Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Markus Ellisaar

Mittelineaarsed regressioonmudelid rakendusega spordiandmetele

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: dots. Imbi Traat

Tartu 2017

Mittelineaarsed regressioonmudelid rakendusega spordiandmetele

Kõiki praktikas esinevaid andmestikke ja tunnuste vahelisi seoseid ei ole võimalik lineaarselt piisavalt täpselt kirjeldada. Käesoleva töö peamiseks eesmärgiks on uurida erinevaid mittelineaarseid mudeleid ja tutvustada nende teooriat ning kasutamist. Illustreeriva materjalina on toodud näidetena ka graafikuid, mis aitavad kõiki regressioonitüüpe paremini mõista. Töö lõpetavad kergejõustiku Berliini maailmameistrivõistluste 100m jooksu mõõtmiste peal tehtud praktilised näited.

Märksõnad: mittelineaarsus, polünoomregressioon, treppfunktsioon, splainid, lokaalne regressioon.

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

Nonlinear regression models with applications to sports data

There are data and variables in practice that cannot be described well enough with linear methods. The aim of this thesis is to study different nonlinear regression models and give an overview of the theory and usage. To understand different approaches and regressions better, several graphics are given. Thesis ends with practical examples made with Berlin athletics world championships 100m sprint measurings.

Keywords: nonlinearity, polynomial regression, step function, splines, local regression.

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

Sisukord

Sissejuhatus 4
1. Lühikirjeldus 5
2. Mittelineaarsed mudelid 6
2.1. Polünoomregressioon 6
2.2. Treppfunktsioonid 8
2.3. Baasfunktsioonid 10
2.4. Regressioonisplainid 10
2.4.1. Tükiti määratud polünoomid 11
2.4.2. Splainide kitsendused 11
2.4.3. Splainide üldine kuju 13
2.4.4. Sõlmede arv ja väärtused 14
2.5. Siluvad splainid 15
2.6. Lokaalne regressioon 15
3. Praktilised näited 18
Kokkuvõte 23
Kasutatud kirjandus 24
Lisa 1 25
Lisa 2 27
Lisa 3 29
Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks 31

Sissejuhatus

Käesolev bakalaureusetöö on jagatud kaheks suuremaks peatükiks. Esimene peatükk algab meetodite lühitutvustusega. Edasi käsitletakse igat meetodit ja sellega kaasaskäivat teooriat juba eraldi ja täpsemalt. Meetodite juures kaasaskäivad illustreerivad joonised on autori reproduktsioonid kasutades avalikuks kasutamiseks antud andmestikku (James jt., 2013), kui pole öeldud teisiti. Uuritavateks on järgnevad regressioonmudelite tüübid:

- polünoomregressioon,
- treppfunktsioonid,
- regressioonisplainid,
- siluvad splainid,
- lokaalne regressioon.

Teises peatükis rakendatakse saadud teadmisi autori poolt valitud andmestikul, milleks on kergejõustiku Berliini maailmameistrivõistluste 100 meetri sprindi mõõtmistulemused (German Athletics Federation, 2009). Rakendatud on polünoomregressioone, treppfunktsiooni ja splaine.

Töö on kirjutatud tekstitöötlusprogrammiga Microsoft Word. Jooniste tegemiseks ja näidete läbiviimiseks on kasutatud statistikatarkvara R. Töös kasutatud R'i koodid on saadaval lisades.

Autor tänab dotsent Imbi Traati bakalaureusetööd puudutavate nõuannete ja täienduste eest.

1. Lühikirjeldus

Käesolevad lühikirjeldused on refereeritud James jt. (2013) põhjal. Anname võrdleva lühiülevaate meetoditest uuritava tunnuse *Y* modelleerimiseks sõltumatu muutuja *X* korral.

Polünoomregressioon laiendab lineaarset mudelit, lisades juurde argumenttunnuseid, mis omakorda saadakse esialgsete tunnuste astmesse tõstmisel. Näiteks on kuupregressioonis kolm muutujat, X, X^2 ja X^3, mida käsitletakse sõltumatutena. Selline lähenemine on üks lihtsamaid mooduseid, kuidas tekitada ja sobitada mittelineaarne mudel andmetele.

Treppfunktsioonid lõikavad tunnuse *X* piirkonna *K* erinevaks piirkonnaks ja nii moodustatakse uus kvalitatiivne muutuja. Selle tulemusena sobitatakse tunnusega konstantne funktsioon igas piirkonnas. Hüpe tekib iga piirkonna otspunktides ning tulemusena tekkivast funktsiooni kujust sõltuvalt on ka meetodi nimi.

Regressioonisplainid on paindlikumad kui polünoomid või treppfunktsioonid. Tegemist ongi kahe eelmise kombinatsiooniga. Sarnaselt jagatakse tunnuse *X* muutumispiirkond *K* osaks. Igas osas leitakse polünoomfunktsioon, mis sobib *Y*-tunnusega kõige paremini. Polünoomidele pannakse kitsendusi, et saavutada sujuv üleminek ühelt piirkonnalt teisele. Eeldusel, et tekkinud piirkondi on piisavalt palju, võib sellise meetodi prognoosivõime olla ülimalt täpne.

Siluvad splainid sarnanevad küll eelmistega, kuid tulevad esile siiski mõnevõrra teises olukorras. Siluvate splainide korral on oluline otsitava funktsiooni sujuvus, mille saavutamiseks lisatakse vastav kitsendus vähimruutude meetodile juurde. Funktsioonist, mida lisatakse ning mis aitab piirata siluvaid splaine, räägitakse vastavas peatükis täpsemalt.

Lokaalne regressioon on sarnane splainidele, kuid oluline erinevus on see, et piirkonnad võivad omavahel kattuda, kuid sellest hoolimata tuleb saavutada sujuv regressioonifunktsioon. Peamiseks ideeks on valida väiksemaid punktiparvesid sobitamiseks. Tulemuseks saame palju hinnanguid, mille põhjal moodustatakse regressioonifunktsioon.

2. Mittelineaarsed mudelid

Selles peatükis kirjeldame valitud mittelineaarseid mudeleid põhjalikumalt. Teemat illustreerime joonistega, mis on reprodutseeritud James jt. (2013) andmestiku ja R-koodide abil või võetud otse internetileheküljelt[1].

2.1. Polünoomregressioon

Kirjeldame polünoomregressiooni kasutades allikat James jt, 2013, lk 266-268.

Lineaarsuse laiendamine mittelineaarseks on tavapäraselt lähtunud valemist, mille määrab seos sõltumatu muutuja *X* ja prognoositava tunnuse Y vahel:

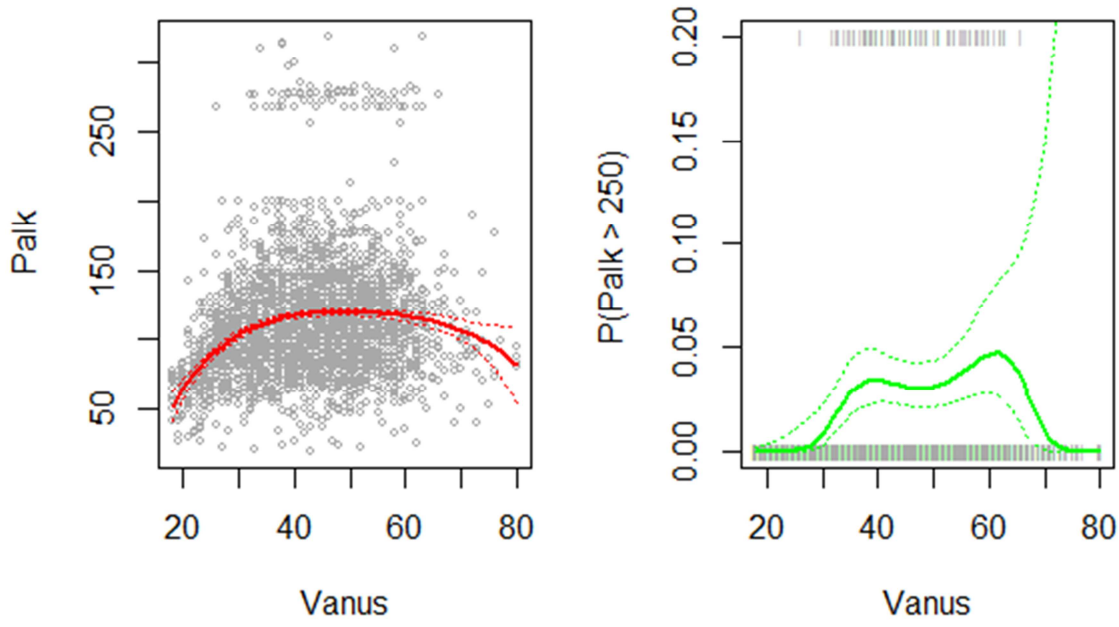
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (1)$$

Tulemuseks on järgnev valem:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \varepsilon_i \quad (2)$$

kus ε_i on juhuslik viga. Seos (2) esitabki polünoomregressiooni. On loomulik, et mida suurem on aste *d*, seda suurem erinevus on polünoomil lineaarse sirgega. Vaadeldes x_i astmeid aga uute tunnustena, näeme valemis (2) lineaarset mudelit. Parameetreid β_i on nüüd lihtne hinnata vähimruutude meetodil, kus muutujateks on vastavalt x_i, x_i^2 x_i^3 jne. Praktikas kasutatakse harva suuremat *d* väärtust, kui 3 või 4, kuna selle tulemusena võivad tekkida joontel väga ootamatud kujud, eriti veel tunnuse *X* väärtuspiirkonna otstes. Järgnevalt uurime kahte näidet.

Joonise 1 vasakpoolsel pildil on pideva joonega tähistatud 4. astme polünoom vanuse funktsioonina palgale, mis on leitud vähimruutude meetodil. Joon näitab keskmise aastapalga muutumist tuhandetes dollarites sõltuvana vanusest. Katkendlike joontega on tähistatud 95% usaldusvahemik. Paremal pool on sama andmestiku peal tehtud binaarne tunnus, mille puhul väärtus 1 omistatakse juhul, kui inimene teenib rohkem kui 250000$ aastas. Sel juhul on kasutatud logistilist regressiooni, samuti 4. astme polünoomiga. Joonisel on pideva joonega esitatud tõenäosus saada suuremat palka kui 250000$ sõltuvalt vanusest. Katkendlike joontega on siin ära märgitud 95% usaldusvahemik.



Joonis 1. Vasakpoolne: Palkade ja vanuse seos kasutades polünoomregressiooni. Parempoolne: Logistiline polünoomregressioon binaarse tunnuse põhjal, mida genereeritakse tingimusest palk > 250 (250000$).

Uurime nüüd lähemalt, kuidas jõuda prognoosi standardveani, mida on vaja usaldusvahemike jaoks. Arvutame konkreetse funktsiooni väärtuse st. prognoosi kohal vanus $= x_0$:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4. \quad (3)$$

Oluline on leida selle funktsiooni dispersioon ehk antud hetkel $Var(\hat{f}(x_0))$. Vähimruutude meetod annab igale parameetrile hinnangu $\hat{\beta}_j,\ j = 0,1,\dots,4$, ja vastava dispersioonihinnangu ja samuti suuruste $\hat{\beta}_j$ vahelised kovariatsioonid. Selle tulemusena tekib kovaratsioonimaatriks **C**, mis antud juhul on 5x5 maatriks. Selle abil saame arvutada dispersiooni hinnangu ka $\hat{f}(x_0)$-le valemiga $Var\left(\hat{f}(x_0)\right) = l_0^T * \mathbf{C} * l_0$, kus $l_0^T = (1, x_0, x_0^2, x_0^3, x_0^4)$. Vastav standardhälbe punkthinnang ehk standardviga on ruutjuur saadud tulemusest. Kahekordse standardvea abil saadakse usalduspiirid $f(x_0)$-le. Seda põhimõtet rakendatakse igas punktis x, mille tulemusena saadakse hinnanguline polünoomi joon koos kahekordse standardveajoonega kummalgi pool. (James jt, 2013, lk 266-268)

Kuna andmestikus on üsna suure erinevusega palkasid, siis moodustame binaarse tunnuse kõrgepalgalistest ja madalapalgalistest, kus aastapalga piiriks tuhandetes on valitud 250.

Vastavat binaarset tunnust saab prognoosida logistilise regressiooni abil, kus sisendiks on endiselt valitud 4. astme polünoom, mudeliks saame:

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}{1+\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)} \quad (4)$$

Joonekesed parempoolse joonise üla- ja alaosas näitavad vastavalt kõrgepalgaliste ja madalapalgaliste vanust. Pidev kõverjoon näitab, millise tõenäosusega saab inimene olla kõrgepalgaline. Nagu näha, on usaldusintervall üsna lai, eriti parempoolses otsas. Kuigi valim on võrdlemisi suur (*n* = 3000), siis on kõrgepalgalisi vaid 79, mistõttu on nende parameetrite dispersiooni hinnangud väga suured, mis viibki laia usaldusvahemikuni.

2.2. Treppfunktsioonid

Kui me ei soovi konstrueerida sellist polünoomi, mis on sama kogu *X* määramispiirkonnas, siis on heaks alternatiiviks treppfunktsioonid. Treppfunktsioonide kirjeldamiseks kasutame materjali James jt., 2013, lk 268-270. Siin jagatakse *X* määramispiirkond osadeks ja igal osal sobitatakse *Y*-tunnusele konstantne väärtus. Sisuliselt tehakse pidevast tunnusest x diskreetne tunnus. Täpsemalt tähendab see, et tekitame punktid $c_1, c_2, \ldots, c_K$ määramispiirkonnas, millest saavad meie piirid. Seejärel konstrueerime $K+1$ uut muutujat:

$$\begin{aligned} C_0(X) &= I(X < c_1), \\ C_1(X) &= I(c_1 \leq X < c_2), \\ C_2(X) &= I(c_2 \leq X < c_3), \\ &\ldots \\ C_{K-1}(X) &= I(c_{K-1} \leq X < c_K), \\ C_K(X) &= I(c_K \leq X), \end{aligned} \quad (5)$$

kus I on indikaatorfunktsioon, mis väljastab väärtuse 1, kui tingimus(ed) on täidetud, vastasel juhul 0. Paneme tähele, et iga *X* korral

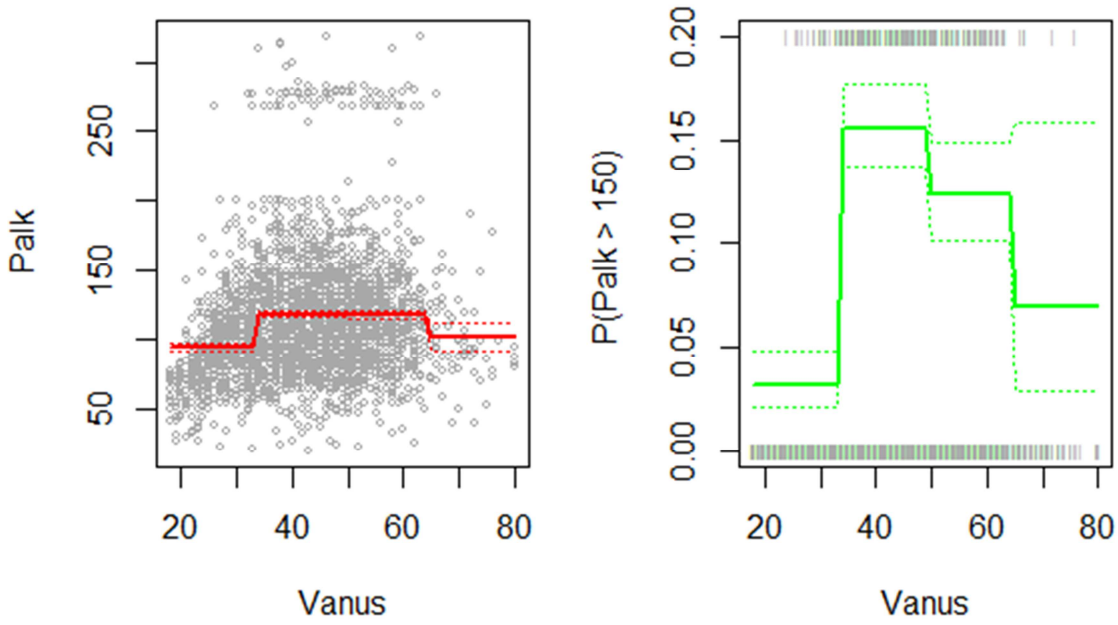
$$C_0(X) + C_1(X) + \cdots + C_K(X) = 1, \quad (6)$$

kuna *X* peab olema ühes moodustatud piirkondadest. Oleme saanud lineaarse mudeli, kasutades $C_1(X), C_2(X) \ldots C_K(X)$ kui sõltumatuid muutujaid. Siinkohal jätame $C_0(X)$ välja, sest see on üleliigne koos vabaliikmega.

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \varepsilon_i. \quad (7)$$

Konkreetse X väärtuse korral saab korraga maksimaalselt üks $C_1, C_2, \dots, C_K$ olla nullist suurem. Juhul, kui $X < c_1$, siis kõik muutujad tulevad nullid, mis tähendab, et β_0 võime vaadata kui Y keskmist piirkonnas $X < c_1$. Lisaks kui võtame $\beta_0 + \beta_j$ juhul $c_j \leq X < c_{j+1}$, saame, et β_j tähistab keskmist kasvu piirkonnast $X < c_1$ piirkonda $c_j \leq X < c_{j+1}$ üleminekul. Parameetrid $\beta_j, j = 0,1, \dots, k$ hinnatakse vähimruutude meetodil.

Uurime sama andmestikku, st. palga sõltuvust vanusest. Nüüd aga modelleerime seda treppfunktsioonidega.



Joonis 2. Palkade ja vanuse seos treppfunktsiooni abil. Vasakul olev joonis kujutab endast regressiooni treppfunktsiooniga ja parempoolne joonis logistilist regressiooni treppfunktsiooniga.

Joonise 2 vasakpoolsel graafikul on treppjoonega tähistatud keskmise aastapalga prognoos, mis on saadud vanuse jagamisel kolmeks piirkonnaks. Katkendlikud jooned on 95% usaldusvahemikeks. Paremal joonisel on sarnaselt varasemaga tekitatud binaarne tunnus, kus kõrgepalgalise aastapalga piiriks tuhandetes dollarites on seekord valitud 150. Logistilise regressiooni modelleerimise tulemusena saame trepi kujulise tõenäosusgraafiku koos katkendlike joontega märgitud 95% usaldusintervallidega. Logistilise regressiooni valem on

$$\Pr(y_i > 150|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i))}. \quad (8)$$

Kahjuks on treppfunktsioonil ka teatavad puudused. Kui puuduvad loomulikud eralduspunktid andmestiku määramispiirkonnas, siis tükihaaval sisestatud konstandid võivad jätta kirjeldamata olulised iseärasused, nagu on näha ka selle näite puhul, kus trepi esimene osa ei kajasta kuidagi esimest tõusu, küll aga kirjeldas selle ära aga polünoomregressioon. Sellele vaatamata on trepid eriti populaarsed näiteks biostatistikas ja epidemoloogias, kus tüüpiline lõikepunkt vanuse jaoks on iga 5 aasta järel.

2.3. Baasfunktsioonid

Senini kirjeldatud polünoom- ja treppregressiooni mudelid on vaadeldavad nn. baasfunktsioonide erijuhtudena. Idee seisneb selles, et meil on olemas etteantud arv baasfunktsioone $b_1, b_2, b_3, \dots, b_K$, mida saame rakendada tunnuse X peal. Tulemusena saame järgmise lineaarse mudeli:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_K b_K(x_i) + \varepsilon_i. \quad (9)$$

Paneme tähele, et baasfunktsioonid b_j on alati teada ja fikseeritud. Eelnevalt vaadatud polünoomide puhul oli $b_j(x_i) = x_i^j$ ja treppide puhul olid need defineeritud kui $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$. Vaadeldes saadud valemit kui üldist lineaarset mudelit, mille tunnusteks on $b_j(x_i)$, saame rakendada vähimruutude meetodit, et leida teadmata regressioonikordajad. Seega erinevad järeldused, mida saame teha lineaarse mudeli puhul nagu näiteks F-statistiku leidmine või standardvead, rakenduvad ka siin (James jt., 2013, lk 270). Järgmisena uurimegi üht väga populaarset valikut baasfunktsioonideks.

2.4. Regressioonisplainid

Regressioonisplainid võtavad endasse head omadused nii polünoomidelt kui ka konstantsetelt treppidelt. Järgnevalt uurime täpsemalt erinevaid aspekte, mis puudutavad just neid splaine, sest lähenemisi ja piiranguid on erinevaid, mida hiljem illustreerivad ka joonised. Materjalina kasutame James jt., 2013, lk 271-274.

2.4.1. Tükiti määratud polünoomid

Selle asemel, et sobitada üks kõrge astmega polünoom kogu määramispiirkonnale, saame jagada piirkonna tükkideks ja igale tükile määrata madalama astmega sobiva polünoomi. Siin on näide ühest kuuppolünoomist

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3} \quad (10)$$

kus kordajad $\beta_0, \beta_1, \beta_2$ ja β_3 erinevad piirkonna erinevates osades. Kohti, kus parameetrid muudavad väärtust, nimetatakse sõlmedeks. Polünoomregressiooni näites polnud ühtegi sõlme, mille tulemusena tekib tavaline polünoom. Kui aga näiteks on üks sõlm *c*, siis võtab polünoom uue kuju:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i, & x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i, & x_i \geq c. \end{cases} \quad (11)$$

Siin oleme jaganud vaatlused kahte gruppi, ühed, kus $x_i < c$ ja teised, kus $x_i \geq c$. Mõlemat gruppi iseloomustab erinevate kordajatega polünoom, mis on hinnatud vähimruutude meetodil.

Mida rohkem sõlmi kasutada, seda paindlikumaks muutub ka hinnang. Teisisõnu kui valida K sõlme üle kogu määramispiirkonna, siis peame ka leidma K+1 kuuppolünoomi. Tegelikult ei ole oluline kasutada just kuuppolünoomi, võib kasutada ka tükiti lineaarset funktsiooni. Treppide juures vaadatud mudelit võib käsitleda ka kui tükiti polünoome, mille aste on 0.

2.4.2. Splainide kitsendused

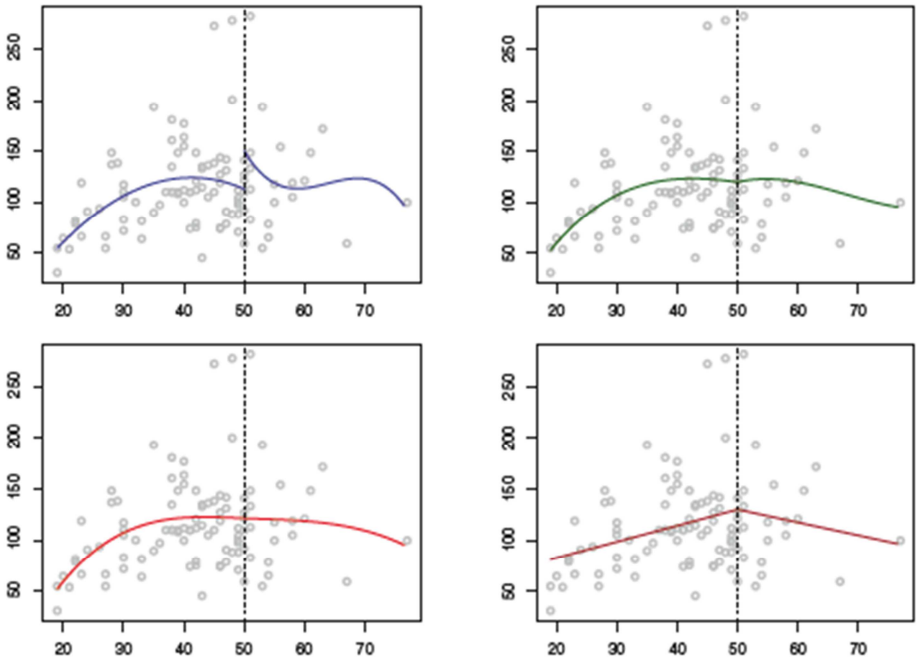
Splainide kasutamisel on väga olulised neile seatud kitsendused, sest vastasel juhul võime saada ebaloogilised tulemused, nagu on näha joonisel 3 vasakul üleval graafikul, kus punktis *x* = 50 prognoositakse kahte erinevat väärtust. Probleemi vastu aitab tingimus, et graafik peab olema pidev, ja ei saa olla hüpet punkti 50 juures.

Tulemus on näha parempoolsel graafikul samuti joonisel 3, mis on küll parem kui eelmine graafik, kuid vajab veel kitsendusi, sest V-kujuline graafiku osa on sees.

Esimest ja teist järku tuletiste võrdumise nõue vaadeldavas punktis aitab siluda meie poolt otsitud regressioonigraafikut. Tingimuseks on, et tuletised oleksid pidevad. Kokkuvõtlikult saame öelda, et iga rakendatud piirang vähendab vabadusastet ühe võrra. Joonise 3 puhul on

vasakul üleval joonisel kaheksa vabadusastet ja rakendades sellele kolme piirangut(pidevus, esimese tuletise pidevus, teise tuletise pidevus), saame tulemuseks kuupsplaini (Joonis 3 all vasakul) ja vabadusastmete arvuks viis.

Kuupsplainid on populaarsed just seetõttu, et raske on visuaalselt tuvastada katkevust sõlmedes. Sõlmede arvust *K* saame ka vabadusastmete arvuks 4 + *K*. (James jt., 2013, lk 271-273)



Joonis 3. Regressioonisplainid sõlmega kohal 50. Üleval vasakul: Kuuppolünoomid on kitsendusteta. Üleval paremal: Tingimuseks on polünoomide pidevus kohal 50. All vasakul: Tingimusteks on polünoomide, nende esimeste ja teiste tuletiste pidevused kohal 50. All paremal: Lineaarsed splainid, mis on ühendatud pidevalt (James jt., 2013, lk 272).

Siit saame ka üldise definitsiooni *d*-astme splainide jaoks (James jt., 2013, lk 273): *d*-astme splainiks nimetatakse tükiti *d*-astme polünoome, mille puhul on kõik kuni *d* - 1 järku tuletised pidevad igas sõlmes.

Järgmises punktis uurime juba üldisemalt, kuidas saavutada nõutud tingimusi.

2.4.3. Splainide üldine kuju

Splainidega tutvumisel kasutame materjali James jt., 2013, lk 273-274. Splainide esitamiseks lähtume valemist (9). Sellest tulenevalt saame modelleerida kuupsplaini K sõlmega järgnevalt:

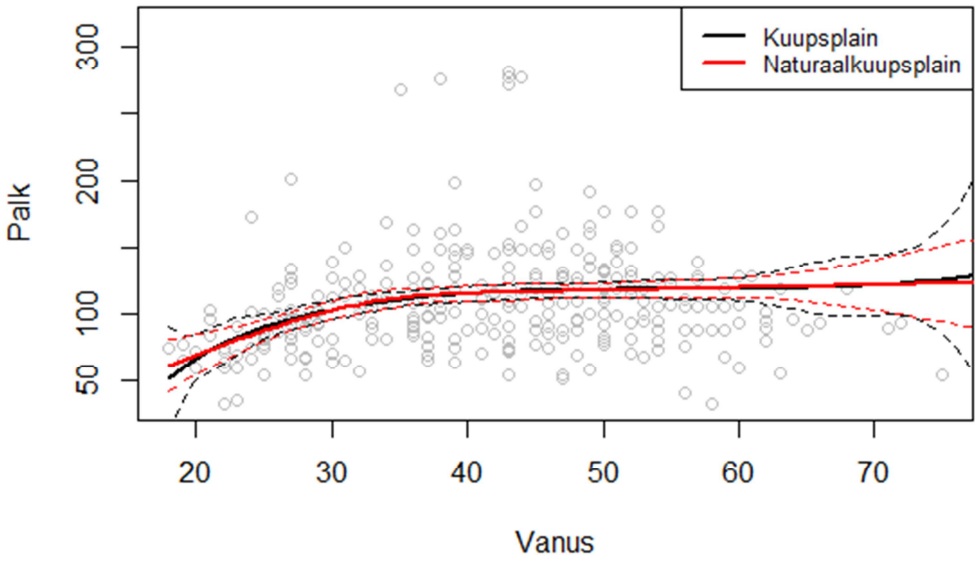
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \varepsilon_i, \qquad (12)$$

kus b_1, b_2,…,b_{K+3} on baasifunktsioonid.

Kuupsplainide korral kasutatakse valemis (12) baasfunktsioonidena x, x^2, x^3 ja lisaks lõigatud astmefunktsioone vastavalt sõlmede arvule K. Lõigatud astmefunktsioon on defineeritud järgnevalt:

$$h(x,\xi) = (x-\xi)_+^3 = \begin{cases} (x-\xi)^3, & \text{kui} \quad x > \xi \\ 0, & \text{mujal,} \end{cases} \qquad (13)$$

kus ξ on sõlm. On ilmne, et kui lisada valemile (10) liidetav $\beta_4 h(x,\xi)$, siis jääb polünoom pidevaks, nagu jäävad pidevaks ka esimene ja teine tuletis. Seega kui soovime kuupsplaini sobitada andmetele, milles on valitud K sõlme, rakendame vähimruutude meetodit hindamaks vabaliiget ja $K + 3$ parameetrit, kasutades tunnustena x^1, x^2, x^3, $h(x, \xi_1)$, $h(x, \xi_2)$,…, $h(x,\xi_K)$, kus ξ_1,…, ξ_K on sõlmed. Meetodi tulemusena saame $K + 4$ ilma kitsendusteta leitavat regressioonikordajate hinnangut, millest omakorda järeldub, et K sõlmega kuupsplaini vabadusastmete arv on $K + 4$.



Joonis 4. Splainid keskmise palga iseloomustamiseks kasutades osakogumit varasemast andmestikust. Mustaga on tähistatud kuupsplain ja punasega on tähistatud naturaalkuupsplain.

Joonisel 4 on näha selliste splainide põhiline probleem: prognoosijoone hajuvus on tunnuse otspunktides võrdlemisi suur. Selle mudeli puhul on kasutatud kolme vanuse sõlme – 25, 40 ja 60 – ja nagu näha, siis on hajuvus eriti suur üle 60-ste osas.

Naturaalkuupsplainid on samuti regressioonisplainid, aga lisakitsendusena peavad olema esimeses ja viimases lõigus funktsioonid lineaarsed. Üldjuhul nende kahe kuupsplaini võrdluses saavutab naturaalkuupsplain parema hinnangu otsmistes piirkondades, mida kinnitab ka väiksem usaldusvahemik.

2.4.4. Sõlmede arv ja väärtused

On selge, et mida rohkem on sõlmesid, seda paindlikum tuleb regressioonmudel, kuna regressioonikordajad muutuvad tihedamini. Seetõttu kui meil on ettekujutus andmestikust, on üks võimalus paigutada rohkem sõlmesid piirkonda, kus muutused on suuremad ja selgemad, vähem sõlmi piirkondadesse, kus funktsioon on stabiilsem. Praktikas on aga enimkasutatavaks mooduseks siiski ühtlaselt jaotuvad sõlmed. Sõlmede valikuks pole üht kindlat parimat viisi ja seetõttu on ka lähenemisi mitmeid. Üks variantidest on katse eksitus meetodil, valides iga kord erinev arv sõlmi ja seejärel visuaalselt otsustades, milline graafik annab parima tulemuse. On selge, et see ei pruugi olla kõige otstarbekam, kui andmestik on väga suur ja arvutamine võtab liialt aega. Konkreetsema tulemuse saame ristvalideerimisega. Esimese sammuna eemaldame mingi osa andmetest, näiteks 10%. Seejärel sobitame splaini allesjäänud andmetele valides mingi arvu sõlmi. Selle splaini abil teeme järeldusi väljajäetud andmete kohta. Seda protsessi tuleb korrata mitu korda, kuni kõik andmestiku punktid on üks kord välja jäetud ja seejärel arvutame jääkide ruutude summa. Protsessi tasub korrata erinevate sõlmede arvuga ja kõigi puhul arvutada jääkide ruutude summa. Parimaks sõlmede arvuks sobib see, mille puhul jääkide ruutude summa on väikseim. (James jt., 2013, lk 274-276)

2.5. Siluvad splainid

Sarnaselt eelmise meetodiga moodustab ka järgnev meetod splaine, kuid siin on kõige olulisemaks aspektiks graafiku sujuvus. Kasutame materjali James jt., 2013, lk 277-278. Otsitakse sellist funktsiooni $g(x)$, mis sobiks hästi andmetega, see tähendab, et tal oleks väike jääkide ruutude summa. Samas ei saa $g(x)$ piiranguteta valida, sest vastasel juhul valiksime $g(x)$ selliselt, et see läbib kõiki y_i väärtusi, aga ei anna meile üldistavat infot. Seega põhilised eesmärgid on splaini sujuvus ning väike jääkide ruutude summa.

Üks põhilisemaid lähenemisi on leida funktsioon g selliselt, et see minimeeriks avaldist

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt, \qquad (14)$$

kus λ on mittenegatiivne seadistusparameeter. Lahendina leitud funktsiooni g nimetatakse siluvaks splainiks.

Valemi (14) esimene summa on vähimruutude summa ning teine liidetav on karistusliige. Kirjutis $g''(t)$ tähistab funktsiooni g teist tuletist kohal *t*. Teise tuletise interpretatsioon on funktsiooni tõusu muut. Teine tuletis on absoluutväärtuselt suur, kui punkti *t* ümber on funktsioon käänuline, teine tuletis on nullilähedane, siis funktsioon on sujuv. Tähelepanekuna tasub välja tuua näide sirgest, mis on maksimaalselt sujuv ja mille teine tuletis on 0. Integraal on oluline selleks, et saaks summeerida üle kogu *t*. Kokkuvõtvalt, kui g on sujuv, siis $g'(t)$ on lähedane konstandile ja $g''(t)$ omab väikest väärtust ja seega $\lambda \int g''(t)^2 dt$ kontrollib funktsiooni g sujuvust. Mida suurem on λ, seda sujuvam tuleb ka g. Kui λ = 0, siis optimiseerimisel saavutame täpse interpoleerimise läbi andmestiku punktide, aga kui $\lambda \to \infty$, siis on g lihtsalt sirge, mis üritab kõikidele punktidele võimalikult lähedal olla. Tulemuseks on vähimruutude meetodil leitud lineaarne regressioon. (James jt., 2013, lk 277-278)

2.6. Lokaalne regressioon

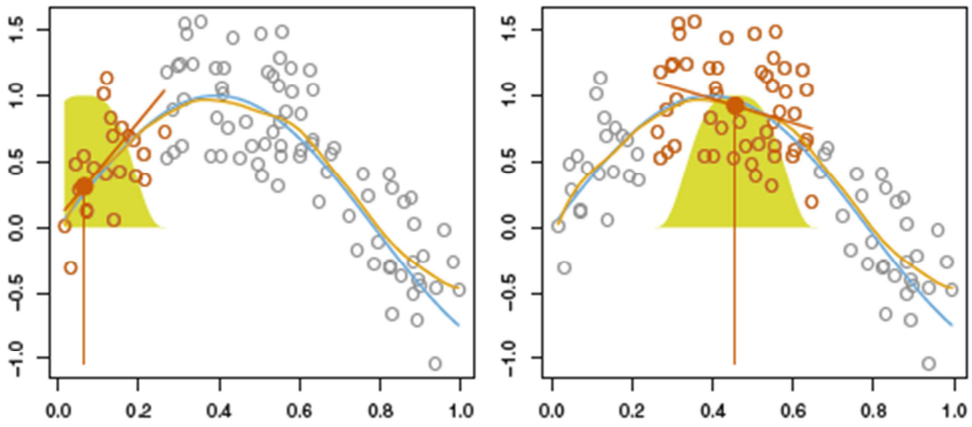
Viimase mittelineaarse mudelina käsitleme lokaalset regressiooni. Kasutame materjali James jt., 2013, lk 280-282. Teistega võrreldes on lokaalne regressioon erinev selle poolest, et genereerides hinnangut, kasutame me ainult osa uurimise all olevate tunnuste punktidest. Kõigepealt kirjeldame algoritmi lokaalse regressiooni leidmiseks, seejärel kirjeldame komponente täpsemalt.

1) Fikseerime lokaalse regressiooni koha $X = x_0$ jaoks.
2) Valime mingi arvu k punkte x_0 ümbruses. Arvu k abil saame defineerida ka lokaalset regressiooni iseloomustava suuruse $s = k/n$, kus n on kõigi vaatluste arv.
3) Edasi anname punktidele kaalud $K_{i0} = K(x_i, x_0)$. Ainsad kaalud, mis on suuremad kui 0, on need, mis satuvad valitud k punktiga määratud ümbrusesse.
4) Järgmisena arvutame kaalutud vähimruutude meetodi abil uuritava tunnuse jaoks kordajad $\hat{\beta}_0$ ja $\hat{\beta}_1$ minimiseerides järgmist valemit:

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \qquad (15)$$

5) Prognoosi punktis x_0 saame kui $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

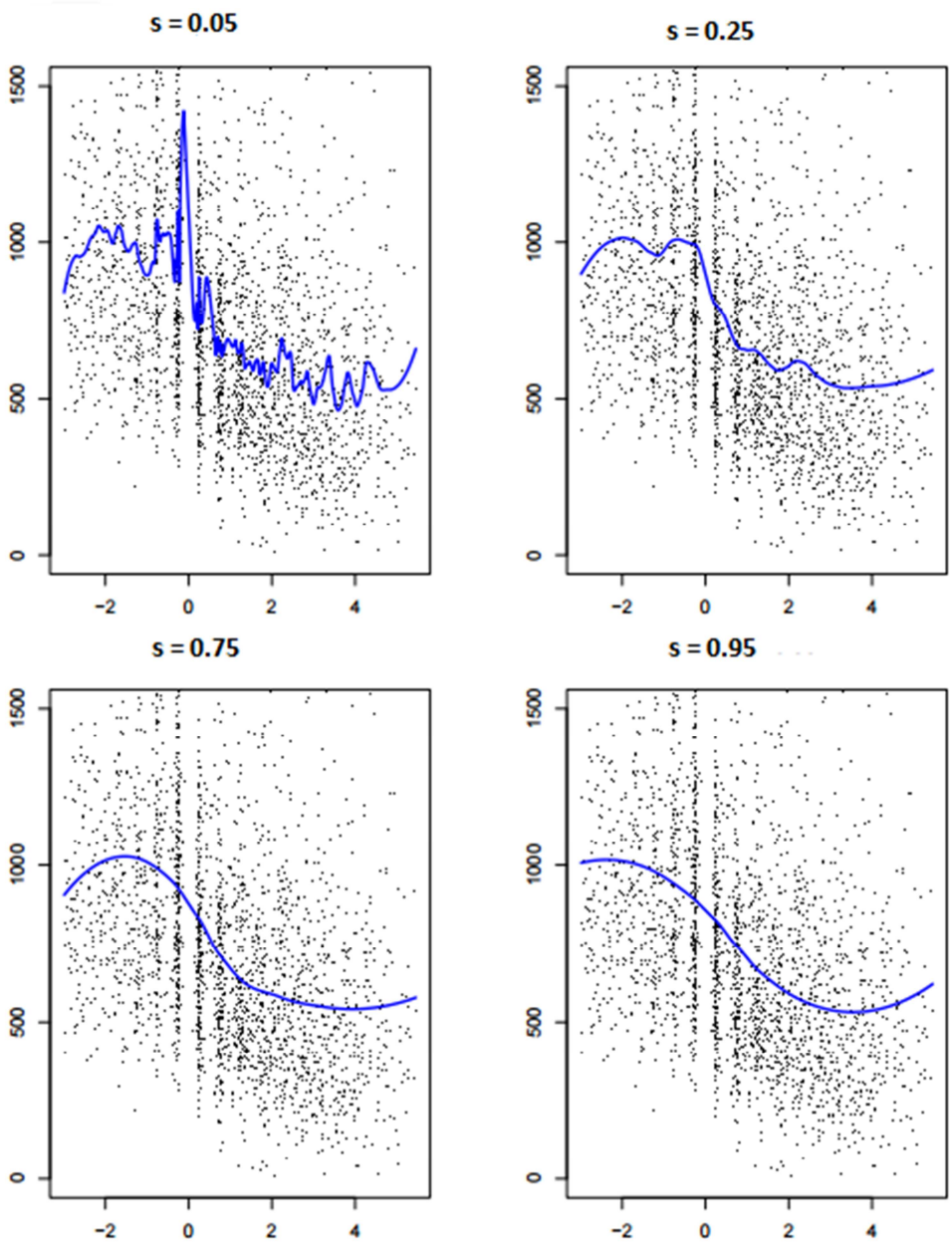
Kaalud K_{i0} sõltuvad alguspunkti x_0 valikust, seega iga uue x_0 korral peame hindama uued kaalud ning uued kordajad $\hat{\beta}_0$ ja $\hat{\beta}_1$. Lokaalse regressiooni tulemust mõjutavad mitmed tegurid, näiteks kaalufunktsiooni K valik või regressioonimudeli valik vastavas punktis (kirjeldatud algoritmis on käsitletud lineaarset regressiooni). Väga olulise tegur on eespool defineeritud kordaja s, mida nimetatakse ka ulatuseks. Lokaalse regressiooni olulisemaid tegureid aitab mõista järgnev Joonis 5:



Joonis 5. Lokaalne regressioon erineva x_0 ümbruses. Joonistel olev sinine joon tähistab $\boldsymbol{f}(\boldsymbol{x})$, millest on punktid genereeritud, oranž joon tähistab selle hinnangut $\hat{\boldsymbol{f}}(\boldsymbol{x})$ lokaalse regressiooniga. Oranž täpp on siin x_0. Sellest lähtuvalt on k oranži seest tühja punkti, mille kaal on > 0.

Normaaljaotust meenutav oranž kelluke mõlemal graafikul kujutab kaalude väärtusi nii, et mida lähemal on punkt algpunktile x_0, seda suurem on ka tema kaal.

Ulatus *s* määrab arvutustesse kaasatava punktide osakaalu. Mida suurem on *s*, seda sujuvam on regressioonijoon. Joonisel 6 on visualiseeritud ulatuse *s* mõju lokaalse regressiooni joone kujule. Ülevalt vasakult alustades on *s* väärtusteks 0.05, 0.25, 0.75 ja 0.95. (Irizarry, 2001)



Joonis 6. Rakkude CD4 arvu sõltuvus serokonversiooni ajast lokaalse regressiooniga. (Irizarry, 2001)

3. Praktilised näited

Näited viime läbi 100m sprindi mõõtmiste peal, mis on tehtud Berliini kergejõustiku MM-l. Andmed on võetud Rahvusvahelise Kergejõstikuliidu kodulehelt, mis on kokku pandud Saksamaa Kergejõustikuliidu poolt (German Athletics Federation, 2009). Valitud on Berliini MM just sellepärast,et seal on joostud tänini kehtiv maailmarekord Usain Bolti poolt – 9.58 sek. Andmestik koosneb jooksja nimest (kui nime ees on SF, tähistab see poolfinaali jooksu, F tähistab finaali jooksu). Vahemikke, mille läbimise aega jooksu kestel on mõõdetud, on viis:

- 0 kuni 20m,
- 20m kuni 40m,
- 40m kuni 60m,
- 60m kuni 80m,
- 80m kuni 100m.

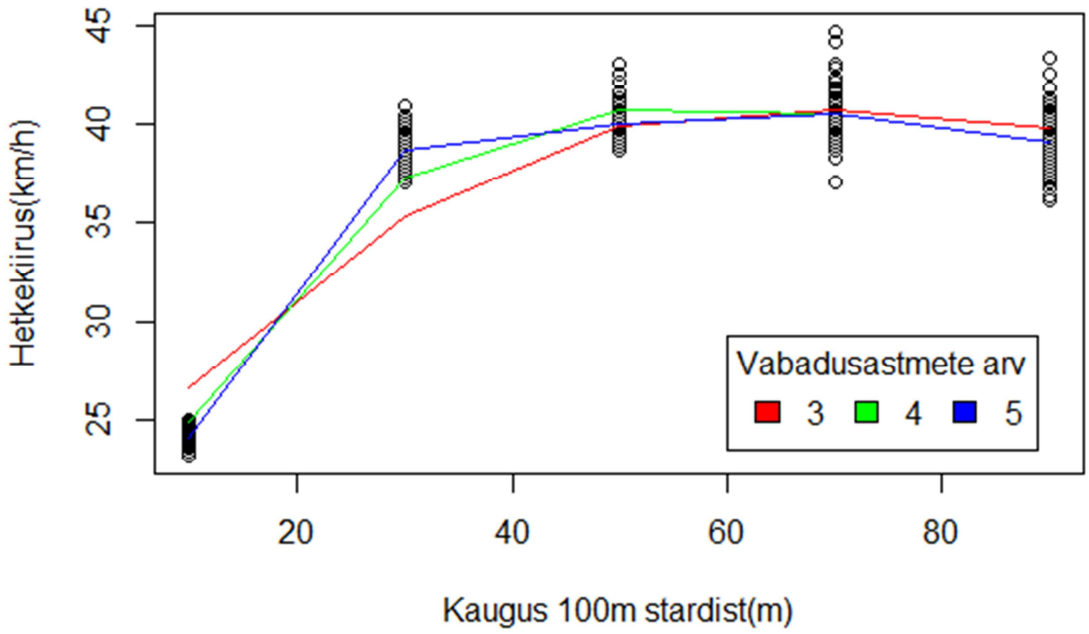
Nende kõigi tunnuste ühikuks on sekund. Andmestikku on täiendatud uute tunnustega kasutades olemasolevaid vastavalt ülesande püstitusele, mida kirjeldatakse täpsemalt vastava näite juures.

Üldine põhimõte ütleb, et esimeses pooles on sportlastel tugev kiirendus, millele järgneb tippkiirus vahemikus 50-70m sõltuvalt sportlasest ja edasi peaks toimuma kiiruse langus. Meie eesmärgiks on näidata taolist kiiruse käitumist 100m sprindi jooksul. Esimeses näites kasutame siluvaid splaine, et saavutada regressioonikõver. Selleks kasutame vabadusastmete arvudena nii 3, 4 kui ka 5. Viiest edasi on kasutu suurendada vabadusastmete arvu, sest lõpuks kasutab algoritm ikkagi viit arvu, kuna rohkem erinevaid väärtusi argumenttunnus x ei oma. Probleem on otseselt mõõtmistega seotud, sest meil ei ole distantsi tunnus pidevana, vaid diskreetsena iga 20m tagant, mistõttu on joonisel olevad punktid automaatselt gruppides. Kuna tunnused on vahemikena, siis määramaks vastava vahemiku keskmist hetkekiirust on sobilik paigutada punktid vahemike keskele, st. vahemikus 0 kuni 20m võtame keskpunktiks 10m jne.

Arvutustes on kasutatud argumendina x läbitud vahemike keskpunkte. Funktsioontunnusena y on uurimise all hetkekiirus, mis on arvutatud vastava vahemiku läbimise jaoks kulunud aja põhjal:

$$\text{kiirus} = \frac{3.6*20}{aeg(20m)} \quad (16)$$

Valemis (16) tähistab 3.6 teisendust, mis on saadud meetrite teisendamisel kilomeetriteks ning sekundid tunniks.



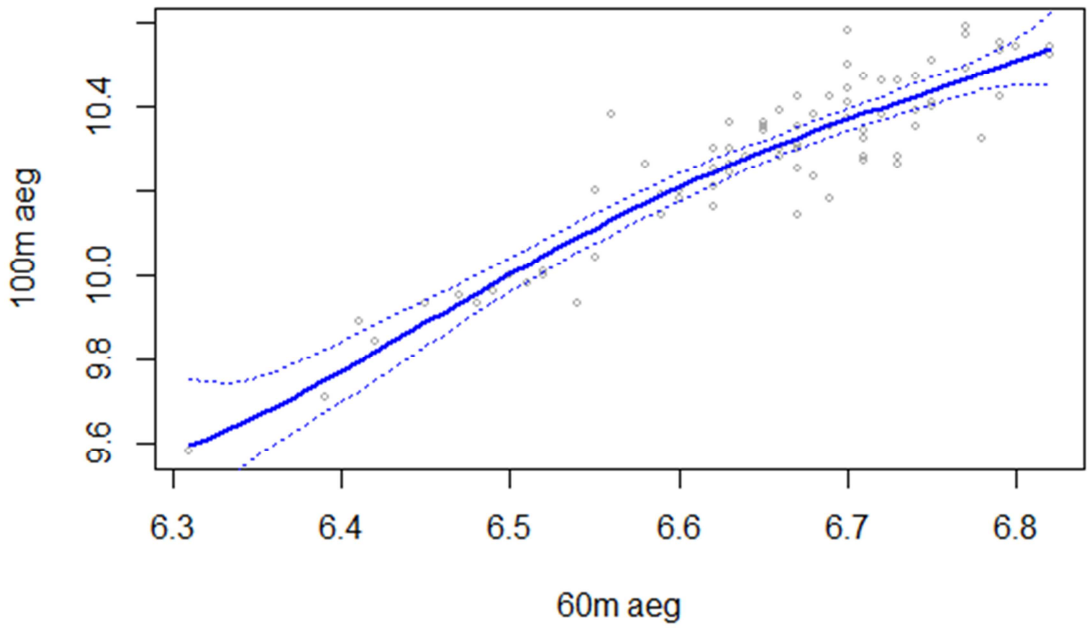
Joonis 7. Sõltuvus läbitud distantsi ja kiiruse vahel.

On näha, et maailma parimate sprinterite puhul ulatub tippkiirus pigem 70m juurde, pärast mida algab ootuspärane langus kiiruses. Kuna vabadusastmeid ei saa palju olla, on üsna mugav kasutada ära maksimaalne arv, milleks on 5, sest see annab kõige täpsema visuaalse tulemuse.

Võrlduseks on toodud välja ka kõigi kolme funktsiooni λ väärtused:

- vabadusastme 3 puhul λ = 0.5835,
- vabadusastme 4 puhul λ =0.0775,
- vabadusastme 5 puhul λ = 0.000146.

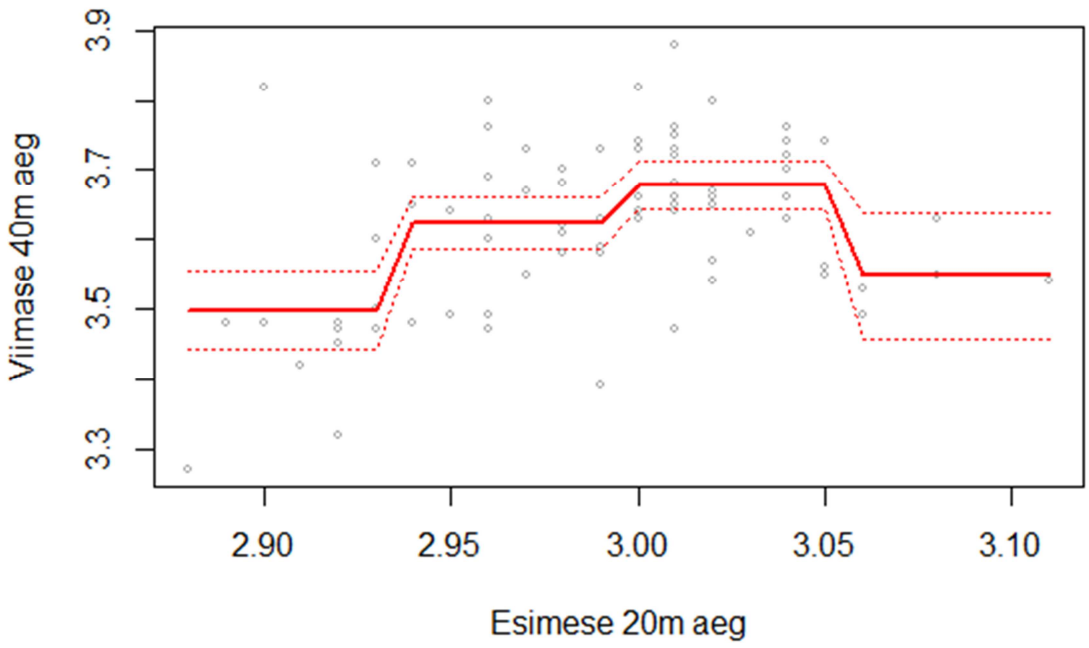
Järgmine näide on näitamaks polünoomregressiooni kasutust. Selle näite jaoks on arvutatud uued tunnused “100m_aeg” ja “60m_aeg”, mis on saadud liites kõik vahemike mõõtmised kokku 100m aja jaoks ning esimesed 3 mõõtmist 60m aja jaoks. Oodatav tulemus on siin paraku vägagi lineaarsuse lähedane, sest mida kiirem on 60m aeg, seda kiirem on üldiselt ka 100m aeg. Tulemust visualiseerib Joonis 8.



Joonis 8. 100m jooksu aja sõltuvus 60m ajast. Kasutatud on 4. astme polünoomi.

Nagu jooniselt näha, siis toimub väike tõusu langus graafiku paremal üleval osas. Seda võib tõlgedada nii, et olenemata kui hea või halb on esimesed 60m, siis jooksu teises pooles tase võrdsustub. Kehvem esimene jooksupool pole midagi ootamatut, sest põhiline ebaõnnestumine, mida nii lühikesel distantsil saab juhtuda, on kehv start. Kehva stardi tulemuseks on kehv 60m aeg, aga seda rohkem parandatakse ennast jooksu teises pooles, mis seletab graafiku tõusu langust.

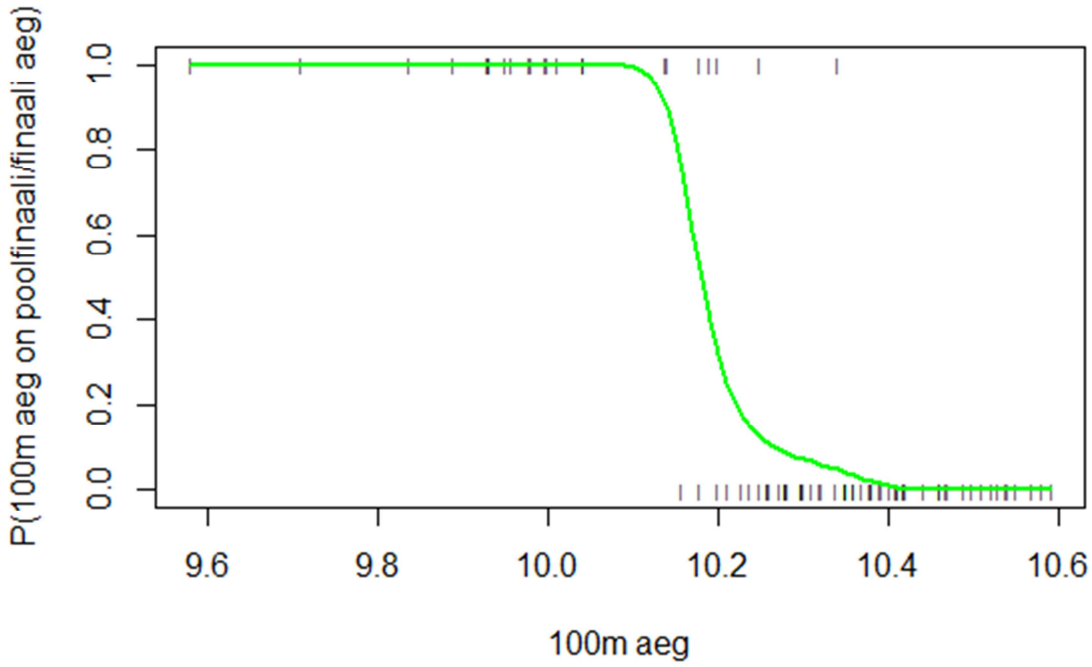
Uurime saadud tulemusi natuke uue nurga alt. Tihtipeale on kehva stardiga inimesed jooksu teises pooles tunduvalt paremad, samas kui loomulik on ka see, et hea start ei välista ka head lõppu. Parim näide on selleks Usain Bolti rekordjooks, kus ta on parim nii stardis kui ka lõpukiiruses. Järgmise joonise jaoks on arvutatud uus tunnus “viimase_40m_aeg”, mida on kõige lihtsam leida juba leitud tunnuste “100m_aeg” ja “60m_aeg” lahutamisel. Uurime siin, kuidas seob treppfunktsioon esimesed 20m ja viimased 40m. Joonisel 9 tuleb juba selgemalt välja eespool mainitud jooksu loogika. Maailma tipud on selgelt kõige esimesel osal, sest suudavad teha nii hea stardi kui ka hea lõpu. Edasi toimub loogiline jätk lineaarses mõttes ehk mida kehvem startija on, siis on tõenäoline, et ta on ka jooksu teises pooles kehvem. Mis puutub aga viimasesse osasse, siis siin eristuvad sportlased, kes pole just parima stardiga, kuid selle tõttu on nende trumbiks jooksu teine pool, millest annab aimu ka meie treppfunktsioon. Tegemist on ootuspäraselt mittelineaarse regressiooniga, kasutatud on treppfunktsiooni 4 piirkonnaga.



Joonis 9. Viimase 40m aja sõltuvus esimese 20m ajast.

Viimase näitena võrdleme eeljooksude aegu poolfinaali ja finaali omadega. Nimetame poolfinaali ja finaali jookse kõrgema taseme jooksudeks. Eesmärk on kinnitada oletust, et teatud piirist alates on üsna selge, millised ajad on joostud kõrgema tasemega jooksudes. Modelleerime tõenäosust, et tegemist on kõrgema taseme jooksu ajaga. Selleks teeme tabelisse Lisa 3 uue binaarse tunnuse “sai_edasi”, kus eeljooksu ajad saavad väärtuseks 0 ja poolfinaali ning finaali ajad saavad väärtuseks 1. Graafiku koostamiseks on kasutatud logistilist 4. astme polünoomregressiooni mudelit.

Joonisel 10 oleva regressiooni puhul on otspunktidel kõige lihtsam interpretatsioon – kui aeg on aeglasem kui 10.4 sekundit, võib olla kindel, et tegemist on eeljooksu ajaga, kui aeg on kiirem kui 10.1, võib olla üsna kindel, et tegemist on kas poolfinaalis või finaalis joostud ajaga. Väike anomaalia on 10.3 sekundi juures, mis tuleneb sellest, et finaalis kas jooksjal juhtus midagi või lihtsalt polnud jõudu enam jäänud viimaseks jooksuks. Kuna üks 10.3 sekundi jooks on tehtud finaalis, siis jätab see ka väikse märke hinnangule.



Joonis 10. Viimane joonis illustreerib 100m aja headust selles kontekstis, et kas tegemist on eeljooksus joostud 100m ajaga või juba hilisemas voorus (poolfinaal, finaal). Üsna tugev üleminek on tõenäosusel vahemikus 10.1-10.2 sekundit.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks oli anda ülevaade erinevatest mittelineaarsetest regressioonmudelitest ning teha praktilised näited spordiandmetel. Ülevaateks kasutati nii teooriat kui ka illustreerivaid graafikuid.

Töö esimeses osas anti kasutatavate regressioonmudelite kohta lühikirjeldus. Teises osas uuriti iga meetodit juba täpsemalt, seletati illustreerivate joonistega ning vajadusel reprodutseeriti varasemaid tulemusi. Olenevalt ülesande püstitusest leidub erinevaid mittelineaarseid mudeleid, mida rakendada valitud tunnustel, mis sobivad kõige paremini.

Kolmandas osas viidi läbi kergejõustiku Berliini maailmameistrivõistluste 100m jooksu mõõtmistulemuste peal erinevaid meetodeid eelmises peatükis tutvustatud mudelitega. Kõige rohkem pakkus huvi kiiruse muutus 100m läbimise jooksul, mille puhul prognoos ütles, et maksimumkiirus saavutatakse umbes jooksu 70m juures. Samuti uuriti jooksu esimese ja viimase osa aegade sõltuvust teineteisest ning võrreldi paremaid ja halvemaid 100m jooksuaegu. Teatavat mittelineaarset seost võis täheldada esimese 20m ja viimase 40m jooksu aja vahel. Tulemused olid ootuspärased ja loogikaga kooskõlas.

Kasutatud kirjandus

1. G. James, D. Witten, T. Hastie, R. Tibshirani.(2013) *An Introduction to Statistical Learning with Applications in R*, New York.
   Kättesaadav: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf (08.05.2017)
2. German Athletics Federation.(2009) *Scientific Research Project. Biomechanical Analysis at the 12th IAAF World Championships in Athletics*, (Berlin).
   Kättesaadav: https://www.iaaf.org/download/download?filename=76ade5f9-75a0-4fda-b9bf-1b30be6f60d2.pdf&urlSlug=1-biomechanics-report-wc-berlin-2009-sprint (13.10.2016)
3. Irizarry, R. A.(2001). *Applied Nonparametric and Modern Statistics*, konspekt, biostatistika osakond, Johns Hopkins University.
   Kättesaadav: http://rafalab.github.io/pages/754/section-03.pdf (05.05.2017)

Lisa 1

Lisas 1 on toodud reprodutseerimise käigus kasutatud R-i koode.[1]

Polünoomregressiooni reproduktsioon

```
library(ISLR)
attach(wage)
fit=lm(wage~poly(age ,4) ,data=Wage)
agelims =range(age)
age.grid=seq (from=agelims [1], to=agelims [2])
preds=predict (fit ,newdata =list(age=age.grid),se=TRUE)
se.bands=cbind(preds$fit +2* preds$se.fit ,preds$fit -2* preds$se.fit)
par(mfrow =c(1,2) ,mar=c(4.5 ,4.5 ,1 ,1) ,oma=c(0,0,4,0))
plot(age ,wage ,xlim=agelims, xlab="Vanus", ylab="Palk" ,cex =.5, col =" darkgrey ")
title ("Neljanda astme polünoom ",outer =T)
lines(age.grid ,preds$fit ,lwd =2, col =" red")
matlines (age.grid ,se.bands ,lwd =1, col =" red",lty =3)
```

Logistilise polünoomregressiooni reproduktsioon

```
library(ISLR)
attach(wage)
fit=glm(I(wage >250)~poly(age ,4) ,data=Wage ,family =binomial )
preds=predict (fit ,newdata =list(age=age.grid),se=T)
pfit=exp(preds$fit )/(1+ exp( preds$fit ))
se.bands.logit = cbind(preds$fit +2*preds$se.fit ,preds$fit-2*preds$se.fit)
se.bands = exp(se.bands.logit)/(1+ exp(se.bands.logit))
plot(age ,I(wage >250),xlab="Vanus", ylab="Palk > 250000$" ,xlim=agelims ,type ="n",ylim=c(0 ,.2)  )
points (jitter (age), I((wage >250) /5) ,cex =.5, pch ="|",col =" darkgrey ")
lines(age.grid ,pfit ,lwd =2, col =" green")
matlines (age.grid ,se.bands ,lwd =1, col =" green",lty =3)
```

Trepifunktsiooni reproduktsioon

```
library(ISLR)
attach(wage)
table(cut (age ,4))
fit=lm(wage~cut (age ,4) ,data=Wage)
agelims =range(age)
age.grid=seq (from=agelims [1], to=agelims [2])
```

```
preds=predict (fit ,newdata =list(age=age.grid),se=TRUE)
se.bands=cbind(preds$fit +2* preds$se.fit ,preds$fit -2* preds$se.fit)
par(mfrow =c(1,2) ,mar=c(4.5 ,4.5 ,1 ,1) ,oma=c(0,0,4,0))
plot(age ,wage ,xlim=agelims, xlab="Vanus", ylab="Palk" ,cex =.5, col =" darkgrey ")
title ("Neljanda astme polünoom ",outer =T)
lines(age.grid ,preds$fit ,lwd =2, col =" red")
matlines (age.grid ,se.bands ,lwd =1, col =" red",lty =3)
```

Logistilise trepifunktsiooni reproduktsioon

```
fit=glm(I(wage >150)~cut(age ,4) ,data=Wage ,family =binomial )
preds=predict (fit ,newdata =list(age=age.grid),se=T)
pfit=exp(preds$fit )/(1+ exp( preds$fit ))
se.bands.logit = cbind(preds$fit +2*preds$se.fit ,preds$fit-2*preds$se.fit)
se.bands = exp(se.bands.logit)/(1+ exp(se.bands.logit))
plot(age ,I(wage >150),xlab="Vanus", ylab="Palk > 150000$" ,xlim=agelims ,type ="n",ylim=c(0 ,.2)  )
points (jitter (age), I((wage >150) /5) ,cex =.5, pch ="|",col =" darkgrey ")
lines(age.grid ,pfit ,lwd =2, col =" green")
matlines (age.grid ,se.bands ,lwd =1, col =" green",lty =3)
```

Kuup- ja naturaalkuupsplainide reproduktsioon osakogumi n=500 peal.

```
library (splines )
uuswage=Wage[sample(nrow(Wage), 500), ]
attach(uuswage)
fit=lm(wage~bs(age ,knots =c(25 ,40 ,60) ),data=uuswage)
pred=predict (fit ,newdata =list(age =age.grid),se=T)
plot(age ,wage ,xlab="Vanus", ylab="Palk",col ="gray")
title(main="Kuupsplainid")
lines(age.grid ,pred$fit ,lwd =2)
lines(age.grid ,pred$fit +2* pred$se ,lty ="dashed")
lines(age.grid ,pred$fit -2* pred$se ,lty ="dashed")
fit2=lm(wage~ns(age ,df =4) ,data=uuswage)
pred2=predict (fit2 ,newdata =list(age=age.grid),se=T)
lines(age.grid , pred2$fit ,col ="red",lwd =2)
lines(age.grid ,pred2$fit +2* pred2$se ,col="red", lty ="dashed")
lines(age.grid ,pred2$fit -2* pred2$se ,col="red", lty ="dashed")
legend ("topright",legend =c("Kuupsplain" ,"Naturaalkuupsplain") ,
       col=c("black","red"),lty =1, lwd =2, cex =.8)
```

Lisa 2

Lisas 2 on autori poolt uuritud spordiandmete analüüsiks vajaminevad R-i koodid.

Hetkekiiruse sõltuvus läbitud teepikkusest 100m jooksu ajal kasutades hindamiseks siluvaid splaine.

```
library(dplyr)
andmed %>% mutate_if(is.factor, as.character) -> andmed
test1 <- cbind(andmed$Nimi,as.numeric(rep(10,81)),3.6*20/as.numeric(andmed$X0.20m))
test2 <- cbind(andmed$Nimi,as.numeric(rep(30,81)),3.6*20/as.numeric(andmed$X20.40m))
test3 <- cbind(andmed$Nimi,as.numeric(rep(50,81)),3.6*20/as.numeric(andmed$X40.60m))
test4 <- cbind(andmed$Nimi,as.numeric(rep(70,81)),3.6*20/as.numeric(andmed$X60.80m))
test5 <- cbind(andmed$Nimi,as.numeric(rep(90,81)),3.6*20/as.numeric(andmed$X80.100m))
tulemus <- data.frame(rbind(test1,test2, test3, test4, test5))
x <- as.numeric(levels(tulemus$X2))[tulemus$X2]
y <- as.numeric(levels(tulemus$X3))[tulemus$X3]

plot(x, y,xlab="Kaugus 100m stardist(m)", ylab="Hetkekiirus(km/h)")
fit=smooth.spline(x, y,df=3)
lines(fit,col="red")
fit1=smooth.spline(x, y,df=4)
lines(fit1,col="green")
fit2=smooth.spline(x, y,df=5)
lines(fit2,col="blue")
legend("bottomright", inset=.05, title="Vabadusastmete arv",
    c("3","4","5"), fill=c("red","green", "blue"), horiz=TRUE)
```

60m läbimiseks läinud aja sõltuvus 100m läbimiseks läinud ajast.

```
fit=lm(X100m_aeg~poly(X60m_aeg ,4) ,data=andmed)
kuuslims =range(X60m_aeg)
kuus.grid=seq(from=kuuslims[1], to=kuuslims[2],by=0.01)
preds=predict (fit ,newdata =list(X60m_aeg=kuus.grid),se=TRUE)
se.bands=cbind(preds$fit +2* preds$se.fit ,preds$fit -2* preds$se.fit)
plot(X60m_aeg ,X100m_aeg ,xlim=kuuslims, xlab="60m aeg", ylab="100m aeg" ,cex=.5, col =" darkgrey ")
title ("Neljanda astme polünoom ",outer =T)
lines(kuus.grid ,preds$fit ,lwd =2, col ="blue")
matlines (kuus.grid ,se.bands ,lwd =1, col ="blue",lty =3)
```

Sõltuvus esimese 20m ja viimase 40m läbimiskiiruse vahel.

```
fit=lm(viimased_40m~cut (X0.20m,4) ,data=andmed)
nulllims =range(X0.20m)
null.grid=seq (from=nulllims[1], to=nulllims[2],by=0.01)
preds=predict (fit ,newdata =list(X0.20m=null.grid),se=TRUE)
se.bands=cbind(preds$fit +2* preds$se.fit ,preds$fit -2* preds$se.fit)
plot(X0.20m ,viimased_40m ,xlim=nulllims, xlab="Esimese 20m aeg", ylab="Viimase 40m aeg" ,cex =.5, col
=" darkgrey ")
title ("Trepifunktsioon: Esimese 20m ja viimase 40m aeg",outer =T)
lines(null.grid ,preds$fit ,lwd =2, col =" red")
matlines (null.grid ,se.bands ,lwd =1, col =" red",lty =3)
```

Poolfinaali ja finaali jooksude võrdlemine eeljooksu aegadega kasutades binaarset tunnust ja logistilist polünoomregressiooni.

```
for (i in 1:58){andmed$sai_edasi[i]=0}
for (i in 59:81){andmed$sai_edasi[i]=1}
attach(andmed)
sadalims=range(X100m_aeg)
fit=glm(sai_edasi~poly(X100m_aeg ,4) ,data=andmed ,family =binomial )
preds=predict (fit ,newdata =list(X100m_aeg=sada.grid),se=T)
pfit=exp(preds$fit )/(1+ exp( preds$fit ))
se.bands.logit = cbind(preds$fit +2*preds$se.fit ,preds$fit-2*preds$se.fit)
se.bands = exp(se.bands.logit)/(1+ exp(se.bands.logit))
plot(X100m_aeg ,sai_edasi,xlab="100m aeg", ylab="Poolfinaali/Finaali ajad" ,xlim=sadalims ,type ="n")
points(jitter(X100m_aeg), sai_edasi ,cex =.5, pch ="|",col ="black")
lines(sada.grid ,pfit ,lwd =2, col =" green")
matlines (sada.grid ,se.bands ,lwd =1, col =" green",lty =3)
```

Lisa 3

Spordiandmete analüüsiks kasutatud andmestik tabelikujul, kuhu on lisatud ka juurde arvutatud tunnused.

Nimi	0-20m	20-40m	40-60m	60-80m	80-100m	60m_aeg	100m_aeg	viimase_40m_aeg	sai_edasi
Chambers	2.99	1.84	1.77	1.75	1.83	6.60	10.18	3.58	0
Francis	2.99	1.84	1.79	1.77	1.82	6.62	10.21	3.59	0
Lemaitre	3.05	1.85	1.78	1.76	1.79	6.68	10.23	3.55	0
Callander	2.98	1.86	1.79	1.77	1.84	6.63	10.24	3.61	0
Rodgers	2.96	1.85	1.81	1.78	1.85	6.62	10.25	3.63	0
Patton	2.98	1.85	1.75	1.77	1.91	6.58	10.26	3.68	0
Bailey	2.99	1.84	1.80	1.79	1.84	6.63	10.26	3.63	0
Martina	3.06	1.88	1.79	1.74	1.79	6.73	10.26	3.53	0
Grueso	3.05	1.87	1.79	1.75	1.81	6.71	10.27	3.56	0
Mbandjock	3.08	1.86	1.79	1.75	1.80	6.73	10.28	3.55	0
Palacios	3.02	1.88	1.81	1.77	1.80	6.71	10.28	3.57	0
Collins	2.98	1.87	1.81	1.78	1.84	6.66	10.28	3.62	0
Tsukahara	2.95	1.87	1.82	1.80	1.84	6.64	10.28	3.64	0
Frater	2.98	1.86	1.78	1.78	1.90	6.62	10.30	3.68	0
Hinds	3.00	1.87	1.80	1.79	1.84	6.67	10.30	3.63	0
Phiri	2.97	1.86	1.80	1.79	1.88	6.63	10.30	3.67	0
Ogho-Oghene	3.00	1.85	1.81	1.79	1.85	6.66	10.30	3.64	0
Fasuba	3.01	1.86	1.80	1.78	1.86	6.67	10.31	3.64	0
Edwards	3.11	1.89	1.78	1.75	1.79	6.78	10.32	3.54	0
Williamson	3.04	1.87	1.80	1.79	1.84	6.71	10.34	3.63	0
Gregorio	3.01	1.85	1.83	1.80	1.86	6.69	10.35	3.66	0
Ndure	3.01	1.86	1.80	1.79	1.89	6.67	10.35	3.68	0
Pognon	3.03	1.86	1.82	1.78	1.83	6.71	10.32	3.61	0
Keller	3.03	1.89	1.82	1.78	1.83	6.74	10.35	3.61	0
Harris	2.98	1.85	1.82	1.81	1.89	6.65	10.35	3.70	0
Thompson	2.97	1.86	1.80	1.79	1.94	6.63	10.36	3.73	0
Cerutti	2.94	1.88	1.83	1.82	1.89	6.65	10.36	3.71	0
Griffith	3.02	1.86	1.82	1.80	1.87	6.70	10.37	3.67	0
Eriguchi	2.98	1.88	1.82	1.82	1.88	6.68	10.38	3.70	0
Metu	3.00	1.89	1.83	1.80	1.86	6.72	10.38	3.66	0
Burns	2.99	1.86	1.81	1.79	1.94	6.66	10.39	3.73	0
Rodriguez	3.01	1.89	1.84	1.80	1.85	6.74	10.39	3.65	0
Ouhadi	3.02	1.90	1.83	1.81	1.84	6.75	10.40	3.65	0
Abrantes	3.02	1.89	1.84	1.80	1.86	6.75	10.41	3.66	0
Al-Harthi	2.93	1.94	1.83	1.82	1.89	6.70	10.41	3.71	0
Meite	3.04	1.88	1.83	1.80	1.86	6.75	10.41	3.66	0
Unger	2.97	1.88	1.84	1.83	1.90	6.69	10.42	3.73	0

Barnett	3.08	1.89	1.82	1.79	1.84	6.79	10.42	3.63	0
Edgar	3.01	1.87	1.79	1.76	1.99	6.67	10.42	3.75	0
Atkins	3.04	1.88	1.78	1.86	1.88	6.70	10.44	3.74	0
Durant	3.01	1.88	1.84	1.83	1.90	6.73	10.46	3.73	0
Kuc	3.00	1.89	1.83	1.84	1.90	6.72	10.46	3.74	0
Kimura	2.96	1.89	1.86	1.85	1.91	6.71	10.47	3.76	0
Gittens	3.00	1.89	1.85	1.84	1.89	6.74	10.47	3.73	0
Collio	3.01	1.92	1.84	1.83	1.89	6.77	10.49	3.72	0
Schwab	2.96	1.88	1.86	1.86	1.94	6.70	10.50	3.80	0
Nabe	3.01	1.89	1.85	1.86	1.90	6.75	10.51	3.76	0
Osovnikar	3.04	1.93	1.85	1.82	1.88	6.82	10.52	3.70	0
Abeypitiyage	3.05	1.89	1.85	1.84	1.90	6.79	10.53	3.74	0
Moraes	3.04	1.92	1.86	1.84	1.88	6.82	10.54	3.72	0
Magakwe	3.05	1.90	1.85	1.84	1.90	6.80	10.54	3.74	0
Moreira	3.04	1.90	1.85	1.83	1.93	6.79	10.55	3.76	0
Zakari	3.02	1.90	1.85	1.82	1.98	6.77	10.57	3.80	0
Moseley	3.01	1.91	1.78	1.94	1.94	6.70	10.58	3.88	0
Hyman	3.00	1.91	1.86	1.86	1.96	6.77	10.59	3.82	0
Gay	3.02	1.83	1.77	1.73	1.81	6.62	10.16	3.54	0
Powell	2.90	1.82	1.84	1.88	1.94	6.56	10.38	3.82	0
Bolt	2.94	1.83	1.78	1.78	1.87	6.55	10.20	3.65	0
SFBailey	2.93	1.81	1.75	1.70	1.77	6.49	9.96	3.47	1
SFPatton	2.96	1.82	1.73	1.70	1.77	6.51	9.98	3.47	1
SFThompson	2.92	1.82	1.77	1.71	1.76	6.51	9.98	3.47	1
SFBurns	2.95	1.81	1.76	1.71	1.78	6.52	10.01	3.49	1
SFChambers	2.96	1.83	1.76	1.71	1.78	6.55	10.04	3.49	1
SFRodgers	2.95	1.84	1.76	1.72	1.77	6.55	10.04	3.49	1
SFEdwards	3.01	1.88	1.78	1.71	1.76	6.67	10.14	3.47	1
SFFrater	2.97	1.84	1.78	1.72	1.83	6.59	10.14	3.55	1
SFMbandjock	3.06	1.86	1.77	1.72	1.77	6.69	10.18	3.49	1
SFPhiri	2.93	1.86	1.80	1.77	1.83	6.59	10.19	3.60	1
SFNdure	2.96	1.85	1.79	1.76	1.84	6.60	10.20	3.60	1
SFTsukahara	2.98	1.88	1.81	1.77	1.81	6.67	10.25	3.58	1
SFBolt	2.89	1.79	1.73	1.70	1.78	6.41	9.89	3.48	1
SFGay	2.99	1.81	1.74	1.67	1.72	6.54	9.93	3.39	1
SFPowell	2.92	1.81	1.74	1.70	1.78	6.47	9.95	3.48	1
FBolt	2.88	1.76	1.67	1.61	1.66	6.31	9.58	3.27	1
FGay	2.92	1.78	1.69	1.63	1.69	6.39	9.71	3.32	1
FPowell	2.91	1.80	1.71	1.68	1.74	6.42	9.84	3.42	1
FBailey	2.92	1.81	1.75	1.70	1.75	6.48	9.93	3.45	1
FThompson	2.90	1.81	1.74	1.72	1.76	6.45	9.93	3.48	1
FBurns	2.94	1.82	1.76	1.72	1.76	6.52	10.00	3.48	1
FChambers	2.93	1.82	1.75	1.72	1.78	6.50	10.00	3.50	1
FPatton	2.96	1.89	1.80	1.77	1.92	6.65	10.34	3.69	1

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Markus Ellisaar, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

„Mittelineaarsed regressioonmudelid rakendusega spordiandmetele“, mille juhendaja on Imbi Traat,

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 08.05.2017