Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Risto Korb

Ülevaade regressioonmudeli diagnostika graafikutest R-i paketis „car“

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja Anne Selart

Tartu 2016

Ülevaade regressioonmudeli diagnostika graafikutest R-i paketis „car“

Käesoleva bakalaureusetöö esimeses osas tutvutakse regressioonmudeli diagnostika selle osaga, mida on võimalik sooritada läbi graafikute. Lisaks selgitatakse, mida täpsemalt saab ja ei saa graafikutest järeldada. Tuuakse graafikute näiteid kirjandusest. Töö teises osas antakse ülevaade R-i paketis „car“ olevate võimalustega sooritada regressiooni diagnostikat: tutvutakse paketis olevate funktsioonide, nende parameetrite ja rakendustega, kasutades andmestikke, mis leiduvad paketis „car“. Antakse hinnang funktsioonide efektiivsusele ja sooritatakse esimeses osas tutvustatud metoodika abil graafiku analüüs.

Regressioonanalüüs, graafilised meetodid, R (programmeerimiskeel)

P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Summary of plots for regression model diagnostics included in R’s package ’car’

In the first part of this bachelor’s thesis regression model diagnostics is introduced focusing on the part of diagnostics which can be carried out by plots. What can and cannot be concluded by these plots is explained. Examples of these plots from literature are presented. The second part gives a summary of the functions, their parameters and applications included in R’s package ’car’ that can be used to carry out regression diagnostics. It is reviewed how effective these plots are and a short analysis is carried out based on the methodic introduced in part one.

Regression analysis, graphical methods, R (programming language)

P160 Statistics, operation research, programming, actuarial mathematics

Sisukord

Sissejuhatus

Statistilise mudeli loomisel kasutatakse regressioonanalüüsi kui statistilist protsessi, et kirjeldada muutujate vahelist seost. See sisaldab endas mitmeid erinevaid tehnilisi võtteid, seal hulgas mudeli kuju määramist, argumentide valikut, parameetrite väärtuste arvutamist statistiliste andmete põhjal ja parameetrite olulisuse kontrolli. Kuigi eelnimetatud võtete abil on võimalik koostada täielik mudel, mis võiks kehtida üldkogumis, ei piirdu regressioonanalüüs ainult mudeli koostamisega. E. Käärik on järjestanud statistilise mudeli leidmise seitsmeks alapunktiks, kus nimetatud võtted katavad vaid esimest kolme punkti [1]. Täieliku analüüsi korral oleks vaja sooritada mudeli diagnostika, mille visuaalsele poolele on käesolev bakalaureusetöö keskendunud.

Regressioonmudeli diagnostika tähendus kirjanduses varieerub. Näiteks E. Käärik on kirjutanud: „Mudeli diagnostika tegeleb mudeli erindite ehk mingis mõttes iseäralike vaatluste väljaselgitamisega“ [1]. Samas S. Weisberg on kirjutanud: „Regressiooni diagnostika eesmärk on kontrollida, kas saadud statistilise mudeli keskväärtusfunktsioon ja eeldused on kooskõlas tegeliku andmestikuga“ [2]. Nii nagu seda on teinud mitmed autorid, ei piirdu antud bakalaureusetöö autor regressiooni diagnostika tähenduse juures vaid erindite välja selgitamisega, vaid hõlmab sellesse ka regressioonimudeli eelduste kontrolli.

Regressiooni diagnostika graafikutega tegelevad tänapäeval mitmed statistilised tarkvarad, sealhulgas SAS, NCSS, Statistica ning erinevad statistikapaketid R-s. Üks statistikapakettidest, mis regressiooni diagnostikat toetab on John Foxi pakett „car“ pealkirjaga „Rakendusregressiooni abiline“. Käesoleva bakalaureusetöö eesmärgiks on anda ülevaade regressiooni diagnostika graafikutest, mida on võimalik genereerida kasutades J. Foxi paketi „car“ funktsioone.

Antud töö esimeses osas tutvutakse regressiooni diagnostikaga ja selle osaga, mida on võimalik sooritada läbi graafikute. Lisaks selgitatakse, mida täpsemalt saab ja ei saa graafikutest järeldada ehk antakse ülevaade visuaalse diagnostika plussidest ja miinustest. Bakalaureusetöö teises osas antakse ülevaade R-i paketi „car“ funktsioonidega , mis aitavad sooritada diagnostikat. Andmestikena kasutatakse erinevaid andmetabeleid, mis leiduvad paketis „car“.

1 Regressioonimudel ja diagnostika

Statistilise mudeli üldkuju on $Y = f(X_1, X_2, \dots, \alpha, \beta, \dots) + \varepsilon$, kus Y on funktsioontunnus ehk sõltuv tunnus, $f(\cdot)$ tähistab mingit funktsiooni (nt lineaarne funktsioon), $X_1, X_2, \dots$ on argumenttunnused ehk sõltumatud tunnused, $\alpha, \beta, \dots$ on mudeli parameetrid ja ε on mudeli juhuslik viga. [1]

Üldjuhul mõeldakse regressioonmudeli all statistilist mudelit, kus funktsioon $f(X_1, X_2, \dots, \alpha, \beta, \dots)$ on mittelineaarse regressiooni puhul mittelineaarne funktsioon ja lineaarse regressiooni puhul vastav lihtsale või mitme argumendiga regressioonmudelile. [3]

Lihtsa regressioonimudeli kuju on $Y = \alpha + \beta X + \varepsilon$, kus spetsiifiliselt regressioonimudelile on argumenttunnus arvuline, α nimetatakse vabaliikmeks ja β nimetatakse regressioonikordajaks. Mudeli parameetrit hinnatakse vähimruutude meetodil ehk minimiseeritakse erinevused tegelikult mõõdetud uuritava tunnuse väärtuse ja mudeli järgi prognoositud väärtuste vahel. [1]

Regressioonimudelid ei piirdu aga ainult ühe argumendiga, vaid neid võib keerukamate mudelite korral olla mitmeid. Mitme argumendiga regressioonimudeli kuju on

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,$$

kus $\beta_0, \beta_1, \dots, \beta_k$ on mudeli parameetrid ja $X_1, X_2, \dots, X_k$ on argumenttunnused. [1]

Mudeli jääk e on mudeli juhusliku vea ε hinnang ($e = \hat{\varepsilon}$). Iga vaatluse i korral saab hinnata mudeli juhuslikku viga võrdusega

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

kus y_i on vaatlusele i vastav tegelik funktsioontunnuse väärtus ja $\hat{y}_i$ on hinnang vaatlusele i vastavale funktsioontunnuse väärtusele jättes mudelist välja juhusliku vea. Standardiseeritud jäägid, saadakse jäägi jagamisel tema standardhälbega, Studenti jäägid saadakse seevastu jäägi jagamisel tema standardhälbega, mille arvutamisel on antud vaatlus välja jäetud. Mõlemad jäägid on ligikaudu *t*-jaotusega. [1]

$$e_i^{stand} = \frac{e_i}{s_{e_i}}, \qquad e_i^{stud} = \frac{e_i}{s_{e_{(i)}}}.$$

Läbi juhuslike vigade on võimalik defineerida regressioonmudeli eeldused, milleks on

- juhuslikud vead on erinevate vaatluste korral sõltumatud;
- juhuslike vigade keskväärtus on null ($E\varepsilon_i = 0, \forall i$);
- juhuslike vigade hajuvus on konstantne ($D\varepsilon_i = \sigma^2, \forall i$);
- juhuslikud vead peavad olema normaaljaotusega.

Lisaks aitab jääkide analüüs mõista mudeli sobivust läbi erindite kontrollimise, millest räägitakse täpsemalt alapeatükis 1.3. [1]

Täielik regressioonanalüüs ei sisalda endas ainult võrduse koostamist sõltuvate ja sõltumatute muutujate vahel, vaid tuleks vastata küsimustele:

- kas mudelit saaks parandada muutes regressioonikordajate väärtusi või kasutades mittelineaarset seost tunnuste vahel;
- kas mudeli eeldused on täidetud ja kui ei ole, siis kuidas saaks olukorda parandada;
- kas leidub erindeid ehk kas mõni vaatlustest avaldab mudeli regressioonikordajatele teistest märgatavalt rohkem mõju. [3]

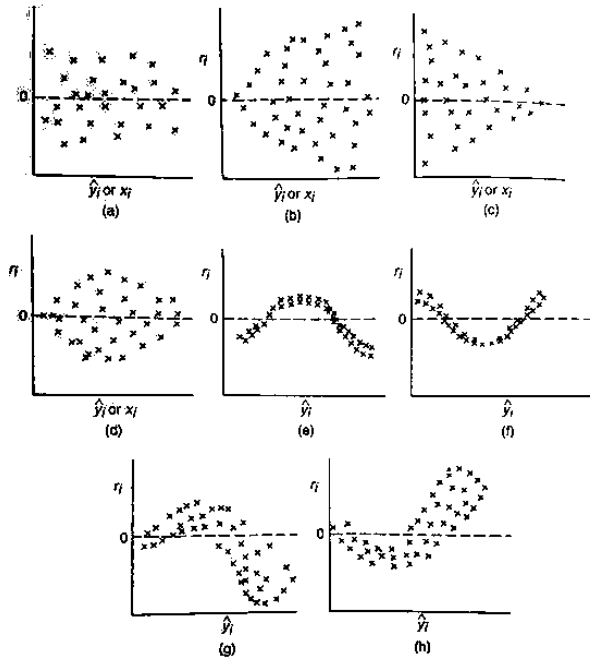
Käesoleva bakalaureusetöö raames on regressioonmudeli diagnostika all mõeldud regressioonmudeli jääkide analüüsi ehk mudeli eelduste kontrolli ning erindite väljaselgitamist. Regressioonmudeli diagnostika graafikute all peetakse silmas graafikuid, mille abil on visuaalselt võimalik sooritada jääkide analüüsi.

Regressiooni diagnostikat kasutatakse pärast seda, kui regressioonmudel on koostatud. Kui mudeliga saadud jäägid ei tundu visualiseerituna eeldustele vastavana, siis muutub mudeli korrektsus küsitavaks. Lisaks võib valmisse sattuda juht, mille ära jätmine mõjutab võrdlemisi palju lõplikku mudelit. Sellistele vaatlustele vastavad jäägid paistavad silma olles jääkide visualisatsioonil tihtipeale isoleeritud ning erinditele vastavate vaatluste õigsuse peaks üle kontrollima. [2]

Alati ei piirduta regressiooni diagnostikat sooritades visuaalse analüüsiga, vaid eelduste kontrollimiseks sooritatakse lisaks statistilisi teste. Statistilised testid tulevad kasuks näiteks olukorras, kus mittelineaarse seose hindamine, kus jääkide mittelineaarne seos pole ilmselge, võib uurija subjektiivne arvamuse tõttu viia eksliku tulemuseni. Seetõttu on visuaalne analüüs alati hea olukorra hindaja, kuid mitte parim viis põhjalikku analüüsi sooritada. [2]

1.1 Vigade keskväärtus ja mittekonstantne dispersioon

Regressioonanalüüsi eelduste kohaselt peab mudeli iga juhuslik viga olema konstantse dispersiooniga ja keskväärtusega 0. Mõlemat eeldust on võimalik visuaalselt kontrollida, kui kujutada kõik vaatlused graafikul, kus x-teljeks on argumenttunnuse väärtus või mudeli prognoos ning y-teljeks on standardiseeritud või Studenti jäägid. Kui mõlemad eeldused on täidetud, siis peaksid vead olema ühtlaselt hajunud ümber $e = 0$ telje. Joonis 1 kujutab ühte eeldusi mitterikkunud ja mitut eeldusi rikkunud jääkide graafikuid lineaarse regressiooni puhul.



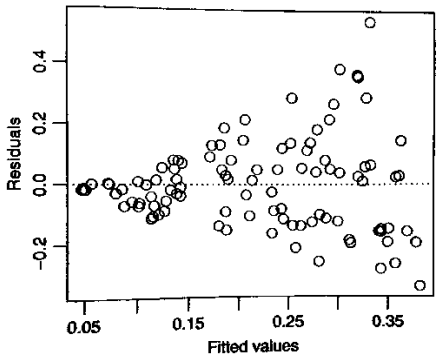
Joonis 1. Eeldustele vastav ja eeldusi rikkunud mudelite võimalikud jääkide graafikud [2]

Joonisel 1 asuv esimene graafik (a) kujutab täidetud eelduste korral võimalikku vaatluste asetust jääkide graafikul. Graafikutel (b) kuni (d) võib näha, et dispersioon ei ole konstantne. Jääkide mittekonstantne dispersioon tähendab seda, et mudel ei kirjelda hajuvust sama hästi igal tunnuste väärtuste korral ja seetõttu on mudeli hinnangu headus sõltuv tunnuste väärtusest. Graafikul (b) on näha, et dispersioon on paremal pool suurem, kui ta on seda graafiku vasakul pool. Graafikul (c) on ebaühtlane dispersioon kujutatud vastupidi ning joonis (d) kujutab suurt dispersiooni graafiku keskel, servades aga väikest. Graafikutel (e) ja (f) nähtava jääkide paiknemise põhjal võib eeldada, et mudeli jäägid ei ole iga tunnuse väärtuse korral keskväärtusega 0. Sellise tulemuse võib saada näiteks lineaarfunktsiooni kasutades, kui peaks kasutama kõrgema astme polünoomi või logaritmfunktsiooni. Mudel, kus vea keskväärtus ei ole igal juhul 0, on seetõttu ka eeldustele mitte vastav. Graafikutel (g) ja (h) sisaldavad mõlemat probleemi: ebakonstantset dispersiooni ja mittenullulist keskväärtust. [2]

Kui mudel sisaldab endas mitut sõltumatut muutujat, siis pole tingimata võimalik säärase graafiku põhjal välja selgitada, milline eeldus on rikutud. Näiteks joonisel 2 on mudeli keskväärtusfunktsioonina kasutatud $E(Y|\boldsymbol{X}=\boldsymbol{x})=\beta_0+\beta_1x_1+\beta_2x_2$. Paremale poole hajuvad jäägid võiksid eeldada mittekonstantset hajuvust. Tegelikkuses on andmestik saadud genereerides funktsiooni

$$E(Y|\boldsymbol{X}=\boldsymbol{x})=\frac{|x_1|}{2+(1.5+x_2)^2} \tag{1.1}$$

väärtusi, millel on ühtlane hajuvus. Siin seisneb probleem hoopis valesti hinnatud keskväärtus-funktsioonis, mida antud graafiku põhjal pole võimalik näha. Seega jääkide graafiku põhjal võib küll arvata, et eeldusi on rikutud, kuid nende põhjal ei saa väita, milline on see konkreetne eeldus, mida peaks üle kontrollima. [2]



Joonis 2. Funktsiooni 1.1 abil genereeritud andmestiku põhjal loodud regressioonmudeli jääkide graafik [2]

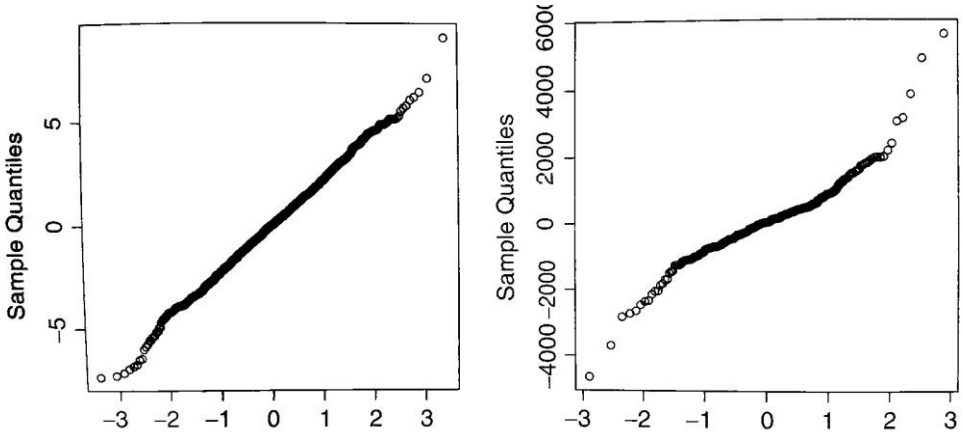
1.2 Juhuslike vigade kuulumine normaaljaotusesse

Graafilised meetodid annavad visuaalse pildi, mille abil on võimalik teha subjektiivseid oletusi tunnuse jaotuse kohta. Tihti pole statistilist testi vajagi tunnuse ligikaudse jaotuse kontrollimiseks, vaid piisab graafikust saadud informatsioonist. [1]

Graafikud, mis aitavad kontrollida, kas tunnus on normaaljaotusega, on näiteks

- histogramm, kus võrreldakse valimi histogrammi normaaljaotuse tihedusfunktsiooni graafikuga;
- kvantiilide graafik (*Quantile-Quantile plot*) või tõenäosuspaber (*Normal Probability Plot*), kus joonistatakse hajuvusgraafik selliselt, et y-teljel on uuritava tunnuse järjestatud väärtused ja x-teljel on normaaljaotuse kvantiilid.

Viimase puhul näitab punktide enam-vähem sirgel asetsemine, et tegemist võiks olla muutujaga, mis on pärit normaaljaotusest. [1]



Joonis 3. Tõenäosuspaberid kahel juhul [2]

Väikese valimi korral ei pruugi tõenäosuspaberil olev hõre punktihulk visuaalsel hindamisel hea tulemuseni viia. Suure valimi korral, nagu seda on kujutatud joonisel 3, saab aga väga hästi subjektiivselt hinnata tunnuse jaotust. Vasakpoolsel joonisel on näha, et väga suur osa punktides asuvad ühel sirgel ning visuaalse analüüsi põhjal ei saa normaaljaotuse kohta kehtivat eeldust nimetada rikutuks. Parempoolsel joonisel aga on näha, et punktid on moodustanud selgelt S-kuju. See tähendab, et tõenäoliselt pole mudelist tulenevad jäägid normaaljaotusest. [2]

1.3 Erindid

Andmestikus võib esineda vaatlusi, kus saadud statistiline mudel ei kehti. Selliseid vaatlusi nimetatakse erinditeks. Erindeid võib jaotada oma olemuselt kolme rühma:

- erindid suure jäägi mõttes, kus mudeli jääk konkreetsel vaatlusel on võrreldes teiste vaatlustega ebatavaliselt suur;
- omapärane vaatlus, mida iseloomustab tema kaugus argumenttunnnuse keskväärtusest;
- mõjus vaatlus, mis muudab liiga palju regressioonikordajat.

Leitud erindid tuleks üle kontrollida, kas tegemist on tõepoolest mingi erilise vaatlusega või on andmetes viga. Erindite andmetest ära jätmisel tuleb olla aga ettevaatlik, sest eesmärk pole saada ilus mudel vaid mudel, mis vastab ka andmestikule. [1]

Aru saamiseks, kas jääk on erind suure jäägi mõttes, leitakse kõigepealt standardiseeritud või Studenti jäägid. Kui selle jäägi väärtus on ligikaudu 3, siis võib seda nimetada erindiks suure jäägi mõttes. [1]

Vaatluse omapära iseloomustab selle kaugus vaatluse argumenttunnuse keskväärtusest. Seda tähistatakse i-nda vaatluse korral h_i (*hat-value*) ning arvutatakse lineaarse regressioonmudeli korral:

$$h_i = (\boldsymbol{H})_{ii},$$

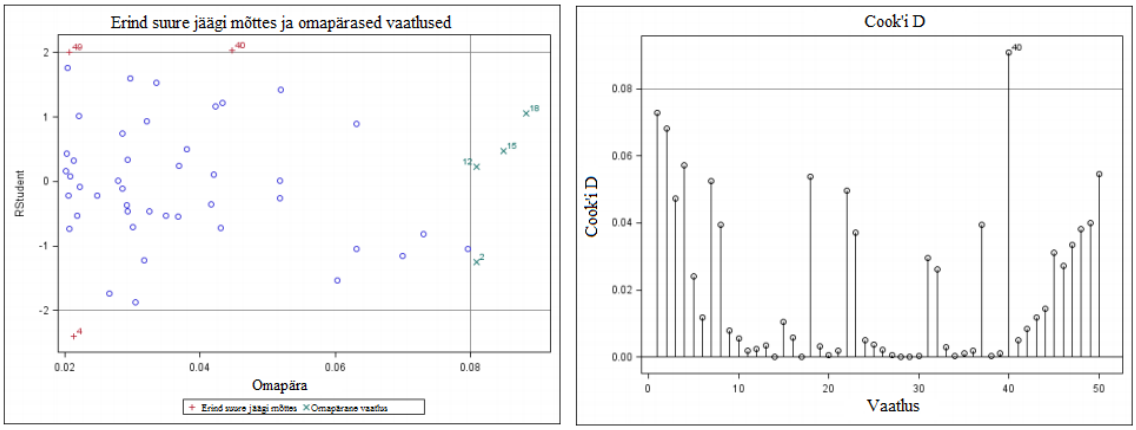
kus $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, milles $\boldsymbol{X}$ on $n \times k$ maatriks, kus element x_{ij} vastab i-ndal vaatlusel oleva j-nda tunnuse väärtusele. Üldjuhul kehtib reegel, kui $h_i > 2(k+1)/n$, siis on tegemist omapärase vaatlusega. [1]

Vaatluse i mõju regressioonikordajatele lineaarsel regressioonil hinnatakse Cook'i D abil. Kasutades *hat-value* väärtust i-ndal vaatlusel, saab arvutada Cook'i D valemiga

$$D_i = \frac{\left(e_i^{stand}\right)^2}{k+1} \cdot \frac{h_i}{1-h_i}.$$

Üldiselt peetakse erinditeks vaatlusi, mille korral $D_i > \frac{4}{n-k-1}$. [4]

Materjalis „Andmeanalüüs II. Loengukonspekt." [1] võib leida joonised, mis kujutavad erindite paiknemist võrreldes teiste jääkidega. Need joonised on genereeritud statistikaprogrammi SAS abil ning asuvad joonisel 4.



Joonis 4. Erindite kuvamine SAS abil [1]

Nagu jooniselt 4 näha, on SAS sümboliga üksteisest eristanud erindid suure jäägi mõttes ja omapärased vaatlused. Lisaks on SAS märkinud kõikide erindite alla kuuluvate jääkide juurde nende tähistused vastavalt andmestikule. See aitab hiljem analüüsida vastavaid erindeid ning määratleda, kas on tegemist veaga või iseäraliku vaatlusega. Eraldi tähistamine aitab täpsemini mõista, milles võiks viga olla. Sama kehtib ka mõjusate vaatluste kohta, mis on ära tähistatud joonise 4 paremal graafikul.

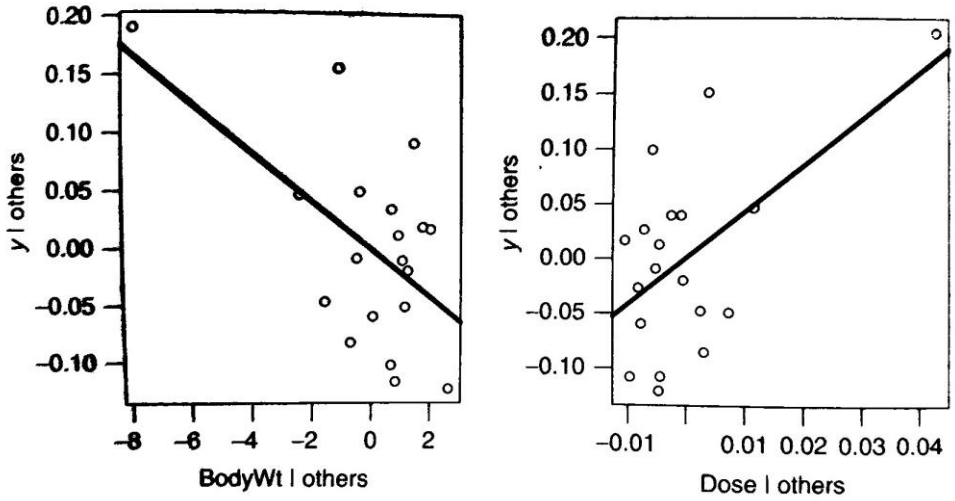
Lisaks eelnevale kahele joonisele aitab paremat regressioonmudelit ja eesotsas erindeid leida lisatud-muutujate graafik (*partial-regression plot*). Lisatud-muutujate graafik sõltumatule tunnusele x_j konstrueeritakse järgnevalt:

- leitakse kõik mudeli jäägid, mis on saadud, kui funktsioontunnuseks on võetud Y ja argumenttunnusteks on võetud kõik tunnused peale x_j;
- leitakse kõik mudeli jäägid, mis on saadud, kui funktsioontunnuseks on võetud x_j ja argumenttunnusteks on võetud kõik ülejäänud argumenttunnused;
- luuakse hajuvusgraafik eelnevatest jääkidest.

Kui mõjusad vaatlused puuduvad, siis graafikul kujutatud punktid paiknevad lähestikku piiratud alal. Mõjusad vaatlused paistavad joonisel silma oma isoleerituse tõttu. [3]

Ühte võimalikku lisatud-muutujate graafikut võib näha joonisel 5. Isoleeritud punktidele vastavad vaatlused mõjutavad tõenäoliselt regressioonikordajat, kui vaatluse all olev argumenttunnus mudelisse lisatakse. Nagu näiteks vasakul graafikul olev punkt, mis

asetseb vasakul üleval nurgas on teistest selgelt isoleeritud ja tõenäoliselt vaatlus, mis sellele punktile vastab, on mõjus. [2]



Joonis 5. Lisatud-muutujate graafikud [2]

Tuleb mainida, et kuigi nii lisatud-muutujate graafik kui ka Cook'i *D* hindavad vaatluse mõjusust, siis tegemist ei ole sama arvutusalgoritmiga ja vastavate graafikute tulemused ei pruugi kattuda.

Lisatud-muutujate graafik tuleb kasuks ka hindamaks kui tugev lineaarne seos on valitud argumenttunnuse ja funktsioontunnuse vahel, kui kasutatakse ülejäänud argumenttunnuseid. [3]

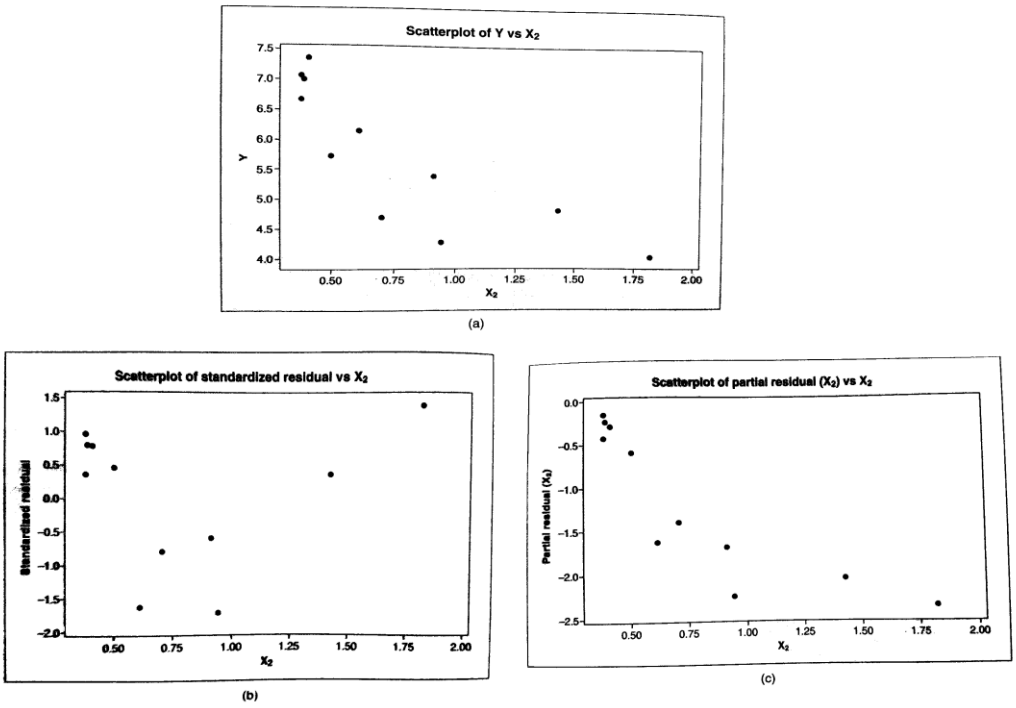
1.4 Lineaarse seose visuaalne kontroll

Nagu alapeatükist 1.1 võis lugeda, on eelduste kontrollimisel suureks abiks sealsed näidetena toodud jääkide graafikud. Sellest hoolimata ei pruugi vastavad jääkide graafikud kõige paremini vastata küsimusele, kas lineaarfunktsioon on parim lahendus andmete kirjeldamiseks. Selle põhjuseks võib olla nii-öelda kirju graafik, kus suure arvu vaatluste korral on kujutatud punktid tihedalt koos ja ilmselge mittelineaarne seos kaob ära ebaselge visualisatsiooni tõttu. Sellises olukorras pole jääkide graafik just kõige parem. Abiks võiks olla osaliste jääkide graafik (*partial residual plot*). [3]

Osaliste jääkide graafiku saamiseks arvutatakse kõigepealt osaline jääk e_i^* argumenttunnuse X_j ja vaatluse i jaoks vastavalt:

$$e_i^* = e_i + \beta_j X_{ji}.$$

Seejärel luuakse graafik, kus x-teljel on argumenttunnuse X_j väärtus ja y-teljeks on osalise jäägi suurus. Igale vaatlusele vastavalt lisatakse graafikule punkt. Ühte võimalikku osaliste jääkide graafikut võib näha joonisel 6.



Joonis 6. (a) Hajuvusgraafik (b) Jääkide graafik (c) Osaliste jääkide graafik [3]

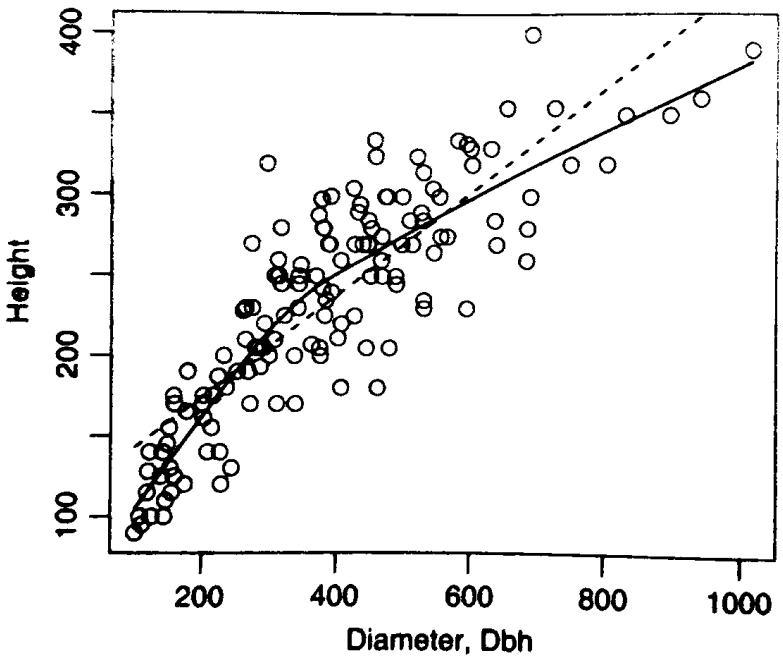
Joonisel 6 olevalt graafikult (a) on näha, et argumenttunnus X_2 ja funktsioontunnus Y ei ole seotud lineaarselt vaid pigem on tegu pöördvõrdelise seosega. Seda seost ei suuda kahjuks standardiseeritud jääke kujutav jääkide graafik (b) kirjeldada. Seevastu osaliste jääkide graafikult (c) on visuaalsel analüüsil väga kerge välja lugeda, et tegemist ei ole lineaarse seosega vaid hoopis millegi muuga. Siit ka erinevus mittelineaarse seose hindamisel jääkide graafikul ja osaliste jääkide graafikul.

Lineaarse või mittelineaarse seose hindamiseks on võimalik kasutada ka CERES graafikut (*combining conditional expectations and residuals*), mis on üldistus osaliste

jääkide graafikutele. CERES graafikud ei ole nii kergesti mõjutatavad lineaarse sõltuvuse puudumisest. Näide CERES graafikust ja selle sisu tutvustus asub peatükis 2.4. [5]

1.5 Mudeli headuse hindamine graafiliselt

Suur osa regressioonmudeli diagnostikast on kindlaks tegemine, kas mudeli eeldusi on rikutud. Väga sarnane sellele ülesandele on kontrollimine, kui hästi sobib saadud mudel andmetega. Selleks võib kasutada marginaal mudeli graafikuid (*marginal model plots*). Sellise graafiku näitega võib tutvuda joonisel 7.



Joonis 7. Marginaal mudeli graafik [2]

Joonisel 7 on katkendjoonega tähistatud lineaarset seost argumenttunnuse ja funktsioon-tunnuse vahel, mis on leitud regressioonanalüüsi käigus. Pideva joonega on kujutatud mitteparameetrilist hinnangut funktsioontunnuse keskväärtusele konkreetse argumenttunnuse väärtusel ($E(Y|X = x)$). Antud juhul on tegemist joonega, mida nimetatakse *loess* jooneks. *Loess*i joone võrdlus mudelist saadud joonega aitab hinnata mudeli headust. Kui katkendjoon sarnaneb pideva joonega, siis on alust arvata, et saadud mudel vastab hästi andmetele. Kui katkendjoon erineb pidevast joonest väga, siis ei kirjelda mudel tegelikku andmestikku kõige paremini.

Samale joonisele on võimalik lisada ka standardhälbe jooned. Need jooned ümbritsevad eelnevalt nimetatud jooni määrates piirkonna, mis asub maksimaalselt ühe standardhälbe kaugusel keskväärtusjoontest.

2 Regressiooni diagnostika graafikute ülevaade paketis „car“

Peatükis 1 tutvuti võimalustega sooritada regressiooni diagnostikat läbi visuaalse analüüsi ja toodud näidete graafikud olid võetud kirjandusest. Selles peatükis püütakse sarnaseid graafikuid genereerida R-i paketi „car“ abil.

Statistikapaketi „car“ pealkirjaga „Rakendusregressiooni abiline“ („*Companion to Applied Regression*“) autoriks on John Fox. Paketi „car“ kõige varasemad versioonid ulatuvad aastasse 2001. Versioon 2.1-2, mida on käesolevas bakalaureusetöös kasutatud, on avaldatud 25. märtsil 2016. Pakett sisaldab endas funktsioone, mis muuseas genereerivad graafikuid ja sooritavad statistilisi teste. Funktsioonidega, mis aitavad läbi viia regressiooni diagnostikat, tutvutakse alljärgnevates alapeatükkides. Lisaks sisaldab pakett endas andmestikke, mille hulgast kasutavad järgnevad peatükid kahte: „Duncan“ ja „Prestige“.

2.1 Keskväärtus, mittekonstantne dispersioon. Funktsioon `residualPlots()`.

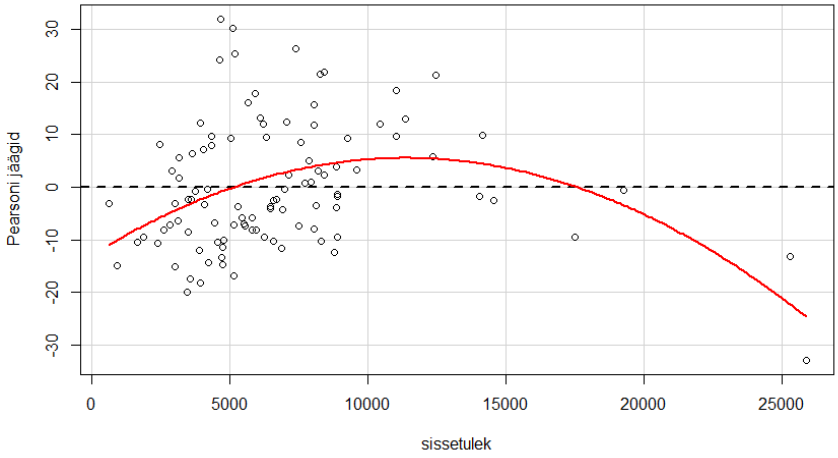
Paketis „car“ leiduv funktsioon `residualPlots()` kujutab peatükis 1.1 kirjeldatud jääkide graafikut. Lisaks sooritab funktsioon testi hindamaks, kui hästi kirjeldab ruutfunktsioon jääke, ning visualiseerib selle, lisades parabooli graafikule. [5]

Funktsiooni `residualPlots()` parameetritest on kõige olulisemad mudeli diagnostika sooritamiseks järgnevad.

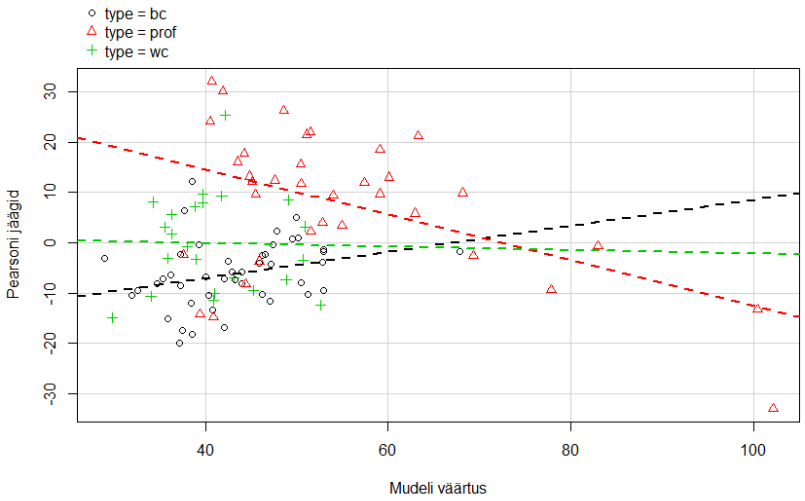
- `model` – võtab väärtuseks regressioonimudeli.
- `terms` – võtab väärtuseks valemi, mille abil on võimalik täpsustada, milliste parameetrite kohta jääkide graafikud kujutatakse. Antud parameetri väärtust muutes on võimalik luua ka grupeeritud jääkide graafik (regressioonimudeli mitme tunnuse jääkide graafikud on kujutatud ühel teljestikul).
- `fitted` – tõese väärtuse korral kuvab jääkide graafiku, kus x-teljeks on mudeli prognoos.
- `type` – võtab väärtuseks jääkide tüübi, mida kuvatakse. Võimalikud väärtused näiteks „rstudent“ ja „rstandard“

- `groups` – võtab väärtuseks grupi indikaatori nimekirja. Parameetri väärtustamisel kuvatakse jäägid spetsiifilises grupis erineva värvi või sümboliga.
- `quadratic` – tõesel väärtusel kuvab graafikul ruutregressiooni ehk parabooli.

Kasutades dokumentatsioonis [5] toodud näidet on võimalik genereerida graafikud, mis asuvad joonisel 8 ja 9. R-i kood graafikute kujutamiseks võib leida lisast 1.



Joonis 8. Andmestiku "Prestige" põhjal loodud jääkide graafik argumenttunnuse „sissetulek“ kohta



Joonis 9. Andmestiku "Prestige" põhjal loodud jääkide graafik sõltuvalt tunnusest "type"

Joonistel 8 ja 9 on kujutatud mudeli jääke, mis hindab seost Kanada ameti prestiiži ja ameti keskmise sissetulekute vahel. Kasutatud andmed võib leida paketiga „car“ kaasatulevast andmestikust „Prestige“. Andmestik „Prestige“ koosneb seitsmest tunnusest: nominaaltunnus määramaks ametit; pidev tunnus „education“ määramaks ametis olevate töötajate keskmist haridust aastates; pidev tunnus „income“ määramaks ametis olevate töötajate keskmist sissetulekut; pidev tunnus „women“ määramaks naiste osakaalu ametis olevate töötajate seas, pidev tunnus „prestige“ määramaks ametile vastavat Pieno-Porteri prestiiži skoori, mis on saadud 1960ndatel aastatel läbi viidud uuringu põhjal; numbriline nominaaltunnus „census“ määramaks ameti koodi Kanada rahvaloenduse põhjal; faktortunnus „type“ määramaks ameti tüüpi, kus „bc“ (*Blue Collar*) tähendab manuaalset tööd (nt tööstuses), „prof“ tähendab professionaal-, hooldus- ja tehnilisi töid ning „wc“ (*White Collar*) tähendab tööd teenindussektoris. [5]

Jääkide graafikule joonisel 8 on lisatud parabool, mis on abiks hindamaks, kas iga jääk on nullilise keskväärtusega või mitte. Antud juhul peaks tegelikult alustama analüüsi erinditest, sest jäägid, mis paiknevad joonise all paremas servas, tunduvad kõverat väga palju mõjutavat. Kui tegemist pole vigadega, siis võiks arvata, et suurema sissetuleku korral on tunduvalt suurem hajuvus ning ka keskväärtus pole 0: tõenäosus, et paremal pool üks suvaline vaatlus paikneb 0-joonest väga kaugel, kui vasakul pool on seda saavutanud

vaid paar vaatlust kümnetest, on eeldusel, et dispersioon on kõikjal sama, väga väike. Seega on alust arvata, et konstantse dispersiooni eeldus ja nullilise keskväärtuse eeldus on rikutud.

Vaadates joonist 9, kus jäägid on üksteisest eristatud sõltuvalt ametitüübist, võib näha, et eelnevalt silma paistvad erindid ei pruugi olla vead. Nimelt tänu funktsiooni parameetri `terms` muutmisele on näha, et sama ametitüübi (tüüp „prof“) alla kuuluvate väärtuste korral on jäägi suurus väike suure sissetuleku korral ja vastupidi, moodustades sellega suhteliselt vähe hajunud jääkide hulga. Vaadates nüüd ainult ameteid, mis kuuluvad tüübi „prof“ alla, siis mittelineaarset seost pole näha. Sama kehtib ka teiste ametitüüpide kohta. Mittelineaarse seose vähendamise eesmärgil võib seetõttu proovida luua mudel kasutades ametitüübi ja sissetuleku koosmõju. Koosmõju lisamisel saadud jääkide graafiku võib leida lisast 1 (joonis 21), millest on näha, et uute tunnuste lisamine tõepoolest aitab saada parema mudeli.

Võib järeldada, et jääkide graafikuid on väga hea genereerida kasutades funktsiooni `residualPlots()`. Kuid kuna mittelineaarse seose hindamiseks on kasutada vaid parabooli, siis ei suuda see kuvada mõnda keerulisemat seost. Kui funktsioontunnus ja argumenttunnus peaks seotud olema mõne muu mittelineaarse funktsiooni kaudu (näiteks kuupfunktsiooni kaudu) jääb funktsioon `residualPlots()` hätta.

2.2 Jääkide jaotus. Funktsioon `qqPlot()`.

Paketis „car“ leiduv funktsioon `qqPlot()` kujutab muutuja või mudeli Studenti jääkide empiiriliste kvantiilide seose soovitud jaotuse teoreetiliste kvantiilidega. Täpsemalt genereerib funktsioon kvantiilide graafiku muutujale või Studenti jääkidele lineaarse mudeli puhul. Studenti jääkide korral kasutatakse sobivat *t*-jaotust. [5]

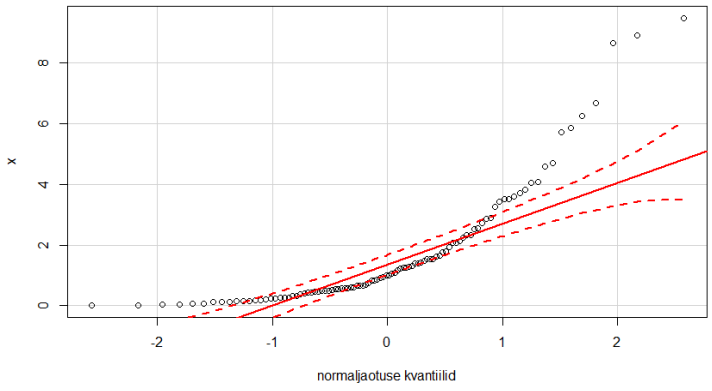
Paketi „car“ dokumentatsioonist [5] võib funktsiooni `qqPlot()` parameetritest välja valida kõige olulisemad, mida võib vaja minna regressiooni diagnostikat sooritades.

- `x` – võtab argumendiks regressioonmudeli või numbrilise vektori.

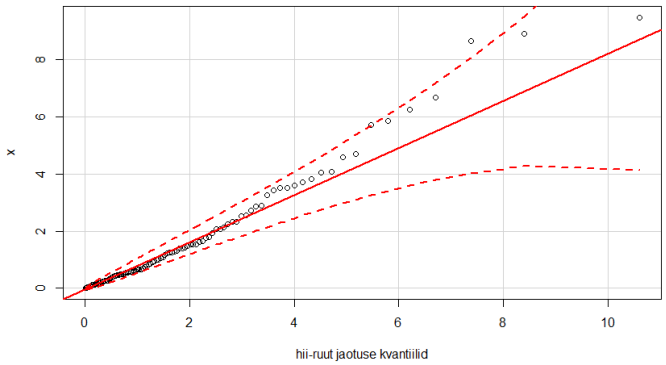
- `distribution` – võtab argumendiks R-ile vastava jaotuse nimetuse, mida funktsioon kasutab teoreetilise jaotusena graafiku kujutamisel.
- `envelope` – võtab argumendiks usaldusnivoo, mis kuvatakse graafikul kui piirkonda, kuhu vastava kindlusega võib väita, et jäägid eelduste kehtimisel kuuluvad.
- `simulate` – tõese väärtuse korral genereerib usalduspiirkonna graafikul kasutades bootstrap meetodit (võimalik vaid, kui kasutada `x`-i argumendina mudelit).
- `reps` – numbriline väärtus, mis tähistab bootstrap valimite hulka.

Kasutades dokumentatsioonis [5] toodud näidet on võimalik genereerida graafikud, mis asuvad joonisel 10, 11 ja 12. R-i koodi graafikute kujutamiseks võib leida lisast 2.

Joonistel 10 ja 11 on esialgu genereeritud 100 liikmeline valim hii-ruut jaotusest (vabadus-astmega 2) ning seejärel on saadud valimit võrreldud joonise 10 puhul normaaljaotuse kvantiilidega, joonise 11 puhul sama parameetriga hii-ruut jaotuse kvantiilidega.



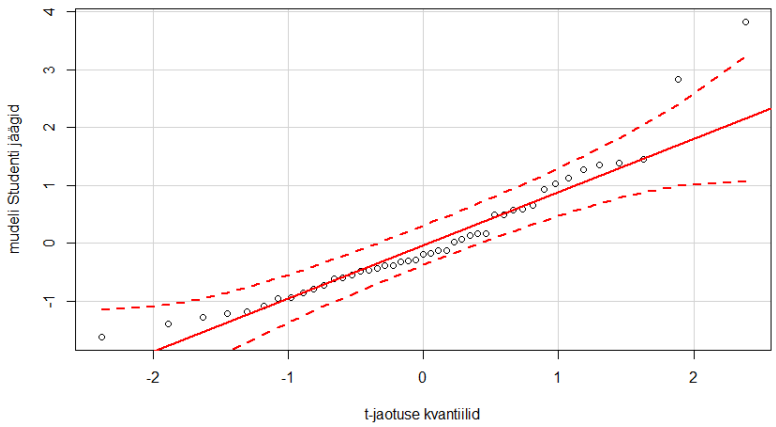
Joonis 10. Hii-ruut jaotusest pärit väärtuste seos normaaljaotuse kvantiilidega



Joonis 11. Hii-ruut jaotusest pärit väärtuste seos hii-ruut jaotuse kvantiilidega

Joonisel 10 on näha, et suur osa punktidest ei paikne pideval joonel ja katkendjoonega piiratud ala ei päästa ka kahtlusest, et tegemist ei ole normaaljaotusega. Seevastu on aga joonisel 11 näha, et enamus punktidest (vähemalt nendest, mis asuvad vasakul allpool nurgas) asuvad pideval joonel. Kui joonisel olevat katkendjoont mitte jälgida, siis võiks arvata, et viimased väärtused asuvad joonest liiga kaugel, et kogu valim kuuluks vastavasse hii-ruut jaotusesse. Pannes tähele aga katkendjoontega piiratud ala, näeme, et pea kõik punktid asuvad nimetatud piirkonnas (vaikimisi on usaldusnivooks 95%). Järelikult, kuigi suur osa punktidest asub pidevast joonest kaugel, ei saa siiski visuaalsel analüüsil piisava kindlusega väita, et tegemist pole valimiga hii-ruut jaotusest.

Joonisel 12. on andmetena kasutatud paketist „car“ pärit andmestikku „Duncan“, mis sisult ja tunnuste poolest sarnaneb alapeatükis 2.1 kirjeldatud andmestikuga „Prestige“, kuid tegemist on 45 ametiga USAst. Mudelist puuduvad tunnused ameti numbrilise tähistuse ja naiste osakaalu määramiseks. Mudeli koostamisel on seekord üritatud hinnata prestiiži suurust sõltuvalt sissetulekust, haridusest ja ametitüübist.



Joonis 12. Andmestiku "Duncan" põhjal loodud kvantiilide graafik

Nagu jooniselt 12 näha võib asub suur osa punkte ühel sirgel, kuid mõlemas servas hakkavad need hajuma liikudes pidevast joonest eemale. Suurte Studenti jääkide korral asuvad mõned punktid ka usalduspiirkonnast väljas. Kuna punktide hulk ei ole väga suur (vaid 45), siis siinkohal võib visuaalse analüüsi tulemus sõltuda uurija subjektiivsest arvamusest. Pigem võiks öelda, et jääkide jaotuse kohta käiv eeldus on täidetud, sest punktide hulk, mis rikuvad graafiku põhjal eeldust on väga väike.

Võrreldes paketi „car“ poolt genereeritud kvantiilide graafikuid graafikutega, mis on esitatud joonisel 3, võib öelda, et palju mugavam on analüüsida graafikuid, kuhu on lisatud usalduspiirkonnad ja ka teoreetiline sirge, kuhu punktid kuuluda võiksid. Kuid, kasutades funktsiooni `qqPlot()`, kasutab programm vaikimisi usalduspiirkonna määramiseks bootstrap meetodit, siis teatud parameetrite väärtusi muutmata võib graafik ja visuaalse analüüsi tulemus sõltuda sellest, millisteks bootstrap valimid osutasid. Seega tuleb antud funktsiooni kasutades ettevaatlik olla.

2.3 Mudeli erindid.

Kuna mudeli erindite tuvastamiseks mõeldud graafikute funktsioone on mitu, siis on jaotatud alapeatükk 2.3 mitmeks osaks vastavalt graafiku tüübile.

2.3.1 Erindid suure jäägi mõttes, omapärased vaatlused. Funktsioon `influencePlot()`.

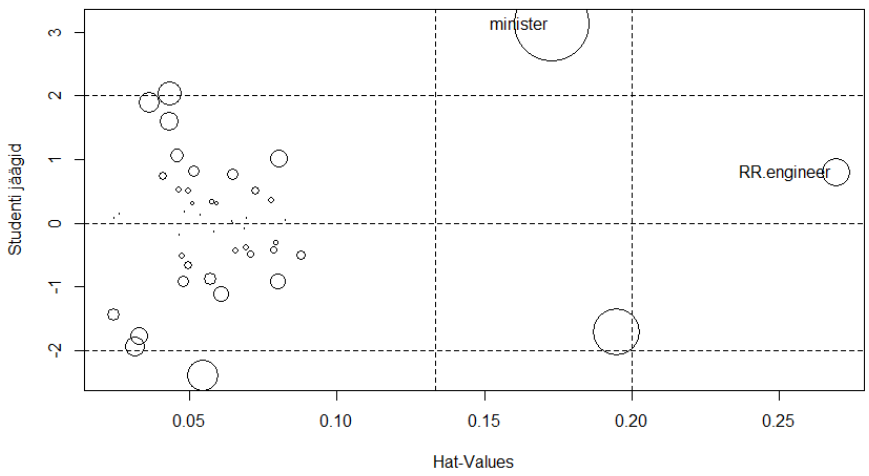
Paketis „car“ leiduv funktsioon `influencePlot()` koostab graafiku, kus y-teljeks on Studenti jäägi suurus ja x-teljeks vaatluse *hat-value* suurus. Graafikule kuvatakse kõikide vaatluste jäägid, kusjuures jääke kujutavate „mullide“ pindala sõltub sellest, kui suur on selle jäägi Cook'i D. Graafikule on lisatud vertikaalsed punktiirjooned x-i väärtustele, mis vastavad kahe- ja kolmekordse vaatluste *hat-value* keskmisega. Horisontaalsed punktiirjooned on lisatud märkimaks kus saavutab y-telg (Studenti jäägi suurus) väärtused -2, 0 ja 2. [5]

Eelnevast järeldub see, et tegelikult on võimalik antud funktsiooni abil tuvastada ka mõjusad vaatlused. Kuid kuna mulli suuruse hindamine silma järgi võib olla petlik, arvab töö koostaja, et tegemist ei ole parima funktsiooniga tuvastamaks mõjusaid vaatlusi.

Funktsioon omab mudeli diagnostika vaatenurgast järgnevaid olulisi parameetreid.

- `model` – võtab argumendiks regressioonmudeli.
- `scale` – võtab argumendiks mullide pindala kasvu mõjutava faktori.
- `labels`, `id.method`, `id.n`, `id.cex`, `id.col` – võtavad argumentideks graafikul olevate vaatluste siltide lisamist mõjutavaid väärtusi, mille tüüp sõltub parameetrist. Vaikimisi on parameetri id.method väärtus “*noteworthy*“, mis tähendab, et kõikide vaatlustele lisatakse silt, mille Studenti jääk, *hat-value* või Cook'i D on suur. Sama parameetri väärtuse muutmine võimaldab interaktiivset andmesiltide lisamist. [5]

Joonistel 13 ja 14 on kujutatud mudeli jääke, mis on saadud andmestiku „Duncan“ abil, hinnates prestiiži läbi sissetuleku ja hariduse. Andmestikku „Duncan“ on kirjeldatud alapeatükis 2.2. R-i koodi, mida on jooniste genereerimiseks kasutatud, võib leida lisast 3.



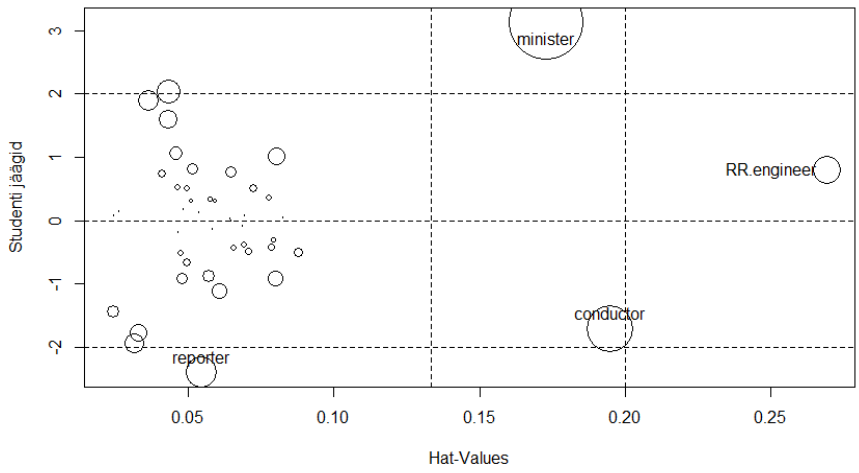
Joonis 13. Erindid andmestikus „Duncan“

Nagu jooniselt 10 näha võib, erineb paketi „car“ funktsiooni `influencePlot()` poolt kuvatud graafik SASi väljundist (joonis 4) selle poolest, et silte ei lisata kõikidele vaatlustele, mille Studenti jäägi absoluutväärtus või *hat-value* on hinnatud piirkonna põhjal suur. Vaadates sellist graafikut, tundub, et mingi osa informatsioonist jääb puudu. Konkreetsemalt puuduvad andmesildid kõige väiksema Studenti väärtusega vaatlusel (joonise kõige alumine punkt) ja puudub andmesilt vaatlusel, mille ligikaudsed koordinaadid on (0.2, -2). Need punktid asetsevad graafikul teistest piisavalt kaugel, et võiks erinditeks lugeda. Tõenäoliselt pole tegemist erinditega, järgides reegleid, mida on kirjeldatud alapeatükis 2.3. Kui see vastab tõele, siis graafikul kuvatud punktiirjooned ei anna seda informatsiooni, mida nad võiksid anda. Seetõttu ei teeks andmete õigsuse kontrollime saadud joonise põhjal isoleeritud punktides halba. Kahjuks pole vaikimisi funktsioon nendele andmesilte lisanud. Loomulikult võib ära sildistada ka kõik isoleeritud punktid, määrates parameetri `id.n` väärtuseks isoleeritud punktide arvu, kuid teatud olukordades, kus isoleeritud punktid on lähestikku, võivad sildid loetamatuks muutuda või ei anna automaatne vaatluste kokku lugemine soovitud tulemust.

Parema tulemuse võib anda parameeter `id.method` ning väärtustades selle sõnega „identify“. Programmilõigu jooksutamise järel jääb R ootama graafikul hiireklõpsu. Pärast hiireklõpsu määrab funktsioon ära piirkonna, mille keskpunktiks on punkt, mida kasutaja

hiirega vajutas, ja raadiuseks 0,25 tolli. Piirkonda sattunud vaatlustest valib funktsioon keskpunktile kõige lähedasema ja lisab sellele andmesildi, kui identifitseerimine on lõppenud. Kui identifitseerimise jooksul soovib kasutaja määrata samale punktile mitu andmesilti, annab programm veateate. Identifitseerimisel tuleb aga ettevaatlik olla, sest andmesilt kuvatakse mitte mulli keskpunkti juurde vaid asukohta, kuhu tehti hiireklõps. Seega, et kõik andmesildid pärast identifitseerimist näha oleks, tuleb mulli keskpunktist natuke eemale vajutada.

Pärast huviäratavate punktide identifitseerimist võib joonisel 14 näha andmesilte ka nendel isoleeritud punktidel, millel joonisel 13 vaatluse nimed puudusid. Sedaviisi identifitseerimine on küllaltki tülikas, sest täieliku pildi saamiseks tuleb identifitseerida ka need punktid, mida funktsioon vaikimisi ära märgib (ilma parameetri id.method väärtust muutmata).



Joonis 14. Erindid andmestikus „Duncan“, kasutades id.method="identify"

Funktsiooni `influencePlot()` kokkuvõtteks võib öelda seda, et genereeritud horisontaal ja vertikaaljooned tekitavad segadust, sest vaatlused, mis asuvad katkendjoonte põhjal erinditele vastavas piirkonnas, on ilma andmesiltideta. Tekib küsimus, kas on tegemist erindiga või ei ole. Võrreldes joonist 13 SASi poolt genereeritud graafikutega, siis saab erindite tuvastamisega paremini hakkama SAS. Küll aga kaasneb funktsiooniga hea

omadus identifitseerida punkte graafikul, mis aitab küsimuse alla sattunud punkte ja vaatlusi kokku viia.

2.3.2 Mõjusad vaatlused. Funktsioon `infIndexPlot()`, funktsioon `avPlots()`.

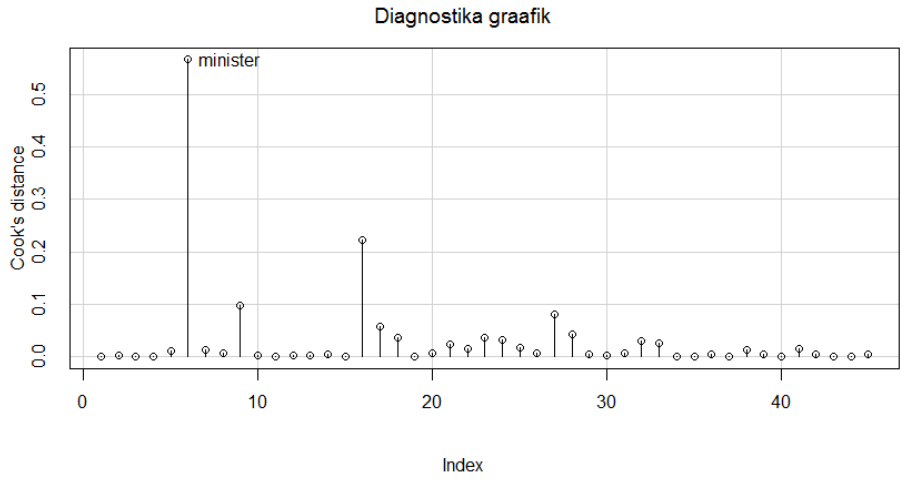
Paketis „car“ sisalduv funktsioon `infIndexPlot()` genereerib graafiku, kus x-teljeks on vaatluse indeks ja y-teljeks on Cook'i D, *hat-value*, Studenti jäägi või Bonferroni p väärtus. [5]

Kuna alapeatükis 2.3.1 sai kaetud *hat-value* ja Studenti jääkide graafikud, siis antud funktsiooni juures pakub huvi vaid graafik Cook'i D kohta.

Funktsioon `infIndexPlot()` võtab mudeli diagnostika vaatenurgast olulisteks parameetriteks järgnevad.

- `model` – võtab väärtuseks regressioonmudeli.
- `vars` – võtab väärtuseks sõne määramaks, milliseid graafikuid kujutatakse. Täpsemalt tähistab sõne „Cook“ Cook'i D, sõne „*Studentized*“ Studenti jääkide, sõne „Bonf“ Bonferroni p ja sõne „*hat*“ *hat-value* kohta käivat graafikut.
- `labels`, `id.method`, `id.n`, `id.cex`, `id.col` – kirjeldatud peatükis 2.3.1.

Joonisel 15 on kujutatud peatükis 2.3.1 kirjeldatud mudeli põhjal vaatluste Cook'i D väärtust. Joonis on genereeritud funktsiooni `infIndexPlot()` abil, mille R-i koodi võib leida lisast 4.



Joonis 15. Graafik Cook'i D kohta andmestikus „Duncan“

Võrreldes saadud joonist SASi väljundiga (joonis 4) on suurimaks erinevuseks ja ka funktsiooni `infIndexPlot()` miinuseks puuduv horisontaaljoon määramaks, millisest Cook'i *D* väärtusest alates võib nimetada vaatlust mõjusaks vaatluseks. Lisaks ei genereeri funktsioon vaikimisi ühtegi andmesilti. Joonisel nähtav andmesilt on lisatud identifitseerimise meetodil, mida on kirjeldatud alapeatükis 2.3.1.

Paketis „car“ sisalduv funktsioon `avPlots()` genereerib lisatud-muutja graafiku(d) lineaarsetele ja üldistatud lineaarsetele mudelistele. [5]

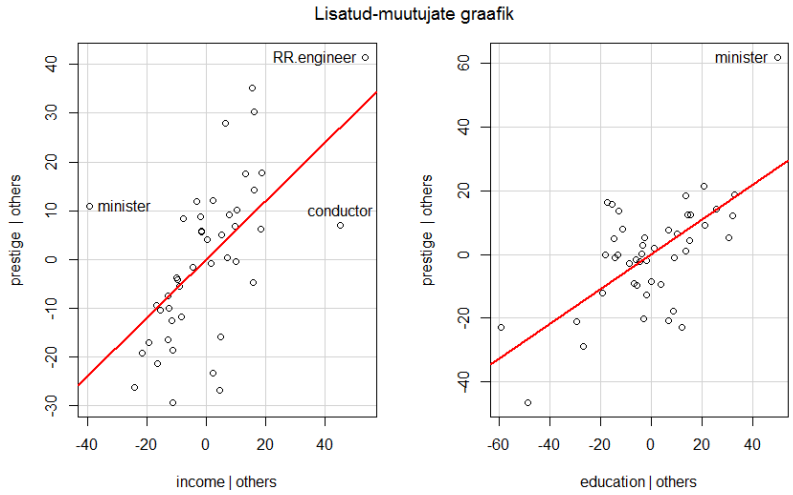
Regressioonmudeli diagnostika vaatenurgast on olulised funktsiooni `avPlots()` parameetrid järgmised.

- `model` – võtab väärtuseks regressioonmudeli.
- `terms` – kirjeldatud peatükis 2.1.
- `intercept` – tõese väärtuse korral lisab graafikute hulka lisatud-muutuja graafiku, kus huvialuseks argumenttunnuseks on vabaliige.
- `variable` – võtab väärtuseks sõnede hulga määramaks argumenttunnused, millele lisatud-muutuja graafik genereeritakse
- `labels`, `id.method`, `id.n`, `id.cex`, `id.col` – kirjeldatud peatükis 2.3.1.

- `ellipse` – tõese väärtuse korral genereerib joonisele ellipsi, mis kujutab, kuhu vaatlused on kõige tihedamini sattunud.

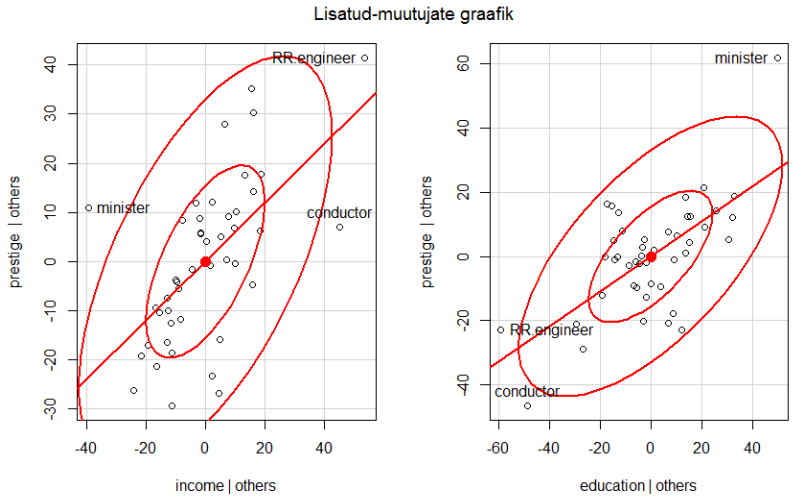
Joonise 16 genereerimiseks on kasutatud paketi „car“ funktsiooni `avPlots()`, kus andmestikuks ja regressioonmudeliks on kasutatud peatükis 2.3.1 kirjeldatud mudelit. R-i koodi joonise genereerimiseks võib leida lisast 5.

Kuna funktsioon vaikimisi andmesilte ei lisa on joonisel 16 identifitseerimismeetodil ära märgitud nende vaatluste kohta käivad punktid, mis töö autori poolt tundusid teistest rohkem isoleeritud olevat. Nagu parempoolsel graafikul näha, paistab vaatlus „minister“ väga selgelt silma sellega, et paikneb teistest vaatlustest eemal. See annab alust arvata, et tegemist on erindiga, mida kinnitas ka Cook'i *D* väärtus joonisel 15. Cook'i *D* järgi aga ei paistnud silma vaatlused „conductor“ ja „RR.engineer“, kuid autori arvates tundusid nende vaatluste kohta käivad punktid olevat isoleeritud. Mõlemad vaatlused osutusid erinditeks ka joonise 14 järgi.



Joonis 16. Lisatud-muutujate graafik andmestikule "Duncan"

Isoleeritud punktide määramine võib olla subjektiivne. Abi võib leida ellipsi genereerimisest joonisele, mis näitab, kuhu piirkonda on punktid kõige tihedamini sattunud. Ellipsiga lisatud-muutuja graafik on kujutatud joonisel 17. R-i kood genereerimiseks paikneb lisas 5.



Joonis 17. Lisatud-muutujate graafik andmestikule "Duncan" lisatud ellipsiga

Joonisel 17 on identifitseerimismeetodil ära sildistatud kõik punktid, mis paiknevad suuremast ellipsist väljaspool. Kuigi varem ei paistnud vaatluste „conductor“ ja „RR.engineer“ isoleeritus parempoolselt jooniselt välja, siis ellips selgelt eristab need teistest. See annab veelgi alust arvata, et tegemist on mõjusate vaatlustega. Sellise konkreetse joone tõmbamine isoleeritud ja mitteisoleeritud punktide vahel aitab vähendada uurija subjektiivset hinnangut visuaalset analüüsi sooritades ning aitab kergemini erindeid üles leida.

Kokkuvõttes ei tundu mõjusate vaatluste leidmine paketi „car“ abil kõige efektiivsem, kui võrrelda graafikuid nendega, mida suudab genereerida statistikaprogramm SAS. Kuigi SASi puhul on tegemist tasuliste teenustega ning pakett „car“ on osa vabavarast R, näitavad tulemused, et antud paketis on siiski veel arenemisruumi. Kõige enam vajaks lihtsustamist andmesiltide lisamise protsess, mille võiks efektiivselt automatiseerida ja keerukuseta kasutajani tuua. Andmesiltideks võib identifitseerimise asemel kasutada ka parameetrit `id.n`, kui on teada, mitut punkti soovitakse ära märkida.

2.4 Lineaarse seose kontroll. Funktsioon `crPlots()`, funktsioon `ceresPlots()`.

Paketi „car“ funktsioon `crPlots()` genereerib osaliste jääkide graafiku ning funktsioon `ceresPlots()` genereerib CERES graafiku lineaarsetele ja üldistatud lineaarsetele mudelitele. [5]

Regressioonmudeli diagnostika vaatenurgast on kõige olulisemad `crPlots()` ja `ceresPlots()` parameetrid järgmised.

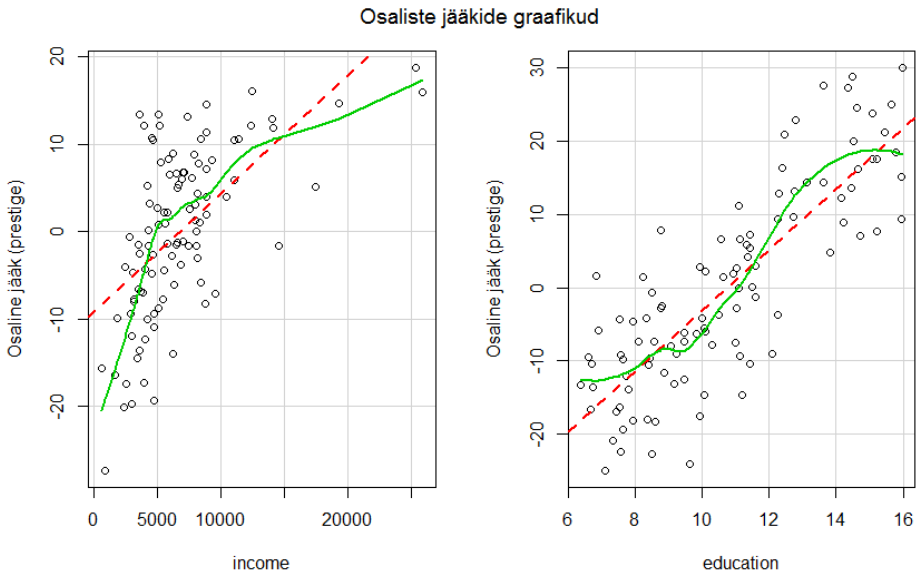
- `model` – võtab väärtuseks regressioonmudeli.
- `terms` – seletatud peatükis 2.1.
- `variable` – võtab väärtuseks sõne või sõnede hulga x-telje argumenttunnuste määramiseks.
- `labels`, `id.method`, `id.n`, `id.cex`, `id.col` – kirjeldatud peatükis 2.3.1.
- `line` – tõese väärtuse korral genereerib vähim-ruutude joone.
- `smoother` – tõese väärtuse korral lisab joonisele mitteparameetrilise kõvera.
- `smoother.args` – võtab väärtuseks kõvera tüübi (nt *loess* joon).

Joonisel 18 on funktsiooni `crPlots()` abil genereeritud osaliste jääkide graafikuid andmestikule „Prestige“, kus on üritatud kirjeldada prestiiži kasutades sissetulekut ja haridust. Andmestik on kirjeldatud peatükis 2.1. R-i koodi graafikute genereerimiseks võib leida lisast 6.

Mõlemal graafikul joonisel 18 on katkendjoonega tähistatud vähimruutude joont, rohelise joonega on tähistatud *loess* joont. Joonise põhjal võib öelda, et kui haridus ja prestiiž on omavahel seotud lineaarselt, siis lineaarfunktsioon sissetuleku ja prestiiži suhet kõige paremini ei kirjelda. Seda võib järeldada nii vaatluste kohta käivate punktide kui ka *loess*i joone põhjal, mis on selgelt kumer.

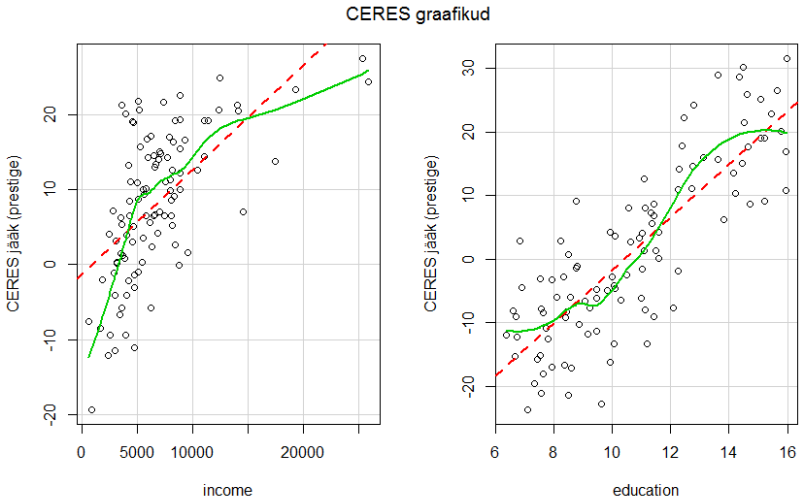
Võrreldes vasakpoolset graafikut joonisel 18 ja joonist 8 (peatükis 2.1), kus on kujutatud jääkide graafik samale seosele, võib kindlalt väita, et rohkem informatsiooni mittelineaarse seose kohta annab edasi osaliste jääkide graafik. Jooniselt 8 võis välja lugeda, et kasutatav lineaarne seos tunnuste vahel ei ole kõige sobivam, kuid joonis 18 informeerib uurijat ka sellest, millise kujuga funktsioon paremini võiks tunnuste vahelist

seost kirjeldada. Antud juhul võiks selleks olla näiteks mingi logaritmfunktsioon. Mudeli, kuhu tunnus „income“ on lisatud mudelisse läbi logaritmfunktsiooni, võib leida lisast 6 (joonis 22). Sellelt on näha, et tõepoolest kirjeldab andmestikku mittelineaarne funktsioon paremini.



Joonis 18. Osaliste jääkide graafikud andmestikule "Prestige"

Joonis 19 kujutab joonisega 18 samal andmestikul ja mudelil koostatud CERES graafikuid, mis on genereeritud funktsiooniga `ceresPlots()`. R-i koodi graafikute genereerimiseks võib leida lisast 6.



Joonis 19. CERES graafikud andmestikule „Prestige“

Nagu näha võib, annavad joonis 18 ja joonis 19 lineaarse seose hindamise vaatenurgast sarnased tulemused. Tuleb aga mainida, et graafikutel pole tegemist samade jääkide arvutamisega, mida võib tunnistada vasakpoolsete graafikute erinevus y-telje kuvatavas vahemikus. Kuna parameetrite väärtused on samad mõlema funktsiooni puhul, siis pole keeruline genereerida nii osaliste jääkide kui ka CERES graafikud, kui üks nendest juba koostatud on. Informatsioon, mida on võimalik erinevate funktsioonide kasutamisel saada, võib erineda, mistõttu on kasulik alati kontrollida mõlemat funktsioon, et mitte teha valesid järeldusi.

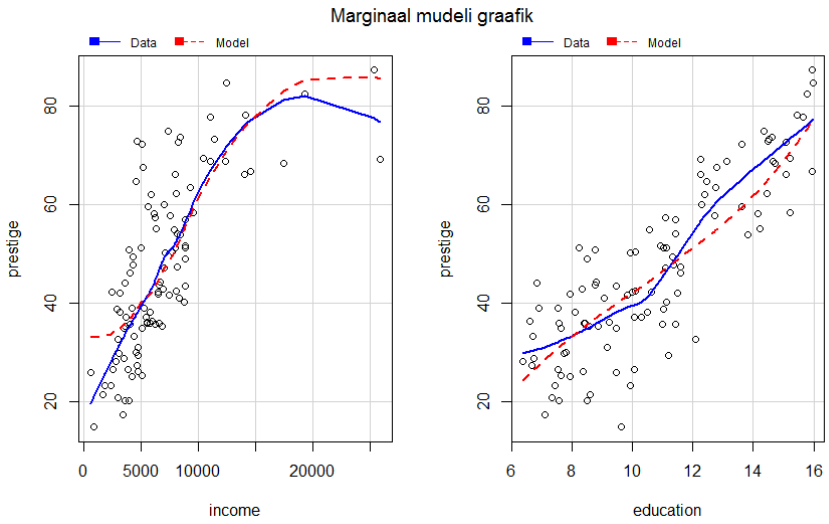
2.5 Mudeli headus. Funktsioon `marginalModelPlots()`.

Paketi „car“ funktsioon `marginalModelPlots()` genereerib graafiku, kus x-teljeks on argumenttunnuse väärtus ja y-teljeks on funktsioontunnuse väärtus. Funktsioon lisab graafikule punkti iga vaatluse kohta. Lisaks lisab funktsioon kaks kõverat: ühe joone tähistamaks mudeli prognoosi, teise joone tähistamaks mitteparameetrilist ennustust funktsioontunnuse keskväärtuse kohta konkreetse argumenttunnuse väärtuse korral (põhineb andmetel). [5]

Regressioonmudeli diagnostika eesmärgil on funktsiooni `marginalModelPlots()` kõige olulisemad parameetrid järgmised.

- `model` - võtab väärtuseks regressioonmudeli.
- `terms` – seletatud peatükis 2.1.
- `fitted` – tõese väärtuse korral lisab graafiku kogu mudeli prognoosi kohta.
- `sd` – tõese väärtuse korral lisab graafikule standardhälve jooned.
- `smoother` – tõese väärtuse korral lisab joonisele mitteparameetrilise kõvera.
- `smoother.args` – võtab väärtuseks kõvera tüübi (nt *loess* joon).
- `groups` – võtab väärtuseks grupi indikaatori nimekirja. Parameetri väärtustamisel kuvatakse jäägid spetsiifilises grupis erineva värvi või sümboliga.
- `labels`, `id.method`, `id.n`, `id.cex`, `id.col` – kirjeldatud peatükis 2.3.1.

Joonisel 20 on kaks graafikut, mis on genereeritud kasutades funktsiooni `marginalModelPlots()`. Andmestikuks on „Prestige“ ning selle põhjal saadud regressioonmudel on kirjeldatud peatükis 2.1. R-i koodi graafikute genereerimiseks võib leida lisast 7.



Joonis 20. Marginaal mudeli graafikud andmestikule "Prestige".

Joonisel 20 on katkendjoonega kujutatud mudeli prognoosi ning pideva joonega kujutatud joont, mis mitteparameetriliselt on hinnanud prestiiži keskväärtust konkreetse sissetuleku ja hariduse väärtuse korral.

Joonise 20 põhjal võib väita, et mudel peegeldab tegelikkust suhteliselt hästi. Erinevused mudeli ja andmestiku vahel võivad tekkida väga väikeste ja väga suurte sissetulekute korral, mil mudel võib ameti prestiiži valesti hinnata. Lisast 7 võib leida ka marginaal mudeli graafiku (joonis 23) samal andmestikul, kus on kasutatud peatükis 2.4 saadud parandatud mudelid. On näha, kui tunnus „income“ siduda logaritmfunktsiooniga, siis saadud marginaal mudeli graafikul kattuvad katkend- ja pidev joon rohkem ehk mudel kirjeldab veelgi paremini tegelikkust.

Üldiselt tundub, et funktsioon `marginalModelPlots()` suudab väga hästi visualiseerida selle, kui hea või halb mingi mudel on. Arvesse ei tohiks võtta ainult marginaal mudeli graafikuid, sest sama andmestiku ja mudeli peal on eelnevates peatükkides kontrollitud ka lineaarse seose olemasolu ja parema mudeli saamiseks tuleks proovida mittelineaarset mudelit.

Kokkuvõte

Käesoleva bakalaureusetöö esimeses osas tutvuti regressioonmudeli diagnostika sooritamise võimalustega läbi graafikute. Täpsemalt kirjeldati regressioonmudeli eelduste kontrollimist läbi jääkide hajuvusgraafiku ja kvantiilide graafiku. Seejärel tutvuti ka meetoditega tuvastada erindeid läbi visuaalse analüüsi. Erindeid aitas tuvastada vaatlustele vastavate *hat-value*, Cook'i D ja Studenti jääkide visualiseerimine hajuvusgraafikul ning lisatud-muutujate graafik. Kontrollimaks, kas tunnuste vahel võib esineda mittelineaarne seos, tutvuti osaliste jääkide graafiku ja CERES graafikuga. Mudeli headuse hindamiseks tutvuti marginaal mudeli graafikuga. Iga graafiku puhul toodi kirjandusest näiteid ning selgitati, mida antud graafikutelt on võimalik lugeda ja mida järeldada ei tohiks.

Bakalaureusetöö teises osas anti ülevaade R-i statistikapaketis „car" graafikuid genereerivatest funktsioonidest `residualPlots()`, `qqPlot()`, `influencePlot()`, `infIndexPlot()`, `avPlots()`, `crPlots()`, `ceresPlots()`, `marginalModelPlots()`. Iga funktsiooni juures kirjeldati regressioonmudeli diagnostika vaatenurgast olulisi parameetreid ja nende võimalikke väärtusi; toodi näide iga funktsiooni poolt genereeritud graafikust kasutades andmetena paketis „car" olevaid andmestikke. Lisaks tõlgendati saadud graafikuid töö esimeses osas kirjeldatud metoodikaga. Kui funktsiooni graafikud ja vastavad näitegraafikud kirjandusest erinesid edastatava informatsiooni poolest, siis hinnati vastavat erinevust diagnostika sooritamise vaatenurgast.

Üldiselt osutusid paketi „car" funktsioonide poolt genereerid graafikud väga headeks diagnostika graafikuteks. Kasutatud funktsioone on kerge rakendada, sest parameetrid on lihtsasti mõistetavad ja soovitud tulemust on lihtne saavutada. Kui graafik sisaldab mitut joont, siis jooned genereeritakse nii värvi kui viisi poolest (katkendjoon vs pidev joon) erinevalt, et visualiseeritu oleks mõistetavam. Keerukaks osutus graafiku punktidele andmesiltide lisamine, mis üldjuhul pole täielikult automaatne. See muudab erindite leidmise tülikamaks, võrreldes mõne muu statistikaprogrammi kasutamisega (näiteks SAS).

Igal juhul tuleb kasuks paketi „car" kasutamine regressioonmudeli diagnostika sooritamiseks, sest pakutavad funktsioonidest on efektiivsed, neist on kerge aru saada ning need on kõigile kättesaadavad.

Kasutatud kirjandus

[1] E. Käärik (2014). *Andmeanalüüs II. Loengukonspekt.* Tartu: Tartu Ülikool, matemaatika ja statistika instituut.
Saadaval: http://dspace.ut.ee/bitstream/handle/10062/35401/AndmeanaluusII.pdf (vaadatud 18.04.2016)

[2] S. Weisberg (2005) *Applied Linear Regression. Third Edition.* Kirjastus Wiley.

[3] T. P. Ryan (2009) *Modern Regression Methods. Second Edition.* Kirjastus Wiley.

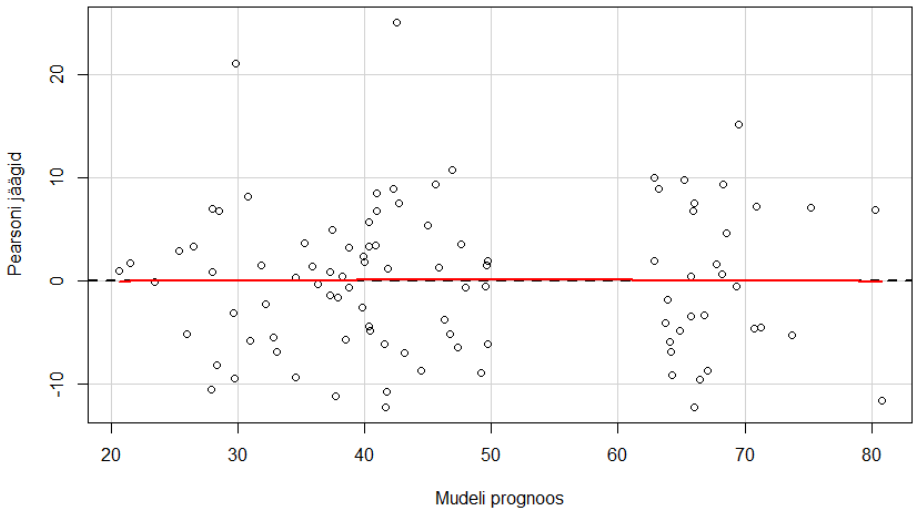
[4] W. Jacoby (2005) *Regression III: Advanced Methods. Lecture 11.* Michigan: Michigan State University.
Saadaval: http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week3/11.Outliers.pdf (vaadatud 10.05.2016)

[5] J. Fox (2016) *Package 'car'. Companion to Applied Regression.*
Saadaval: https://cran.r-project.org/web/packages/car/car.pdf (vaadatud 18.04.2016)

Lisad

Lisa 1.

```
library(car)


m1 <- lm(prestige ~ income, data=Prestige)

residualPlots(m1, xlab=c("sissetulek"), ylab=c("Pearsoni jäägid"), fitted=F)

residualPlots(m1, terms= ~ 1 | type, ylab=c("Pearsoni jäägid"), xlab=c("Mudeli väärtus"))

m2 <- lm(prestige ~ income + income:type + type, data=Prestige)

residualPlots(m2, terms= ~1, ylab=c("Pearsoni jäägid"), xlab=("Mudeli prognoos"))
```



Joonis 21. Mudeli jääkide graafik pärast koosmõju lisamist andmestikul „Prestige“

Lisa 2.

```
library(car)

seed(1)

x<-rchisq(100, df=2)

qqPlot(x, xlab="normaljaotuse kvantiilid")

qqPlot(x, dist="chisq", df=2, xlab="hii-ruut jaotuse kvantiilid", envelope=0.95)

qqPlot(lm(prestige ~ income + education + type, data=Duncan), envelope=.95, simulate=FALSE, xlab="t-jaotuse kvantiilid", ylab="mudeli Studenti jäägid")
```

Lisa 3.

```
library(car)

influencePlot(lm(prestige ~ income + education, data=Duncan), ylab="Studenti jäägid")

influencePlot(lm(prestige ~ income + education, data=Duncan), ylab="Studenti jäägid", id.method="identify")
```

Lisa 4.

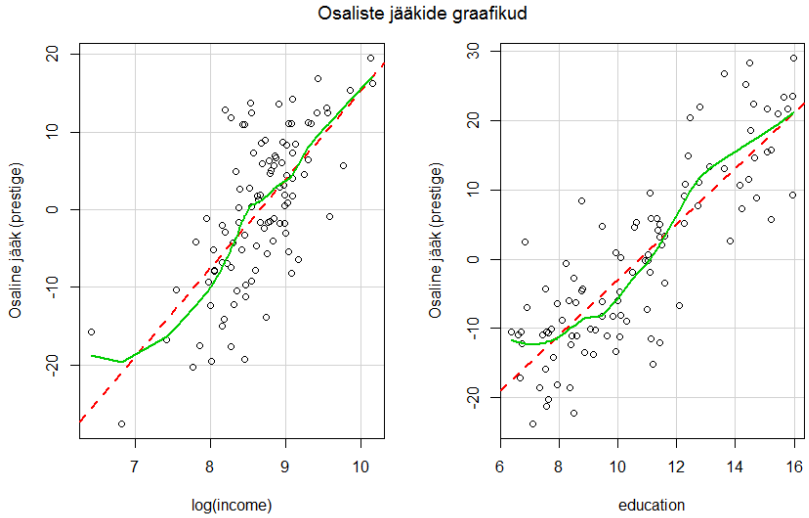
```
library(car)


m1 <- lm(prestige ~ income + education, Duncan)

infIndexPlot(m1, vars="cook", id.method="identify", main="Diagnostika graafik")
```

Lisa 5.

```
library(car)

avPlots(lm(prestige~income+education, data=Duncan), main="Lisatud-muutujate graafik", id.method="identify")

avPlots(lm(prestige~income+education, data=Duncan), ellipse=T, main="Lisatud-muutujate graafik", id.method="identify")
```

Lisa 6.

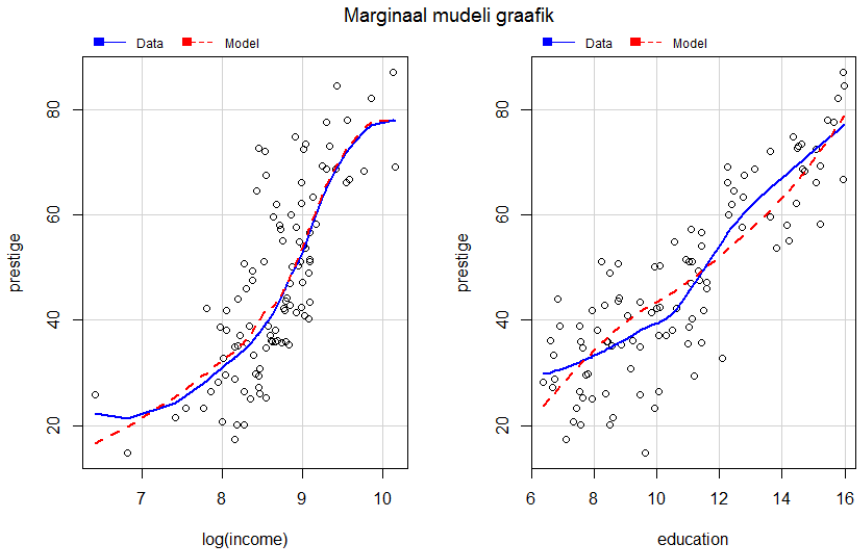
```
library(car)

crPlots(m<-lm(prestige~income+education, data=Prestige), main="Osaliste jääkide graafikud", ylab="Osaline jääk (prestige)")

crPlots(m<-lm(prestige~log(income)+education, data=Prestige), main="Osaliste jääkide graafikud", ylab="Osaline jääk (prestige)")

ceresPlots(m<-lm(prestige~income+education, data=Prestige), main="CERES graafikud", ylab="CERES jääk (prestige)")
```



Joonis 22. Osaliste jääkide graafikud andmestikule "Prestige" pärast mittelineaarse seose lisamist.

Lisa 7.

```
library(car)


c1 <- lm(prestige~income+education, data=Prestige)
mmps(c1,  main="Marginaal mudeli graafik", fitted=F)


c2 <- lm(prestige~log(income)+education, data=Prestige)
mmps(c2,  main="Marginaal mudeli graafik", fitted=F
```



Joonis 23. Marginaal mudeli graafikud andmestikule "Prestige" pärast mittelineaarse seose lisamist.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, _____Risto Korb_____,

(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose _Ülevaade regressioonmudeli diagnostika graafikutest R-i paketis „car“ _____ _____ _____,

(*lõputöö pealkiri*)

mille juhendaja on _____ Anne Selart _____,

(*juhendaja nimi*)

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

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

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

Tartus, **13.05.2016**