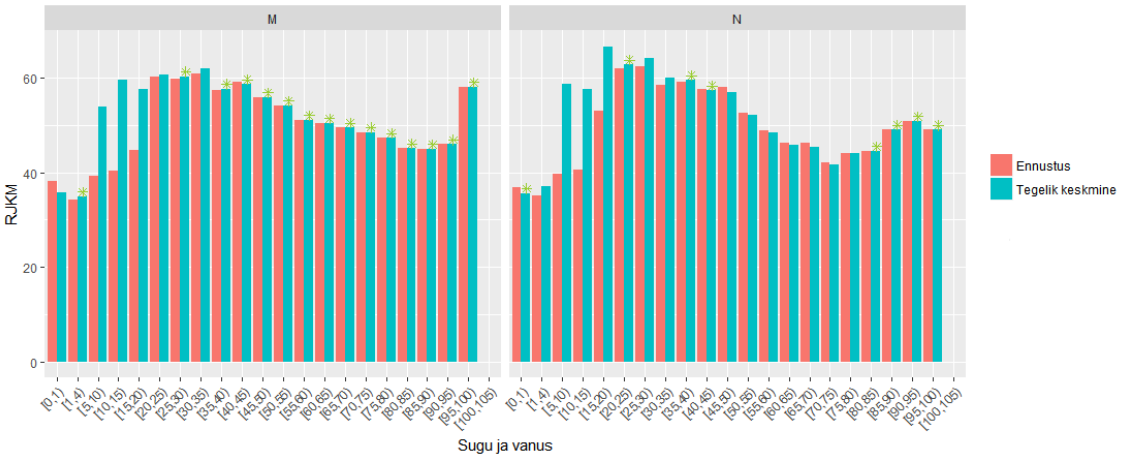
Eelnevalt saime kahtlaselt suure RJKM-e Tallinna Lastehaiglas vanusegrupis $[20, 25)$, mis arvatavasti on tingitud üksikutest suurtest ravijuhu maskumustest. Selle väite kinnitamiseks vaatame, kas suudame ennustust parandada, kui moodustame mudeli andmete pealt, kust on eriti suured arved eemaldatud. Uue andmestiku pealt saadud mudeli täpsus on 46 € ja mudel kirjeldab 2.7% uuritava tunnuse RJKM hajuvusest. Seega võrreldes eelmise mudeliga peaks antud mudel täpsemalt ennustama. Kui kaasata mudelisse kõik argumenttunnused, siis selline mudel kirjeldab 70% uuritava tunnuse hajuvusest.

Jooniselt 12 näeme, et ennustus on oluliselt paranenud. Kui enne oli tegeliku ja ennustatava keskmise raviarve vahe keskmiselt 269.5 €, siis nüüd on suudetud see vahe viia 19.3 €-ni. Samuti võime jooniselt näha, et eelnevalt saadud kõrge keskmine raviarve vanusegrupis $[20, 25)$ oli tõepoolest tingitud nendest üksikutest väga kõrgetest ravijuhtude maksumustest. Permutatsioonitesti järgi statistiliselt oluliseks osutunud kõrged keskmised vanuses 45-80 naiste hulgas võivad aga taaskord olla tingitud üksikutest suurtest raviarvetest. Silma paistab testi järgi mitteoluliseks osutunud RJKM 35-60 aastate meeste hulgas. Seega selles grupis on ravijuhu maksumuste jaotus haiglate vahel sama.



Joonis 12: Tallinna Lastehaigla tegelik keskmine ja ennustatud keskmine soo ja vanusegrupi järgi 0.9-kvantiili andmestiku põhjal)

Tartu Ülikooli Kliinikumi ennustust võib näha jooniselt 13. Kui ennustus oli juba algse andmestiku pealt üsna täpne, siis nüüd on suudetud ennustus viia veelgi lähemale tegelikkusele. Samas permutatsioonitest näitab taaskord, et tulemuseks saadud keskmised, mida mudel ei suuda kuigi täpselt ennustada (patsiendid vanuses 5-20), on olulised ning võivad olla tingitud üksikutest suurtest raviarvetest.



Joonis 13: Tartu Ülikooli Kliinikumi tegelik keskmine ja ennustatud keskmine soo ja vanusegrupi järgi 0.9-kvantiili andmestiku põhjal

Vaadates uue mudeli parameetreid jooniselt 14 näeme, et lisandunud on soo üksikmõju, mis ei ole kuigi suur aga on siiski oluline. Täpsemalt RJKM on 1.5 € suurem,

kui tegemist on meessoost patsiendiga. Suuremate raviarvete eemaldamise tulemusel on saadud mudel, mis nii konkreetselt enam RJKM-e tõusu vanuse kasvades ei näita. Seega võib väita, et RJKM tõus vanuse kasvades oli tingitud suuremate raviarvete suuremast esinemissagedusest vanuse kasvades. Väiksema andmestiku pealt saadud mudeli parameetrite pealt saame välja lugeda, et kõige madalam on RJKM vanusegrupi $[1, 4)$ korral. Vanusegruppide jaoks arvutatud parameetrid ei erine üksteisest kuigi palju, mis haigekassa jaoks võib isegi tähendada tunnuse vanusegrupp mitteolulisust.

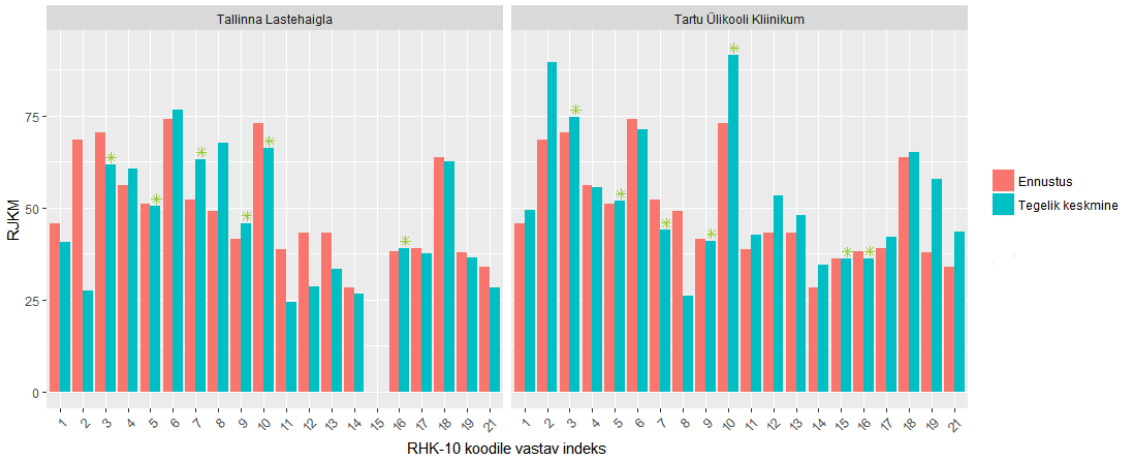
Joonis 14: Mudeli parameetrid (vanusegrupp, sugu ja nende koosmõju) 0.9-kvantiili andmestiku põhjal

Kokkuvõttes võime öelda, et suuremate raviarvete eemaldamine andis hoopis teise pildi võrreldes algse andmestikuga. Nimelt tuli välja, et vanusegrupi mõju RJKM-le ei olegi nii suur, kui algselt tundus. Lisaks saadi mudel, mille näitajad (täpsus ja headus) olid paremad. Kuna ennustuse tulemus muutus niivõrd palju võrreldes esialgse andmestiku pealt saadud ennustusega, siis oleks vajalik vaadata suuri ravijuhu maksumusi teistest eraldi. Selles töös ei keskenduta erandlikult suurte ravijuhu maksumuste analüüsimisele, vaid vaatame edaspidi tunnuste mõju RJKM-le andmestiku pealt, kust suuremad ravijuhu maksumused on eemaldatud 90% kvantiili abil.

4.2 RHK-10 mõju ravijuhtude keskmisele maksumusele

Moodustame nüüd uue mudeli, mis ennustaks RJKM-st RHK-10 koodide põhjal. Saame mudeli, mille ennustuse täpsust kirjeldav vea standardhälve on 45€ ja mudel kirjeldab 5.6% uuritava tunnuse RJKM hajuvusest. Seega mudel, kuhu on kaasatud ainult RHK-10 koodid, on taaskord natuke parem kui eelmine mudel, mis koosnes patsiendi eristuskategooriatest ja nende koosmõjudest.

Jooniselt 15 näeme tegeliku RJKM-e ja ennnustatud RJKM-e võrdlust. Väga suur erinevus tegeliku ja ennustatud keskmise vahel on Tallinna Lastehaigla RHK-10 koodile vastava indeksi 2 juures. Antud indeks käib kokku diagnoosiga kasvaja. Samal ajal Tartu Ülikooli Kliinikumis on kasvaja ravimisele keskmiselt kulunud rohkem raha, kui Tallinna Lastehaiglas. Kuna permutatsioonitesti tulemusel on haiglate vahelised jaotused selles grupis erinevad, siis ilmselt on Tartu Ülikooli Kliinikumis kõrge keskmine summa kasvaja ravimiseks tingitud üksikutest kõrgetest raviarvetest. Selle tõttu on ka Tallinna Lastehaiglas ennustus tulnud niivõrd kõrge ja erinev tegelikust keskmisest. Täiesti kindel selles järelduses siiski olla ei saa ja saadud erinevus vajab edasist analüüsimist. Lisaks on permutatsioonitesti läbiviimisel osutunud eriti olulisteks RHK-10 koodid, mida võib näha tabelist 3. Nende gruppide puhul saadi testi läbi tegemisel p-väärtus nullilähedane ehk olulisusenivoo 0.05 korral on nullhüpotees kindlalt kummutatud. Seega jaotus nende RHK-10 koodide korral on haiglati erinev.

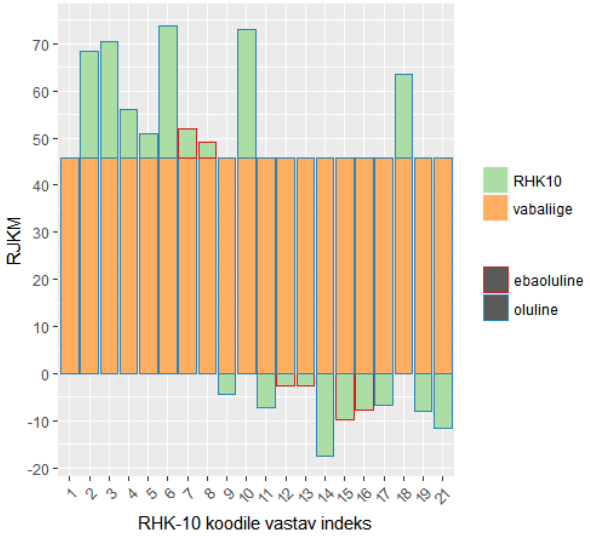


Joonis 15: Tallinna Lastehaigla ja Tartu Ülikooli Kliinikumi tegelik ja ennustatud RJKM RHK-10 koodide kaupa

Tabel 3: Permutatsioonitesti tulemusel oluliseks osutunud RHK-10 koodide tähendused

Indeks	**Kood**	**Nimetus**
6	G00-G99	Närvisüsteemihaigused
11	K00-K93	Seedeelundite haigused
12	L00-L99	Naha- ja nahaaluskoe haigused
13	M00-M99	Lihasluukonna ja sidekoehaigused
14	N00-N99	Kuse-suguelundite haigused
17	Q00-Q99	Kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad
19	S00-T98	Vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed
21	Z00-Z99	Terviseseisundit mõjustavad tegurid ja kontaktid teriseteenistusega

Jooniselt 16 võime näha koostatud mudeli parameetreid. Kollasega on märgitud vabaliige ja rohelisega erinevate RHK-10 koodile vastavate indeksite kordajad mudelis. Kui argumenttunnuse kordaja on y teljestikul positiivne, siis tähendab see selle RHK-10 koodile vastava indeksi positiivset mõju RJKM-le.



Joonis 16: Üle RHK-10 klasside koostatud mudeli parameetrid

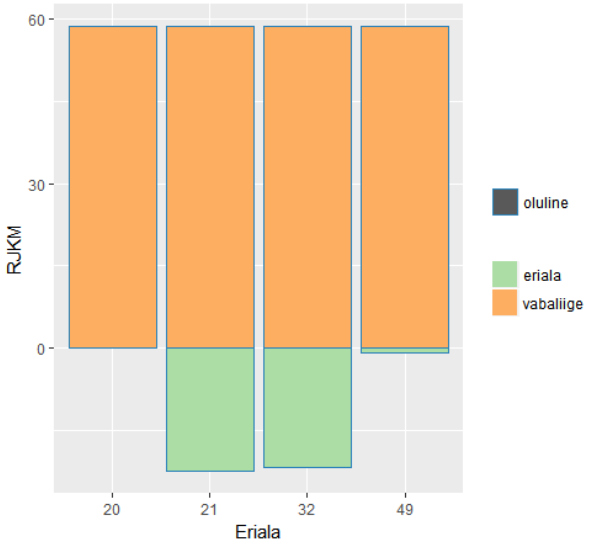
Kokkuvõttes võib öedla, et keskmine ravijutude maksumus põhidiagnooside lõikes on haiglati erinev, sest permutatsioonitesti tulemus näitas ainult mõnda üksikut RHK-10 koodi jaotuse sarnasust haiglate vahel.

Kui lisada argumenttunnus RHK-10 kood eelnevalt vaadatud mudelisse, mis sisaldas argumenttunnuseid sugu, vanus ja nende koosmõju, siis sugu ja vanus muutuvad mudelis ebaolulisteks. See näitab seda, et teades RHK-10 koodi on sellel tunnusel tunduvalt suurem mõju RJKM kujunemisel.

4.3 Meditsiinieriala mõju ravijuhtude keskmisele maksumusele

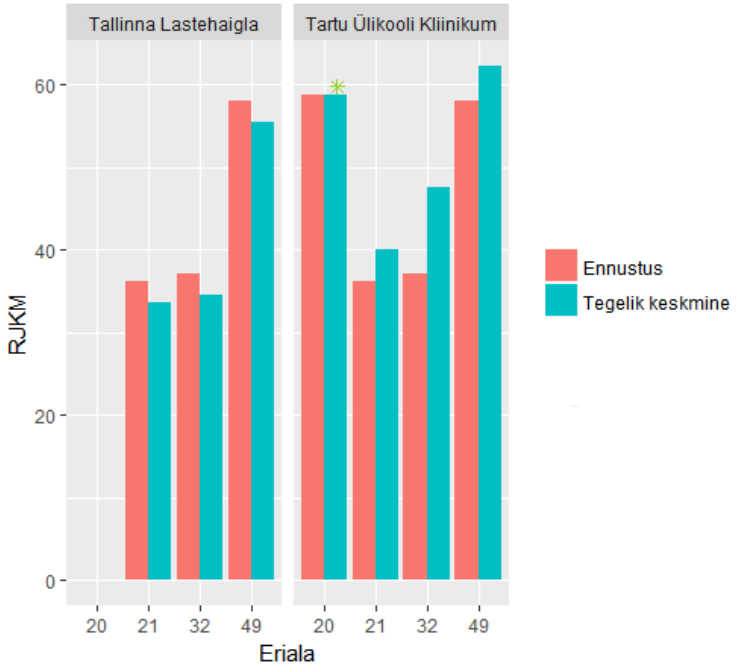
Koostame mudeli, mis näitaks seost RJKM ja meditsiinieriala vahel. Saadud mudeli täpsuseks on 45.5 € ja mudel kirjeldab 4.3% uuritava tunnuse RJKM hajuvusest.

Jooniselt 17 võime näha saadud mudeli parameetreid. Kõige kulukamad on meditsiinierialade 20-üldkirurigia ja 49-pediaatria alla kuuluvad haigusjuhtumid. Kõige vähem nõuavad aga raha meditsiinivaldkondade 21-lastekirurgia ja 32-ortopeedia alla kuuluvate haiguste ravimine. Kõik mudeli parameetrid on olulised.



Joonis 17: Üle meditsiinierialade koostatud mudeli parameetrid

Vaadates erialade keskmisi haiglate lõikes jooniselt 18, võime näha, et permutatsioonitesti tulemus viitab sellele, et jaotused erialade lõikes on kahe haigla vahel erinevad. Ainult Tartu Ülikooli Kliinikumis esinev meditsiinieriala 20-üldkirurgia olulisust on kontrollitud bootstrap-meetodiga. Tulemuseks on saadud, et RJKM erinevus üldkeskmisest ei ole statistiliselt oluline. Seega uuritav tunnus käitub sarnaselt üldkogumile ja võime tulemust üldistada ka teistele haiglatele.

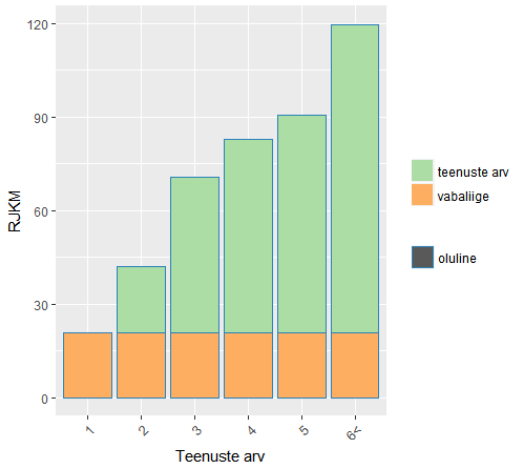


Joonis 18: Tallinna Lastehaigla ja Tartu Ülikooli Kliinikumi tegelik ja ennustatud RJKM meditsiinierialade kaupa

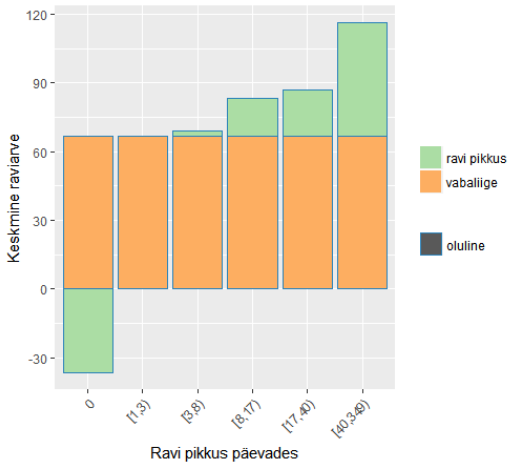
4.4 Teenuste arvu ja ravi pikkuse mõju ravijuhtude keskmisele maksumusele

Koostame eraldi mudelid selleks, et kirjeldada sõltuvust teenuste arvu ja RJKM-e ning ravi pikkuse ja RJKM-e vahel. Esimesena mainitud mudeli täpsuseks saadakse 36.8 € ja mudel kirjeldab 37% uuritava tunnuse RJKM hajuvusest. Ravi pikkuse ja RJKM-e vahelist seost kirjeldava mudeli täpsus on 39.5 € . Mudel kirjeldab 28% uuritava tunnuse RJKM hajuvusest.

Jooniselt 19 võib näha teenuste arvu ja RJKM-e vahelist seost. Selgelt on näha, et mida rohkem teenuseid üks ravijuht endas sisaldab, seda suurem on ka selle ravijuhu maksumus. Samasuunalist seost on näha ka kõrvaljoonisel 20 ehk mida rohkem päevi on patsiendi ravimisele kulunud, seda suurem on ka selle ravijuhu maksumus. Mõlemal joonisel on kõik mudeli parameetrid osutunud olulisteks.

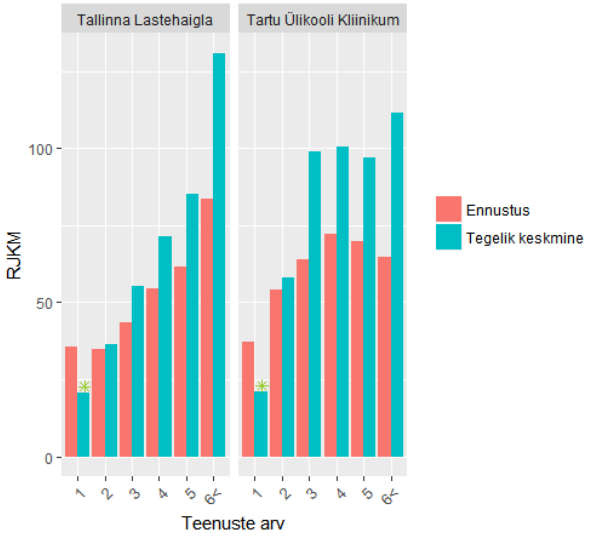


Joonis 19: Üle tunnuse "teenuste arv"koostatud mudeli parameetrid

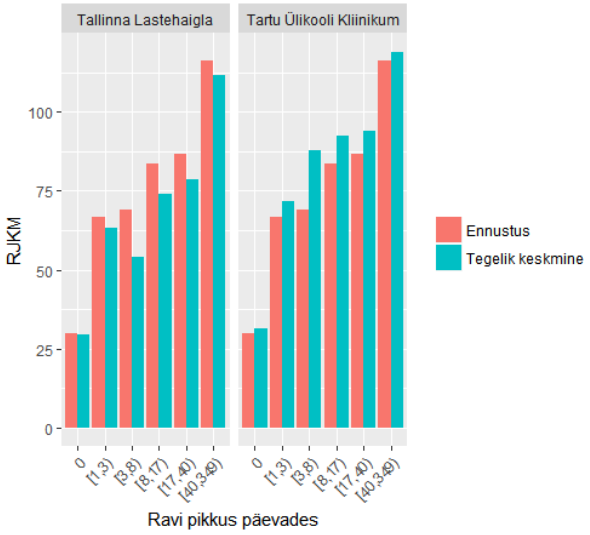


Joonis 20: Üle tunnuse "ravi pikkus"koostatud mudeli parameetrid

Järgnevalt on toodud joonised Tallinna Lastehaigla ja Tartu Ülikooli Kliinikumi tegelikud ja ennustatud RJKM-ed nii teenuste arvu (Joonis 21), kui ka ravi pikkuse (Joonis 22) erinevate tasemete kaupa. Mõlema joonise pealt näeme, et permutatsioonitesti tulemusel enamus juhtudel nullhüpotees kummutatakse. See tähendab seda, et samade argumenttunnuse tasemete vahel erinevates haiglates pole jaotus sama, mille tõttu tuleb ka ennustus ebatäpne.



Joonis 21: Tallinna Lastehaigla ja Tartu Ülikooli Kliinikumi tegelik ja ennustatud RJKM teenuste arvu kaupa



Joonis 22: Tallinna Lastehaigla ja Tartu Ülikooli Kliinikumi tegelik ja ennustatud RJKM ravi pikkuse kaupa

4.5 Ülevaade

Eelnevalt koostatud mudelid ei ennusta argumenttunnuse tegelikku väärtust kuigi täpselt haigla siseselt. Kui aga võtta mudelisse sisse kõik 11 argumenttunnust ja nende koosmõjud, saaksime väga pika mudeli (täpsemalt 2 199 parameetriga mudeli), mida on äärmiselt keeruline interpreteerida. Hea on siiski teada, et selline mudel ennustaks RJKM-st palju täpseamalt. Nimelt on taolise mudeli täpsus 25.5 € ja mudel kirjeldab 70% uuritava tunnuse RJKM hajuvusest.

Eelnevates peatükkides saadud tulemusi, mis tulid permutatsioonitesti või bootstrap-meetodiga statistiliselt ebaolulised (tärniga märgitud tulbad), võime üldistada ka teistele haiglatele. Kui aga olulisuse määramisel selgus, et jaotus on haiglate vahel vaadeldavas tunnuse tasemes erinev, siis üldistust teiste haiglate peale teha ei saa. Nimelt on sellisel juhul ennustuse ja tegeliku keskmise vahe tulnud haigla eripärast ja on seega oluline viga.

Analüüsi raames uuriti ka RJKM muutust kolme aasta vältel. Kuna aasta mõju RJKM tuli väga väike ±2 €, ei pakkunud teiste argumenttunnuste ajas muutumine töö koostajale erilist huvi. See aga ei tähenda seda, et tunnuste RJKM ajas

oleks muutumatu ja analüüsi läbi viimist tegema ei peaks.

Kokkuvõte

Eesmärk oli teada saada, kas erinevate haiglate ravijuhtude keskmine maksumus (RJKM) erineb oluliselt, kui võtta arvesse patsiendi eristuskategooriaid ja ravi iseloomu. Selleks töötati välja analüüsiks vajalik metoodika, mis aitaks paremini toime tulla suurte andmekogustega.

Viidi läbi regressioonanalüüs diskreetsete argumentidega. Vaadati mudelit, kus uuritav tunnus oli pidev ja kõik argumenttunnused diskreetsed. Selleks, et sellise mudeli leidmine arvutuslikult liiga mahukas ei oleks, jõuti lahenduseni, kus mudel on võimalik koostada teades ainult argumenttunnuse keskmist vaadeldavas tasemes ja vastavat kaalu, mis on võrdne selle taseme alla kuuluvate objektide arvuga. Kuna argumenttunnustel oli tasemeid kohati päris palju, siis parema interpreteeritavuse huvides anti mudeli poolt saadud ennustused ja mudeli parameetrid lugejale edasi visuaalsete joonistega. Joonistega mudeli sisu edasi andmine osutus väga heaks meetodiks, kuid on äärmiselt töömahukas. Selle tõttu vaadati lähemalt eraldi viite mudelit, mis endas eriti palju tunnuseid ei sisaldanud.

Mudelite ennustusvõimet uuriti eraldi Tallinna Lastehaigla ja Tartu Ülikooli Kliinikumi peal. Võrreldi joonise abil ennustatud RJKM ja tegelikku RJKM mõlemas haiglas. Rakendades permutatsioonitesti või bootstrap-meetodit uuritavate argumenttunnuste kõigi tasemete peal saadi sedasi teada, kas saadud erinevus tegeliku ja ennustatava keskmise vahel on statistiliselt oluline. Statistilise olulisuse määramine nende meetodite abil oli õigustatud. Sedasi saadi aimust, kas tulemus sobib üldistada ka teistele haiglatele või mitte.

Analüüsi läbiviimisel avastati, et mudel ennustab märgatavalt täpsemalt, kui eemaldada 90% kvantiili abil kõrgemad ravijuhu maksumused. Kõrgete raviarvete eemaldamise vajadusele vihjas Tallinna Lastehaigla soo ja vanuse põhjal tehtud joonis, kus ühes vanusegrupis tuli väga suur erinevus tegeliku ja ennustatud keskmise vahel. Samal ajal permutatsioonitest ütles, et erinevus on statistiliselt oluline ehk jaotus selles vanuseklassis on haiglati erinev.

Töös tehti enamus järeldused andmestiku pealt, kust olid eemaldatud erandlikult suured raviarved. Täieliku ülevaate saamiseks peaks kindlasti eraldi uurima ka väl-

ja jäetud suuri ravijuhu maksumusi. Lisaks võiks täiendava analüüsina vaadata, kuidas tunnuste tasemed muutuvad ajas.

Kasutatud kirjandus

[1] Bikienga, S. (2016). Introduction to simple Linear Regression. [www] https://bookdown.org/sbikienga/Intro_to_stat_book/ (15.03.2018). introduction-to-simple-linear-regression.html

[2] Käärik, E. (2017). Loengukonspekt Andmeanalüüs II.

[3] Med24. (2016). RHK-10. [www] https://www.med24.ee/andmebaasid/rhk10 (28.04.2018).

[4] Pennsylvania State University. (2015). Linear Regression. [www] https://onlinecourses.science.psu.edu/stat857/book/export/html/12 (15.03.2018).

[5] Pennsylvania State University. (2015). Permutation Principle. [www] https://onlinecourses.science.psu.edu/stat464/node/35 (15.03.2018).

[6] Shalizi, C. (2015). Sufficient Statistics. http://bactra.org/notebooks/sufficient-statistics.html (03.05.2018).

[7] World Health Organization. (2016). ICD-10 Classifications. [www] http://www.who.int/classifications/icd/icdonlineversions/en/ (28.04.2018).

Lisad

Lisa 1. RHK-10 koodide tähendused

Tabel 4: RHK-10 (Rahvusvaheline Haiguste Klassifikatsioon) [3]

Indeks	Kood	Nimetus
1	A00-B99	Teatavad nakkus-ja parasiithaigused
2	C00-D48	Kasvajad
3	D50-D89	Vere-ja vereloomeelundite haigused ning teatavad immuunmehhanismidega seotud haigusseisundid
4	E00-E90	Sisesekretsiooni-, toitumis-ja ainevahetushaigused
5	F00-F99	Psüühika- ja käitumishäired
6	G00-G99	Närvisüsteemihaigused
7	H00-H59	Silma-ja silmamanuste haigused
8	H60-H95	Kõrva- ja nibujätkehaigused
9	I00-I99	Vereringeelundite haigused Morbi systematis circulatorii
10	J00-J99	Hingamiselundite haigused
11	K00-K93	Seedeelundite haigused
12	L00-L99	Naha- ja nahaaluskoe haigused
13	M00-M99	Lihasluukonna ja sidekoehaigused
14	N00-N99	Kuse-suguelundite haigused
15	O00-O99	Rasedus, sünnitus ja sünnitusjärgne periood
16	P00-P96	Perinataal- e sünniperioodis tekkivad teatavad seisundid
17	Q00-Q99	Kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad
18	R00-R99	Mujal klassifitseerimata sümptomid, tunnused ja kliinilised ning laboratoorsete leidude hälbed
19	S00-T98	Vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed
20	V01-Y98	Haigestumise ja surma välispõhjused
21	Z00-Z99	Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega
22	U00-U99	Koodid spetsiifiliste eesmärkide jaoks

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kristiina Uusna,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Statistilise analüüsi rakendamine Eesti Haigekassaraviarvetele", mille juhendaja on Sven Laur,

   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, 08.05.2018