Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Perttu Narvik

Kant- ja lassoregressioon ning nende rakendamine müügiskoori loomiseks Creditinfo Eesti AS andmetel

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendaja Taavi Unt, MSc

Tartu 2017

Kant- ja lassoregressioon ning nende rakendamine müügiskoori loomiseks Creditinfo Eesti AS andmetel

Käesoleva bakalaureusetöö eesmärgiks on tutvustada kant- ja lassoregressiooni ning rakendada logistilist regulariseeritud regressiooni müügiskoori loomiseks Creditinfo Eesti AS andmetel. Töö esimeses osas antakse ülevaade lineaarsest regressioonist, lineaarsest kant- ja lassoregressioonist, nende omadustest ning tavalisest ja regulariseeritud logistilisest regressioonist. Töö teises osas konstrueeritakse müügiskoor, mille põhjal on võimalik prognoosida, kui suure tõenäosusega võiks mingist ettevõttest saada uus klient.

Märksõnad: *kantregressioon, lassoregressioon, ristvalideerimine*
P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Ridge and Lasso Regression and Their Application in Developing a Purchase Score Based on Data from Creditinfo Estonia AS

The aim of this thesis is to introduce ridge and lasso regression and to apply regularized regression in developing a purchase score based on data from Creditinfo Estonia AS. In the first section an overview is given about ordinary linear regression, linear ridge and lasso regression, their properties and both ordinary and regularized logistic regression. In the second section a purchase score is developed to predict the probability of a company becoming a new client.

Keywords: *ridge regression, lasso regression, cross-validation*
P160 Statistics, operation research, programming, actuarial mathematics

Sisukord

Sissejuhatus

Potentsiaalsete klientide leidmiseks tuleb teada, mille poolest erinevad olemasolevad kliendid teistest ettevõtetest. Müügiskoori ideeks on koondada kliente eristavad tunnused ühte muutujasse, mida nimetatakse ostupotentsiaaliks. Selle alusel saab müügimeeskond otsustada, missugustele ettevõtetele on mõistlik oma aega pühendada. Müügiskoori loomiseks hinnatakse logistiline regressioonimudel.

Logilistine ning lineaarne regressioon on väga laialdaselt kasutatud statistilised meetodid, mille abil hinnatakse mitmete tunnuste mõju ühele tunnusele. Üldjuhul leitakse mudeli parameetrite hinnangud lineaarse regressiooni korral vähimruutude meetodil ning logistilise regressiooni korral suurima tõepära meetodil. Klassikalised meetodid ei anna alati parimaid tulemusi ja mõningatel juhtudel ei toimi üldse, näiteks juhul, kui hinnatavate parameetrite arv on suurem kui vaatluste arv. Sellisel juhul tuleks kasutada parameetrite hindamiseks teisi meetodeid, näiteks regulariseeritud regressiooni.

Käesoleva töö eesmärgiks on tutvustada kant- ja lassoregressiooni ning rakendada logistilist regulariseeritud regressiooni müügiskoori konstrueerimiseks Creditinfo Eesti AS andmetel.

Töö on liigendatud kaheks peatükiks. Esimene peatükk on teoreetiline ning jaguneb omakorda kaheksaks alajaotuseks. Esmalt tutvustatakse tavalist lineaarset regressiooni, seejärel lineaarset kant- ja lassoregressiooni ning nende omadusi. Peatüki lõpus kirjeldatakse nii tavalist kui ka regulariseeritud logistilist regressiooni. Teises peatükis tutvustatakse mudeli loomiseks kasutatavat R-i paketti „glmnet”, antakse ülevaade kasutatavatest andmetest ning kirjeldadakse mudeli hindamist.

Kuigi praktilises osas kasutatakse logistilist regressioonimudelit, on teoreetilises

osas kant- ja lassoregressiooni omadusi tutvustatud lineaarse regressioonimudeli põhjal. Seda on tehtud lihtsal põhjusel - regulariseeritud regressiooni omadused kehtivad enamjaolt nii lineaarse kui ka logistilise regressioonimudeli korral, kuid lineaarse regressioonimudeli korral on ülevaate saamine mõnevõrra lihtsam.

Käesolev bakalaureusetöö on vormistatud tekstitöötlusprogrammi *LaTeX* veebiversiooniga *Overleaf*. Andmete simuleerimiseks, jooniste tegemiseks ning müügiskoori konstrueerimiseks on kasutatud statistikatarkvara *R* versiooni 3.3.3.

Autor tänab juhendajat Taavi Unti suunamise, rohkete täpsustuste ning pühendatud aja eest.

1 Regulariseeritud regressioon

1.1 Lineaarne regressioon

Käesolev alajaotus tugineb teosel „Introduction to Linear Regression Analysis" (Montgomery et al., 2013, lk 70-73, 79-81, 587). Sageli huvitab uurijat, kuidas kirjeldada ühte tunnust mitmete teiste tunnuste abil. Kui sõltuv tunnus Y on pidev ning sõltumatuid tunnuseid $X_1, X_2,..., X_p$ on p tükki, siis saab kasutada lineaarset regressioonimudelit, mis avaldub kujul

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i. \tag{1}$$

Antud valemis on y_i i-nda objekti sõltuva tunnuse väärtus ($i = 1,...,n$; n on valimimaht), β_0 on vabaliige, $\beta_1,...,\beta_p$ on regressioonikordajad, x_{ij} on i-nda objekti j-nda tunnuse väärtus ($j = 1,...,p$) ning ε_i on juhuslik viga. Vigade puhul eeldatakse, et need on sõltumatud, keskväärtusega 0 ning dispersiooniga σ^2. Tavapäraselt eeldatakse veel, et juhuslikud vead on normaaljaotusega, kuid antud töö kontekstis pole see oluline.

Lineaarset regressioonimudelit on võimalik väljendada ka maatrikskujul:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

kus

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Lineaarse regressiooni parameetrite hinnangud leitakse vähimruutude meetodil. See tähendab, et vabaliikme ja regressioonikordajate hinnangud $\hat{\beta}_0,...,\hat{\beta}_p$ saadakse

minimeerides jääkide ruutude summat:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 . \tag{3}$$

Jääkide ruutude summa valem maatrikskujul on

$$RSS = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{4}$$

Vähimruutude hinnangu leidmiseks tuleb võtta suurusest RSS tuletis parameetervektori $\boldsymbol{\beta}$ järgi ning saadud tulemus võrdsustada nulliga. Saadud lahend ongi soovitud hinnanguks. Seega,

$$\left. \frac{\partial RSS}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{0},$$

mis lihtsustub kujule

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}. \tag{5}$$

Vähimruututе hinnang $\hat{\boldsymbol{\beta}}$ avaldub kujul

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{6}$$

kui leidub pöördmaatriks $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. See eksisteerib juhul, kui maatriksi X veerud on lineaarselt sõltumatud.

Gauss-Markovi teoreemist tuleneb, et vähimruutute meetodil saadud hinnang on parim lineaarsete nihketa hinnangute seast (BLUE - *best linear unbiased estimator*). Seda nimetatakse parimaks, kuna saadakse täpseim ehk väikseima dispersiooniga hinnang. Parameetervektori $\boldsymbol{\beta} = (\beta_0,..,\beta_p)$ hinnangu $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0,..,\hat{\beta}_p)$ nihkeks nimetatakse suurust $B = E\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ (Traat, 2006, lk 31). Kui $B = \mathbf{0}$, siis on tegemist nihketa hinnanguga.

On lihtne tõestada, et vähimruutude hinnang $\hat{\boldsymbol{\beta}}$ on nihketa:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = \\ &= E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = \\ &= E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon})] = \boldsymbol{\beta}, \end{aligned} \tag{7}$$

kuna $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = I$ ning $E(\boldsymbol{\varepsilon}) = 0$.

Parameetrite vähimruutute hinnangu $\hat{\boldsymbol{\beta}}$ kovariatsioone kirjeldatakse kovariatsioonimaatriksiga, mille peadiagonaalil on parameetri hinnangu $\hat{\beta}_j$ dispersioon ning väljaspool peadiagonaali i-nda rea j-nda veeru elemendiks on kovariatsioon hinnangute $\hat{\beta}_i$ ja $\hat{\beta}_j$ vahel. Kuna $Cov(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$, siis $\hat{\boldsymbol{\beta}}$ kovariatsioonimaatriks avaldub järgmiselt:

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}) &= Cov[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov\boldsymbol{y}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T = \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

Mudeli vigade dispersiooni hinnang avaldub järgmise valemiga:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2}{n-(p+1)} = \frac{RSS}{n-(p+1)}. \tag{8}$$

Mudeli vigade dispersioon on suur, kui hinnatavaid parameetreid on ligikaudu sama palju kui valimis objekte. Sellisel juhul võib vähimruutude meetodi asemel kasutada alternatiivseid parameetrite hindamise meetodeid, mida kirjeldatakse järgnevates alapeatükkides.

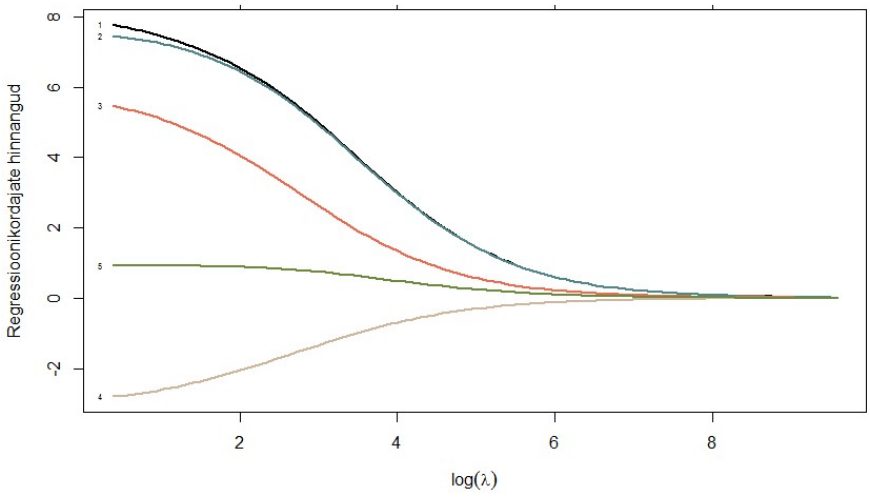
1.2 Kantregressioon

Kantregressioon (*ridge regression*) on lineaarse regresiooni edasiarendus ning mudeli üldkuju on mõlemal juhul samasugune, erinevus seisneb parameetrite hindamises. Lineaarse regressioonimudeli parameetrite hindamisel minimeeritakse jääkide ruutude summat, mis on antud valemiga 3. Kantregressiooni puhul on pa-

rameetrite hinnanguteks suurused $\hat{\beta}^R_{j,\lambda}$, mille korral on minimeeritud suurust

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{9}$$

kus $\lambda \geqslant 0$. Suurus λ, mida nimetatakse karistusparameetriks (*tuning parameter*), tuleb eraldi määrata. Kui vähimruutude meetodil saadakse parameetritele vaid üks hinnangute komplekt, siis kantregressiooni korral saadakse üks hinnangute komplekt iga λ väärtuse korral. Sobiva λ leidmist kirjeldadakse alajaotuses 1.5. Suurust $\lambda\sum_j \beta_j^2$ nimetatakse karistusliikmeks (*shrinkage penalty*) ning selle väärtus on väike, kui β_1,...,β_p on nullilähedased. Seega kantregressiooni korral saadakse parameetrite hinnangud lähtudes kahest aspektist: need peavad sobima andmetega võimalikult hästi, kuid samas nende väärtused ei saa olla suured. Karistusparameetri väärtusest sõltub, millisel aspektil on minimeerimisel suurem mõju. Kui $\lambda = 0$, siis karistusliikmel mõju puudub ning saadakse tavalised vähimruutude hinnangud. Kui $\lambda \to \infty$, siis karistusliikme mõju kasvab ning hinnangud lähenevad nullile. (James et al., 2015, 215-217)



Joonis 1. Kantregressiooni kordajate hinnangud

Joonisel 1 on kasutatud simuleeritud andmeid viie seletava tunnusega ning sellel on kujutatud kantregressiooniga saadud regressioonikordajate hinnanguid vastavalt λ väärtusele. Jooniselt on näha, et karistusparameetri kasvades parameetrite hinnangud lähenevad nullile.

Kantregressiooni kasutamiseks tuleb seletavate tunnuste väärtused standardiseerida. Tavalise vähimruutude meetodi puhul pole see vajalik, kuna tunnuse X_j skaala muutmine c korda põhjustab hinnangu $\hat{\beta}_j$ muutuse $\frac{1}{c}$ korda ning suurus $X_j\hat{\beta}_j$ sellest ei muutu. Karistusliikme tõttu kantregressiooni korral see nii ei ole, mistõttu $X_j\hat{\beta}^R_{j,\lambda}$ sõltub nii tunnuse X_j skaalast kui ka λ valikust. Skaala mõju vastava tunnuse parameetri hinnangule on võimalik elimineerida, kui tunnuse väärtused standardiseerida valemi

$$x_{ij} = \frac{x^*_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x^*_{ij} - \bar{x}^*_j)^2}} \tag{10}$$

abil, kus x^*_{ij} on i-nda objekti j-nda tunnuse tegelik väärtus ning $\bar{x}^*_j = \frac{1}{n}\sum_{i=1}^{n} x^*_{ij}$. (James et al., 2015, lk 217)

Selleks, et leida parameetri β_0 hinnang, tuleb standardiseeritud sisendmaatriksi $\boldsymbol{X}$ veerud tsentreerida: iga x_{ij} asendatakse suurusega $x_{ij} - \bar{x}_j$. Kuna iga j korral $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = 0$, siis avaldise 9 põhjal saame β_0 hinnanguks

$$\hat{\beta}^R_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{11}$$

Ülejäänud parameetrite hinnangud saadakse hinnates kantregressiooniga vabaliikmeta mudel. Kantregressiooni minimeerimisülesanne on võimalik viia maatrikskujule:

$$RSS^R_\lambda = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta},$$

kus maatriksi $\boldsymbol{X}$ dimensioon on $n \times p$, välja on jäetud vabaliikme veerg ja teised veerud on normeeritud, $\boldsymbol{y}$ on tsentreeritud ning parameetervektoriks on

$\boldsymbol{\beta} = (\beta_1,...,\beta_p)$. Edaspidi eeldame, et $\boldsymbol{X}$ on normeeritud ja $\boldsymbol{y}$ on tsentreeritud, misõttu mudelite vabaliikmeid ei hinnata. (Hastie et al., 2013, lk 64) Kantregressiooni parameetrite hinnang $\hat{\boldsymbol{\beta}}^R_\lambda = (\hat{\beta}^R_{1,\lambda},...,\hat{\beta}^R_{p,\lambda})$ peab rahuldama võrdust

$$\left.\frac{\partial RSS^R_\lambda}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}^R_\lambda} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}^R_\lambda + 2\lambda\hat{\boldsymbol{\beta}}^R_\lambda = \boldsymbol{0},$$

mis lihtsustub kujule $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})\hat{\boldsymbol{\beta}}^R_\lambda = \boldsymbol{X}^T\boldsymbol{y}$. Sellest järeldub, et kantregressiooni parameetrite hinnang $\hat{\boldsymbol{\beta}}^R_\lambda$ avaldub kujul

$$\hat{\boldsymbol{\beta}}^R_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{12}$$

Kui leidub pöördmaatriks $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, siis kantregressiooni parameetrite hinnangu $\hat{\boldsymbol{\beta}}^R_\lambda$ saab avaldada vähimruutude hinnangu $\hat{\boldsymbol{\beta}}$ kaudu,

$$\hat{\boldsymbol{\beta}}^R_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

Kuna vähimruutute meetodil saadud parameetrite hinnang on nihketa avaldise 7 põhjal, siis kantregressiooni parameetrite hinnangu keskväärtus avaldub

$$E[\hat{\boldsymbol{\beta}}^R_\lambda] = E[(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{\beta},$$

kui $\lambda > 0$. Seega kantregressiooni parameetrite hinnang on nihkega. Kovariatsioonimaatriks

$$\begin{aligned} Cov[\hat{\boldsymbol{\beta}}^R_\lambda] &= Cov((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}) = \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{y})[(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T]^T = \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} \end{aligned}$$

kirjeldab kantregressiooni parameetrite hinnangute dispersiooni. (Montgomery et al., 2013, lk 306)

Tabelis 1 on simuleeritud andmetelt leitud keskmised kantregressiooni parameetrite hinnangud nelja erineva λ väärtuse korral. On näha, et suurema λ korral parameetrite hinnangute hajuvus on väiksem, kuid kaugus tegelikust väärtusest on

suurem. Simuleeritud andmestikus on seletavad tunnused X_1 ning X_2 tugevalt seotud, tunnused X_3 ja X_4 teistest seletavatest tunnustest ei sõltu. Kuigi tegelikud parameetrite β_1 ja β_2 väärtused on vastavalt 10 ja 6, siis tugeva korrelatsiooni tõttu on nii $\lambda = 5$, $\lambda = 100$ kui ka $\lambda = 200$ korral keskmised parameetrite hinnangud sarnased.

Tabel 1. Kantregressiooni parameetrite hinnangute keskväärtused ja dispersioonid korduval simuleerimisel

		β_1	β_2	β_3	β_4
Tegelik väärtus		10	6	6	-3
Aritmeetiline keskmine	λ=0	9,8411	6,1493	6,0013	-2,9938
	λ=5	7,0078	6,8812	4,5441	-2,2255
	λ=100	1,9792	1,9622	0,8176	-0,4114
	λ=200	1,1275	1,1182	0,4395	-0,2239
Dispersioon	λ=0	3,2834	3,2569	0,0393	0,0410
	λ=5	0,0133	0,0130	0,0244	0,0238
	λ=100	0,0020	0,0020	0,0010	0,0008
	λ=200	0,0007	0,0007	0,0003	0,0002

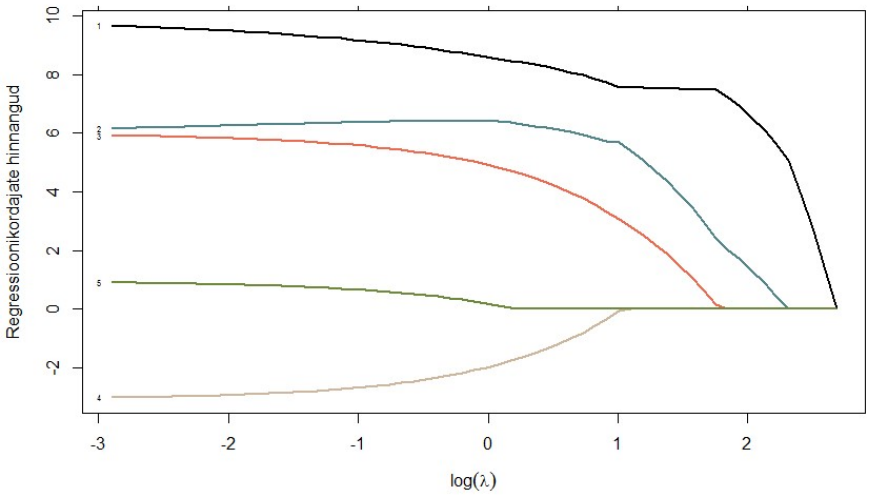
1.3 Lassoregressioon

Järgnev alajaotus tugineb teosel „An Introduction to Statistical Learning with Applications in R" (James et al., 2015, lk 219). Lassoregressioon (*Least Absolute Shrinkage and Selection Operator*) sarnaneb idee poolest kantregressiooniga. Samuti on hinnatava mudeli üldkuju ühine lineaarse regressioonimudeliga, kuid

parameetrite hinnangud $\hat{\beta}^L_{j,\lambda}$ saadakse minimeerides järgnevat suurust:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|, \qquad (13)$$

kus $\lambda \geqslant 0$. Seega kant- ja lassoregressioon erinevad parameetrite hindamise protsessis karistusliikme poolest, kus ühel juhul kasutatakse summeerimisel väärtusi β_j^2 ning teisel juhul $|\beta_j|$. Sellise karistusparameetri tõttu lassoregressiooni korral puudub analüütiline lahend (Hastie et al., 2013, lk 68). Ka lassoregressiooni puhul on väga oluline leida karistusparameetrile sobiv väärtus ning see saadakse samamoodi nagu kantregressiooni puhul, mida tutvustatakse lähemalt alajaotuses 1.5.



Joonis 2. Lassoregressiooni kordajate hinnangud

Kui $\lambda = 0$, siis saadakse parameetritele vähimruututе hinnangud ning λ kasvades parameetrite hinnangud lähenevad nullile. Erinevalt kantregressioonist, piisavalt suur λ võib lassoregressiooni korral anda tulemuse, kus mõne parameetri hinnang on võrdne nulliga. Seega lassoregressiooni korral selgub, missugused tunnused tuleks mudelisse kaasata, mis lihtsustab mudeli tõlgendamist. Joonisel 2 on ku-

jutatud lassoregressiooni parameetrite hinnangud sõltuvalt λ väärtusest. Mudelite loomiseks on kasutatud simuleeritud andmeid ($p = 5$). Jooniselt on näha, et hinnangute väärtused vähenevad λ kasvades, mitte küll tingimata monotoonselt, ning võivad saada teatud karistusparameetri väärtusest alates võrdseks nulliga.

Tabel 2. Lassoregressiooni parameetrite hinnangute keskväärtused ja dispersioonid korduval simuleerimisel

		β_1	β_2	β_3	β_4
Tegelik väärtus		10	6	6	-3
Aritmeetiline keskmine	λ=0	9,998	5,940	6,021	-2,997
	λ=1	8,428	6,485	4,953	-1,914
	λ=5	6,635	3,996	1,017	-0,002
	λ=10	3,308	1,755	0,000	0,000
Dispersioon	λ=0	106,646	104,963	1,021	1,015
	λ=1	37,324	36,096	1,019	0,963
	λ=5	20,796	19,777	0,720	0,001
	λ=10	6,030	5,334	0,000	0,000

Tabelis 2 on simuleeritud andmetelt leitud keskmised lassoregressiooni parameetrite hinnangud nelja erineva λ väärtuse korral. Tabeli 1 ja 2 loomiseks on kasutatud sama sisendmaatriksit ning tegelikud regressioonikordajate väärtused on samuti samad. Tabeli 2 põhjal võib järeldada, et karistusparameetri suurenedes dispersioon kahaneb ning nihe suureneb, nagu ka kantregressiooni korral.

1.4 Geomeetriline interpretatsioon

Kant- ja lassoregressiooni minimeerimisülesannet on võimalik ka teisiti väljendada, vastavalt

$$\min_{\beta_0,\ldots,\beta_p} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ tingimusel } \sum_{j=1}^{p} \beta_j^2 \leqslant s \qquad (14)$$

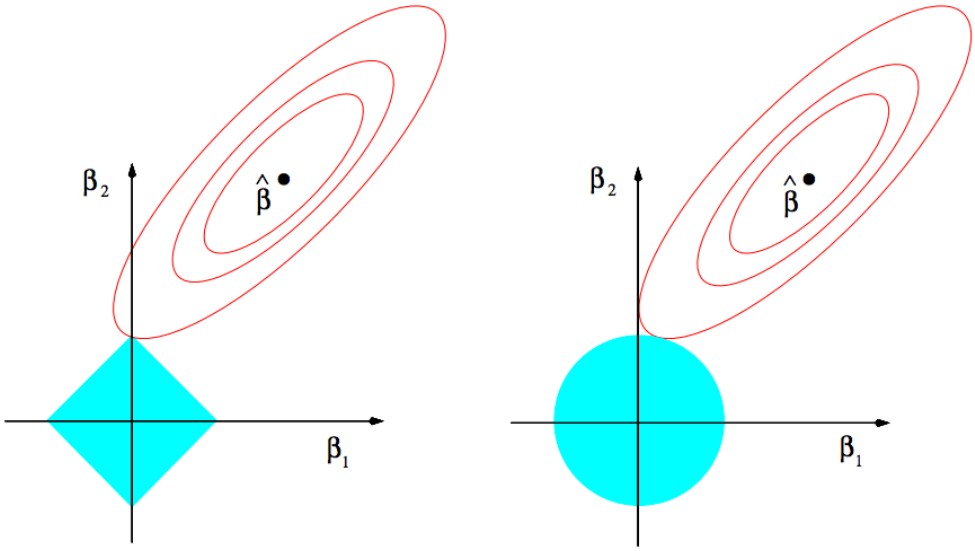
ning

$$\min_{\beta_0,\ldots,\beta_p} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ tingimusel } \sum_{j=1}^{p} |\beta_j| \leqslant s. \qquad (15)$$

Valemeid 9 ja 13 nimetatakse Lagrange'i vormideks ning on ekvivalentsed vastavalt valemitega 14 ja 15. See tähendab, et iga λ väärtuse jaoks leidub s, mille korral 9 ning 14 annavad samasuguse kantregressiooni parameetrite hinnangu. Samuti iga λ väärtusele vastab mingi s, mille korral 13 ning 15 annavad samasuguse lassoregressiooni parameetrite hinnangu. (James et al., 2015, lk 220-221)

Kui mudelis on kaks seletavat tunnust, siis saadud parameetrite hinnangud on väikseima vigade ruutude summaga piirkonnas $\beta_1^2 + \beta_2^2 \leqslant s$ kantregressiooni korral ning piirkonnas $|\beta_1| + |\beta_2| \leqslant s$ lassoregressiooni korral. Joonisel 3 on kujutatud vasakul lassoregressiooni ning paremal kantregressiooni, sinise kujundiga on märgitud vastavad piirkonnad. Sümbol $\hat{\beta}$ tähistab vähimruutude hinnangut parameetritele ning ellipsid selle ümber tähistavad punktide hulka, mille korral saadakse sama suurusega jääkide ruutude summa. Regulariseeritud regressiooni parameetrite hinnangud saadakse punktis, kus ellips ning tähistatud piirkond esimesena kokku langevad. Kuna kantregressiooni korral on selleks piirkonnaks ring, siis selle esmane kattumine ellipsiga ei ole üldiselt teljel ning seetõttu kantregressiooni kordajate väärtused ei ole võrdsed nulliga. Lassoregressiooni korral on piirkond rombikujuline, seega on sellel n-ö teravaid nurki ning esmane kattumine ellipsiga võib tihti olla just teljel. Kui see juhtub, siis üks hinnangutest on

võrdne nulliga. Sama idee kehtib ka mitmemõõtmelises ruumis, lassoregressiooni korral on lubatud piirkonnal n-ö teravaid nurki ning seega võib mõne hinnangu väärtus olla võrdne nulliga. (Hastie et al, 2013, lk 69-72)



Joonis 3. Jääkide ruutude summa kontuurid ning lasso- ja kantregressiooni kordajate piirkonnad kahe seletava tunnuse korral (Hastie et al., 2013, lk 71)

1.5 Karistusparameetri valimine

Parima karistusparameetri väärtuse valimiseks kasutatakse ristvalideerimist, mis on meetod mudeli täpsuse testimiseks. Selle rakendamiseks tuleb andmestikus olevad objektid juhuslikult jagada nii, et tekiks k ligikaudu sama mahuga gruppi, mida nimetatakse partitsioonideks. Esimesse partitsiooni kuuluvad objektid eemaldatakse andmestikust ning ülejäänud objektide pealt konstrueeritakse mudel. Eemaldatud objekte kasutatakse testandmestikuna, et uurida, kui hästi töötab mudel nende andmete peal, mida treenimiseks ei kasutatud. Lineaarse regressiooni korral saab mudeli sobivust mõõta näiteks ruutkeskmise vea abil, mis on partit-

siooni l pealt arvutatav kui

$$MSE_l = \frac{1}{n_l} \sum_{i=1}^{n_l} \left(y_i^l - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}^l \right)^2,$$

kus n_l on vaatluste arv partitsioonis l. Eemaldatud testandmestiku pealt arvutatakse ruutkeskmine viga MSE_1. Seejärel eemaldatakse kogu andmestikust teise partitsiooni kuuluvad objektid ning protsessi kordamisel saadakse MSE_2. Nii saadakse k ruutkeskmist viga ning ristvalideerimise hinnanguks on saadud vigade aritmeetiline keskmine,

$$CV_{(k)} = \frac{1}{k} \sum_{l=1}^{k} MSE_l.$$

Praktikas kasutatakse tihti ristvalideerimisel gruppide arvuna $k = 5$, $k = 10$ või $k = n$, nendest viimane kannab ingliskeelses kirjanduses nimetust *leave-one-out cross-validation*. (James et al., 2015, lk 181)

Karistusparameetri valikul lähtutakse ristvalideerimise tulemustest. Esmalt valitakse hulk λ võimalikke väärtusi ning seejärel arvutatakse iga λ korral ristvalideerimise viga. Karistusparameetriks valitakse selline λ, mille korral ristvalideerimise viga on kõige väiksem. (James et al., 2015, lk 227)

1.6 Regulariseeritud regressiooni eelised

Lineaarse sõltuvuse olemaolu korral tagab vähimruutude meetod teoreetiliselt nihketa hinnagu, kuid ruutkeskmine viga võib olla suur. Seda esineb sageli juhtudel, kui hinnatavate parameetrite arv on samas suurusjärgus valimimahuga. Seevastu regulariseeritud regressiooni korral on hinnang nihkega, kuid selle arvelt võib ruutkeskmine viga olla palju madalam. Idee seisneb ruutkeskmise vea lahutuses. (James et al., 2015, lk 127-128)

Olgu fikseeritud punkt $\boldsymbol{x}_0 = (x_{01},...,x_{0p})$ ja olgu funktsioonid f ja $\hat{f}$ defineeritud järgmiselt:

$$f(\boldsymbol{x}_0) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + ... + \beta_p x_{0p},$$

$$\hat{f}(\boldsymbol{x}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + ... + \hat{\beta}_p x_{0p}.$$

Seega $f(\boldsymbol{x}_0)$ väljendab tegelikku seost punktis $\boldsymbol{x}_0$ (eeldusel, et tegelik seos avaldub lineaarselt) ja $\hat{f}(\boldsymbol{x}_0)$ prognoositud väärtust antud punktis. Uurimaks, kui palju prognoositud väärtus tegelikust seosest fikseeritud punktis teoreetiliselt erineb, saab kasutada ruutkeskmist viga:

$$\begin{aligned}
MSE_f(\boldsymbol{x}_0) &= E[f(\boldsymbol{x}_0) - \hat{f}(\boldsymbol{x}_0)]^2 = \\
&= E[(f(\boldsymbol{x}_0))^2 - 2f(\boldsymbol{x}_0)\hat{f}(\boldsymbol{x}_0) + (\hat{f}(\boldsymbol{x}_0))^2] = \\
&= (f(\boldsymbol{x}_0))^2 - 2f(\boldsymbol{x}_0)E\hat{f}(\boldsymbol{x}_0) + E(\hat{f}(\boldsymbol{x}_0))^2 = \\
&= (f(\boldsymbol{x}_0))^2 - 2f(\boldsymbol{x}_0)E\hat{f}(\boldsymbol{x}_0) + (E\hat{f}(\boldsymbol{x}_0))^2 + \\
&+ E(\hat{f}(\boldsymbol{x}_0))^2 - (E\hat{f}(\boldsymbol{x}_0))^2 = \\
&= (f(\boldsymbol{x}_0) - E\hat{f}(\boldsymbol{x}_0))^2 + E(\hat{f}(\boldsymbol{x}_0))^2 - (E\hat{f}(\boldsymbol{x}_0))^2 = \\
&= [B(\hat{f}(\boldsymbol{x}_0))]^2 + D[\hat{f}(\boldsymbol{x}_0)].
\end{aligned}$$

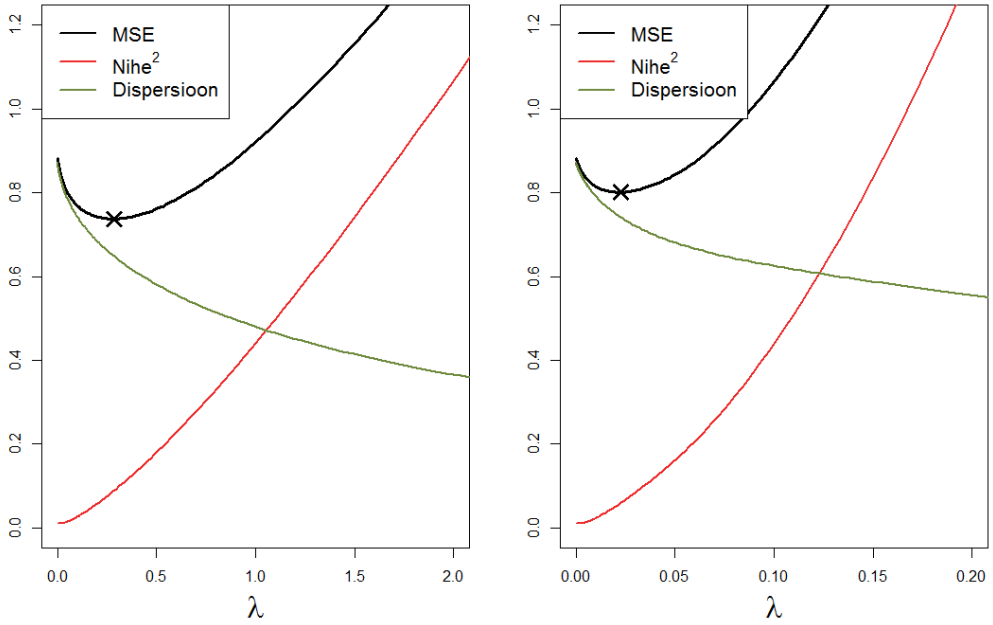
Ruutkeskmine viga punktis $\boldsymbol{x}_0 = (x_{01},...,x_{0p})$ avaldub seega suuruse $\hat{f}(\boldsymbol{x}_0)$ nihke ruudu ning dispersiooni summana. (Rojas, 2015)

Kogu valimile vastav toereetiline ruutkeskmine viga on leitav kui aritmeetiline keskmine üle valimi punktide, st

$$MSE_f = \frac{1}{n}\sum_{i=1}^{n} MSE_f(\boldsymbol{x}_i).$$

Joonisel 4 kujutatakse ruutkeskmise vea lahutust sõltuvalt karistusparameetrist simuleeritud andmetel (n=50, p=45). Punane joon tähistab nihke ruutu, roheline

joon dispersiooni, must joon ruutkeskmist viga ning punkt „×” märgib selle minimaalset punkti. Kui $\lambda = 0$, siis on tegemist vähimruutude hinnangul saadud ruutkeskmise veaga. Jooniselt on näha, et mõlema regulariseeritud regressiooni korral leidub $\lambda > 0$, nii et ruutkeskmine viga on väiksem kui tavalise lineaarse regressiooni korral. Seega väikest nihet kompenseerib oluliselt madalam dispersioon.



Joonis 4. Ruutkeskmise vea lahutus kant- (vasakul) ja lassogregressiooni (paremal) korral

Üldiselt annab lassoregressioon paremaid tulemusi, kui vaadeldavatest tunnustest omavad mõju vähesed. Kantregressiooni tulemused on paremad juhul, kui uuritav tunnus sõltub paljudest seletavatest tunnustest. Kuna praktikas pole teada, kui paljudest seletavatest tunnustest uuritav tunnus sõltub, siis kasutatakse ristvalideerimist, et otsustada, kumba lähenemist kasutada. Kui mudelisse kaasatakse väga palju tunnuseid, siis on kantregressiooni väljundit keeruline interpreteerida. Lassoregressioon teostab argumentide valikut ning see on oluline eelis nii tavalise kui

ka kantregressiooni ees. (James et al., 2015, lk 223-224)

Vähimruutude hinnang on avaldatud valemis 6 ning see eksisteerib vaid juhul, kui leidub $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Kui $p > n$, siis maatriksi X veerud ei saa olla lineaarselt sõltumatud, mistõttu maatriks X on singulaarne ning sellist pöördmaatriksit ei eksisteeri. Maatriksi $\boldsymbol{X}^T\boldsymbol{X}$ peadiagonaali elementidele $\lambda > 0$ lisamisel saadakse pööratav maatriks. Just selline pöördmaatriks $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$ on kantregressiooni hinnangu leidumise eelduseks ning seetõttu on võimalik leida kantregressiooni hinnangud parameetritele ka siis, kui $p > n$. Singulaarsusest tulenev probleem vähimruutude hinnangute leidmisel oli esialgne põhjus, miks kantregressioon kasutusele võeti. (Hastie, 2013, lk 64)

Regulariseeritud regessioonidel on arvutuslikud eelised parima mudeli leidmise jaoks. Vähimruutude meetodi puhul on parima mudeli saamiseks võimalik valida 2^p mudeli seast, mis arvutuslikult on väga ajamahukas juba üsna väikese parameetrite arvu puhul. Regulariseeritud regressiooni korral on iga λ jaoks vaid üks mudel ning parameetrite hindamise ajakulu on väga väike. (James et al., 2015, lk 218-219)

1.7 Logistiline regressioon

Järgnev alapeatükk põhineb töödel „Applied Logistic Regression" (Hosmer, Lemeshow, 2000, lk 6-8, 31-32) ning „Categorical Data Analysis" (Agresti, 2002, lk 192-193).

Sageli soovitakse konstrueerida mudelit, kus sõltuval tunnusel Y on vaid kaks võimalikku väärtust: jah/ei, on/ei ole. Enamasti kodeeritakse tunnuse Y huvipakkuv sündmus väärtusega 1 ning vastandsündmus väärtusega 0. Eesmärgiks on hinnata huvipakkuva sündmuse esinemise tõenäosust $\pi_i = P(Y_i = 1)$ ning seega

peab prognoositav väärtus olema mitte suurem kui üks ning mitte väiksem kui null.

Sellises olukorras kasutatakse $logit$-seosefunktsiooni, mis on defineeritud kui

$$g(\pi_i) = \ln \frac{\pi_i}{1-\pi_i}.$$

Logistilise regressiooni puhul on $logit$-seosefunktsioon võrdne seletavate tunnuste lineaarkombinatsiooniga ehk mudeli üldkuju on

$$\ln \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \tag{16}$$

kus β_0,...,β_p on mudeli parameeterid ning x_{ij} on i-nda objekti j-nda sõltumatu tunnuse väärtus ($j = 0$,...,p; $i = 1$,...,n; p on seletavate tunnuste arv, n on valimimaht). Valemist (16) avaldub sündmuse esinemise tõenäosus

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}} = \left(1+e^{-(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})}\right)^{-1}. \tag{17}$$

Logistilise regressiooni korral leitakse parameetrite hinnangud suurima tõepära meetodiga. Selle ideeks on maksimeerida tõepärafunktsiooni, mis avaldub kujul

$$L(\theta) = \prod_{i=1}^{n} p(y_i;\theta),$$

kus $p(y_i;\theta)$ on tunnuse Y tõenäosusfunktsioon ning θ on tundmatu parameeter. Sageli kasutatakse suurima tõepära hinnangu leidmisel tõepärafunktsiooni logaritmi

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln p(y_i;\theta).$$

Kui sõltuv tunnus Y on kodeeritud $1/0$, siis i-nda objekti sõltuv tunnus Y_i on Bernoulli jaotusega, $Y_i \sim Bern(\pi_i)$. Bernoulli jaotuse tõenäosusfunktsiooniks on $P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$. Seega logistilise regressiooni korral on tõepärafunktsiooniks

$$L(\beta_0,...,\beta_p) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} \prod_{i=1}^{n}(1-\pi_i)$$

ning tõepärafunktsiooni logaritm avaldub kujul

$$\begin{aligned}\ell(\beta_0,...,\beta_p) = \ln L(\beta_0,...,\beta_p) = \sum_{i=1}^{n} \ln \left(\frac{\pi_i}{1-\pi_i} \right)^{y_i} + \sum_{i=1}^{n} \ln(1-\pi_i) = \\ = \sum_{i=1}^{n} y_i \ln \frac{\pi_i}{1-\pi_i} + \sum_{i=1}^{n} \ln(1-\pi_i).\end{aligned} \tag{18}$$

Valemist (17) järeldub, et $1-\pi_i = \left(1 + e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}\right)^{-1}$. Selle järelduse ning valemi (16) põhjal saadakse logaritmiliseks tõepärafunktsiooniks

$$\begin{aligned}\ell(\beta_0,...,\beta_p) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) + \\ - \sum_{i=1}^{n} \ln \left(1 + e^{\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}}\right).\end{aligned} \tag{19}$$

Seega logistilise regressiooni parameetrite hinnangud saadakse suurust (19) maksimeerides.

1.8 Regulariseeritud logistiline regressioon

Kant- ja lassoregressiooni on võimalik rakendada ka logistilisele regressioonimudelile. Lineaarse regressiooni puhul lisatakse karistusparameeter vähimruutude meetodil minimeeritavale suurusele ning logistilise regressiooni korral on minimeeritav suurus sarnane. Log-tõepärafunktsiooni maksimeerimine on samaväärne negatiivse log-tõepärafunktsiooni minimeerimisega ning seega regulariseerimiseks liidetakse karistusliige negatiivsele log-tõepärafunktsioonile ning saadud suurus minimeeritakse. Seega suuruse

$$-\ell(\beta_0,...,\beta_p) + \lambda \sum_{i=1}^{p} \beta_j^2 \tag{20}$$

minimeerimisel saadakse parameetrite hinnangud logistilise kantregressiooni korral (Elkan, 2014, lk 11-12) ning suuruse

$$-\ell(\beta_0,...,\beta_p) + \lambda \sum_{i=1}^{p} |\beta_j| \tag{21}$$

minimeerimisel saadakse parameetrite hinnangud logistilise lassoregressiooni korral (Hastie et al, 2013, lk 125).

Karistusparameetri valimiseks kasutatakse ristvalideerimist, mida on kirjeldatud alajaotuses 1.5. Erinevus seisneb selles, et logistilise regressiooni korral hinnatakse mudeli headust mitte ruutkeskmise vea, vaid hälbimuse (*deviance*) põhjal (Friedman et al., 2010, lk 17-18). Hälbimus avaldub valemiga (Hosmer, Lemeshow, 2000, lk 13)

$$D = -2\,\ell(\hat{\beta}_0,...,\hat{\beta}_p), \tag{22}$$

kus $\hat{\beta}_0,...,\hat{\beta}_p$ on vastavalt meetodile saadud parameetrite hinnangud. Mida väiksem on hälbimus, seda parem on mudel.

2 Müügiskoor

Müügiskoori eesmärgiks on leida üles ettevõtted, kes võiksid vajada Creditinfo AS tooteid ja teenuseid, ning seeläbi lihtsustada müügimeeskonna tööd. Müügiskoori abil soovitakse hinnata, kui suure tõenäosusega võiks ettevõte olla Creditinfo Eesti AS klient. Mudelist saadavat prognoosi nimetatakse ostupotentsiaaliks. Müügiskoori loomiseks kaasatakse tunnuseid, mis iseloomustavad ettevõtete suurust ja eripära ning on praktikas kasutusel potentsiaalsete klientide leidmiseks.

Müügiskoori loomiseks kasutatakse regulariseeritud regressiooni, kuna soovitakse, et parameetrite hinnangud ei oleks väga suured. Kuna mitmed tunnused kirjeldavad ettevõtete eripära, siis võib esineda andmestikus multikollineaarsust, mille korral võiks samuti regulariseeritud regressioon anda paremaid tulemusi. Müügiskoori konstrueerimiseks kasutatakse statistikatarkvara R paketti „glmnet", mida tutvustatakse järgmises alapeatükis.

2.1 R-i pakett „glmnet"

Statistikapaketi „glmnet" pealkirjaga „*Lasso and Elastic-Net Regularized Generalized Linear Models*" on loonud J. Friedman, T. Hastie, N. Simon ja R. Tibshirani ning järgnev tutvustus tugineb paketi dokumentatsioonile (Friedman et al., 2017). Pakett „glmnet" sisaldab väga tõhusaid protseduure nii lineaarse, logistilise, multinomiaalse, Poissoni kui ka Coxi regressioonimudeli hindamiseks regulariseeritud regressiooni korral. Pakett koosneb viiest funktsioonist, neist lähemalt tutvustatakse käesolevas töös funktsioone `glmnet` ja `cv.glmnet`.

Funktsiooni `glmnet` kasutatakse mudeli hindamiseks ning selle tähtsamad argumendid on:

- `x` - seletavate tunnuste maatriks: $n \times p$;

- `y` - sõltuv tunnus;
- `family` - hinnatava regressioonimudeli tüüp, väärtusteks `"gaussian"`, `"binomial"`, `"poisson"`, `"multinomial"`, `"cox"`, `"mgaussian"`;
- `lambda` - karistusparameetrite vektor;
- `standardize` - tõese väärtuse korral maatriksis `x` olevad tunnused standardiseeritakse;
- `alpha` - karistusliiget määrav parameeter, $0 \leqslant \alpha \leqslant 1$. Karistusliige on defineeritud kui
$$\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|,$$
seega on võimalik kasutada kant- ja lassoregressiooni karistusliiget korraga ühe mudeli parameetrite hindamisel. Antud töös käsitletakse kant- ja lassoregressiooni vaid eraldi, $\alpha = 0$ annab tulemuseks kantregressiooni mudeli ning $\alpha = 1$ annab tulemuseks lassoregressiooni mudeli.

Funktsiooni `cv.glmnet` teostab k partitsiooniga ristvalideerimist `glmnet` objekti jaoks ning selle olulisemad argumendid on:

- `x` - seletavate tunnuste maatriks: $n \times p$;
- `y` - sõltuv tunnus;
- `family` - hinnatava regressioonimudeli tüüp(vt funktsiooni `glmnet`);
- `lambda` - karistusparameetrite vektor;
- `nfolds` - partitsioonide arv;
- `type.measure` - ristvalideerimisel hinnatav suurus, võimalikud väärtused on `"deviance"` (hälbimus), `"mse"` (ruutkeskmine viga), `"mae"` (keskmine absoluutne viga), `"class"` (valesti klassifitseerimise viga) või `"auc"` (ROC-kõvera alune pindala).

2.2 Andmestik

Valimisse kuulub 4737 ettevõtet ning valim on moodustatud nii, et tulemusi oleks võimalik üldistada uute klientide leidmiseks. Seletavaid tunnuseid on andmestikus 15. Creditinfo Eesti AS soovil on antud töös tunnuste nimetused varjatud, mistõttu on seletavad tunnused tähistatud tähe T ja järjenumbri abil. Tunnused jaotuvad tüübi alusel järgmiselt:

- pidevad tunnused - T2, T4, T5, T6, T7, T8, T9, T10, T14;
- järjestustunnused - T13;
- binaarsed tunnused - T1, T3, T11, T12, T15.

Sõltuva tunnuse tähiseks on Y ning tegemist on binaarse tunnusega, mis on kodeeritud järgnevalt:

$$y = \begin{cases} 1, & \text{kui ettevõte on klient} \\ 0, & \text{kui ettevõte ei ole klient.} \end{cases}$$

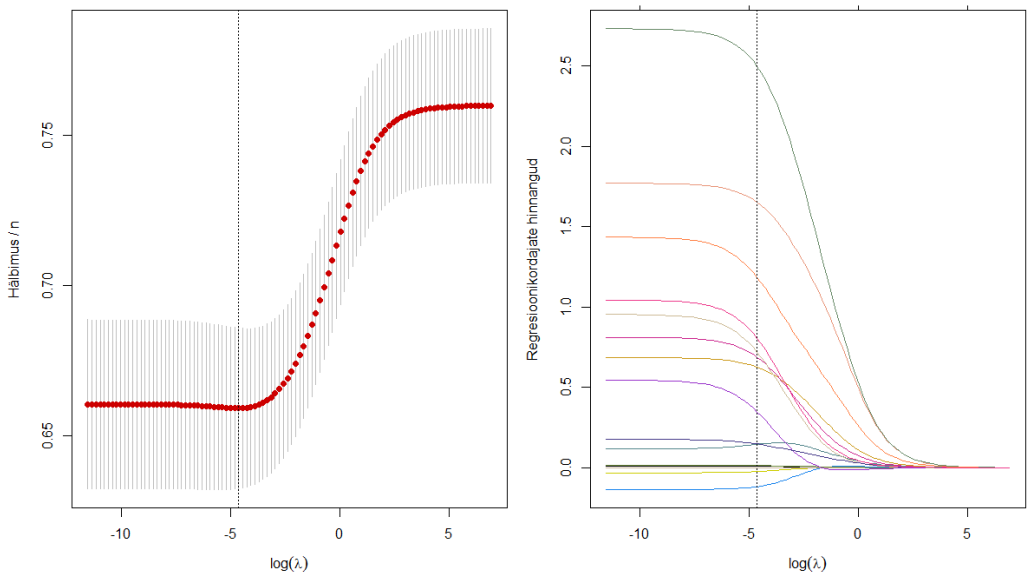
Andmestikus on klientide osakaal 12,6%. Mitteklientideks on ettevõtted, kes on teatud ajaperioodil Creditinfo toodetega tutvunud, kuid neist ei saanud Creditinfo kliente.

2.3 Müügiskoori konstrueerimine

Müügiskoori loomiseks hinnatakse logistiline regressioonimudel nii kant- kui ka lassoregressiooni meetodil. Andmestik jagatakse esmalt lihtsa juhuvaliku põhjal kaheks osaks - treeningandmestikuks võetakse 75% valimist ning ülejäänud objektid kuuluvad testandmestikku. Treeningandmestikul leitakse parim karistusparameeter ning seejärel leitakse testandmestikul hälbimus. Kant- ja lassoregressiooni parameetrite hindamiseks kasutatakse kogu andmestikku ning parima mu-

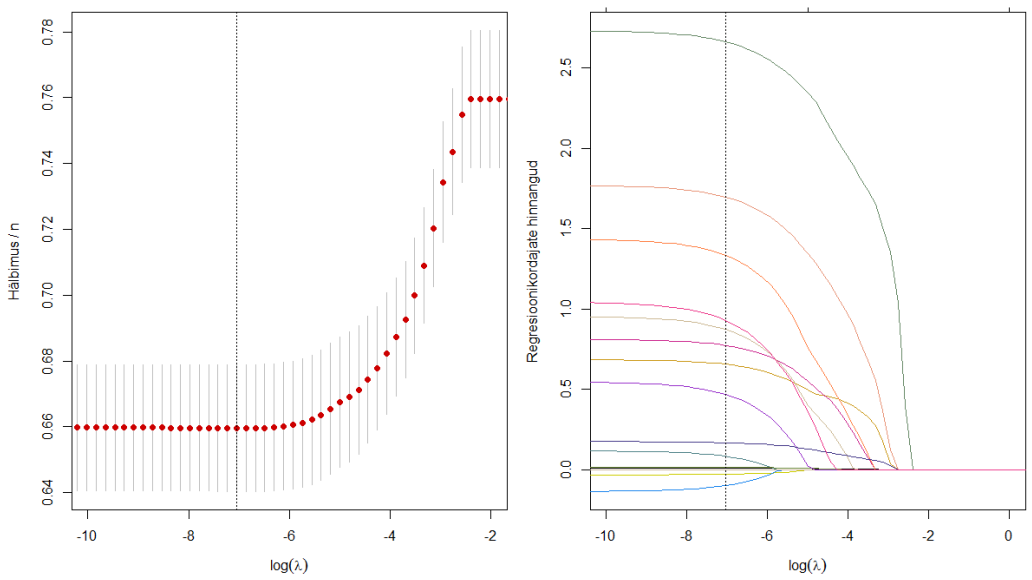
deli valimiseks leiatakse uuesti parim karistusparameeter kogu andmestikku kasutades.

Joonisel 5 on punasega tähistatud hälbimus vastavalt karistusparameetrile, halliga on tähistatud standardviga ning vertikaalne joon tähistab minimaalse karistusparameetri kohta mõlemal graafikul. Kantregressiooni korral saavutatakse minimaalne hälbimus, kui $\lambda = 0{,}0098$. Parempoolselt jooniselt võib järeldada, et regressioonikordajate hinnangud on parima karistusparameetri korral teatud määral väiksemad kui suurima tõepära meetodil saadavad hinnangud.



Joonis 5. Vasakul hälbimus ning paremal regressioonikordajad vastavalt karistusparameetrile kantregressiooni korral

Joonisel 6 on loodud samade tähistustega kui joonis 5. Jooniselt on näha, et parim mudel saavutatakse, kui karistusparameetri väärtus on väga väike. See on mõistetav, kuna lassoregressiooni korral lähenevad kordajate hinnangud kiiremini nullile kui kantregressiooni korral ning võivad saada väärtuseks ka nulli. Kuna mudelisse on kaasatud muutujad, mida juba praktikas kasutatakse, siis on oodatav,



Joonis 6. Vasakul hälbimus ning paremal regressioonikordajad vastavalt karistusparameetrile lassoregressiooni korral

et regressioonikordajate hinnangud pole võrdsed nulliga. Minimaalne hälbimus saavutatakse, kui $\lambda = 0{,}0013$.

Mudeleid võib võrrelda testandmestikul leitud hälbimuse põhjal. Kantregressioonimudeli korral on hälbimus $D = 818{,}3$ ning lassogressioonimudeli korral $D = 819{,}0$. Oodatavalt ka hälbimused on sarnased ning ei mõjuta kindlat mudelit eelistama. Lisaks on leitud hälbimus $\lambda = 0$ korral. Hälbimus on sel juhul $D = 820{,}0$, mis on küll antud juhul väga väiksel määral suurem, kuid üldiselt antud andmete puhul ei ole kant- ega lassoregressioonil olulist eelist võrreldes suurima tõepära meetodiga.

Tabelis 3 on toodud nii kant- kui ka lassoregressiooni kõigi parameetrite hinnangud tavalisel ning standardiseeritud kujul parimate karistusparameetrite korral. Hinnangud on sarnased ning ühelgi juhul pole hinnang võrdne nulliga. Sellest tulenevad ka sarnased prognoosid ettevõtete ostupotentsiaalile ning seega pole olu-

list põhjust eelistada ühte mudelit teisele. Autor otsustab kasutusele võtta kantregressiooniga hinnatud mudeli. Standardiseeritud parameetrite põhjal on võimalik hinnata, kui suurt mõju avaldab mingi tunnus sõltuvale tunnusele. Suurimat mõju avaldavad tunnused T3, T12 ning T13. Suurem mõju on veel ka tunnustel T1, T4, T14 ning T15.

Tabel 3. Kant- ja lassoregressiooni mudelite parameetrite hinnangud parima karistusparameetri korral

Tunnus	$\hat{\beta}_j^R$	$\hat{\beta}_j^L$	$\hat{\beta}_{j,st}^R$	$\hat{\beta}_{j,st}^L$
Vabaliige	–4,0031	-4,0962	-2,147	-2,158
T1	0,5582	0,5570	0,220	0,219
T2	-0,0160	-0,0161	-0,096	-0,097
T3	0,6829	0,7151	0,312	0,326
T4	0,0070	0,0070	0,229	0,231
T5	0,0006	0,0002	0,016	0,007
T6	0,2574	0,2140	0,056	0,047
T7	0,0001	0,0001	0,156	0,155
T8	-0,1564	-0,1282	-0,049	-0,040
T9	0,0004	0,0004	0,085	0,081
T10	0,0133	0,0132	0,129	0,128
T11	1,5474	1,5366	0,146	0,145
T12	2,5999	2,6622	0,287	0,294
T13=2	0,4752	0,5428	0,233	0,266
T13=3	0,8270	0,9087	0,387	0,424
T13=4	1,3472	1,4283	0,234	0,248
T14	0,1630	0,173	0,205	0,219
T15	0,7020	0,7107	0,184	0,186

Kokkuvõte

Antud bakalaureusetöö eesmärgiks oli tutvustada kant- ja lassoregressiooni ning rakendada neid müügiskoori loomiseks. Müügiskoor konstrueeriti kasutades Creditinfo Eesti AS andmeid. Töö esimeses osas anti ülevaade lineaarsest regressioonist, lineaarsest kant- ja lassoregressioonist, logistilisest regressioonist ning logistilisest kant-ja lassoregressioonist. Töö teises osas tutvustati R-i paketti ,,glmnet", kasutatavat andmestikku ning seejärel koostati mudelid.

Kant- ja lassoregressiooni kasutatakse sageli juhul, kui hinnatavate parameetrite arv on suurem kui vaatluste arv, kuna klassikalisi meetodeid pole alati võimalik sellisel juhul kasutada. Kant- ja lassoregressiooni parameetrite hinnangud on nihkega, kuid nende abil on võimalik parameetrite hinnangute hajuvust vähendada. Lassoregressioon teostab ka tunnuste kaasamise valikut.

Müügiskoori loomisel andsid kant- ja lassoregressioon sarnaseid tulemusi. Kuna andmestikku kuulusid tunnused, mida praktikas kasutatakse ettevõtete ostupotentsiaali hindamiseks, siis oodatavalt lassoregressiooni parimasse mudelisse kaasati kõik tunnused.

Kasutatud kirjandus

[1] Agresti, A., (2002), *Categorical Data Analysis*, Second Edition, Wiley.

[2] Elkan, C., (2014), *Maximum Likelihood, Logistic Regression and Stochastic Gradient Training*, San Diego: University of California. Kasutatud 28.04.2017. http://cseweb.ucsd.edu/~elkan/250B/logreg.pdf

[3] Friedman, J., Hastie, T., Simon, N., Tibshirani, R., (2017), *Package 'glmnet', Lasso and Elastic-Net Regularized Generalized Linear Models*. Kasutatud 02.05.2017. https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

[4] Friedman, J., Hastie, T., Tibshirani, R., (2010), Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, **33**(1), 1-22. Kasutatud 12.04.2017. https://www.jstatsoft.org/article/view/v033i01/v33i01.pdf

[5] Hastie, T., Tibshirani, R., Friedman, J., (2013), *The Elements of Statistical Learning*, Second Edition, 10th printing, New York: Springer.

[6] Hosmer, D. W., Lemeshow, S., (2000), *Applied Logistic Regression*, Second Edition, New York: Wiley.

[7] James, G., Witten, D., Hastie, T., Tibshirani, R., (2015), *An Introduction to Statistical Learning with Applications in R*, 6th printing, New York: Springer.

[8] Montgomery, D. C., Peck, E. A., Vining, G. G., (2013), *Introduction to Linear Regression Analysis*, 5th Edition, Hoboken: Wiley.

[9] Rojas, R., (2015), *The Bias-Variance Dilemma*. Kasutatud 05.05.2017. https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/bias.pdf

[10] Traat, I., (2006), *Matemaatilise statistika põhikursus*, Tartu: Tartu Ülikool, matemaatilise statistika instituut.

Lisad

```
#regressioonikordajate joonised
library(glmnet)
#genereerin normaaljaotusega arvud
set.seed(1)
n=100
x1 = rnorm(n); x2 = x1 + rnorm(n, sd=0.1)
x3 = rnorm(n); x4 = rnorm(n); x5 = rnorm(n)
y = 2 + 10*x1 + 6*x2 + 6*x3 - 3*x4 + x5+ rnorm(n)
X = cbind(x1, x2, x3, x4, x5)
#sobitan kantreg mudeli
kantreg=glmnet(X, y, alpha=0)
par(mfrow=c(1,1))
plot(kantreg, xvar="lambda",
  xlab=expression(paste(log(lambda))), label=TRUE,
  col=c("black","cadetblue4","coral2","bisque3",
  "darkolivegreen4"), lwd=2,
  ylab="Regressioonikordajate hinnangud")
#sobitan lassoreg mudeli
lassoreg=glmnet(X, y, alpha=1)
plot(lassoreg, xvar="lambda", label=TRUE,
  xlab=expression(paste(log(lambda))), lwd=2,
  col=c("black","cadetblue4","coral2","bisque3",
  "darkolivegreen4"), ylab="Regressioonikordajate hinnangud")

#disp ja kv tabelid
set.seed(1)
n=100
x1 = rnorm(n); x2 = x1 + rnorm(n, sd=0.1)
x3 = rnorm(n); x4 = rnorm(n)
mu = 2 + 10*x1 + 6*x2 + 6*x3 - 3*x4
X = cbind(x1, x2, x3, x4)
R=1000
lam=c(0, 5, 100, 200); laml=c(0, 1, 5, 10)
```

```
nlam=length(lam)
beeta1<-matrix(0,R,nlam); beeta2<-matrix(0,R,nlam)
beeta3<-matrix(0,R,nlam); beeta4<-matrix(0,R,nlam)
lbeeta1<-matrix(0,R,nlam); lbeeta2<-matrix(0,R,nlam)
lbeeta3<-matrix(0,R,nlam); lbeeta4<-matrix(0,R,nlam)
for(i in 1:R){
  y.k=mu+rnorm(n,sd=2)
  y.l=mu+rnorm(n,sd=10)
  kantreg<-glmnet(X,y.k,alpha=0,lambda=lam)
  beeta1[i,]<-coef(kantreg)[2,]
  beeta2[i,]<-coef(kantreg)[3,]
  beeta3[i,]<-coef(kantreg)[4,]
  beeta4[i,]<-coef(kantreg)[5,]
  lassoreg<-glmnet(X,y.l,alpha=1,lambda=laml)
  lbeeta1[i,]<-coef(lassoreg)[2,]
  lbeeta2[i,]<-coef(lassoreg)[3,]
  lbeeta3[i,]<-coef(lassoreg)[4,]
  lbeeta4[i,]<-coef(lassoreg)[5,]
}
b1.var<-apply(beeta1,2,var); b2.var<-apply(beeta2,2,var)
b3.var<-apply(beeta3,2,var); b4.var<-apply(beeta4,2,var)
b1.m<-apply(beeta1,2,mean); b2.m<-apply(beeta2,2,mean)
b3.m<-apply(beeta3,2,mean); b4.m<-apply(beeta4,2,mean)
b1<-cbind(rev(b1.m),rev(b1.var))
b2<-cbind(rev(b2.m),rev(b2.var))
b3<-cbind(rev(b3.m),rev(b3.var))
b4<-cbind(rev(b4.m),rev(b4.var))
lb1.var<-apply(lbeeta1,2,var); lb2.var<-apply(lbeeta2,2,var)
lb3.var<-apply(lbeeta3,2,var); lb4.var<-apply(lbeeta4,2,var)
lb1.m<-apply(lbeeta1,2,mean); lb2.m<-apply(lbeeta2,2,mean)
lb3.m<-apply(lbeeta3,2,mean); lb4.m<-apply(lbeeta4,2,mean)
lb1<-cbind(rev(lb1.m),rev(lb1.var))
lb2<-cbind(rev(lb2.m),rev(lb2.var))
lb3<-cbind(rev(lb3.m),rev(lb3.var))
lb4<-cbind(rev(lb4.m),rev(lb4.var))
```

```
#mse lahutuse joonis, aluseks voetud kood:
#http://www.stat.cmu.edu/~ryantibs/datamining/lectures/16-modr1.R
set.seed(1)
n<-50;p<-45
x<-matrix(rnorm(n*p),nrow=n)
bstar<-runif(p,-1,1)
mu<-as.numeric(x%*%bstar)
R<-100;nlam<-600
lam<-10**seq(-5,5,length=nlam)
fit.rid<-array(0,dim=c(R,nlam,n))
fit.lasso<-array(0,dim=c(R,nlam,n))
for (i in 1:R) {
  y = mu + rnorm(n)
  aa1<-glmnet(x,y,lambda=lam,alpha=0,intercept = FALSE,
  thresh = 10e-12)
  fit.rid[i,,]<-t(predict(aa1,s=lam,type="response",newx=x))
  aa2<-glmnet(x,y,lambda=lam,alpha=1,intercept = FALSE,
  thresh = 10e-12)
  coef(aa2)
  fit.lasso[i,,]<-t(predict(aa2,s=lam,type="response",newx=x))
}
bias.rid = (rowSums(scale(apply(fit.rid,2:3,mean),
  center=mu,scale=F)^2)/n)
var.rid = rowSums(apply(fit.rid,2:3,var))/n
bias.lasso = (rowSums(scale(apply(fit.lasso,2:3,mean),
  center=mu,scale=F)^2)/n)
var.lasso = rowSums(apply(fit.lasso,2:3,var))/n
mse.rid = bias.rid + var.rid
mse.lasso = bias.lasso + var.lasso
par(mfrow=c(1,2))
par(mar=c(4.5,4.5,0.5,0.5))
plot(lam,mse.rid,type="l",ylim=c(0,1.2),xlim=c(0,2),
  xlab=expression(paste(lambda)),ylab="",lwd=3,cex.lab=2)
lines(lam,bias.rid,col="firebrick2",lwd=2)
lines(lam,var.rid,col="darkolivegreen4",lwd=2)
```

```
points(lam[mse.rid==min(mse.rid)],min(mse.rid),lwd=3,
  cex=2,pch=4)
legend("topleft",lty=c(1,1,1), lwd=2,cex=1.3,
  legend=expression(MSE,paste(Nihe**2),Dispersioon),
  col=c("black","firebrick2","darkolivegreen4"))
par(mar=c(4.5,4.5,0.5,0.5))
plot(lam,mse.lasso,type="l",ylim=c(0,1.2),
  xlab=expression(paste(lambda)),
  ylab="",lwd=3,xlim=c(0,0.2), cex.lab=2)
lines(lam,bias.lasso,col="firebrick2",lwd=2)
lines(lam,var.lasso,col="darkolivegreen4",lwd=2)
points(lam[mse.lasso==min(mse.lasso)],min(mse.lasso),
  lwd=3,cex=2,pch=4)
legend("topleft",lty=c(1,1,1), lwd=2,cex=1.3,
  legend=expression(MSE,paste(Nihe**2), Dispersioon),
  col=c("black","firebrick2","darkolivegreen4"))
#muugiskoor
library(readxl)
library(dismo)
andmestik<-read_excel(
  "C:/Users/pertt/Documents/myygiskoor_dataset_noreg.xlsx",
  sheet="mudeliks")
veerud<-c(3,4,7,8,12,16,19,22,23,24,25,26,30,31,32,35,36,54)
andmed<-andmestik[,veerud];andmed[,5]<-andmed[,5]/1000
andmed[,16]<-andmed[,16]/1000
andmed<-as.data.frame(andmed)
andmed.mtrks<-as.matrix(andmed[,-ncol(andmed)])
y<-andmed[,ncol(andmed)]
#jagan andmestiku treening- ja testandmestikuks
set.seed(454588)
treening<-sample (1: nrow(andmed), nrow(andmed)/4*3)
test=(-treening)
lam=c(0,10**seq(-5,3,length=100))
par(mfrow=c(1,2))
```

```
#KANTREGRESSIOON
rv.kant<-cv.glmnet(andmed.mtrks[treening,],y[treening],
  alpha=0,lambda=lam,family="binomial",
  type.measure = "deviance")
plot(log(rv.kant$lambda),rv.kant$cvm,type="p",pch=16,
  ylim=c(min(rv.kant$cvlo),max(rv.kant$cvup)),col="red3",
  xlab=expression(paste(log(lambda))),
  ylab="Halbimus / n",xlim=c(-10,2))
segments(x0=log(rv.kant$lambda),
  y0=rv.kant$cvlo,y1=rv.kant$cvup,col="gray75")
points(log(rv.kant$lambda),rv.kant$cvm,pch=16,
  ylim=c(min(rv.kant$cvlo),max(rv.kant$cvup)),col="red3")
abline(v=log(rv.kant$lambda.min),lty=3 )
lam.kant<-rv.kant$lambda.min
#kantreg mudelid
kant.mudelid<-glmnet(andmed.mtrks[treening,],y[treening],
  alpha=0, lambda=lam,family="binomial")
plot(kant.mudelid,xvar="lambda", label=TRUE,
  xlab=expression(paste(log(lambda))),
  ylab="Regresioonikordajate hinnangud",
  col=c("goldenrod3","yellow3","maroon3","black",
  "grey","cadetblue4","coral2", "dodgerblue2", "bisque3",
  "darkolivegreen4","darksalmon","darkseagreen4",
  "darkorchid","wheat3","sienna1","slateblue4","violetred2"))
abline(v=log(lam.kant),lty=3)
#leian parima lambda korral halbimuse
#treening-ja testandmestikul
tr.kant.d<-calc.deviance(y[treening],predict(kant.mudelid,
  s=lam.kant,type="response",newx=andmed.mtrks[treening,]),
  calc.mean=FALSE)
test.kant.d<-calc.deviance(y[test],predict(kant.mudelid,
  s=lam.kant,type="response",
  newx=andmed.mtrks[test,]),calc.mean=FALSE)
#LASSOREGRESSIOON
#ristvalideerimine, parima lambda leidmiseks halbimuse pohjal
```

```
rv.lasso<-cv.glmnet(andmed.mtrks[treening,],y[treening],
  alpha=1, lambda=lam,family="binomial",
  type.measure = "deviance")
plot(log(rv.lasso$lambda),rv.lasso$cvm,
  ylim=c(min(rv.lasso$cvlo), max(rv.lasso$cvup)),
  type="p",pch=16,col="red3",
  xlab=expression(paste(log(lambda))),
  ylab="Halbimus / n",xlim=c(-10,-2))
segments(x0=log(rv.lasso$lambda),
  y0=rv.lasso$cvlo,y1=rv.lasso$cvup,col="gray75")
points(log(rv.lasso$lambda),rv.lasso$cvm,pch=16,
  ylim=c(min(rv.lasso$cvlo), max(rv.lasso$cvup)),col="red3")
  abline(v=log(rv.lasso$lambda.min),lty=3 )
lam.lasso<-rv.lasso$lambda.min
#lassoreg mudelid
lasso.mudelid<-glmnet(andmed.mtrks[treening,],y[treening],
  alpha=1, lambda =lam,family="binomial",thresh=1e-12)
plot(lasso.mudelid,xvar="lambda",label=TRUE,xlim=c(-10,0),
  xlab=expression(paste(log(lambda))),
  ylab="Regresioonikordajate hinnangud",
  col=c("goldenrod3", "yellow3", "maroon3", "black",
  "grey", "cadetblue4", "coral2","dodgerblue2",
  "bisque3","darkolivegreen4","darksalmon",
  "darkseagreen4","darkorchid","wheat3","sienna1",
  "slateblue4","violetred2"))
abline(v=log(lam.lasso),lty=3)
#leian parima lambda korral halbimuse
#treening-ja testandmestikul
tr.lasso.d<-calc.deviance(y[treening],predict(lasso.mudelid,
  s=lam.lasso,type="response",newx=andmed.mtrks[treening,]),
  calc.mean=FALSE)
test.lasso.d<-calc.deviance(y[test],predict(lasso.mudelid,
  s=lam.lasso, type="response",newx=andmed.mtrks[test,]),
  calc.mean=FALSE)
#leian halbimuse testandmestikul, kui lambda=0
```

```
#(VRM, vahet pole, kas kasutan lasso- v kantreg mudelit)
test.lin.d<-calc.deviance(y[test],predict(lasso.mudelid,s=0,
  type="response",newx=andmed.mtrks[test,]),calc.mean=FALSE)
test.kant.d;test.lasso.d;test.lin.d
#loplik mudel koigi andmete pealt
rv.kant.k<-cv.glmnet(andmed.mtrks,y,alpha=0,lambda=lam,
  family="binomial",type.measure = "deviance")
rv.lasso.k<-cv.glmnet(andmed.mtrks,y,alpha=1,lambda=lam,
  family="binomial",type.measure = "deviance")
lam.kant.k<-rv.kant.k$lambda.min
lam.lasso.k<-rv.lasso.k$lambda.min
sds<-apply(andmed.mtrks,2,sd)*sqrt((n-1)/n)
nrm.andmed<-scale(andmed.mtrks,center=TRUE,scale=FALSE)
nrm.andmed<-t(t(nrm.andmed)/sds)
kant.mudel<-glmnet(andmed.mtrks,y,alpha=0,
  lambda =lam.kant.k, family="binomial",thresh=1e-12)
lasso.mudel<-glmnet(andmed.mtrks,y,alpha=1,
  lambda=lam.lasso.k, family="binomial",thresh=1e-12)
sd.kant.mudel<-glmnet(nrm.andmed,y,alpha=0,
  lambda=lam.kant.k, family="binomial",thresh=1e-12,
  standardize = F)
sd.lasso.mudel<-glmnet(nrm.andmed,y,alpha=1,
  lambda =lam.lasso.k, family="binomial",
  thresh=1e-12,standardize = F)
round(coef(sd.kant.mudel),3)
round(coef(sd.lasso.mudel),3)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Perttu Narvik,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
„Kant- ja lassoregressioon ning nende rakendamine müügiskoori loomiseks Creditinfo Eesti AS andmetel“, mille juhendaja on Taavi Unt,

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, **09.05.2017**