TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Annegrete Peek

Üldistatud aditiivne mudel

Bakalaureusetöö (6 EAP)

Juhendaja: Märt Möls, PhD

Tartu 2014

Üldistatud aditiivne mudel

Käesoleva töö eesmärgiks on anda ülevaade üldistatud aditiivsest mudelist. Esimene peatükk käsitleb mudeli kuju ja omadusi. Kirjelduse paremaks mõistmiseks on teises peatükis vaadeldud kolme näidet. Järgnevas peatükis on esitatud ülevaade üldistatud aditiivsete mudelite võimalustest rakendustarkvaras R. Viimases peatükis on kasutatud üldistatud aditiivset mudelit põllulindude andmestikul. Eesmärgiks on teada saada, kas mahepõldudel on rohkem põllulinde kui põldudel, kus keskkonnanõudeid ei ole.
Märksõnad: matemaatiline statistika, andmeanalüüs, statistilised mudelid, statistilised meetodid, mitteparameetrilised meetodid, R (programmeerimiskeel)

Generalized additive model

The purpose of this research is to give an overview of generalized additive model. First section describes form and characteristics of the model. For better understanding three examples of generalized additive models are presented in second section. Next section shows opportunities for using generalized additive model in program R. In last section generalized additive model is used for bird dataset. The purpose is to find out if organic farming fields have more birds than farm fields without any regulations.
Keywords: mathematical statistics,data analysis, statistical models, statistical methods, nonparametric methods, R (programming language)

Sisukord

Sissejuhatus

Regressioonmudelid mängivad olulist rolli andmeanalüüsis, pakkudes prognoose ja klassifitseerimisreegleid ning vahendeid, et mõista erinevate sisendite olulisust. Lineaarne mudel tihti ei sobi sellistes olukordades, kuna päris elus seosed pole sageli lineaarsed. ([3], lk 257) Jerome Friedman ja Werner Stuetzle pakkusid 1981. aastal välja ühe mitteparameertrilise regressioonimudeli - aditiivse mudeli. T. Hastie ja R. Tibshirani kohandasid 1986. ja 1987. aastal aditiivse mudeli tehnoloogia üldistatud lineaarsetele mudelitele ja nimetasid neid üldistatud aditiivseteks mudeliteks. ([4], lk 102) Üldistatud aditiivne mudel on kompromiss lineaarsete ja täielikult mitteparameetriliste mudelite vahel ([5], lk 290).

Käesoleva töö eesmärgiks on anda ülevaade üldisest aditiivsest mudelist ning kasutada seda andmestikul.

Bakalaureusetöös annab autor ülevaate üldistatud aditiivse mudeli kujust, hindamisest ja omadustest ning selle paremaks mõistmiseks toob autor omalt poolt mõned näited. Järgnevalt kirjeldab töö autor mudeli kasutusvõimalusi statistikapaketis R. Töö lõpus kasutatakse üldistatud aditiivset mudelit reaalsete vaatlusandmete korral.

Töö on kirjutatud tekstitöötlusprogrammiga LaTeX. Näidete koostamiseks ja saadud tulemuste graafiliseks esitamiseks on kasutatud statistikapaketti R.

Kasutatud allikatele on töös viidatud nurksulgude abil. Esimene pool näitab allika numbrit töö lõpus asuvas kirjanduse loetelus ja teine pool lehekülge või lehekülgi, kus viidatud faktist juttu on.

Autor tänab Riho Marja põllulindude andmete kasutamise loa ja lektor Märt Mölsi arvukate paranduste ja täienduste eest.

1 Üldistatud aditiivne mudel

1.1 Lokaalne regressioon

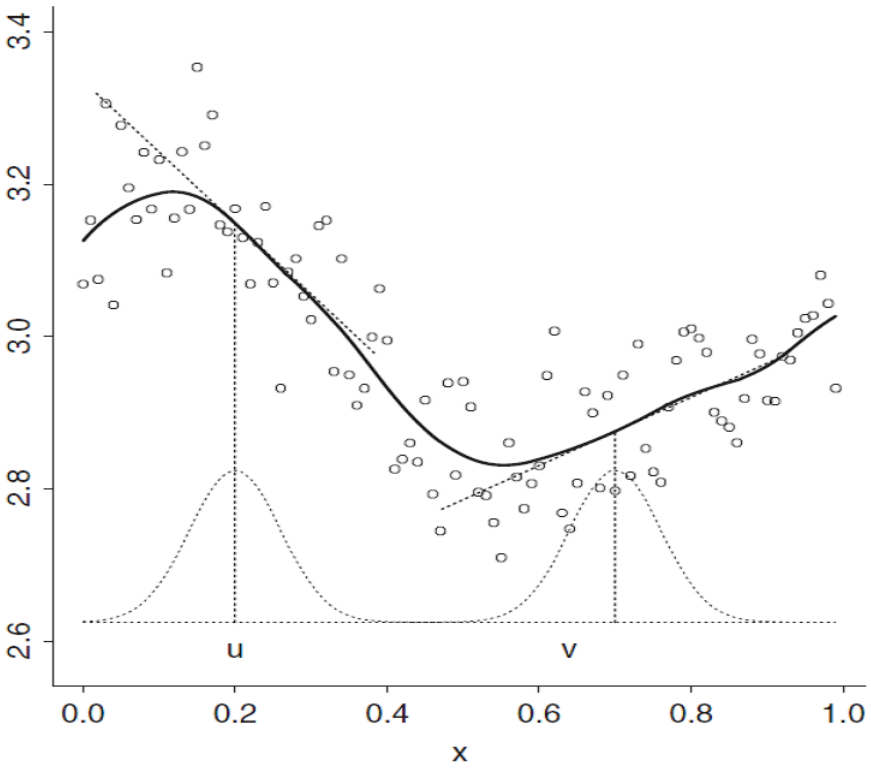
See alampeatükk põhineb autorite D. Ruppert, M. Wand ja R. Carroll ([9], lk 84-86) teosel. Üks kõige populaarsemaid hajuvusdiagrammi silumise meetodeid on lokaalne regressioon. Lokaalse regressiooni ideed illustreerib joonis 1.1. Regressioonsirge paiknemist punktis u hinnatakse nii, et vaatlused, mille x tunnuse väärtus on lähedal väärtusele u, saavad suurema kaalu (regressioonisirge hinnatakse kaalutud vähimruutude meetodil). Regressioonsirge hindamiseks punktides $x = u$ ja $x = v$ kasutatud kaalud on näidatud ära joonise allosas. Kaalud leitakse kasutades tuumafunktsiooni $K(x)$. Vaatluse x_i kaal regressioonsirge hindamisel punktis u valitakse proportsionaalselt funktsiooni $K(u-x_i)$ väärtustele. Tuumafunktsioonina kasutatakse enamasti keskväärtusega 0 tihedusfunktsiooni. ([8], lk 100) Hinnang kohas $x = v$ on leitud samamoodi ja on ka välja toodud joonisel 1.1. Kui seda protseduuri rakendada üle kogu x määramispiirkonna, saame pideva kõverjoonelise tulemuse.

1.2 Aditiivne mudel

Käesolev alampeatükk põhineb Sam Efromovichi ([1], lk 245-249) teosel. Klassikaline lineaarse regressiooni mudel

$$Y = f_L(x_1, \ldots, x_d) + \varepsilon = \beta_0 + \sum_{k=1}^{d} \beta_k x_k + \varepsilon$$

eeldab, et regressioonifunktsioon on nii lineaarne kui ka aditiivne argumenttunnuste suhtes. Kui me kaotame lineaarsuse eelduse, kuid jätame alles adi-



Joonis 1.1: Lokaalne regressioon (allikas: [9], lk 85)

tiivsuse eelduse, saame aditiivse mudeli

$$Y = f_A(X_1, \ldots, X_d) + \varepsilon := \beta + \sum_{k=1}^{d} f_k(X_k) + \varepsilon.$$

Siin on Y uuritav tunnus, $X_1, \ldots, X_d$ argumenttunnused, $f_k(x), k = 1, \ldots, d$, on tundmatud ühe argumendiga funktsioonid ja ε mudeli juhuslik viga. Eesmärgiks on hinnata regressioonifunktsiooni f_A ja selle aditiivseid osasid f_k, kui hindamine põhineb n sõltumatul ja samast jaotusest realisatsioonil $\{(Y_l, X_{1l}, \ldots, X_{dl}), l = 1, 2, \ldots, n\}$. Hindamisprotessi vaatleme lähemalt peatükis 1.4. Kui regressioonimudeli aluseks on aditiivne mudel, saab mudeli komponente suhteliselt hästi hinnata isegi siis, kui valimi suurus on suhteliselt väike. Ka siis kui regressioonimudeli aluseks ei ole aditiivne mudel, võib aditiivne mudel anda aimu tegeliku mudeli ligikaudsest kujust.

1.3 Üldistatud aditiivne mudel

Aditiivne mudel üldistab lineaarset mudelit, kuna lubatakse mittelineaarseid funktsioone iga argumenttunnuse jaoks, samas jääb alles aditiivsuse eeldus. Üldistatud aditiivne mudel on kompromiss lineaarsete ja täielikult mitteparameetriliste mudelite vahel. ([5], lk 287-290)
Üldistatud lineaarse mudeli kuju on

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k, \tag{1.1}$$

kus $\beta_0, \beta_1, \ldots, \beta_k$ on mudeli tundmatud parameetrid, $X_1, X2, \ldots, X_k$ on seletavad tunnused, μ on uuritava tunnuse keskväärtus ja g on seosefunktsioon, mille kuju sõltub tavaliselt uuritava tunnuse jaotusest. Seosefunktsiooniks nimetatakse mingit funktsiooni uuritava suuruse keskväärtusest ja tavaliselt tähistatakse $\eta := g(\mu)$. Mõned näited seosefunktsioonidest:

- $g(\mu) = \mu$ on identsusseos, mida kasutatakse lineaarsete ja aditiivsete mudelite korral,
- $g(\mu)$ =logit(μ) on *logit* seosefunktsioon, mida kasutatakse binaarsete juhuslike suuruste keskväärtuse modelleerimisel,
- $g(\mu) = \log(\mu)$ on *log* seosefunktsioon, mida kasutatakse Poissoni jaotuse korral. ([7], lk 107)

Kui loobume valemis 1.1 lineaarsuse nõudest η ja X_i vahel, siis avaldub üldistatud aditiivne mudel

$$g(\mu) = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p), \tag{1.2}$$

kus f_j on sile (mitteparameetriline) funktsioon ([3], lk 257). Siledaks funktsiooniks nimetatakse funktsiooni, mille määramispiirkonna iga punkt on sile-

duspunkt. Sileduspunktiks nimetatakse ühe muutuja funktsiooni f argumendi x väärtust, mille korral on täidetud tingimus

$$\lim_{|h|\to 0} \frac{|f(x+h) + f(x-h) - 2f(x)|}{|h|} = 0$$

([6], lk 227). Kuna hindame iga arugumenttunnuse X_j mõju individuaalselt ja summeerime kõik mõjud, nimetame valemis 1.2 nähtavat mudelit aditiivseks mudeliks ([5], lk 288). Funktsioone f_j hinnatakse paindlikul viisil kasutades algoritmi, mis põhineb hajuvusdiagrammi silujal. Hinnatud funktsioon f_j võib näidata tunnuse X_j võimalikku mittelineaarset mõju, aga kõik funktsioonid f_j ei pea olema mittelineaarsed. ([3], lk 259)

1.4 Aditiivse mudeli hindamine

See peatükk põhineb T. Hastie ja R. Tibshirani ([4], lk 87-91) teosel. Kõige üldisem meetod aditiivse mudeli hindamiseks lubab meil igat funktsiooni $f_j(X_j)$ hinnata suvalise silujaga. Selles töös kasutame silujana lokaalset regressiooni, kuid võib kasutada ka teisi silujaid nagu näiteks splaine. *Backfitting* algoritm on üldine algoritm, mis võimaldab hinnata aditiivset mudelit.

Algoritm 1.1. *Backfitting* algoritm ([3], lk 260)

1. Määrame: $\hat{\alpha} = \frac{1}{N}\sum_1^N y_i; \hat{f}_j = 0, \forall j.$

2. Tsükkel: $j = 1, 2, \ldots, p, 1, 2, \ldots, p, \ldots,$

$$\tilde{f}_j \leftarrow S_j[\{y_i - \hat{\alpha} - \sum_{k\neq j} \hat{f}_k(x_{ik})\}_1^N], \tag{1.3}$$

   kus S_j tähistab hajuvusdiagrammi silujat, mis hindab seost $y_i - \hat{\alpha} - \sum_{k\neq j} \hat{f}_k(x_{ik})$ ja x_i vahel, indeks i, $i = 1, \ldots, N$ määrab vaatluse ning

indeksid j ja k tähistavad tunnuste järjekorranumbreid.

$$\hat{f}_j \leftarrow \tilde{f}_j - \frac{1}{N}\sum_{i=1}^{N} \tilde{f}_j(x_{ik}), \tag{1.4}$$

kuni funktsioon $\hat{f}_j$ muutus on väiksem kui etteantud lävend.

Kuna soovime, et funktsioonid oleksid hinnatud samaaegselt, on loogiline hinnata iga tunnuse mõju individuaalselt. Eemaldame kõikide teiste tunnuste mõjud y-le, leiame y-i osajäägi ja siis hindame tunnuse x_j mõju y-i osajäägile. Kuna me ei oska kohe hinnata korrektselt teiste tunnuste mõju y-le, siis tuleb kasutada iteratiivset hindamisprotsessi.
Soovime, et mudelis oleks ainult üks vabaliige ja selleks peame kasutama valemit 1.4. Valem 1.4 tagab, et iga funktsiooni f_j keskmine on null ja sellega eemaldab iga funktsiooni vabaliikme.

1.5 Üldistatud aditiivse mudeli hindamine

See peatükk põhineb T. Hastie ja R. Tibshirani ([4], 140-141) teosel. Kui tegemist on üldistatud aditiivse mudeliga, siis lisaks *backfitting* algoritmile tuleb kasutada ka lokaalskooringu protseduuri, mis on esitatud algoritmis 1.2. Kui seosefunktsiooniks on identsusseos, siis $z_i = y_i$ ja protseduur on lihtsalt y_i iteratiivne kaalutud aditiivne hinnang, mil kaalud muutuvad ainult välimises tsüklis. Kui vead on normaaljaotusega, siis kaalud ei muutu ning protsetuur on lihtsalt ühekordne aditiivne hinnang.

Algoritm 1.2. Lokaalskooringu protseduur (*The local scoring procedure*)

1. Määrame: $\alpha = g(\sum_1^n \frac{y_i}{n})$; $f_1^0 = \ldots = f_p^0 = 0$.

2. Tsükkel: Moodustame kohandatud sõltuva muutuja

$$z_i = \eta_i^0 + (y_i - \mu_i^0)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0 \tag{1.5}$$

kus $\eta_i^0 = \alpha^0 + \sum_{j=1}^p f_j^0(x_{ij})$ ja $\mu_i^0 = g^{-1}(\eta_i^0)$.

Moodustame kaalud

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0^2 (V_i^0)^{-1}, \tag{1.6}$$

kus V_i^0 on Y dispersioon punktis μ_i^0 ([4], lk 138).

Hindame kaalutud aditiivse mudeli z_i jaoks, leiame hinnatavad funktsioonid f_j^1, aditiivse argumenttunnuse η^1 ja hinnatud väärtused μ_i^1.

Arvutame koondumiskriteeriumi

$$\Delta(\eta^1,\eta^0) = \frac{\sum_{j=1}^p \|f_j^1 - f_j^0\|}{\sum_{j=1}^p \|f_j^0\|}. \tag{1.7}$$

$\|f\|$ loomulik kandidaat on $\|$ f $\|$, mis on vektori pikkus ja koosneb n punktis leitud f hinnangutest.

3. Korrata sammu 2., mil η^0 asendame η^1-ga, kuni $\Delta(\eta^1,\eta^0)$ on väiksem kui etteantud lävend.

1.6 Mudeli testimine ja mudeli valik

See alapeatükk põhineb T. Hastie ja R. Tibshirani ([4], lk 155-158) teosel. Tähistame uuritava tunnuse keskväärtuste vektorile μ mudeli abil leitud hinnangut sümboliga $\hat{\mu}$ ja selle hälbimuseks nimetatakse

$$D(y;\hat{\mu}) = 2\{l(\hat{\mu}_{\max};y) - l(\hat{\mu};y)\}, \tag{1.8}$$

kus $\hat{\mu}_{\max}$ on parameetri väärtus, mis maksimeerib logaritmilist tõepärafunktsiooni $l(\mu;y)$ väga rikkaliku mudeli korral (näiteks mudel, kus iga vaatluse

jaoks on eraldi parameeter). Vahel vaadatakse hälbimust ka kui $\hat{\eta}$ funktsiooni. Üldistatud lineaarsete mudelite korral kasutatakse hälbimust nii mudeli sobivuse kirjeldamiseks kui ka mudelite võrdlemiseks. Ka mitteparameetrilisete ja aditiivsete mudelite korral kasutatakse hälbimust mudelite kvaliteedi ja mudelite erinevuste hindamiseks. Mida väiksem on hälbimus, seda parem on mudel ([7], lk 109). Kuid üldistatud aditiivse mudeli jaoks ei ole hälbimuse jaotust ilmutatud kujul leitud, kuid arvatakse, et nullhüpoteesi kehtides on hälbimuse jaotust võimalik lähendada χ^2-jaotusega (või F-jaotusega). Kuna χ^2-jaotusega juhusliku suuruse korral on juhusliku suuruse keskväärtus võrdne vabadusastmete arvuga, siis kasutatakse hälbimuse vabadusastmete arvu df^{err} määramisel hälbimusele leitud lähendi keskväärtust. Kuigi üldistatud aditiivse mudeli hälbimus ei ole täpselt χ^2-jaotusega, isegi mitte asümptootiliselt, on simulatsioonid (kaasa arvatud töö autori enda simulatsioonid) näidanud, et χ^2-jaotus on siiski kasutuskõlblik lähend mudelite sõelumiseks (vt joonis 1.2).

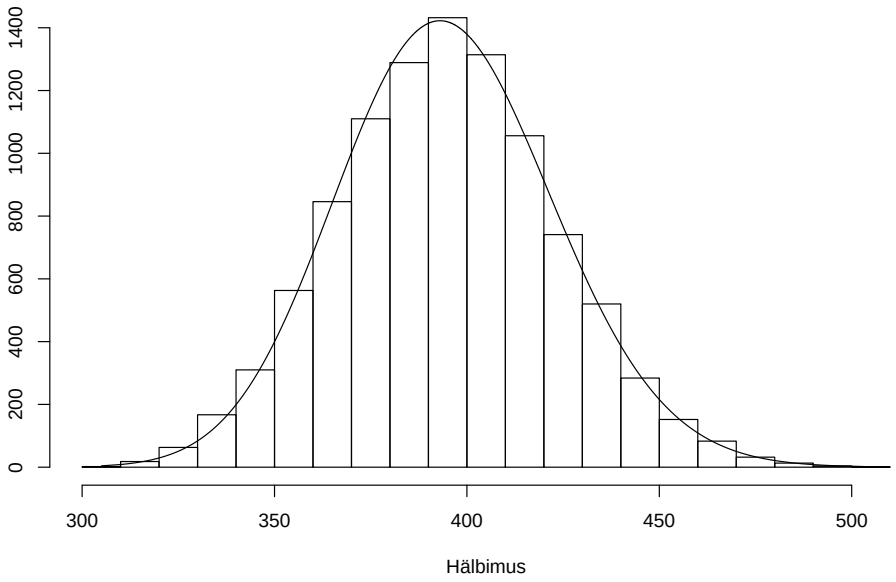
Hälbimuse leidmiseks kasutame asümptootilist lähendit

$$D(y;\mu) \approx (y-\hat{\mu})^T A^{-1}(y-\hat{\mu}) \approx (z-\hat{\eta})^T A(z-\hat{\eta}), \tag{1.9}$$

kus A on hinnatud informatsioonimaatriks. Leides viimase avaldise keskväärtuse, saavad Hastie ja Tibshirani ligikaudse valemi vabadusastmete arvu leidmiseks

$$df^{err} = n - \mathrm{tr}(2R - R^T ARA^{-1}), \tag{1.10}$$

kus R on selline kaalutud aditiivse hinnangu operaator, et $\hat{\eta} = Rz$. Kui mudel on korrektne, siis $E(D) \approx df^{err}\phi$. Oletame, et $\hat{\eta_1}$ ja $\hat{\eta_2}$ erinevad ainult ühe tunnuse poolest, näiteks tunnuse X_j mitteparameetrilise mõju poolest. Kui



Joonis 1.2: Selle simulatsiooni jaoks iga mudel modelleeriti 400 vaatluse põhjal ja modelleeriti 10 000 mudelit. Hälbimuse keskmine oli 394,63 ja valitud χ^2 jaotuse vabadusastmete arv on 395.

väiksem mudel $\hat{\eta_1}$ on korrektne, siis

$$\begin{aligned} ED(\hat{\eta_1};\hat{\eta_2})/\phi &= E\{D(y;\hat{\eta_1}) - D(y;\hat{\eta_2})\}/\phi \\ &\approx \mathrm{tr}(2R_1 - R_1^T A_1 R_1 A_1^{-1}) - \mathrm{tr}(2R_2 - R_2^T A_2 R_2 A_2^{-1}) \\ &= df^{err}(\hat{\eta_1}) - df^{err}(\hat{\eta_2}) = df_j^{err}. \end{aligned}$$

Kui dispersiooni parameeter ϕ on teada, siis $D(\hat{\eta_2};\hat{\eta_1})$ jaotus on asümptootiliselt ligikaudu $\chi^2_{df_j^{err}}$ jaotusega. Kui dispersiooni parameeter ϕ ei ole teada, siis on võimalik tema jaotust lähendada F-jaotusega.
Üldistatud aditiivse mudeli korral on Hastie ja Tibshirani Akaike informatsioonikriteerimi AIC statistik defineerinud kujul

$$AIC = D(y;\hat{\mu})/n + 2df\phi/n, \tag{1.11}$$

kus vabadusastmete arv (erinevalt eelnevast) leitakse valemiga $df = \mathrm{tr}(R)$, n on vaatluste arv ja ϕ on dispersiooni parameeter. Mida väiksem on AIC statistiku väärtus, seda parem on mudel ([7], lk 109).

2 Näited

2.1 Üks argumenttunnus

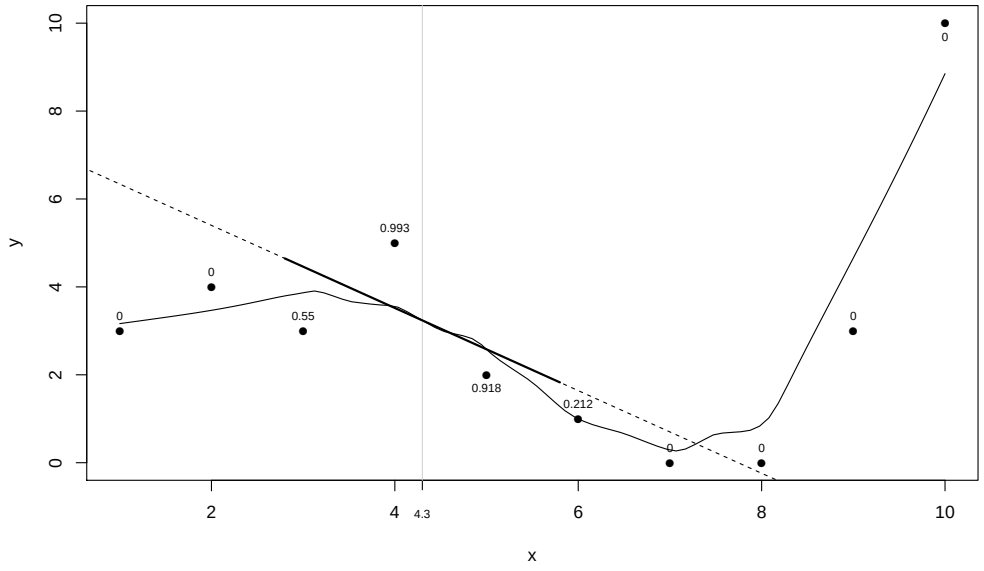
Olgu meil argumenttunnus x ja uuritav tunnus y. Olgu x väärtused 1, 2, 3, 4, 5, 6, 7, 8, 9 ja 10 ning neile vastavad y väärtused 3, 4, 3, 5, 2, 1, 0, 0, 3 ja 10 (vt joonis 2.1). Leiame hinnangu punktis $x = 4{,}3$. Regressioonisirge leidmiseks kasutame punktile $x = 4{,}3$ nelja lähimat x tunnuse väärtusega punkti. Seega vaatlustele, kus x on 3, 4, 5 ja 6, leiame kaalud, millega nad regressioonisirget mõjutavad. Järgnevates näidetes kasutatakse kaalude leidmiseks nn *tricubic*-tuumafunktsiooni:

$$k_i = \max\left[0, \left(1 - \left(\frac{\mid x - x_i \mid}{r}\right)^3\right)^3\right], \tag{2.1}$$

kus i-inda vaatluse kaal on $\frac{k_i}{\sum_{j=1}^{N} k_j}$, N on vaatluste arv, x on punkt, kus regressioonisirget arvutame, x_i on i-nda vaatluse x-tunnuse väärtus ja r on x kaugus mõõdetud x-tunnuse suunas kõige lähemast punktist, mida regressioonisirge leidmisel ei kasutatud. Punkti $x = 4{,}3$ jaoks on $r = |2-4{,}3| = 2{,}3$. Tuumafunktsiooni väärtused on vastavalt 0,550, 0,993, 0,918 ja 0,212 ning vastavad kaalud 0,206, 0,371, 0,434 ja 0,079. Kuna selles näites on ainult üks muutuja, siis valemist 1.3 saame

$$\tilde{f} \leftarrow S[\{y_i\}_1^{10}]$$

ehk leiame lokaalse regressioonikõvera. Punktis $x = 4{,}3$ leitud regressioonisirge on joonisel kujutatud jämeda joonega. Saadud regressioonsirge väärtus kohal $x = 4{,}3$ määrabki üldistatud aditiivse mudeli väärtuse kohal $x = 4{,}3$. Ühe tunnuse korral üldistatud aditiivne mudel, mis kasutab silujana lokaalset regressiooni, ja tavaline lokaalne regressioonimudel kattuvad.



Joonis 2.1: Ühe argumenttunnusega üldistatud aditiivne mudel

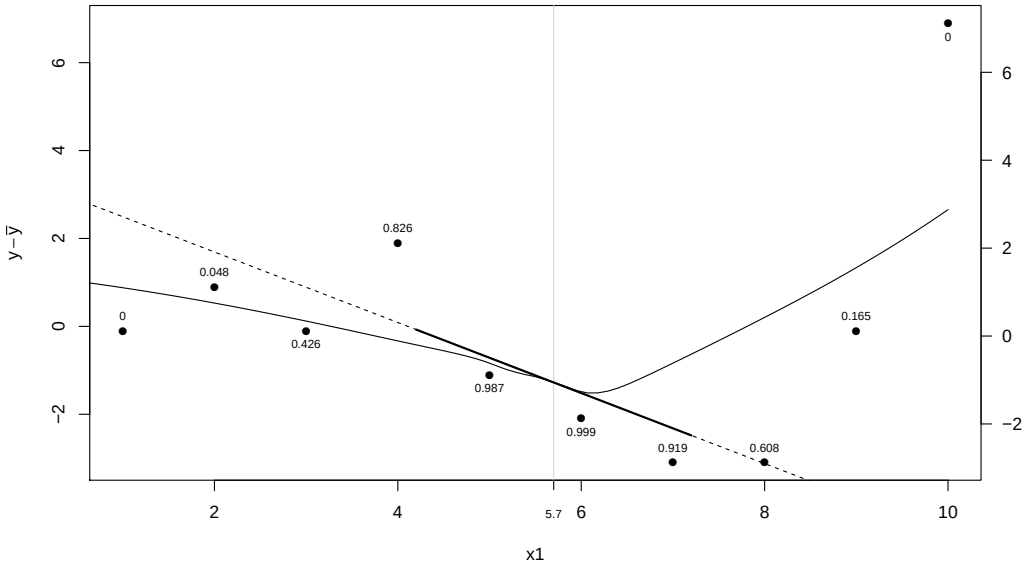
2.2 Kaks argumenttunnust

Olgu meil argumenttunnused $x1$ ja $x2$ ning uuritav tunnus y. Olgu tunnuse $x1$ väärtused 1, 2, 3, 4, 5, 6, 7, 8, 9 ja 10, tunnuse $x2$ väärtused 2, 3, 4, 2, 5, 1, 0,5, 0, 2,5 ja 8 ning neile vastavad y väärtused 3, 4, 3, 5, 2, 1, 0, 0, 3 ja 10. Kuna nüüd on rohkem kui üks argumenttunnus, siis tuleb y-tunnuse väärtusest lahutada y-tunnuse keskmine ehk 3,1. Regressioonisirge leidmiseks kasutame vaadeldavale punktile lähimat kaheksat punkti. Saadud y-tunnuse transformeeritud väärtusi hakkame esmalt hindama $x1$ kaudu. Joonisel 2.2 on $x1$ ja transformeeritud y-tunnuse väärtused. Uurime, mis toimub punktis $x1 = 5{,}7$. Valem 1.3 on algul

$$\tilde{f}_1 \leftarrow S_1[\{y_i - \overline{y}\}_1^{10}].$$

Leiame lähima 8 vaatluse kaalud (vt valem 2.1) ning regressioonisirge punktis $x1 = 5{,}7$. Leitud regressioonisirge on kujutatud joonisel jämeda joonega.

Nüüd korrigeerime leitud hinnangut valemi 1.4 järgi, selleks lahutame temast maha hinnangu keskmise $\hat{f}_1(x1) = \tilde{f}_1(x1) - \overline{\tilde{f}_1(x1)} = \tilde{f}_1(x1) - 0{,}22$. Joonise 2.2 paremal küljel on korrigeeritud hinnangu skaala.



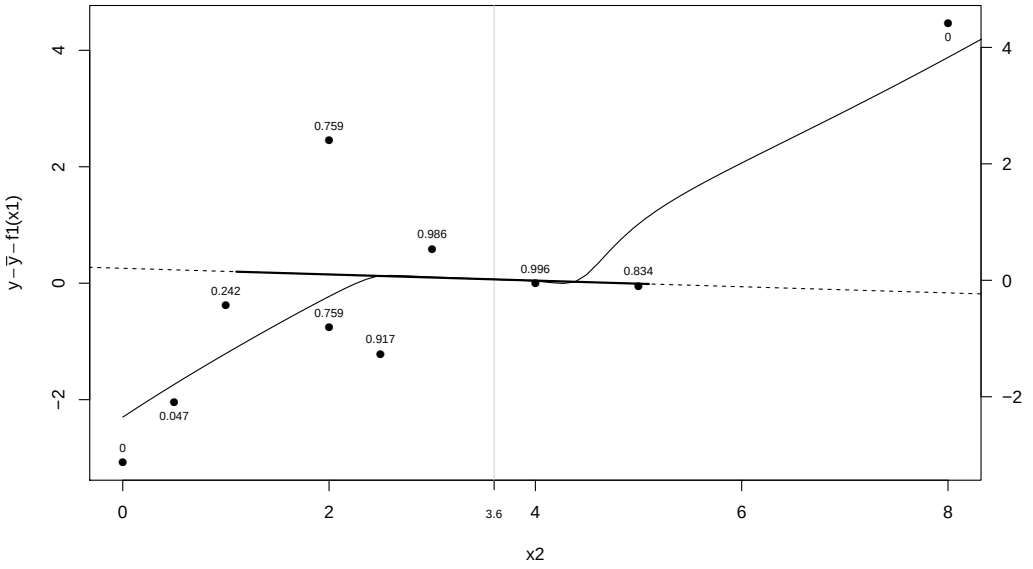
Joonis 2.2: Esialgne seos $x1$ ja y vahel

Hindamaks argumenttunnuse $x2$ mõju tunnusele y eemaldame esmalt tunnuse $x1$ hinnatud mõju. Algsetest y-tunnuse väärtustest lahutame y-tunnuse keskmise ning eemaldame tunnuse $x1$ hinnatud mõju, saame nn y-tunnuse osajäägi. Joonisel 2.3 vaatleme, kuidas hinnata tunnuse $x2$ mõju y-tunnuse osajäägile punktis $x2 = 3{,}6$. Prognoosime $y_i - \overline{y} - \hat{f}_1(x1_i)$ väärtuseid kasutades argumenttunnuse $x2_i$ vaatluseid lokaalse regressiooni abil saamaks funktsiooni $\tilde{f}_2$. Siin esineb valem 1.3 järgmisel kujul

$$\tilde{f}_2 \leftarrow S_2[\{y_i - \overline{y} - \hat{f}_1(x1_i)\}_1^{10}].$$

Leiame kaalud valemi 2.1 järgi ning leiame lokaalse regressiooni. Siin leiame ka korrigeeritud $x2$ mõju hinnangu $\hat{f}_2(x2)$ valemi 1.4 järgi, kus $x2$ mõju

hinnangu keskmine on $-0{,}05$. Leiame uued y-tunnuse osajäägid, mille põhjal hinnata uuesti tunnuse $x1$ mõju. Joonise 2.3 paremal küljel on korrigeeritud hinnangu skaala.

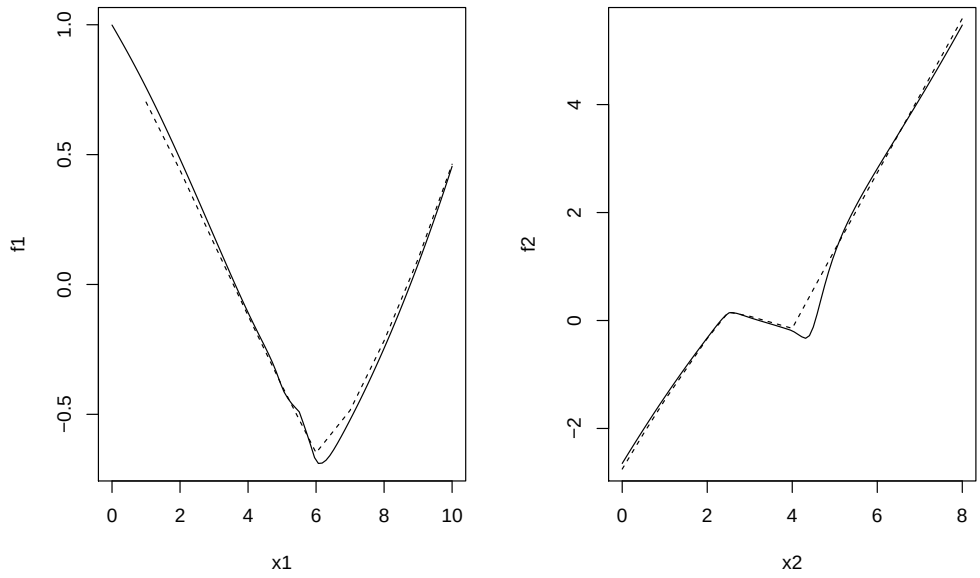


Joonis 2.3: Esialgne seos $x2$ ja y vahel

Eelnimetatud samme tehakse nii kaua kuni eelmise ringi $x1$ ja y vaheline seos erineb väga vähe selle ringi $x1$ ja y vahelisest seosest ning sama tingimus kehtib ka $x2$ ja y seose jaoks. Joonisel 2.4 on näha selle näite lõpptulemus. Samamoodi saab talitada ka siis, kui argumenttunnuseid on kolm või rohkem.

2.3 *Log* seosefunktsioon

Olgu meil argumenttunnused $x1$ ja $x2$ ning uuritav tunnus y, mis on Poissoni jaotusega. Olgu tunnuse $x1$ väärtused 1, 2, 3, 4, 5, 6, 7, 8, 9 ja 10, tunnuse $x2$ väärtused 2, 3, 4, 2, 5, 1, 0,5, 0, 2,5 ja 8 ning y väärtused 1, 4, 3, 5, 1, 4, 1, 0, 3, 5. Tähistme $\hat{\alpha} = \log(\sum_1^{10} \frac{y_i}{10})$, määrame esialgsed $\hat{\eta}_i = \hat{\alpha}$ ja $\hat{\mu}_i = \exp \hat{\eta}_i$.



Joonis 2.4: Lõpptulemus, kus pideva joonega on kujutatud autori hinnangud ja katkendjoonega paketi '`gam`' hinnangud silujatele $\hat{f}_1$ ja $\hat{f}_2$.

Selle näite korral on tuletis

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i}$$

ja seega valem 1.5 avaldub kujul

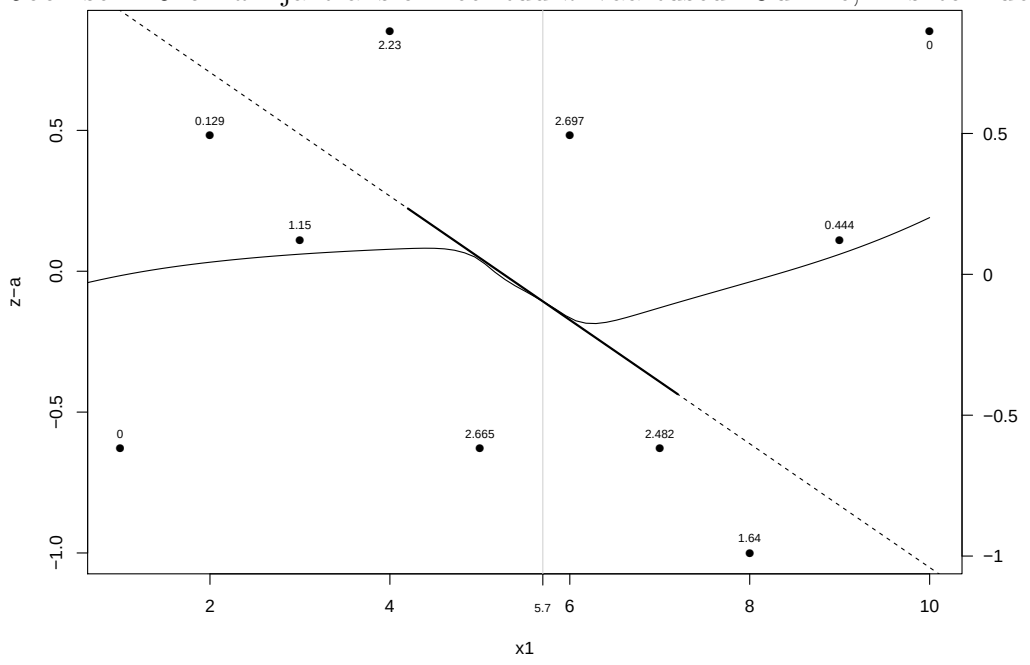
$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}.$$

Valem 1.6 avaldub $w_i = \hat{\mu}_i$, sest

$$\frac{\partial \mu_i}{\partial \eta_i} = \mu_i$$

ja uuritava tunnuse dispersioon on $V = \mu_i$. Nüüd me ei otsi seoseid y ja $x1$ ning $x2$ vahel, vaid y-tunnuse asendame z-ga ja modelleerime $x1$ ja $x2$ mõju z-le. Kuna on rohkem kui üks argumenttunnus, siis tuleb z väärtustest lahutada z-tunnuse keskmine ehk 2,99. Uuritava punkti regressioonisirge arvutamisel kasutame sellele punktile lähimat kaheksat punkti mõõdetud $x1$-tunnuse suunas. Transformeeritud z väärtusi hakkame hindama $x1$ kaudu.

Joonisel 2.5 on $x1$ ja transformeeritud z väärtused. Uurime, mis toimub



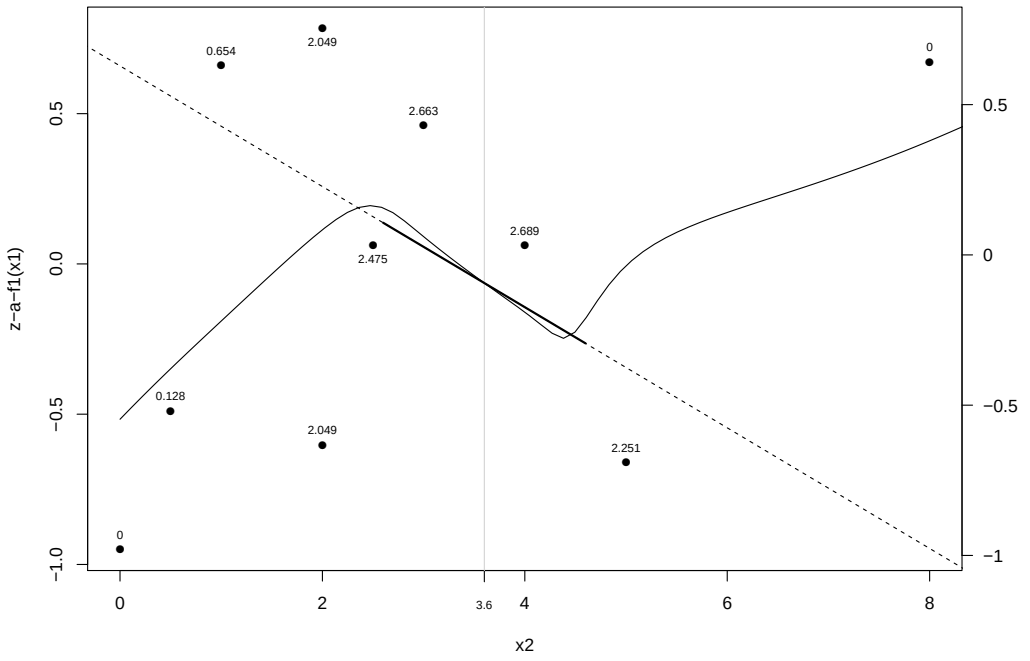
Joonis 2.5: Esialgne seos $x1$ ja z vahel

punktis $x1 = 5{,}7$. Valem 1.3 on algul

$$\tilde{f}_1 \leftarrow S_1[\{z_i - \overline{z}\}_1^{10}].$$

Leiame lähima kaheksa vaatluse kaalud (vt valem 2.1). Kuna igal punktil on nii vaatluse dispersioonist tulenev kaal proportsionaalne w_i-ga kui ka vaatluse kaugusest punktist $x1 = 5{,}7$ tulenev kaal proportsionaalne k_i-ga, siis regressioonisirge hindamisel kasutame kaale, mis on proportsionaalsed $w_i \cdot k_i$-iga. Leiame regressioonisirge, mis on joonisel kujutatud jämeda joonega, punktis $x = 5{,}7$. Leiame lokaalse regressiooni abil regressioonikõvera üle kogu määramispiirkonna $x1$. Nüüd leiame valemi 1.4 järgi korrigeeritud $x1$ mõju hinnangu. Selleks lahutame $x1$ mõju hinnangust $x1$ mõju hinnangu keskmise ehk 0,011. Joonise 2.5 paremal küljel on korrigeeritud hinnangu skaala.

Hindamaks $x2$-tunnuse mõju tunnusele z, tuleb leida z osajäägid, millele



Joonis 2.6: Esialgne seos $x2$ ja z vahel

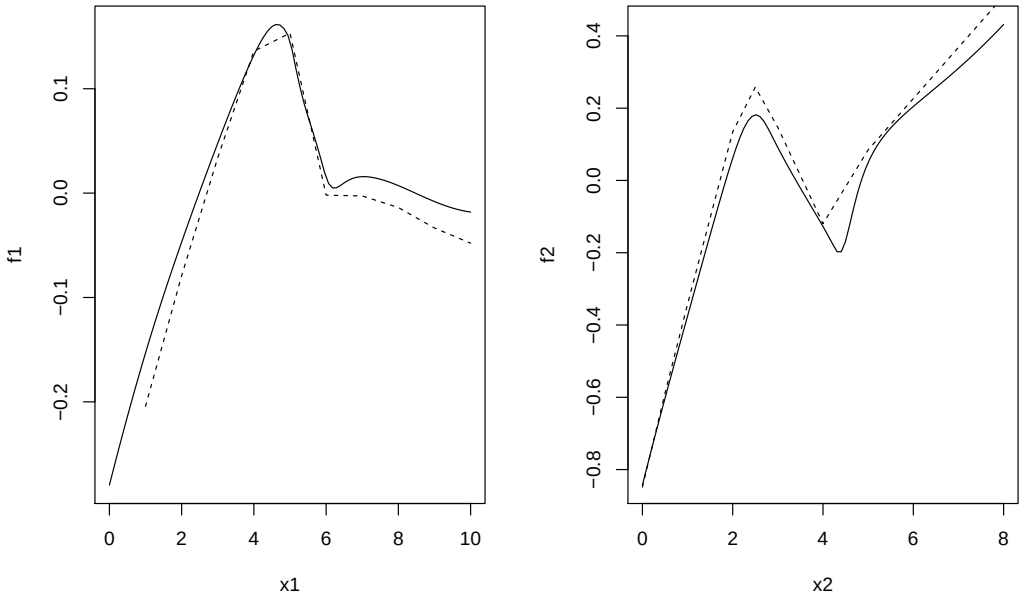
hakkame mõju modelleerima. Algsetest z-tunnuse väärtustest lahutame z-tunnuse keskmise ning eemaldame $x1$-tunnuse hinnatud mõju. Joonisel 2.6 vaatleme, kuidas leida z osajäägi hinnang punktis $x2 = 3{,}6$ argumenttunnuse $x2$ kaudu. Siin esineb valem 1.3 järgmisel kujul

$$\tilde{f}_2 \leftarrow S_2[\{z_i - \overline{z} - \hat{f}_1(x1_i)\}_1^{10}].$$

Saame punktite kaalud proportsionaalselt w_i ja k_i korrutisega ning leiame regressioonisirge. Hindame regressioonikõvera üle kogu määramispiirkonna $x2$. Siin leiame ka korrigeeritud $x2$ mõju hinnangu valemi 1.4 järgi, kus $x2$ mõju hinnangu keskmine on $-0{,}03$. Leiame korrigeeritud z väärtused, mille põhjal hindame uuesti $x1$-tunnuse mõju z osajäägile. Joonise 2.6 paremal küljel on korrigeeritud hinnangu skaala. Järgmisena leitakse uued $\hat{\alpha}$, $\hat{\eta}_i$ ja $\hat{\mu}_i$ väärtused, kus uus

$$\hat{\alpha} = \log\left(\frac{\sum_{i=1}^{10} y_i \mu_i}{\sum_{i=1}^{10} \mu_1}\right).$$

Uute väärtustega leitakse uued $\hat{f}_1$ ja $\hat{f}_2$. Eelnimetatud samme tehakse nii kaua kuni eelmise ringi $x1$-tunnuse mõju z-tunnusele erineb väga vähe selle ringi $x1$-tunnuse mõjust z-tunnusele ning sama tingimus kehtib ka $x2$-tunnuse mõjule. Joonisel 2.7 on näha selle näite lõpptulemus. Samamoodi saab talitada kolme ja enama argumenttunnuse korral.



Joonis 2.7: Silujate $\hat{f}_1$ ja $\hat{f}_2$ lõplikud hinnangud Poissoni jaotusega sõltuva tunnuse korral. Pideva joonega on kujutatud autori hinnang ja katkendjoonega paketi '`gam`' hinnangud.

3 Üldistatud aditiivne mudel R-is

Selles peatükis vaatame lähemalt kahte paketti R-is, mida saame kasutada üldistatud aditiivse mudeli modelleerimiseks. Kõigepealt tutvume Trevor Hastie paketiga `gam` ([2]) ja järgmisena vaatame Simon Wood'i paketti `mgcv` ([12]).

3.1 Pakett `gam`

Üldistatud aditiivse mudeli ingliskeelne nimetus on *generalized additive model* ja siit ka paketi lühend *gam*. Peale lisamooduli `gam` sisselugemist võime näiteks hinnata järgmise mudeli:

```
mudel1=gam(y ~ lo(x1, span=0.8)+s(x2)+x3+factor(x4), family
   = Poisson())
```

Siin näites y on uuritav tunnus ja mudel modelleeritakse nelja argumenttunnse $x1$, $x2$, $x3$ ja $x4$ pealt, kusjuures $x4$ on kvalitatiivne tunnus. Siin `lo(x1,span=0.8)` tähendab, et tunnuse $x1$ mõju y-le modelleeritakse lokaalse regressiooniga, kus igas punktis regressioonisirge leidmiseks kasutatakse lähimaid 80% punktidest. Järgmine osa valemist `s(x2)` tähistab, et tunnuse $x2$ mõju tunnusele y leidmisel kasutatakse splaini. Splainiks nimetatakse tükiti polünomiaalset funktsiooni; lõigul $[a,b]$ määratud funktsiooni, mis teatava alajaotuse $a = x_0 < x_1 < \ldots < x_n = b$ korral on igas vahemikus (x_k,x_{k+1}) esitatav polünoomina (enamasti nõutakse täiendavalt veel funktsiooni ja selle teatavat järku tuletiste pidevust jaotuspunktides x_k)([6], lk 232). Valemi osa `x3` tähendab, et tunnuse $x3$ mõju tunnusele y on hinnatud parameetriliselt lineaarse regressiooni abil. Kuna tunnus $x4$ on kvalitatiivne tunnus, siis see tuleb võtta valemisse faktorina.

Argumendi `family` abil saab ette anda, millisesse jaotuste perre kuulub y-tunnuse jaotus. Vaikimisi väärtuseks on normaaljaotuste pere. Selles näites on jaotustüübiks valitud Poissoni jaotuste pere.

Mudeli headuse kontrollimiseks saab kasutada käsku `anova(mudel1)` ([2], lk 2). See käsk näitab, kas mitteparameetriliselt hinnatud mõju ja parameetriliselt hinnatud mõju erinevus on statistiliselt oluline. Kui p-väärtus on väike, siis pole seos η ja vastava argumenttunnuse vahel lineaarne. Olgu modelleeritud teine mudel veel

```
mudel2=gam(y ~ lo(x1, span=0.8)+x2+x3+factor(x4), family
   = Poisson())
```

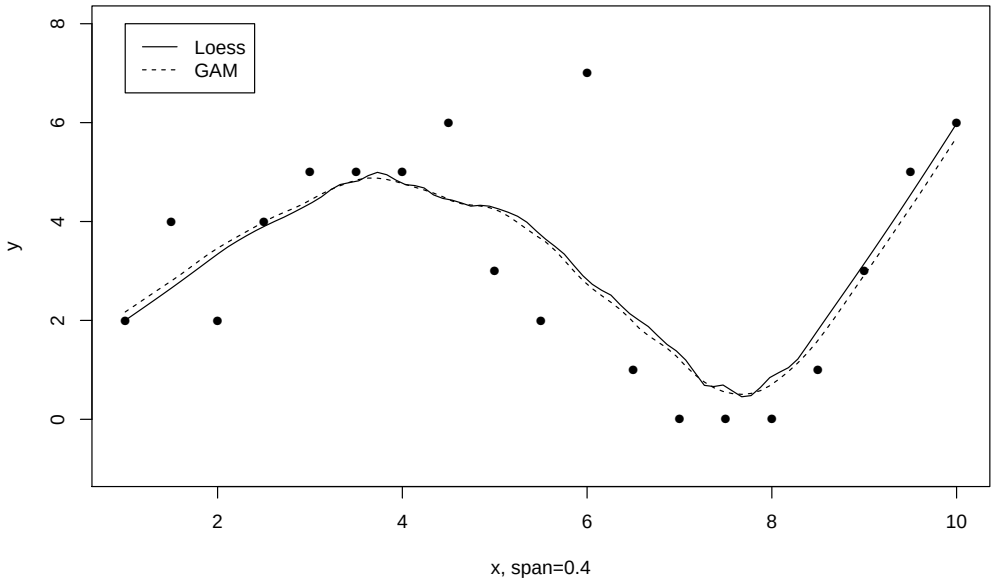
Kui käsus `anova(mudel1, mudel2)` kasutada kahte argumenti, mis mõlemad on üldistatud aditiivsed mudelid (üks erijuht teisest), siis saame teada, kas mudelite erinevus on statistiliselt oluline. Vaikimisi kasutatakse mudelite võrd-lemiseks χ^2 testi, kui sooviksime F-testi, siis kirjutame `anova(mudel1, mudel2, test="F")`. Kui väljastatav p-väärtus on väike, siis rikkam mudel on parem. Antud töös nimetatakse rikkamaks mudeliks mudelit, kus on rohkem tunnuseid või rohkemate tunnuste mõjud on mitteparameetriliselt hinnatud. Siin on rikkamaks mudeliks `mudel1`, sest selles mudelis on tunnuse $x2$ mõju hinnatud mitteparameetriliselt. Kui p-väärtus on suur, siis ei õnnestu tõestada, et vaesem mudel oleks vale ja mõistlik on kasutada lihtsamat mudelit.

Lisaks eeltoodud argumentidele on võimalik `gam` käsule ette anda ka teisi argumente (vt [2], lk 3).

Olgu meil uuritav tunnus y ja argumenttunnus x, mille väärtused on märgitud joonisel 3.1. Kui üldistatud aditiivses mudelis on üks argumenttunnus ja uuritava tunnuse jaotuseks on normaaljaotus, siis mudeli modelleerimisel käsuga

```
gam(y lo(x, span=0.4)
```

peaksime saama sama tulemuse, kui kasutaksime lokaalse regressiooni paketti `loess`. Joonisel 3.1 on pideva joonega tähistatud paketi `loess` modelleeritud mudel ja katkendjoonegaa on tähistatud paketi `gam` modelleeritud mudel. Tulemuseks on kaks veidi erinevat kõverat. Autor arvab, et pakett `gam` ei hinda mõju nii täpselt kui pakett `loess` kuna mitme argumenttunnuste korral võib see võtta liiga palju aega.



Joonis 3.1: Pakettide `loess` ja `gam` võrdlus samal andmestikul, kui paketis `gam` kasutame `lo` käsku.

3.2 Pakett `mgcv`

Üldistatud aditiivse mudeli modelleerimine paketis `mgcv` defineeritud funktsiooni `gam` abil toimub sarnaselt paketile `gam`. Alljärgnev näide hindab kahte aditiivset mudelit:

```
mudel1=gam(y ~ s(x1)+x2+factor(x3), family = Poisson())
```

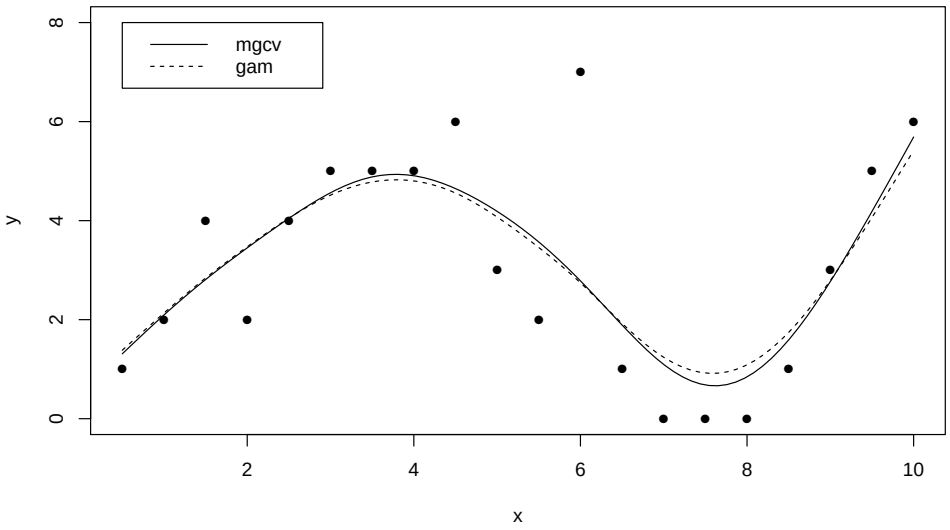
```
mudel2=gam(y ∼ s(x1)+factor(x3), family = Poisson())
```

Ka siin paketis on võimalik mudelile rohkem argumente ette anda (vt [12], lk 32). Erinevalt paketist `gam` ei saa siin valemi argumendis kasutada lokaalset regressiooni. Mitteparameetriliste mõjude hindamiseks saab kasutada ainult splaine ehk `s(x)`. Kui argumenti `family` ette ei anta, siis eeldakse, et uuritava tunnuse jaotus kuulub normaaljaotuste perre.

Ühe mudeli korral käsk `anova(mudel)` käitub samamoodi kui paketis `gam`. Kahe mudeli korral p-väärtuste saamiseks tuleb lisada argument `test` ehk käsk oleks `anova(mudel1,mudel2, test="Chisq")`, kui soovime χ^2 testi, või `anova(mudel1,mudel2, test="F")`, kui soovime F-testi.

Olgu meil uuritav tunnus y ja argumenttunnus x nagu näha joonisel 3.2. Modelleerime nendele andmetele üldistatud aditiivse mudeli kasutades pakette `gam` ja `mgcv`, kus mõlemas on kasutatud splaine. Joonisel 3.2 on katkendjoonega paketi `gam` mudel ja pideva joonega on paketi `mgcv` mudel.

Kuna paketid `mgcv` ja `gam` kasutavad mõlemad sama käsku `gam`, siis mõlemaid pakette ei saa samaaegselt kasutada. Kui R-s kasutatakse ühte eelnimetatud paketitest ja tahetakse hakata kasutama teist, siis tuleb R vahepeal sulgeda ja uuesti käivitada.



Joonis 3.2: Pakettide `mgcv` ja `gam` võrdlus samal andmestikul kui paketis `gam` kasutame `s` käsku.

4 Üldistatud aditiivne mudel andmestikul

4.1 Andmestiku kirjeldus

Põllulindude andmestik on kogutud 2010. ja 2011. aastal Kesk- ja Lõuna-Eesti põllumajandusmaastikul. Mõlemas piirkonnas on kokku 33 lindude seireala, millest

- 11 on mahepõllumajandusega tegelevat põllumajandustootjat,
- 11 keskkonnasõbraliku majandamise põhi- ja lisategevusega tegelevat põllumajandustootjat,
- 11 referentspõllumajandustootjat (põllumajandustootjad, kes ei pea järgima eelnevate toetuste saamiseks rangeid keskkonnanõudeid).

Analüüsi eesmärgiks on välja selgitada, kas mahepõllumajanduse ja keskkonnasõbraliku põllumajanduse seirealadel on rohkem linde kui referentspõllumajanduse seirealadel.
Uuritavaks tunnuseks on pesitsevate paaride arv transektil põllulõokesteta (dominantliik). Selles andmestikus transekt on 1 km pikkune ja 100 m laiune põlluala, kus linde loendati. Uuritav tunnus jääb 0 ja 15 vahele. Argumenttunnusteks on piirkond, loenduskord, toetustüüp, maastik, vili ja hein, millest kolm esimest on kvalitatiivsed ja kolm viimast kvantitatiivsed tunnused.
Tunnusel piirkond on kaks väärtust:

- kesk - Kesk-Eesti loenduspiirkond,
- louna - Lõuna-Eesti loenduspiirkond.

Loenduskorral on kuus väärtust:

- 1 - esimene loenduskord (aprill 2010),

- 2 - teine loenduskord (mai 2010),
- 3 - kolmas loenduskord (juuni 2010),
- 4 - neljas loenduskord (aprill 2011),
- 5 - viies loenduskord (mai 2011),
- 6 - kuues loenduskord(juuni 2011).

Toetustüübil on kolm väärtust:

- mahe - mahetoetus (kõige rangemad keskkonnanõuded),
- ksm - keskkonnasõbralik majandamine (keskmised nõuded),
- ypt - ühtne pindalatoetus (keskkonnanõuded otseselt ei ole).

Tunnus maastik näitab maastikuelementide pindala loendustransektil (näiteks kivihunnikud, hekid), väärtused jäävad 0 ja 0,81 ha vahele. Tunnus hein näitab rohumaa pindala loendustransektil. Tunnuse hein väärtused jäävad 0 ja 10,79 ha vahele. Tunnus vili näitab teravilja pindala loendustransektil, väärtused jäävad 0 ja 10,80 ha vahele. Vaatlusi on kokku 396.

4.2 Analüüsikäik

Autor alustas analüüsi argumenttunnuste omavaheliste korrelatsioonide kontrollimisest. Tunnuste vili ja hein vahel on tugev negatiivne seos, seega mudelis kasutati ainult tunnust vili. Kuna uuritava tunnuse (pesitsevate paaride arv) väärtus ei saa olla negatiivne, siis autor vaatas uuritavat tunnust kui Poissoni jaotusega tunnust, seega seosefunktsiooniks valis $\log(\mu)$. Viit argumenttunnust kasutades koostati üldistatud aditiivne mudel, kus kvalitatiivsed tunnused on lisatud faktortunnustena ja pidevate tunnuste mõju

hinnatakse lokaalse regressiooni abil.

R-is tuleb (lisamooduli `gam`) `gam` käsku kasutades määrata mitut lähimat vaatlust lokaalse regresiooni jaoks kasutatakse. Parima vaatluste arvu saamiseks kasutas autor ristvalideerimist. Ristvalideerimise abil on võimalik hinnata, kui täpselt suudab kasutatav mudel prognoosida uusi vaatluseid. Andmestikust jäetakse üks vaatlus välja ja vaadatakse kui täpselt suudab järele jäänud andmete põhjal hinnatud mudel prognoosida välja jäetud vaatlust. Seejärel eemaldatakse andmestikust järgmine vaatlus ja nii korratakse kogu protseduuri kõigi andmestikus olevate vaatlustega. Prognoosimisel tehtud vead võetakse kokku saamaks ristvalideerimisviga. Mida väiksem on ristvalideerimisviga, seda parem on mudel. ([10]) Autor kasutas ristvalideerimisvea leidmiseks valemit

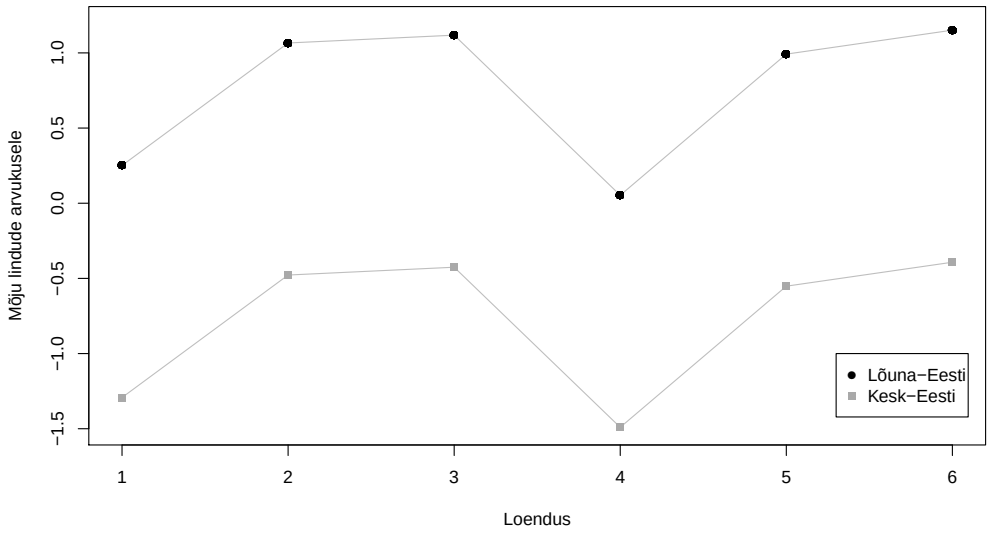
$$CV = \frac{1}{N}\sum_{i=1}^{N} -log(f(x_i)),$$

kus $f(x_i)$ on Poissoni jaotusfunktsioon kohal x_i, mis on seda väiksem, mida tõenäolisemad on vaadeldud vaatlused hinnatud mudeli arvates ([11]). Maastiku mõju hindamisel lokaalse regressiooni jaoks kasutasin 100% vaatlustest ja vilja mõju jaoks 52% vaatlustest. Tunnuste vili ja maastik mõju modelleerimisel osutusid mitteparameetriliselt hinnatud mõjud vajalikuks. Kõik mudelis olnud argumenttunnused osutused statistiliselt oluliseks, seega parimaks mudeliks oli üldistatud aditiivne mudel, kus kvalitatiivsed tunnused on faktortunnustena ning maastiku ja vilja mõju on hinnatud mitteparameetriliselt.

4.3 Tulemused

Järgnevas alampeatükis on antud erinevate argumenttunnuste mõju lindude arvukusele ilma põldlõokesteta. Iga tunnuse mõju juures tuleb arvestada, et

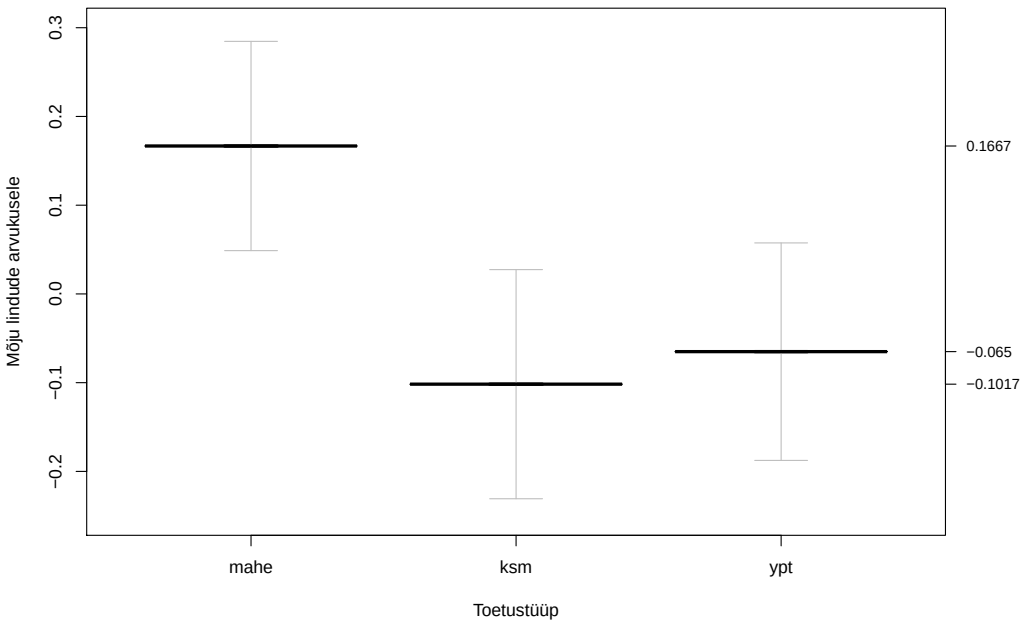
seosefunktsiooniks oli $\log(\mu)$.



Joonis 4.1: Loenduse mõju lindude arvukusele

Joonisel 4.1 on näha, et aprillis on kõige vähem lindude paare, mais rohkem ja juunis kõige rohkem. Kusjuures erinevate aastate, aga samade kuude erinevus ning ka mai ja juuni erinevus ei ole statistiliselt oluline. Seega statistiliselt on aprillis vähem linnupaare kui mais või juunis olenemata aastast. Kui vaadata Lõuna- ja Kesk-Eesti erinevust, siis Lõuna-Eestis on keskmiselt $e^{0,7718-(-0,7718)} = e^{1,5436} \approx 4{,}7$ korda rohkem pesitsevaid linnupaare kui Kesk-Eestis, kui teised argumenttunnused on võrdsed. Kui vaadelda sarnaseid põlde kahel erineval ajal, näiteks esimene kord aprillis 2010 ja teine kord mais 2010, siis maikuus on keskmiselt $e^{1,0568-0,2511} = e^{0,8057} \approx 2{,}24$ korda rohkem linnupaare kui aprillis.

Joonisel 4.2 on näha, et mahepõllul on rohkem lindude paare kui keskkonnasõbralikul põllul või tavapõllul, kusjuures keskkonnasõbraliku põllu ja tavapõllu erinevus ei ole statistiliselt oluline. Kui vaatleme kahte põldu, mille

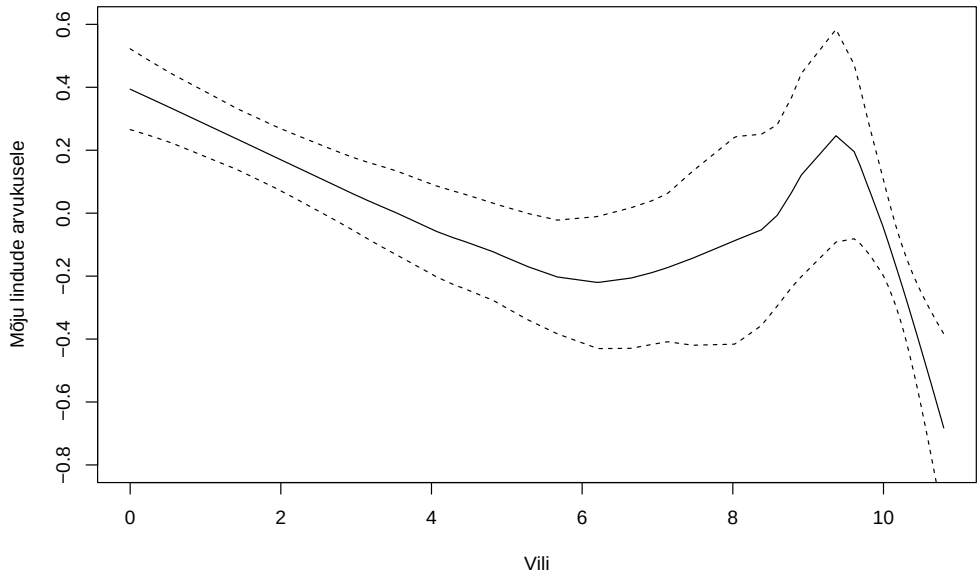


Joonis 4.2: Toetustüübi mõju lindude arvukusele

kõik teised näitajad on võrdsed, ainult esimene põld on mahepõld ja teine on keskkonnasõbraliku majandamisega põld, siis mahepõllul on keskmiselt $e^{0,1667-(-0,1017)} = e^{0,2684} \approx 1{,}3$ korda rohkem linnupaare kui keskkonnasõbralikul põllul. Heledate joontega on ära märgitud hinnangust kahe standardvea kaugusele jäävad väärtused.

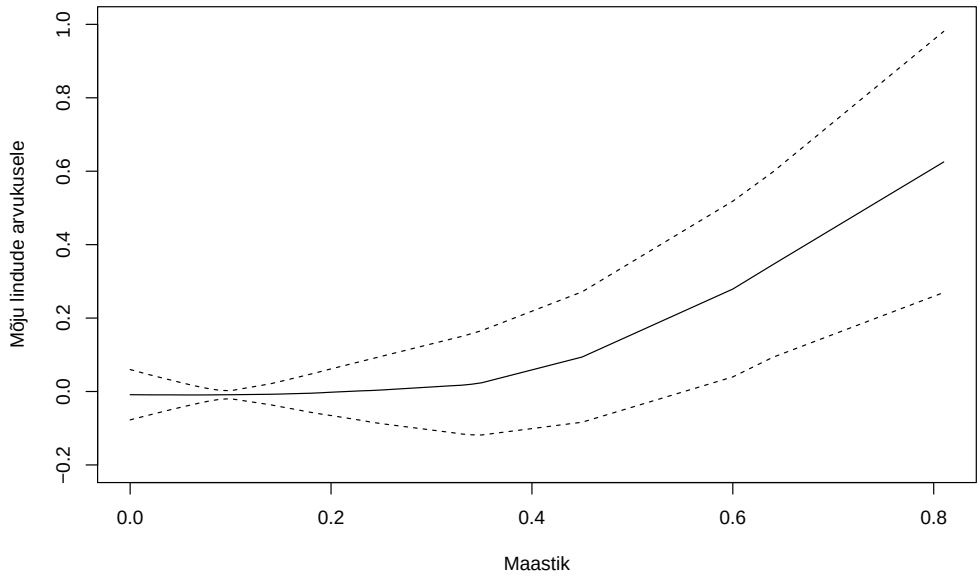
Joonisel 4.3 on pideva joonega tunnuse vili hinnatud mõju ja katkendjoonega on tähistatud kahe standardvea kaugusele hinnangust jääv vahemik. Jooniselt on näha, et vilja mõju on negatiivne välja arvatud vahemikus, kus vilja pindala jääb 6 ha ja 9,5 ha vahele. Mida rohkem on vilja, seda rohkem on põllul väetamist ja inimtegevust ja sellest võib olla linde ka vähem. Seda, miks 6 ha ja 9,5 ha vahel viljal on positiivne mõju, autor seletada ei oska. Autor arvab, et 9,5 ha suuremate väärtuste korral on mõju tugevalt negatiivne, sest siis on loendustransekt peaaegu ainult viljapõld ja lindudel ei ole ala, kus nad saaksid segamatult olla.

Joonisel 4.4 on pideva joonega märgitud tunnuse maastik hinnatud mõju ja



Joonis 4.3: Tunnuse vili mõju lindude arvukusele

katkendjoonega on tähistatud kahe standardvea kaugusele hinnangust jääv vahemik. Tunnuse maastik mõju sisuliselt puudub kui maastikuelementide pindala jääb alla 0,3 ha. Kui pindala jääb 0,3 ha ja 0,8 ha vahele, siis suurem maastiku pindala soosib rohkem linnupaare. Seda saab seletada sellega, et siis on lindudel rohkem ala, kus inimtegevus neid ei sega. Suurema väärtuse mõju hinnangu korral on standardviga suurem, sest selliste väärtustega vaatlusi on vähe.



Joonis 4.4: Tunnuse maastik mõju lindude arvukusele

Kokkuvõte

Käesolevas bakalaureusetöös tutvustatakse üldistatud aditiivset mudelit. Esimeses osas antakse ülevaade aditiivse ja üldistatud aditiivse mudeli kujust ja omadustest. Esitatakse algoritmid, mille järgi on võimalik hinnata nii aditiivset kui ka üldistatud aditiivset mudelit.

Teine peatükk täiendab esimest kolme näitega. Näited on läbiviidud kasutades esimeses peatükis väljatoodud algoritme ja silumismeetodina on kasutatud lokaalset regressiooni. Alustatud on kõige lihtsamast mudelist, kus on ainult üks argumenttunnus ja seosefunktsiooniks on identsusseos. Järgnevalt on selgitatud kahe argumenttunnusega mudeli hindamisprotsessi. Kolmandas näites on hinnatud kahe argumenttunnusega ja seosefunktsiooni $\log(\mu)$ kasutav üldistatud aditiivne mudel. Teises ja kolmandas näites on autori saadud lõpptulemusi võrreldud lisamooduli `gam` abil saadud hinnangutega.

Kolmandas peatükis on antud ülevaade kahest paketist statistikaprogrammis R. Esimesena vaadeldakse Trevor Hastie paketti `gam`. Kuna paketis `gam` saab kasutada mõju hindamiseks ka lokaalset regressiooni, siis on ära toodud ka graafiline võrdlus lisamooduli `gam` abil saadud hinnangu ja käsu `loess` abil saadud hinnangute vahel. Järgmisena tutvustatakse Simon Wood'i paketti `mgcv`, võrreldakse ka pakettide `gam` ja `mgcv` abil saadud hinnanguid.

Töö viimases osas on kasutatud üldistatud aditiivset mudelit põllulindude andmestiku analüüsil. Analüüsi eesmärgiks oli välja selgitada, kas mahepõllumajanduse ja keskkonnasõbraliku majandamisega põldudel on rohkem pesitsevaid linnupaare kui referentspõllumaal. Saadud tulemus näitas, et mahepõldudel on keskmiselt umbes 1,3 korda rohkem linnupaare kui keskkonnasõbralikul põllul ning keskkonnasõbraliku ja referentspõllu mõju erinevus lindude arvukusele ei olnud statistiliselt oluline. Segavad tunnused nagu maastikuelementide ja viljaalune pindala võeti arvesse mitteparameetriliselt.

Kasutatud kirjandus

[1] Efromovich, Sam. 1999. *Nonparametric Curve Estimation: Methods, Theory, and Applications*. New York: Springer.

[2] Hastie, Trevor. 2013. *Package 'gam'*. [Internet]. Saadaval: http://cran.r-project.org/web/packages/gam/gam.pdf [Allalaetud 20.aprill 2014].

[3] Hastie, Trevor, Robert Tibshirani ja Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

[4] Hastie, Trevor J., Robert J. Tibshirani. 1995. *Generalized Additive Models*. London: Chapman & Hall

[5] James, Gareth, Daniela Witten, Trevor Hastie ja Robert Tibshirani. 2013. *An Intorduction to statistical Learning with Applications in R*.

[6] Kaasik, Ülo. 1992. *Matemaatikaleksikon*. Tallinn: Eesti Entsüklopeediakirjastus.

[7] Käärik, Ene. 2013 *Loengukonspekt: Andmeanalüüs II (MTMS.01.007)*. Tartu Ülikool.

[8] Möls, Märt. 2012. *Loengukonspekt: Mitteparameetriline statistika (MTMS.01.037)*. Tartu Ülikool.

[9] Ruppert, David, M. P. Wand ja R. J. Carroll. 2003. *Semiparametric Regression*. New York: Cambridge University Press.

[10] Wikipedia. 2013. *Cross-validation (statistics)*. [Internet]. Saadaval: http://en.wikipedia.org/wiki/Cross-validation_(statistics) [Allalaetud 27.aprill 2014].

[11] Wikipedia. 2013. *Scoring rule*. [Internet]. Saadaval: http://en.wikipedia.org/wiki/Scoring_rule#Logarithmic_scoring_rule [Allalaetud 27.aprill 2014].

[12] Wood, Simon. 2014. *Package 'mgcv'*. [Internet]. Saadaval: http://cran.r-project.org/web/packages/mgcv/mgcv.pdf [Allalaetud 20.aprill 2014].

Lisad

Lisa 1. Hälbimuse jaotuse simuleerimine

```
deviance=c(rep(0,10000)) #teeb muutuja deviance
for (i in c(1:10000)){ #protsess tehakse 10 000 labi
x=rpois(400, 200) #genereerib 400 poissoni jaotusest Po(200) suurust
e=rnorm(400)      #genereerib 400 normaal jaotusest N(0,1) suurust
y=0.3*x+e         #genereerib 400 y-tunnuse vaartust
mudel=gam(y~lo(x)) #hindab uldistatud aditiivse mudeli
deviance[i]=mudel$deviance #halbimus lisab muutujasse
}
mean(deviance) #leiab muutuja keskmise
hist(deviance, ylab="", xlab="Halbimus", main="") #Teeb habimustest histogrammi
#tombab chi^2 jaotuse joone
lines(sort(deviance), 100000*dchisq(sort(deviance), df=395) )
```

Lisa 2. Ühe argumenttunnuse näide

```
x=c(1:10) #argumenttunnus
y=c(3,4,3,5,2,1,0,0,3,10) #uuritav tunnus

#Leiab igale punktile kaalu regressioonisirge arvutamiseks punktis xx
kaalud=function(x,xx){
kaal=c(1:length(x))
maxdist=sort(abs(x-xx))[5] #maxdist on 5 koige lahema vaatluse kaugus x-telje
#suunas
for(i in 1:length(x)){
kaal[i]=max(0, (1-(abs(xx-x[i])/maxdist)^3)^3)
}
return(kaal)
}

#Leiab mudeli vaartuse xx punktides
mudl=function(x, y, xx){
a=c()
for (i in c(1:length(xx))){
valem=lm(y~x, weight=kaalud(x, xx[i])) #regressioonisirge
a[i]=predict(valem, data.frame(x=xx[i])) #mudeli vaartus kohal xx[i]
}
return(a) #tagastab hinnangud
```

```
}

xx=seq(1,10, length=90)
yy2=mudl(x, y, xx) #leiab mudeli joone
#joonistab graafiku
plot(xx,yy2, type="l", lwd=1, xlab="x", ylab="y", ylim=c(0,10))
points(x,y, pch=16) #lisab punktid graafikule
text(x,y,round(kaalud(x, 4.3),3), cex=0.7, pos=c(rep(3,4),1,rep(3,4),1))
axis(1, at=c(4.3),labels=c(4.3), las=0, cex.axis=0.7, tck=-.02)

#lisab sirged x=4,3
abline(v=4.3,col = "lightgray")

#regressioonisirge punkti x=4,3 jaoks
valem=lm(y~x, weight=kaalud(x, 4.3))

#lisab regressioonisirge, mis leidakse punkti x=4,3 jaoks
xx3=seq(0.5, 12, length=100)
yy3=predict(valem, data.frame(x=xx3))
lines(xx3, yy3, lty=2)

#lisab jameda sirge, mis on eelmisest luhem
x3=seq(2.8, 5.8, length=100)
yy3=predict(valem, data.frame(x=xx3))
lines(xx3, yy3, lwd=2)
```

Lisa 3. Kahe argumenttunnuse näide

```
x1=c(1:10) #esimene argumenttunnus
x2=c(2,3,4,2,5,1,0.5,0,2.5,8) #teine argumenttunnus
y=c(3,4,3,5,2,1,0,0,3,10) #uuritav tunnus

xx1=seq(0,10, length=90)
xx2=seq(0,8, length=90)

kaalud=function(x,xx){ #Kaalud, mida kasutatakse regressioonsirge leidmisel
kaal=c(1:length(x))
maxdist=sort(abs(x-xx))[9] #maxdist on kaugus uheksandast lahimast punktist
# x-telje suunas
for(i in 1:length(x)){
kaal[i]=max(0, (1-(abs(xx-x[i])/maxdist)^3)^3)
```

```
}
return(kaal)
}

mudl=function(x,y,xx){
a=c(1:length(xx))
for (i in c(1:length(xx))){
valem=lm(y~x, weight=kaalud(x, xx[i])) #regresioonisirge leidmine
a[i]=predict(valem, data.frame(x=xx[i])) #hinnang punktis xx[i]
}
return(a)
}

yld=function(x1,x2,y, xx1, xx2){
a=mean(y) #uuritava tunnuse keskmine
f1=rep(0, length(x1))
f2=rep(0, length(x1))
for (i in 1:100){
yuus = y-a-f2 #eemaldab y vaartustelt y keskmise ja x2 hinnatud moju
f1algne=mudl(x1,yuus,x1) #x1 moju hinnang vaatluste kohtadel
f1=f1algne-mean(f1algne) #korrigeeritakse tulemusi, et keskmine oleks 0
f1alg=mudl(x1,yuus,xx1)  #x1 moju kogu maaramispiirkonnas
f1kogu=f1alg-mean(f1algne) #korrigeeritakse tulemusi
yuus = y-mean(y)-f1 #eemaldab y vaartustelt y keskmise ja x1 hinnatud moju
f2algne=mudl(x2,yuus,x2) #x2 moju hinnang vaatluste kohtadel
f2=f2algne-mean(f2algne) # korrigeeritakse tulemusi, et keskmine oleks 0
f2alg=mudl(x2,yuus,xx2)  #x2 moju hinnang kogu maaramispiirkonnas
f2kogu=f2alg-mean(f2algne) #korrigeeritud hinanngud
}
return(list(a=a, f1=f1kogu, f2=f2kogu))
}
tul=yld(x1,x2,y, xx1,xx2)

#Esimene joonis x1 jaoks
y1=y-mean(y)
yy1=mudl(x1, y1, xx) #leidakse esimese ringil x1 moju hinnang
valem=lm(y1~x1, weight=kaalud(x1, 5.7))
valem

plot(x1, y1,pch=16, ylab=expression("y"-bar(y))) #graafik punktidest
text(x1,y1,round(kaalud(x1, 5.7),3), cex=0.7, pos=c(3,3,1,3,1,1,3,3,3,1))
```

```
axis(1, at=5.7,labels=5.7, las=0, cex.axis=0.7, tck=-.02)
abline(v=5.7,col = "lightgray")
lines(xx, yy1, type="l", lwd=1) #lisab graafikule hinnangu moju
#lisab korrigeeritud hinnangu skaala
axis(4, at=c(5.78, 3.78, 1.78, -0.22, -2.22),labels=c(6,4,2,0,-2), las=2)

#lisab regressioonisirge, mis leidakse punkti x=5,7 jaoks
xx3=seq(0.5, 12, length=100)
yy3=predict(valem, data.frame(x1=xx3))
lines(xx3, yy3, lty=2)

#lisab sirge eelmisest, mis on luhem
xx3=seq(4.2, 7.2, length=100)
yy3=predict(valem, data.frame(x1=xx3))
lines(xx3, yy3, lwd=2)

#Esimene joonis x2 jaoks
yy12=mudl(x1, y1, x1)
ypar=yy12-mean(yy12)
y2=y-mean(y)-ypar
yy2=mudl(x2,y2,xx) #leiab esimese ringi x2 moju hinnangu
valem=lm(y2~x2, weight=kaalud(x2, 3.6))
valem

#joonis esimese ringi x2 moju hinnangust
plot(x2, y2, pch=16, ylab=expression("y"-bar(y)-"f1(x1)"))
text(x2,y2,round(kaalud(x2, 3.6),3), cex=0.7, pos=c(rep(3,6),1,3,3,1))
axis(1, at=3.6,labels=3.6, las=0, cex.axis=0.7, tck=-.02)
abline(v=3.6,col = "lightgray")
lines(xx, yy2, type="l", lwd=1)
#lisab korrigeeritud hinnangu skaala
axis(4, at=c(4.05,2.05,0.05, -1.95),labels=c(4,2,0,-2), las=2)

#lisab regressioonisirge, mis leidakse punkti x=3,6 jaoks
xx3=seq(-0.5, 12, length=100)
yy3=predict(valem, data.frame(x2=xx3))
lines(xx3, yy3, lty=2)

#lisab sirge eelmisest, mis on luhem
xx3=seq(1.1, 5.1, length=100)
yy3=predict(valem, data.frame(x2=xx3))
```

```
lines(xx3, yy3, lwd=2)

#Samad hinnangud R-i paketiga 'gam'
library(gam)
mudel=gam(y~lo(x1, span=9/10)+lo(x2, span=9/10))

#Uus aken lopptulemuste jaoks
windows()
par(mfrow=c(1,2))
#teeb graafiku autori funktsioonist 'tul' saadud x1 moju hinnangust
plot(sort(xx1), tul$f1[order(xx1)], xlab="x1", ylab="f1", type="l")
#lisab 'gam' paketi saadud hinnangud katkendjoonega
a=preplot(mudel)$'lo(x1, span = 9/10)'$y
lines(sort(x1), a[order(x1)], lty=2)
#teeb graafiku autori funktsioonist 'tul' saadud x2 moju hinnangust
plot(sort(xx2), tul$f2[order(xx2)], xlab="x2", ylab="f2", type="l")
#lisab 'gam' paketi saadud hinnangud katkendjoonega
b=preplot(mudel)$'lo(x2, span = 9/10)'$y
lines(sort(x2), b[order(x2)], lty=2)
```

Lisa 4. *Log* seosefunktsiooni näide

```
x1=c(1:10) #esimene argumenttunnus
x2=c(2,3,4,2,5,1,0.5,0,2.5,8) #teine argumenttunnus
y=c(1,4,3,5,1,4,1,0,3,5) #uuritav tunnus

xx1=seq(0,10, length=90)
xx2=seq(0,8, length=90)


kaalud=function(x,xx){ #kaalud, mida kasutatakse regressioonsirge leidmisel
kaal=c(1:length(x))
maxdist=sort(abs(x-xx))[9] #maxdist on kaugus uheksandast lahimast punktist
# x-telje suunas
for(i in 1:length(x)){
kaal[i]=max(0, (1-(abs(xx-x[i])/maxdist)^3)^3)
}
return(kaal)
}

#leiab muutuja n
```

```
ni=function(a,f1,f2){
n=c(1:10)
for (i in c(1:10)){
n[i]=a+f1[i]+f2[i]
}
return(n)
}

#leiab muutuja z
zi=function(y,n){
m=exp(n)
z=n+(y-m)/m
return(z)
}

#leiab hinnangud
mudl=function(n,z,x,xx){
a=c(1:length(xx))
for (i in c(1:length(xx))){
#kasutatakse nii regressioonsirge leimdise kaale kui ka dispersioonist
# tulenevad kaalud
valem=lm(z~x, weight=exp(n)*kaalud(x, xx[i]))
a[i]=predict(valem, data.frame(x=xx[i])) #hinnang punktis xx[i]
}
return(a)
}

#teeb tsukli, kus hinnatakse x1 ja x2 moju 100 korda
yld=function(x1,x2,y, xx1, xx2){
a=log(mean(y))
f1=rep(0, length(x1))
f2=rep(0, length(x1))
for (i in 1:100){
n=ni(a,f1,f2)
z=zi(y,n)
zuus = z-a-f2
f1algne=mudl(n,zuus,x1,x1) #hindab mudeli punktides
f1=f1algne-mean(f1algne) #korrigeerib lahustades hinnangu keskmise
f1alg=mudl(n, zuus, x1, xx1) #hindab mudeli ule kogu maaramispiirkonna
f1kogu=f1alg-mean(f1algne)
zuus = z-a-f1
```

```
f2algne=mudl(n, zuus,x2,x2)
f2=f2algne-mean(f2algne)
f2alg=mudl(n, zuus, x2, xx2)
f2kogu=f2alg-mean(f2algne)
a=log(sum(exp(n)*y)/sum(exp(n)))
}
return(list(a=a, f1=f1kogu, f2=f2kogu))
}
tul=yld(x1,x2,y, xx1, xx2)

#Esimene joonis x1 jaoks
n=rep(log(mean(y)),10)
z=zi(y,n)
z1=z-mean(z)
yy1=mudl(n,z1,x1,xx)
valem=lm(z1~x1, weight=exp(n)*kaalud(x1, 5.7))
valem

#Parandatud skaala leidmiseks
zz1=mudl(n,z1,x1,x1)
mean(zz1)

par(mar=c(4,4,1,3))
plot(x1, z1,pch=16, ylab="z-a")
text(x1,z1,round(exp(n)*kaalud(x1, 5.7),3), cex=0.7,
pos=c(rep(3,3),1,rep(3,5),1))
axis(1, at=5.7,labels=5.7, las=0, cex.axis=0.7, tck=-.02)
abline(v=5.7,col = "lightgray")
lines(xx, yy1, type="l", lwd=1)
axis(4, at=c(0.489, -0.011, -0.511, -1.011), labels=c(0.5,0,-0.5,-1.0),
las=2)

#lisab regressioonisirge, mis leidakse punkti x=5,7 jaoks
xx3=seq(0.5, 12, length=100)
yy3=predict(valem, data.frame(x1=xx3))
lines(xx3, yy3, lty=2)
#col="lightblue3",

#lisab sirge eelmisest, mis on luhem
xx3=seq(4.2, 7.2, length=100)
yy3=predict(valem, data.frame(x1=xx3))
```

```
lines(xx3, yy3, lwd=2)
#col="blue",

#Esimene joonis x2 jaoks
zpar=zz1-mean(zz1)
z2=z-mean(z)-zpar
yy2=mudl(n,z2,x2,xx)
valem=lm(z2~x2, weight=exp(n)*kaalud(x2, 3.6))
valem

#Parandatud hinnangu skaala jaoks
zz2=mudl(n,z2,x2,x2)
mean(zz2)

plot(x2, z2, pch=16, ylab="z-a-f1(x1)")
text(x2,z2,round(exp(n)*kaalud(x2, 3.6),3), cex=0.7,
pos=c(rep(3,3),1,rep(3,4),1,3))
axis(1, at=3.6,labels=3.6, las=0, cex.axis=0.7, tck=-.02)
abline(v=3.6,col = "lightgray")
lines(xx, yy2, type="l", lwd=1)
axis(4, at=c(0.53,0.03,-0.47,-0.97),labels=c(0.5, 0, -0.5,-1), las=2)

#lisab regressioonisirge, mis leidakse punkti x=3,6 jaoks
xx3=seq(-0.5, 12, length=100)
yy3=predict(valem, data.frame(x2=xx3))
lines(xx3, yy3, lty=2)
#col="lightblue3",

#lisab sirge eelmisest, mis on luhem
xx3=seq(2.6,4.6, length=100)
yy3=predict(valem, data.frame(x2=xx3))
lines(xx3, yy3, lwd=2)
#col="blue",


#Uus aken lopptulemuste jaoks
windows()
library(gam)
mudel=gam(y~lo(x1, span=9/10)+lo(x2,span=9/10), family=poisson(link="log"),
start=c(0,0,0))
```

```
par(mfrow=c(1,2))
plot(sort(xx1), tul$f1[order(xx1)], xlab="x1", ylab="f1", type="l")
a=preplot(mudel)$'lo(x1, span = 9/10)'$y
lines(sort(x1), a[order(x1)], lty=2)
plot(sort(xx2), tul$f2[order(xx2)], xlab="x2", ylab="f2", type="l")
b=preplot(mudel)$'lo(x2, span = 9/10)'$y
lines(sort(x2), b[order(x2)], lty=2)
```

Lisa 5. Võrdlused peatükis 3

```
#Loessi ja Gam(lo) vordlus
library('gam')
x=c(0.5,1,1.5,2,2.5,3,3.5,4,4.5,5,5.5,6,6.5,7,7.5,8,8.5,9,9.5,10) #argumenttunnus
y=c(1,2,4,2,4,5,5,5,6,3,2,7,1,0,0,0,1,3,5,6) #uuritav tunnus

xx=seq(0.5,10, length=90)
#hindab loessi mudeli ja gam mudeli
mudel1=loess(y~x, span=0.4, surface = "direct", degree=1)
mudel2=gam(y~lo(x,span=0.4))

yy1=predict(mudel1, data.frame(x=xx))
yy2=predict(mudel2, data.frame(x=xx))

#graafik molemast hinnangust
plot(xx,yy1, type="l", lwd=1, xlab="x, span=0.4", ylab="y", ylim=c(-1,8))
points(x,y, pch=16)
lines(xx, yy2, lty=2)
legend(0.5 ,8, c("Loess", "GAM"), lty=c(1,2))

#Gam(s) vs mgcv(s)

library('gam')
mudel3=gam(y~s(x), df=2) #hindab mudeli paketiga 'gam'
yy3=predict(mudel3, data.frame(x=xx))
#kirjutab mudeli andmestikuna
assign("data1", data.frame(xx = xx, yy = yy3))
write.table(data1, "C:/data.txt", sep="\t")

#et kasutatada teist paketti, tuleb R vahepeal kinni panna ja uuesti avada
library('mgcv')
#hindab mudeli paketiga 'mgcv'
```

```
mudel4=gam(y~s(x), df=2)
yy4=predict(mudel4, data.frame(x=xx))

#loeb pakei 'gam' andmed sisse
gam=read.table("C:/data.txt", header=T, sep="\t")

#graafik kahe hinnantud mudelist
plot(xx,yy4, type="l", lwd=1, xlab="x", ylab="y", ylim=c(0,8))
points(x,y, pch=16)
lines(gam$xx, gam$yy, lty=2)
legend(0.5 ,8, c("mgcv", "gam"), lty=c(1,2))
```

Lisa 6. Ristvalideerimine antud andmestiku jaoks

```
#Ristvalideerimine

#loeb andmestiku sisse
tabel=read.table("C:/pmk.txt", header=T, sep="\t", dec=",")

library(gam)
x=tabel$maastik #muutes x saab sama teha koikide pidevate tunnuste jaoks
y=tabel$sum2
ntrial <- 61
#span saab olla 40% kuni 100%
span1 <- seq(from = 0.4, by = 0.01, length = ntrial)

span_hinnang <- function(x, y, span){
cv=function(x,y,sp){
a=length(x)
mu=c(rep(0, a))
for (i in c(1:a)){
#eemaldab i-nda vaatluse
xx=x[-i]
yy=y[-i]
#hindab mudeli ilma i-nda vaatluseta
mudel=gam(yy~lo(xx, span=sp), family=poisson)
#annab hinnangu vaatlusele i
if (i<a){
mu[i]=predict(mudel, data.frame(x=x[-(i+1)]), type="response")[i]
}
else{mu[i]=predict(mudel, data.frame(x=x[-(i-1)]), type="response")[i-1]}
```

```
}
#leiab keskmise ristvalideerimise vea Poisooni jaotuse tihedusfunktsioonis
cv1= (1/a)*sum(-log(dpois(y,lambda=mu)))
return(cv1)
}
#teeb listi koikidest keskmistest ristvalideerimisvigadest
kokku=lapply(as.list(span1), cv, x = x, y = y)
return(kokku)
}
cv=span_hinnang(x,y,span1)

#gaafik keskmistest ristvalideerimisvigadest,
#koige vaiksem viga annab parima spani
plot(span1, cv, type = "n", xlab = "span", ylab = "CV")
points(span1, cv, pch = 3)
lines(span1, cv, lwd = 2)
```

Lisa 7. Üldistatud aditiivne mudel andmestikul

```
#andmete sisse lugemine
tabel=read.table("C:/pmk.txt", header=T, sep="\t", dec=",")
#ulevaade andmetest
summary(tabel)
str(tabel)

#tunnuste loomine
loendus=tabel$loendus
maastik=tabel$maastik
hein=tabel$hein
vili=tabel$vili
toetustyyp=factor(tabel$toetustyyp, levels=c("mahe", "ksm", "ypt"))
piirkond=tabel$piirkond
y=tabel$sum2

#toetustyyp vaartuste kokku panemine erinevuste leidmiseks
toetustyyp2=c(rep(0, length(toetustyyp)))
toetustyyp3=c(rep(0, length(toetustyyp)))
toetustyyp4=c(rep(0, length(toetustyyp)))
for (i in c(1:length(toetustyyp))){
if (toetustyyp[i]=='mahe'){
toetustyyp2[i]=1
```

```
toetustyyp3[i]=2
toetustyyp4[i]=2
} else if (toetustyyp[i]=='ksm'){
toetustyyp2[i]=2
toetustyyp3[i]=1
toetustyyp4[i]=2
} else {
toetustyyp2[i]=2
toetustyyp3[i]=2
toetustyyp4[i]=1}
}

#loenduse vaartuste kokku panemine erinevuste leidmiseks
loendus2=c(rep(0, length(toetustyyp)))
for (i in c(1:length(loendus))){
if (loendus[i]==4){
loendus2[i]=1
} else if (loendus[i]==5){
loendus2[i]=2
} else if (loendus[i]==6){
loendus2[i]=2
} else if (loendus[i]==3){
loendus2[i]=2
}else {
loendus2[i]=loendus[i]
}}

#Korrelatsioonid (hein,vili)=-0.95, (vili,maastik)=-0.20, (hein,maastik)=0.18
cor(maastik, hein)

library(gam)
#esialgne mudel
mudel1=gam(y~factor(piirkond)+factor(loendus)+lo(vili, span=0.52)
+lo(maastik, span=1)+factor(toetustyyp), family=poisson)
#mudel uhe argumenttunnuse valja jatmisel
mudel2=gam(y~factor(loendus)+lo(vili, span=0.52)+lo(maastik, span=1)
+factor(toetustyyp), family=poisson)
mudel3=gam(y~factor(piirkond)+lo(vili, span=0.52)+lo(maastik, span=1)
+factor(toetustyyp), family=poisson)
mudel4=gam(y~factor(piirkond)+factor(loendus)+lo(maastik, span=1)
+factor(toetustyyp), family=poisson)
```

```
mudel5=gam(y~factor(piirkond)+factor(loendus)+lo(vili, span=0.52)
+factor(toetustyyp), family=poisson)
mudel6=gam(y~factor(piirkond)+factor(loendus)+lo(vili, span=0.52)
+lo(maastik, span=1), family=poisson)

#Parima mudeli leidmine, erinevad kombinatsioonid
anova(mudel1, mudel6)

#kontrollimine, millised toetustuubid voib kokku panna
mudel1b=gam(y~factor(piirkond)+factor(loendus)+lo(vili, span=0.52)
+lo(maastik, span=1)+factor(toetustyyp2),family=poisson)
mudel1c=gam(y~factor(piirkond)+factor(loendus)+lo(vili, span=0.52)
+lo(maastik, span=1) +factor(toetustyyp3), family=poisson)
mudel1d=gam(y~factor(piirkond)+factor(loendus)+lo(vili, span=0.52)
+lo(maastik, span=1) +factor(toetustyyp4), family=poisson)
anova(mudel1, mudel1b)
#ksm ja ypt voib kokku panna.

#kontrollimine, millised loenduskorrad voib kokku panna
mudel1e=gam(y~factor(piirkond)+factor(loendus2)+lo(vili, span=0.52
+lo(maastik, span=1) +factor(toetustyyp), family=poisson)
#loenduses 1 ja 4, 2 ja 5, 3 ja 6 voib kokku panna.

#leiab maastiku moju y-tunnusele
#muutes argumenttunnust, saab leida koikide argumenttunnuste moju
a=preplot(mudel1)$'lo(maastik, span = 1)'$y
#naiteks a=preplot(mudel1)$'factor(toetustyyp)'$y
#leiab standardhalbe maastiku moju hinnangule
b=preplot(mudel1)$'lo(maastik, span = 1)'$se.y
#naiteks b=preplot(mudel1)$'factor(toetustyyp)'$se.y
c=a+2*b
d=a-2*b

#Tunnuste piirkond ja loendus uhele graafikule panemine
e=preplot(mudel5)$'factor(loendus)'$y
f=preplot(mudel5)$'factor(piirkond)'$y
e1=e+0.7717804
e2=e-0.7717804

plot(sort(loendus), e1[order(loendus)],type="l", col="grey",
ylab="Moju lindude arvukusele", xlab="Loendus", ylim=c(-1.5,1.2))
```

```
lines(sort(loendus), e2[order(loendus)], col="grey")
points(loendus, e2,pch=15, col="darkgrey")
points(loendus, e1, pch=16)
legend(5.3, -1, c("Louna-Eesti", "Kesk-Eesti"), pch=c(16,15),
col=c("black", "darkgrey"))

#Toetustuubi graafiku tegemine
par(mar=c(4,4,1,4))
plot(toetustyyp, a,  xlab="Toetustuup", ylab="Moju lindude arvukusele",
ylim=c(-0.25, 0.3))
axis(4, at=c(0.1667, -0.0650, -0.1017),labels=c(0.1667, -0.0650, -0.1017),
las=2, cex.axis=0.8, tck=-.02)
arrows(1, 0.2846, 1, 0.0488, angle=90, code=3, col="gray")
arrows(2, 0.0274, 2,-0.2308, angle=90, code=3, col="gray")
arrows(3, 0.0575, 3, -0.1876, angle=90, code=3, col="gray")

#Maastiku moju graafik, kui maastik asendada viljaga, saab vilja moju graafiku
plot(maastik, a,  xlab="Maastik", ylab="Moju lindude arvukusele",
ylim=c(-0.2, 1))
lines(sort(maastik), c[order(maastik)], lty=2)
lines(sort(maastik), d[order(maastik)], lty=2)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Annegrete Peek,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Üldistatud aditiivne mudel", mille juhendaja on Märt Möls,

    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, 05.05.2014