Lithuanian Journal of Statistics *2018, vol. 57, No 1, pp. 56-69*
Lietuvos statistikos darbai *2018, 57 t., Nr. 1, 56-69 p.*
www.statisticsjournal.lt

REGRESINĖS ANALIZĖS TAIKYMAS DIDIESIEMS DUOMENIMS

Indrė Baltušninkaitė[1], Nomeda Bratčikovienė[2]

Vilniaus Gedimino technikos universiteto matematinės statistikos katedra.
Adresas: Saulėtekio al. 11, 10223Vilnius, Lietuva.
E-paštas: [1]baltusninkaite.i@gmail.com, [2]nomeda.bratcikoviene@vgtu.lt

Gauta: 2018 m. birželis Pataisyta: 2018 m. lapkritis Atnaujinta: 2019 m. gegužė

Santrauka. Šiame straipsnyje nagrinėjamos didžiųjų duomenų regresinės analizės galimybės ir galimi sunkumai. Straipsnyje išskirtos ir paaiškintos pagrindinės juos nusakančios charakteristikos, nustatyti galimi iššūkiai, kylantys didžiųjų duomenų analitikoje. Atsižvelgiant į tai, pasiūlyta keletas didžiųjų duomenų regresinėje analizėje naudojamų metodų, kurie leidžia sumažinti skaičiavimų naštą ir atrinkti nepriklausomus kintamuosius, geriausiai nusakančius priklausomą kintamąjį, bei pasiekti didesnį modelio tikslumą. Vienas iš darbo tikslų – metodų pritaikymas realiems didiesiems duomenims, todėl didelis dėmesys skiriamas tiriamajai daliai. Realių duomenų regresijos modelių sudarymui ir parametrų vertinimui naudojami išskaidytos ir stebinių įtakos indeksu paremtos regresijos metodai, o geriausiai priklausomąjį kintamąjį nusakančių nepriklausomų kintamųjų atrinkimui naudojama LASSO ir LARS regresija. Straipsnyje taip pat pateikiami atlikti modelių tinkamumo ir tikslumo vertinimai, jų tarpusavio rezultatų palyginimai.

Reikšminiai žodžiai: didieji duomenys, regresinė analizė, stebinių įtakos indeksu pagrįsta regresija, LASSO, LARS, RMSLE.

1. Įvadas

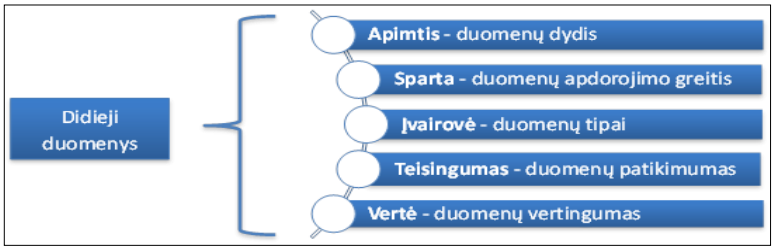
Sparti mokslo ir technologijų pažanga per pastarąjį dešimtmetį sąlygojo nepaprastai didelius duomenų kiekius, kurie vis dar auga milžinišku tempu. Ši didžiųjų duomenų era atveria naujas, beprecedentes galimybes tiek šiuolaikinei visuomenei, tiek verslui ir vyriausybėms, sprendžiant įvairius uždavinius, dėl kurių rezultatų galima padaryti visiškai naujas įžvalgas, kurios nebuvo įmanomos anksčiau. Kita vertus, tai kelia ir nemažai unikalių iššūkių mokslininkams ir analitikams dėl išaugusios skaičiavimų naštos ir klasikinės statistikos apribojimų.

Regresinė analizė – vienas populiariausių klasikinėje statistikoje taikomų duomenų analizės metodų. Jos pritaikymas visai didžiųjų duomenų aibei susiduria su sunkumais dėl atsiliekančių kompiuterinių pajėgumų ir dėl žymiai didesnio nepriklausomų kintamųjų skaičiaus. Todėl tokiose situacijose pagrindinį vaidmenį atlieka kintamųjų atrinkimo procedūros. Siekiant išspręsti šias problemas ir rasti balansą tarp modelio tikslumo ir skaičiavimų efektyvumo, būtina ieškoti naujų galimybių didžiųjų duomenų regresinei analizei atlikti.

Šio straipsnio tikslas yra apžvelgti regresinės analizės metodų taikymo didiesiems duomenims galimybės, kylančius iššūkius ir galimus sprendimo būdus, pritaikyti keletą didžiųjų duomenų regresinėje analizėje siūlomų metodų realiems duomenims, atlikti jų analizę, gauti statistines išvadas ir palyginti taikomų metodų tikslumą.

2. Literatūros apžvalga

Nors didžiųjų duomenų sąvoka pastaruoju metu girdima gana dažnai, unikalios apibrėžties ji neturi ir suprasti, ką ji iš tikrųjų reiškia yra ganėtinai sudėtinga. Viena paprasčiausių apibrėžčių, nusakančių didžiuosius duomenis – labai dideli duomenų kiekiai, dėl kurių apimties tradicinės saugojimo, apdorojimo ir analizės priemonės tampa netinkamos. Didžiuosius duomenis geriausiai apibūdina penkios charakteristikos, kurios pateiktos 1 paveiksle.



1 pav. Didžiųjų duomenų charakteristikos

Kalbant apie iššūkius, kylančius konkrečiai didžiųjų duomenų analitikoje, reikia pabrėžti tai, kad yra dirbama su labai didelės apimties duomenimis (tiek stebinių, tiek kintamųjų skaičiaus atžvilgiu) ir tai lemia šiuos sunkumus:

1) didelis kintamųjų skaičius sukelia triukšmo duomenyse kaupimąsi, klaidingas koreliacijas ir atsitiktinį homogeniškumą;

2) didelis kintamųjų skaičius kartu su dideliu stebinių skaičiumi didina skaičiavimų naštą ir kainą, taip pat skatina naudojamų algoritmų nestabilumą [5].

Visi šie sunkumai rodo išaugusį naujų technologijų, įrankių ir šiuolaikinių analizės metodų poreikį darbui su didžiaisiais duomenimis.

Nors Lietuvoje vis daugiau verslo įmonių ir viešojo sektoriaus institucijų stengiasi išnaudoti neišmatuojamas didžiųjų duomenų analitikos galimybes, vis dėlto mokslinių straipsnių ar mokomosios medžiagos, parašytos lietuvių kalba, beveik nėra. Todėl visa šio darbo literatūros apžvalga yra paremta būtent užsienio literatūra.

2015 m. vieno iš Pietų Korėjos universitetų mokslininkai Sunghae′as Jun′as, Seung-Joo′as Lee ir Jea-Bok Ryu‘as straipsnyje „Išskaidyta regresinė analizė didiesiems duomenims“ (angl. *A Divided Regression Analysis for Big Data*) pasiūlė naują metodologiją didžiųjų duomenų regresinei analizei atlikti, skirtą skaičiavimų naštos sumažinimui. Jų teigimu, viena didžiausių problemų, kylančių didžiųjų duomenų analizėje – statistinių metodų pritaikymas iš karto visiems duomenims. Tokiu atveju susiduriama tiek su klasikinės statistikos, tiek su skaičiavimų atlikimo apribojimais. Autorių pasiūlyta metodologija pagrįsta visų duomenų išskaidymu į pogrupius ir regresinės analizės pritaikymu kiekviename pogrupyje atskirai. Pogrupių sudarymas gali būti atliekamas taikant įvairius imčių išrinkimo metodus. Galutinis rezultatas gaunamas apskaičiavus visų pogrupių regresinių modelių parametrų įverčių vidurkį. Analizės rezultatai, atlikus šimtą modeliuotų duomenų iteracijų, parodė, kad visų duomenų regresijos modelio parametrai pateko į pogrupių regresijos modelių parametrų pasikliautinuosius intervalus, taip pat nustatyta, kad parametrų įverčių vidurkis yra labai artimas visų duomenų parametrų įverčiams. Galiausiai tie patys veiksmai buvo pritaikyti realiems duomenims. Nustatyta, kad nors pogrupių regresijos modelio parametrų įverčiai buvo gana skirtingi, tačiau jų vidurkis vis dėlto buvo labai artimas visų duomenų modelio parametrų įverčiams. Be to, kaip ir modeliuotų duomenų atveju, visi parametrų įverčiai pateko į pogrupių regresijos modelio parametrų pasikliautinuosius intervalus [7].

Tsai-Hung Fan, Dennis′o K. J. Lin′o ir Kuang-Fu′o Cheng′o straipsnyje „Didžiųjų duomenų regresinė analizė“ (angl. *Regression analysis for massive datasets*) vadovavosi ta pačia duomenų išskaidymo į pogrupius idėja, tik galutiniams rezultatams gauti naudojo svorinį vidurkį, kurio apskaičiavimui naudojami svoriai, minimizuojantys galutinę visų pogrupių regresijos modelio parametrų įverčių dispersiją. Svorių skaičiavimui pasiūlyti keli variantai, atsižvelgiant į tai, ar yra tenkinamos duomenų pasiskirstymo pagal normalųjį skirstinį ir lygių dispersijų pogrupiuose sąlygos ar ne. Taip pat autoriai straipsnyje pateikė savo siūlomų metodų teorinį pagrindimą su visais įrodymais, kurie parodo metodų efektyvumą ir pagrįstumą. Atliktos analizės rezultatai parodė, kad didėjant pogrupių skaičiui didėja ir aprėpties tikimybė, taip pat reikia mažiau kompiuterio atminties ir skaičiavimai atliekami greičiau. Taip pat nustatyta, kad svorinis parametrų vidurkis išties duoda šiek tiek geresnius rezultatus nei paprastas vidurkis. Autorių teigimu, duomenų išskaidymas į pogrupius yra efektyvus būdas darbui su didžiaisiais duomenimis ir gali būti naudojamas ne tik regresinės analizės atveju, bet ir kitų statistinių metodų pritaikymui [8].

Ping Ma ir Xiaoxiao Sun savo darbe „Stebinių įtakos indeksu pagrįstų regresijos metodų panaudojimas didžiųjų duomenų regresinėje analizėje“ (angl. *Leveraging for big data regression*) pasiūlė unikalią didžiųjų duomenų analizės techniką. Autoriai nagrinėja stebinių įtakos indeksu pagrįstus metodus, kurie yra paremti vienos imties išrinkimu. Išrinktoje imtyje atliekamos regresinės analizės procedūros ir vėliau iš gautų rezultatų daromos išvados apie visą duomenų aibę. Pagrindinis šių metodų privalumas – nevienodos patekimo į imtį tikimybės, sudarytos atsižvelgiant į stebinių įtakos indekso reikšmes. Tokiu būdu įtakingesni stebiniai turi didesnę tikimybę patekti į nagrinėjamą imtį. Savo darbe autoriai stebinių įtakos indeksu pagrįstos regresijos pritaikymą išskaidė į dvi grupes: svorinę stebinių įtakos indeksu paremtą regresiją ir nesvorinę, priklausomai nuo to, ar naudojamas svorinis mažiausių kvadratų metodas ar paprastas mažiausių kvadratų metodas. Savo tyrime autoriai naudojo tris imties išrinkimo variantus: kai patekimo į imtį tikimybės yra lygios, kai patekimo į imtį tikimybės yra paremtos stebinių įtakos indeksu ir kai naudojamos pakoreguotos stebinių įtakos indeksu paremtos tikimybės. Autoriai atliko teorinę metodų analizę, suformulavo lemas ir pateikė jų įrodymus, taip pat pritaikė stebinių įtakos indeksu pagrįstos regresijos metodą realiems duomenims, kartodami algoritmą tūkstantį kartų įvairiems imčių išrinkimo ir imčių dydžių atvejams. Tyrimo rezultatai parodė, kad stebinių įtakos indeksu paremta regresija yra žymiai greitesnė ir duoda tikslesnius rezultatus, taip pat nustatė, kad nesvorinis mažiausių kvadratų metodas yra tikslesnis [6].

Norint sudaryti lengvai interpretuojamą modelį, kuriame nebūtų nereikalingų ir neinformatyvių kintamųjų, kuris veiktų greitai ir efektyviai, siekiant išvengti modelio persimokymo problemos, naudojami įvairūs kintamųjų atrinkimo metodai.

Valeria Fonti savo darbe „Kintamųjų atrinkimas naudojant LASSO“ (angl. *Feature Selection using LASSO*) atliko tyrimą, kurio tikslas išanalizuoti kintamųjų atrinkimo procesą ir pristatyti bent vieną metodą. Autorė pristatė detalią LASSO (angl. *Least Absolute Shrinkage and Selection Operator*) regresijos teorinę analizę. Atlikusi koreliacinę analizę ir sudariusi regresinius modelius su geriausiai su priklausomu kintamuoju koreliuojančiais kintamaisiais, autorė teigė, kad apie kintamųjų reikšmingumą vien iš ANOVA lentelės būtų netikslinga spręsti, todėl tokiais atvejais LASSO regresija yra išties tinkamas metodas. Darbe pristatomi tyrimo rezultatai, kurie parodė, kad LASSO regresija tikrai atrenka svarbiausius kintamuosius. Tačiau šis metodas turi ir apribojimų, kurie kyla dėl multikolinearumo ir atvejų, kai

kintamųjų yra daugiau nei stebinių. Norint išspręsti minėtus apribojimus, autorė siūlo naudoti LASSO ir RIDGE regresijų kombinaciją – elastinius tinklus (angl. *Elastic Net*) [2].

Kita mokslininkų grupė, kurią sudaro Eric Iturbide, Jaime Cerda ir Mario Graff, savo darbe „LASSO ir LARS palyginimas autoregresinių laiko eilučių modelių prognozavimui“ (angl. *A Comparison between LARS and LASSO for Initialising the Time-Series Forecasting Auto-RegressiveEquations*) lygino paprasto mažiausių kvadratų metodo, LASSO ir LARS (angl. *Least Angle Regression*) regresijų veikimą skirtingų laiko eilučių tiesinių modelių prognozavimui. Tyrimo tikslas – lengvai interpretuojamų ir tiksliai prognozuojančių modelių sudarymas. Autoriai teigia, kad LARS metodas yra efektyvus ir skaičiavimai reikalauja tiek pat resursų, kiek mažiausių kvadratų metodas. Taip pat darbe pabrėžta, kad kintamųjų atrinkimo metodų naudojimas nėra lengva užduotis, tačiau puiki priemonė išvengti modelio persimokymo problemos, kai įvertintas modelis aprašo ne tik analizuojamą priklausomybę, bet ir atsitiktinį triukšmą. Pristatydami LARS regresiją, minėti mokslininkai ją apibūdina kaip žymiai greitesnę pažingsninio kintamųjų įtraukimo regresijos (angl. *forward step wise regression*) versiją. Rezultatai parodė, kad LARS modeliai daugiausia kartų turėjo mažiausią prognozavimo paklaidą tikrinimo aibėje. Taip pat atlikti tyrimai parodė, kad paprastas mažiausių kvadratų metodas buvo tiksliausias apmokymo aibėje, o tai įrodo persimokymo problemos egzistavimą. Antri pagal prognozės tikslumą tikrinimo aibėje buvo LASSO modeliai [4].

3.Metodinė dalis

3.1. Išskaidyta regresinė analizė

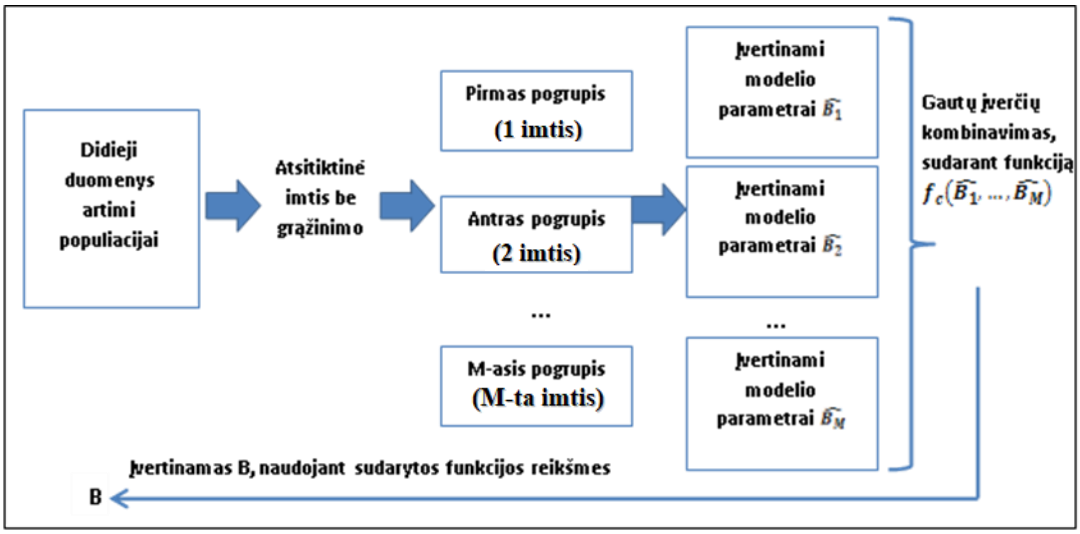
Aproksimacija nedideliu didžiųjų duomenų aibės kiekiu leidžia atlikti skaičiavimus greitai ir gauti pakankamą rezultatų tikslumą, todėl buvo pasiūlyta duomenis skaidyti į pogrupius ir atlikti skaičiavimus kiekviename pogrupyje atskirai. Tokiu būdu didieji duomenys yra apibrėžiami kaip duomenų aibė, o pogrupiai kaip imtys. Imčių išrinkimui gali būti naudojami įvairūs metodai. Šiame darbe pogrupių išrinkimui bus naudojama paprastoji atsitiktinė imtis be grąžinimo, kurioje visi duomenų aibės elementai turi lygias pradines tikimybes patekti į imtį. Antrame paveiksle pateikta išskaidytos regresinės analizės įgyvendinimo schema [7]. Regresinės analizės atveju, kiekviename pogrupyje gauti parametrų įverčiai yra kombinuojami ir užrašomi kaip funkcija, kuri gali būti įvairaus pavidalo:

$$\hat{B}_c = f_c\left(\hat{B}_1, \hat{B}_2, ..., \hat{B}_M\right). \quad (1)$$

Čia M – pogrupių skaičius, c – naudojamos funkcijos tipo žymėjimas. Šiame darbe galutiniams visų duomenų modelio parametrų įverčiams gauti bus naudojamas aritmetinis vidurkis. Tuomet galutiniai parametrų įverčiai bus gaunami naudojant formulę:

$$\hat{B}_C = \frac{1}{M}\sum_{i=1}^{M}\hat{B}_i. \quad (2)$$

Tarkime, kad $B = (\beta_o, \beta_1, ..., \beta_p)$ yra nežinomo modelio tikrieji parametrai. Atlikti tyrimai parodė, kad, jei pogrupiuose įvertinti parametrai yra nepaslinktieji tikrųjų parametrų įverčiai, tuomet ir $\hat{B}_c$ bus nepaslinktieji B įverčiai. Taip pat, kai pogrupiai yra tarpusavyje nepriklausomi ir vienodai pasiskirstę, $\hat{B}_c$ konverguos į B, jei ir pogrupiuose įvertinti parametrai konverguoja į B [7].



2 pav. Išskaidytos regresinės analizės veikimo schema

Kitas būdas galutiniams parametrų įverčiams gauti – svorinis kiekvieno pogrupio geriausių nepaslinktųjų parametrų įverčių vidurkis. Optimalūs svoriai turėtų būti pasirenkami taip, kad būtų minimizuojama galutinė parametrų įverčių dispersija. Tarkim, kad $i = 0,1,..., p-1$ yra vertinamo kintamojo parametro modelyje indeksas, o $j = 1,2,...M$ yra imties numeris. Tuomet mažiausių kvadratų parametrų įverčių $\hat{\beta}_{ij}$ dispersija bus apskaičiuojama taip:

$$\sigma_{ij}^2 = a_{ij}^{-1}\sigma_j^2. \tag{3}$$

Čia a_{ij}^{-1} yra *i*-asis matricos $\left(X_j^T X_j\right)^{-1}$ diagonalinis elementas, kai X_j yra *j*-ojo pogrupio plano matrica, o σ_j^2 yra *j*-ojo pogrupio priklausomojo kintamojo dispersija. Svorius $0 \le w_{ij} \le 1$, su sąlyga, kad $\sum w_{ij} = 1$, turime pasirinkti taip, kad (4) lygybė turėtų minimalią reikšmę.

$$\mathrm{var}\left(\tilde{\beta}_i\right) = \sum_{j=1}^{M} w_{ij}^2 \sigma_{ij}^2 \tag{4}$$

Čia $\tilde{\beta}_i = \sum_{j=1}^{M} w_{ij}\hat{\beta}_{ij}$ yra pogrupiuose įvertinti mažiausių kvadratų modelio parametrų įverčiai. Optimalūs svoriai gaunami naudojant Lagrandžo daugiklių metodą ir yra lygūs:

$$w_{ij}^2 = \frac{\left(\sigma_{ij}^2\right)^{-1}}{\sum_{j=1}^{M}\left(\sigma_{ij}^2\right)^{-1}}. \tag{5}$$

Laikantis prielaidų dėl duomenų pasiskirstymo pagal normalųjį skirtinį ir lygių dispersijų pogrupiuose, nustatyta, kad svorinio vidurkio metodu gauti modelio parametrų įverčiai turi Stjudento skirstinį [7]. Dispersijų lygybės nustatymui dažniausiai naudojamas Bartletto kriterijus, kuris tikrina hipotezę apie dispersijų lygybę skirtingose populiacijose.

3.2. Stebinių įtakos indeksu pagrįsta regresija

Kitas regresinės analizės taikymo didžiųjų duomenų analizėje metodas yra tik dalies duomenų išrinkimas. Šiam tikslui vis dažniau naudojami stebinių įtakos indeksu (angl. *leveraging*) pagrįsti statistiniai metodai. Šių metodų pagrindinis privalumas – stebinių patekimo į imtį nelygių tikimybių, kurios priklauso nuo stebinio įtakos, konstravimas. Tokiu būdu didelę įtaką turintys stebiniai turi didesnę tikimybę patekti į imtį. Kai plano matrica X yra pilno rango, tada mažiausių kvadratų parametrų įverčiai išreiškiami taip:

$$\hat{\beta}_{MKM} = \left(X^T X\right)^{-1} X^T y. \tag{6}$$

Čia $y = \left(y_1,..., y_N\right)^T$. Tačiau pasitaiko atvejų, kai matrica X yra nepilno rango, tuomet $X^T X$ yra neapverčiama. Tokiu atveju dažnai naudojama apibendrinta atvirkštinė matrica. Prognozuojamas priklausomas kintamasis gaunamas taip:

$$\hat{y} = Hy. \tag{7}$$

Čia matrica $H = X\left(X^T X\right)^{-1} X^T$, kurios *i*-asis diagonalinis elementas $h_{ii} = x_i^T\left(X^T X\right)^{-1} x_i$, vadinamas stebinio įtakos indeksu ir rodo *i*-ojo stebinio įtaką mažiausių kvadratų įverčiams, kai $i = 1,..., N$. Kuo šis stebinio įtakos indeksas artimesnis vienetui, tuo prognozuota priklausomo kintamojo reikšmė yra artimesnė tikrajai [6].

3.2.1. Stebinių įtakos indeksu pagrįstos regresijos algoritmas

1. Iš duomenų aibės *D* išrenkama atsitiktinė grąžintinė imtis, kurios dydis $n\left(n \le N\right)$, kai visų duomenų aibės elementų patekimo į imtį tikimybės $\pi = \left(\pi_1, \pi_2,..., \pi_N\right)$ gali būti apskaičiuojamos keliais būdais:

1) $\pi_i = \frac{1}{N}$, $i = 1,..., N$;

2) $\pi_i = \frac{h_{ii}}{\sum_i h_{ii}}, i = 1,..., N$, čia h_{ii} – diagonalinis matricos H elementas, kuris lygus $h_{ii} = x_i^T\left(X^T X\right)^{-1} x_i$;

3) $\pi_i = \lambda \frac{h_{ii}}{\sum_i h_{ii}} + (1-\lambda)\frac{1}{N}, i = 1,...,N$ λ – reguliarizavimo parametras, kuris gali kisti intervale $[0;1]$.

2. Sudaroma stebinių patekimo į imtį tikimybių matrica $\Phi^* = diag\{\pi_p\}$.

3. Gautai imčiai su kintamaisiais (X^*, y^*) įvertinami kintamųjų parametrai naudojant paprastą svorinį mažiausių kvadratų metodą:

1) $\tilde{\beta}^u = \arg\min \left\| y^* - X^* \beta \right\|^2$;

2) $\tilde{\beta} = \left(X^T W X \right)^{-1} X^T W y$. Svorių matrica yra diagonalinė matrica $W = diag\left(w_1, w_2, ..., w_n\right)$, kur $w_i = \frac{k_i}{n\pi_i}$, o skaičius k_i nurodo, kiek kartų *i*-asis elementas pateko į imtį, kurios dydis n. Taip pat svoriams gali būti naudojamos stebinių patekimo į imtį tikimybės π_i.

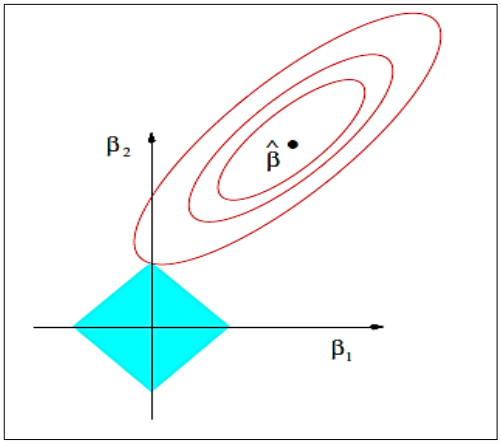
Nors $\tilde{\beta}^u$ aproksimuoja paslinktuosius $\tilde{\beta}_{MKM}$ įvertinius, tyrimais įrodyta kad $\tilde{\beta}^u$ yra nepaslinktasis tikrojo β įvertinys. Taip pat, naudojant modeliuotus duomenis, nustatyta, kad nesvoriniu mažiausiu kvadratu metodu gauti parametrų įverčiai $\tilde{\beta}^u$ turi mažesnę dispersiją nei gauti naudojant svorinį mažiausių kvadratų metodą [6].

3.3. LASSO regresija

LASSO 1996 m. suformulavo Robert′as Tibshirani′is, siekdamas pagerinti regresijos modelių interpretavimą ir prognozavimo tikslumą. LASSO yra efektyvus metodas, kuris atlieka du pagrindinius uždavinius: reguliarizaciją ir kintamųjų atrinkimą. LASSSO regresijos modelio koeficientai sumažinami, nustatant tam tikrą baudą dėl modelio sudėtingumo. Gauti LASSO koeficientai minimizuoja paklaidų kvadratų sumą, atsižvelgiant į nustatytą baudą. LASSO parametrų įverčių išraišką galima užrašyti taip:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_o - \sum_{k=1}^{p} x_{ik}\beta_k)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right\}, \lambda \geq 0, \quad (8)$$

čia λ yra reguliarizavimo parametras, kuris kontroliuoja baudos dydį. Kuo didesnė λ reikšmė, tuo didesnis regresijos modelio parametrų sumažinimo mastas. 3 paveiksle pavaizduotas LASSO metodo taikymo pavyzdys, kai modelio parametrų skaičius *p* = 2, geometrinė interpretacija, kuri leidžia geriau suprasti LASSO regresijos veikimą.



3 pav. LASSO regresijos parametrų vertinimas [3]

Elipsės vaizduoja paklaidų kvadratų sumos funkcijas (RSS), kurių centre yra mažiausių kvadratų metodo parametrų įvertis. Paveiksle matomas LASSO regresijos atvejis, kuriame rombas vaizduoja apribojimus $|\beta_1| + |\beta_2| \leq t$. Analizuojant paveikslą galima įsitikinti, kad LASSO parametrai gali įgyti reikšmes lygias nuliui [3].

3.4. LARS regresija

LARS (angl. *Least Angle Regression*) metodo procedūra gali būti lyginama su pažingsninio kintamųjų įtraukimo procedūra (angl. *forward stepwise regression*) tiesinėje regresijoje. Be to, LARS glaudžiai susijęs su LASSO regresija. Trumpai pristatysime LARS algoritmą [3]:

1. Standartizuojame nepriklausomus kintamuosius. Pirmame modeliavimo etape visi parametrų įverčiai $\beta_1, \beta_2 ..., \beta_p = 0$, o apskaičiuota paklaida lygi $r = y - \bar{y}$.

2. Atrenkame stipriausiai su paklaida *r* koreliuojantį nepriklausomą kintamąjį x_k.

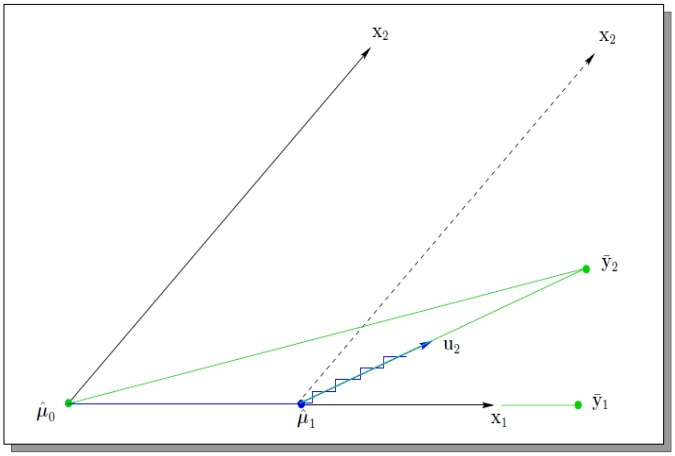
3. Keičiame parametro β_k reikšmę į $\beta_k \leftarrow \beta_k + \delta_k$, kur δ_k – pakankamai mažas, tuomet perskaičiuojame paklaidą $r \leftarrow r - \delta_k x_k$.

4. Surandame kitą nepriklausomąjį kintamąjį x_r kuris pakankamai stipriai koreliuoja su naujai įvertinta paklaida.

5. Keičiame nepriklausomų kintamųjų parametrų β_k ir β_r reikšmes tol, kol rasime kitą nepriklausomą kintamąjį x_l ,kuris pakankamai koreliuos su apskaičiuota paklaida, naudojant prieš tai įvertintus parametrus.

6. Procedūra kartojama, kol visi regresoriai bus įtraukti į modelį, arba likę kintamieji nekoreliuos su paklaida.

Kiekviename žingsnyje į modelį yra įtraukiamas vienas regresorius, taigi po k žingsnių, kai $k = 1, ..., p$, turime modelį su k nenulinių koeficientų. 4 paveiksle grafiškai pavaizduotas LARS algoritmas, kai $p = 2$.



4 pav. LARS regresijos geometrinė interpretacija [3]

Šiek tiek modifikavus LARS algoritmą, galima gauti rezultatus analogiškus LASSO regresijos rezultatams. Pagrindinis pakeitimas atsiranda penktame algoritmo punkte, kurį reiktų papildyti sąlyga, kad, jei kažkuris nenulinis koeficientas pasiekia nulį, tuomet jį reikia išmesti iš atrinktų kintamųjų grupės. Šis algoritmas LARS-LASSO yra išties efektyvus, nes skaičiavimų našta ta pati, kiek mažiausių kvadratų metodo pritaikymas, turint p regresorių. LARS regresijai visada reikia p žingsnių, kad gautų visus mažiausių kvadratų įverčius. Tuo tarpu LASSO gali reikti ir daugiau nei *p* žingsnių [3].

4.Tiriamoji dalis

Pastaraisiais metais transporto spūsčių problema itin aktuali daugelyje didžiųjų pasaulio miestų. Specialistai vertina vietoves, kuriose susidaro automobilių spūstys, analizuoja esamą eismo situaciją, rengia planus, kaip to išvengti, vertina laiką, kurį praranda vairuotojai automobilių vilkstinėse. Tačiau didžiausia problema kyla dėl vis didėjančios transporto taršos miestuose.

Vienas iš šių problemų sprendimo būdų yra dviračių naudojimas. Dviratis yra praktiška, aplinką tausojanti ekologiška ir triukšmo nekelianti susisiekimo priemonė, ne tik gerinanti sveikatą ir fizinę būklę, bet ir galinti išspręsti spūsčių, taršos, parkavimo, maršrutų nebuvimo bei kitas susisiekimo problemas miestuose.

Šiame darbe naudojamų realių duomenų šaltinis yra gana populiari mašininio mokymo duomenų talpykla – duomenys iš dviračių nuomos sistemos, kurios šiuo metu itin populiarėja ir Lietuvoje. Tačiau negavus Lietuvos dviračių nuomos sistemos duomenų, buvo nuspręsta naudoti duomenis, apimančius 2011–2012 m. amerikiečių dviračių nuomos sistemos surinktą informaciją valandos periodiškumu, kurią sudaro beveik 18 000 stebinių [1]. Duomenų matricoje yra 53 kintamieji (savaitės dienos, valandos, metų laiko, mėnesio dienos ir kitos charakteristikos, nusakančios oro sąlygas).

Šiame darbe bus modeliuojamas dviračių nuomos klientų skaičius, atsižvelgiant į oro ir sezoniškumo charakteristikas. Pradedant darbą su šiais duomenimis, pirmiausia buvo atliktas multikolinearumo tikrinimas, tuomet klientų skaičiaus išsiskiriančių reikšmių identifikavimas ir logaritminė transformacija, taip pat praleistų reikšmių įrašymas.

Esant multikolinearumo problemai, atrinkti nepriklausomi kintamieji gali būti nebūtinai tie, kurie geriausiai nusako priklausomąjį kintamąjį, taip pat gali kilti modelio parametrų vertinimo problemos. Todėl, norint išvengti šių problemų, pirmiausia buvo apskaičiuota koreliacija tarp kiekybinių nepriklausomųjų kintamųjų. Koreliacinės analizės metu buvo nustatytas stiprus tiesinis ryšys tarp oro temperatūros ir jutiminės oro temperatūros, todėl siekiant išvengti multikolinearumo problemos, į analizę nebuvo įtrauktas kintamasis nusakantis jutiminę oro temperatūrą. Multikolinearumo tikrinimui buvo įvertintas ir dispersijos mažėjimo daugiklis VIF (angl. *Variance Inflation Factor*) [3], kuris kiekvienam nepriklausomam kintamajam sudaro daugialypės regresijos modelius su kitais nepriklausomais kintamaisiais ir, naudodamas determinacijos koeficiento reikšmes, apskaičiuoja VIF kriterijų. Naudojant šį kriterijų multikolinearumas tarp likusių kintamųjų nebuvo nustatytas.

Išsiskiriančių reikšmių identifikavimui buvo skaičiuojami intervalai, gauti naudojant Q_1 ir Q_3 (imties pirmąjį ir trečiąjį kvartilius) ir $IQR = Q_3 - Q_1$ – tarpkvartilinį skirtumą. Jei reikšmė $y$ra mažesnė už Q_1-3·IQR arba didesnė už Q_3+3·IQR, tai ji buvo vadinama išskirtimi. Analizuojamame duomenų rinkinyje gautos 196 išskirtys, kurios buvo pašalintos.

Logaritminė duomenų transformacija buvo atlikta, nes nustatyta, kad po logaritminės transformacijos gaunami tikslesni regresiniai modeliai, be to, logaritminė duomenų transformacija padeda užtikrinti klasikinių tiesinės regresijos prielaidų tenkinimą.

Praleistos reikšmės buvo įrašyto vėjo stiprumą nusakantiems kintamiesiems, kurie bus naudojami taikant LARS ir LASSO regresijas. Buvo įrašyta beveik 15 proc. praleistų reikšmių, naudojant daugiamatę įrašymo procedūrą [9]. Įrašius praleistas reikšmes, duomenų skirstinys reikšmingai nepasikeitė. Kiti kintamieji praleistų reikšmių neturėjo.

4.1. Išskaidytosios regresijos modeliavimo rezultatai dviejų kintamųjų atveju

Išskaidytos regresijos veikimo analizei atlikti bus naudojami du kintamieji, geriausiai koreliuojantys su dviračių nuomos klientų skaičiumi. Šie kintamieji yra temperatūra ir drėgnumas. Pažymime klientų skaičių y, temperatūrą x_1 ir drėgnumą x_2. Tada analizuojamas regresijos modelis gali būti užrašytas taip:

$$\mathrm{y} = e^{\beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 x_2} \tag{9}$$

Siekiant nustatyti, ar pogrupių skaičius turi įtakos regresijos lygties koeficientų įverčiams, atlikti modeliavimai, kai pogrupių skaičius $M = 3,5,10.$ Kiekvieną pogrupį sudaro 1 718 stebinių. 1 lentelėje pateikti gauti rezultatai.

Šiame darbe modelių tikslumui nustatyti naudojama šaknis iš vidutinės kvadratinės logaritminės paklaidos (angl. *Root Mean Squared Logarithmic Error, RMSLE*), kuri dažnai taikoma mašininio apmokymo srityje ir yra apibrėžta taip:

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\ln\left(P_i+1\right)-\ln\left(T_i+1\right)\right]^2}, i = 1,\ldots,N. \tag{10}$$

Čia N – stebinių skaičius, P_i – prognozuotos reikšmės, T_i – faktinės reikšmės.

1 lentelė. Regresijos parametrų įverčių ir modelio tikslumo priklausomybė nuo pogrupių skaičiaus M

	$M = 3$				$M = 5$				$M = 10$			
Iteracijos	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	RMSLE	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	RMSLE	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	RMSLE
1	4,600	2,815	-2,354	1,222	4,619	2,703	-2,308	1,227	4,675	2,688	-2,377	1,223
2	4,612	2,763	-2,343	1,234	4,639	2,727	-2,324	1,220	4,583	2,807	-2,346	1,230
3	4,615	2,733	-2,335	1,237	4,688	2,698	-2,367	1,212	4,617	2,801	-2,365	1,226
4	4,679	2,702	-2,403	1,225	4,591	2,766	-2,306	1,230	4,638	2,770	-2,385	1,220
5	4,717	2,645	-2,424	1,232	4,716	2,647	-2,388	1,213	4,618	2,751	-2,302	1,217
6	4,584	2,794	-2,314	1,226	4,654	2,732	-2,401	1,232	4,626	2,738	-2,309	1,221
7	4,650	2,728	-2,359	1,224	4,636	2,719	-2,322	1,220	4,591	2,848	-2,371	1,221
8	4,666	2,698	-2,376	1,228	4,683	2,660	-2,361	1,212	4,656	2,707	-2,364	1,235
9	4,657	2,686	-2,364	1,227	4,668	2,703	-2,359	1,224	4,642	2,816	-2,445	1,225
10	4,667	2,704	-2,371	1,226	4,635	2,715	-2,353	1,227	4,587	2,784	-2,310	1,233

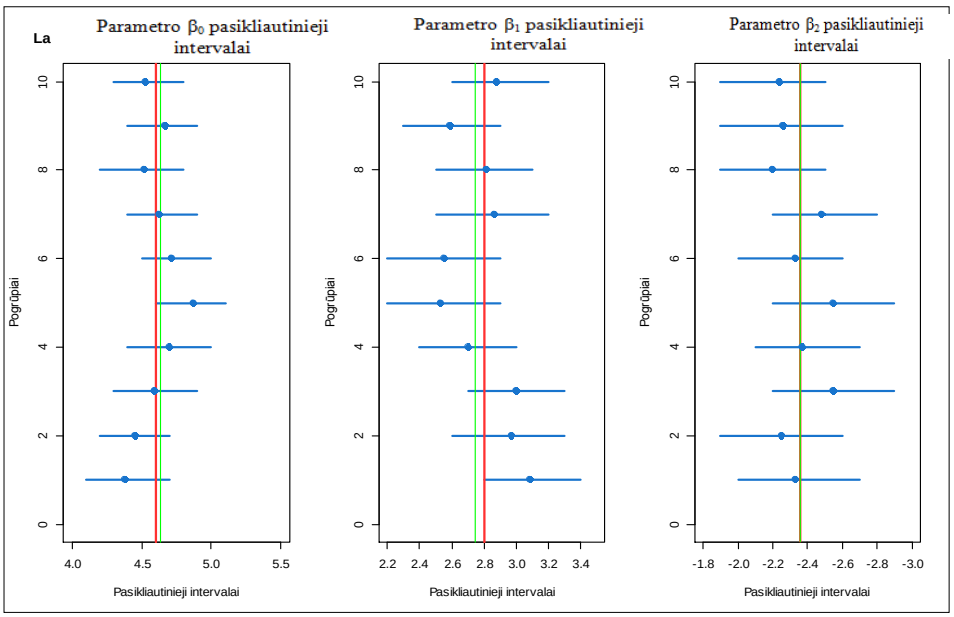
Vidurkis $\hat{B}_C$	4,645	2,727	-2,364	1,228	4,653	2,707	-2,349	1,222	4,623	2,771	-2,357	1,225

Lentelėje matyti, kad skirtumai tarp parametrų įverčių kiekvienos iteracijos metu yra minimalūs. Taip pat sunku įžvelgti žymius rezultatų pakitimus, keičiantis pogrupių skaičiui. Kai pogrupių skaičius $M = 10$, rezultatai yra artimiausi visų duomenų modelio parametrų reikšmėms. Tačiau, kai $M = 5$, parametrų įverčiai labiausiai skiriasi, todėl galima daryti išvadą, kad pogrupių skaičius tikslumui įtakos neturi. Tokią pačią analizę atliekame ir svorinio vidurkio atveju. Visais atvejais nulinė hipotezė apie dispersijų lygybę pogrupiuose buvo priimta. Rezultatai pateikti 2 lentelėje.

2 lentelė. Regresijos koeficientų įverčių ir modelio tikslumo priklausomybė nuo pogrupių skaičiaus M, svorinių vidurkių atveju

	$M = 3$				$M = 5$				$M = 10$			
Iteracijos	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	RMSLE	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	RMSLE	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	RMSLE
1	4,600	2,813	-2,353	1,226	4,620	2,699	-2,308	1,226	4,677	2,683	-2,375	1,225
2	4,611	2,764	-2,342	1,226	4,639	2,723	-2,325	1,225	4,578	2,808	-2,342	1,226
3	4,616	2,733	-2,336	1,226	4,687	2,698	-2,366	1,225	4,618	2,800	-2,365	1,226
4	4,677	2,703	-2,407	1,226	4,589	2,767	-2,304	1,226	4,637	2,770	-2,382	1,226
5	4,717	2,643	-2,423	1,226	4,720	2,647	-2,389	1,225	4,617	2,753	-2,299	1,225
6	4,585	2,794	-2,314	1,226	4,655	2,732	-2,400	1,226	4,627	2,738	-2,310	1,225
7	4,651	2,726	-2,360	1,225	4,636	2,720	-2,323	1,225	4,590	2,846	-2,371	1,226
8	4,667	2,697	-2,376	1,226	4,683	2,663	-2,362	1,225	4,659	2,707	-2,365	1,225
9	4,656	2,687	-2,363	1,226	4,669	2,705	-2,357	1,225	4,642	2,815	-2,445	1,226
10	4,664	2,703	-2,374	1,226	4,636	2,714	-2,355	1,226	4,586	2,783	-2,307	1,226
$\tilde{\beta}_i$	4,644	2,726	-2,365	1,226	4,653	2,707	-2,349	1,225	4,623	2,770	-2,356	1,226

Rezultatai yra beveik analogiški paprasto vidurkio rezultatams. Taip pat rezultatai nėra tendencingi, didėjant pogrupių skaičiui. 5 paveiksle pavaizduoti atvejo, kai pogrupių skaičius $M = 10$, parametrų pasikliautinieji intervalai. Vertikali raudona linija vaizduoja parametrų vidurkio reikšmę, vertikali žalia – visų duomenų modelio parametro reikšmę. Taškai vaizduoja pogrupiuose įvertintų modelių parametrų reikšmes. Matyti, kad reikšmės labai artimos ir visi parametrai, išskyrus $\hat{\beta}_1$ pirmame pogrupyje, patenka į pasikliautinųjų intervalų ribas.



5 pav. Regresijos lygties parametrai

Galima daryti išvadą, kad duomenų išskaidymą į pogrupius yra tikslinga naudoti, nes gauti rezultatai yra labai artimi visų duomenų modelio rezultatams, o skaičiavimų našta nėra tokia didelė.

4.2. Stebinių įtakos indeksu pagrįstos regresijos modeliavimo rezultatai

Naudojant stebinių įtakos indeksu pagrįstą regresinę analizę, sudarytas modelis (9) išraiška). Siekiant nustatyti, ar imties dydis turi įtakos rezultatams ir modelio tikslumui, išbandyti trys imčių dydžiai, gauti dalinant visą duomenų aibę į 2, 3, 10 dalių. Taip pat, norint padaryti tikslesnes išvadas, kiekvienu atveju atlikta dešimt iteracijų ir pateiktas visų iteracijų rezultatų vidurkis ir standartinis nuokrypis. Rezultatai su parametrų įverčiais ir RMSLE paklaidomis, kai imtį sudaro 8 592 stebiniai, o λ= 0,5, pateikti 3 lentelėje.

3 lentelė. Stebinių įtakos indeksu pagrįstos regresijos rezultatai, kai imties dydis lygus 8 592

Patekimo į imtį tikimybės / MKM svoriai		$\pi_i = \frac{1}{N}$		$\pi_i = \frac{h_{ii}}{\sum_i h_{ii}}$		$\pi_i = \lambda \frac{h_{ii}}{\sum_i h_{ii}} + (1-\lambda)\frac{1}{N}$	
		Vidurkis	Stand. nuokrypis	Vidurkis	Stand. nuokrypis	Vidurkis	Stand. nuokrypis
$w_i = \frac{1}{N}$	$\hat{\beta}_0$	4,60	0,052	4,33	0,047	4,45	0,045
	$\hat{\beta}_1$	2,76	0,069	2,80	0,044	2,79	0,054
	$\hat{\beta}_2$	-2,33	0,063	-2,04	0,083	-2,17	0,076
	RMSLE	1,222	0,012	1,204	0,01	1,217	0,012
$w_i = \frac{k_i}{n\pi_i}$	$\hat{\beta}_0$	4,606	0,038	4,47	0,074	4,53	0,061
	$\hat{\beta}_1$	2,78	0,097	2,81	0,068	2,81	0,08
	$\hat{\beta}_2$	-2,31	0,075	-2,20	0,109	-2,28	0,087
	RMSLE	1,222	0,012	1,205	0,011	1,218	0,012
$w_i = \pi_i$	$\hat{\beta}_0$	-	-	3,96	0,053	4,23	0,05
	$\hat{\beta}_1$	-	-	2,86	0,038	2,82	0,056
	$\hat{\beta}_2$	-	-	-1,62	0,108	-1,91	0,086
	RMSLE	-	-	1,211	0,011	1,220	0,012

Apibendrinant gautus rezultatus, naudojant skirtingas stebinių patekimo į imtį tikimybes, matyti, kad tiksliausias modelis gaunamas, kai yra naudojamas stebinių įtakos indeksu paremtas tikimybių skaičiavimas. Lyginant mažiausių kvadratų metodo skirtingas svorių realizacijas, galima daryti išvadą, kad visgi svorių įtraukimas tikslesnių rezultatų nedavė. Toliau atliekame analogiškus veiksmus, kai imtį sudaro 3 437 stebiniai, o λ = 0,5. Rezultatai pateikti 4 lentelėje.

4 lentelė. Stebinių įtakos indeksu pagrįstos regresijos rezultatai, kai imties dydis lygus 3 437

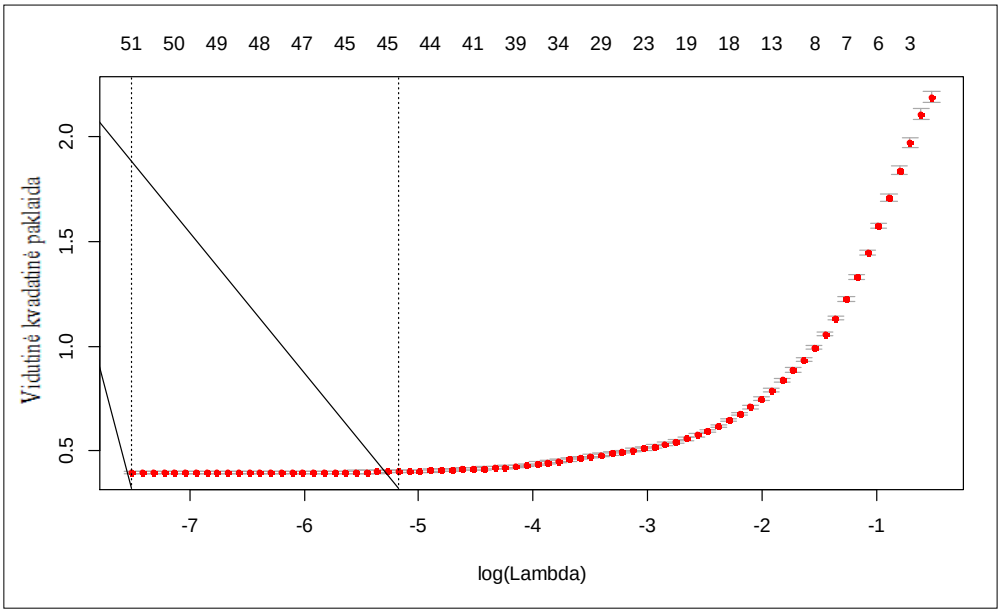
Patekimo į imtį tikimybės / MKM svoriai		$\pi_i = \frac{1}{N}$		$\pi_i = \frac{h_{ii}}{\sum_i h_{ii}}$		$\pi_i = \lambda \frac{h_{ii}}{\sum_i h_{ii}} + (1-\lambda)\frac{1}{N}$	
		Vidurkis	Stand. nuokrypis	Vidurkis	Stand. nuokrypis	Vidurkis	Stand. nuokrypis
$w_i = \frac{1}{N}$	$\hat{\beta}_0$	4,60	0,10	4,34	0,09	4,46	0,09

	$\hat{\beta}_1$	2,74	0,09	2,78	0,11	2,76	0,08
	$\hat{\beta}_2$	-2,32	0,11	-2,05	0,07	-2,16	0,09
	RMSLE	1,233	0,019	1,206	0,020	1,220	0,020
$w_i = \frac{k_i}{n\pi_i}$	$\hat{\beta}_0$	4,58	0,13	4,55	0,09	4,59	0,09
	$\hat{\beta}_1$	2,75	0,13	2,78	0,15	2,77	0,12
	$\hat{\beta}_2$	-2,29	0,16	-2,29	0,07	-2,32	0,09
	RMSLE	1,233	0,019	1,208	0,021	1,220	0,021
$w_i = \pi_i$	$\hat{\beta}_0$	-	-	3,96	0,09	4,24	0,12
	$\hat{\beta}_1$	-	-	2,86	0,10	2,80	0,09
	$\hat{\beta}_2$	-	-	-1,62	0,09	-1,90	0,13
	RMSLE	-	-	1,215	0,021	1,223	0,021

Lentelėje matyti, kad rezultatai mažiau tikslesni nei didesnės imties atveju. Tačiau geriausius rezultatus taip pat davė paprastas mažiausių kvadratų metodas, kai stebinių patekimo į imtį tikimybės yra paremtos stebinių įtakos indeksu. Atlikus skaičiavimus su dar mažesne imtimi, padaryta išvada, kad tiksliausi modeliai gaunami, kai imtis yra kiek įmanoma didesnė. Kadangi naudojant stebinių įtakos indeksu pagrįstą regresiją išauga tikimybė į imtį išrinkti išsiskiriančias reikšmes, būtina atlikti modelio diagnostiką išskirtims nustatyti. Šiam tikslui buvo panaudotas Kuko matas, kuris atsižvelgia ir į standartizuotąją paklaidą, ir į stebinio įtakos indeksą. Nustatyta, kad išsiskiriančių reikšmių nebuvo.

4.3. LASSO regresijos modeliavimo rezultatai

Darbui su LASSO regresija naudojamas R paketas „glmnet“. Šioje dalyje, atlikus logaritminę transformaciją, naudojami klientų skaičiaus duomenys. Visų duomenų aibė buvo padalyta į apmokymo ir testavimo aibes, siekiant nešališkai įvertinti modelio veikimą kitoje duomenų aibėje. Kadangi reguliarizavimo parametro λ parinkimas yra gana sudėtinga užduotis, tam buvo naudojama kryžminio patvirtinimo (angl. *cross-validation*) procedūra, kai grupių skaičius $K = 10$. 6 paveiksle pavaizduotas logaritmuotų λ reikšmių ir modelio vidutinės kvadratinės paklaidos kitimas.



6 pav. Logaritmuotų reikšmių ir paklaidų priklausomybės kitimas

Pirmoji vertikali brūkšninė linija vaizduoja λ reikšmę, su kuria gaunama minimali kryžminio patvirtinimo paklaida, o antroji vaizduoja per vieną standartinį nuokrypį nuo minimalios paklaidos nutolusią parametro λ reikšmę. Šiame darbe bus naudojami abu minėti λ variantai. Abiem atvejais reguliarizavimo parametrai yra gana maži ir lygūs $\lambda_{\min} = 0{,}0005521097$, $\lambda_{\min+1se} = 0{,}00620198$. Toliau yra sudaromas LASSO regresijos modelis su dviem λ reikšmių variantais. Modelio tinkamumui testavimo aibėje nustatyti, įvertinamos, naudojant apmokymo aibėje atrinktų kintamųjų parametrų įverčius, sudaryto LASSO modelio priklausomo kintamojo reikšmės, tuomet apskaičiuojamos modelių RMSLE paklaidos. Rezultatai pateikti 5 lentelėje.

5 lentelė. LASSO regresijos rezultatai

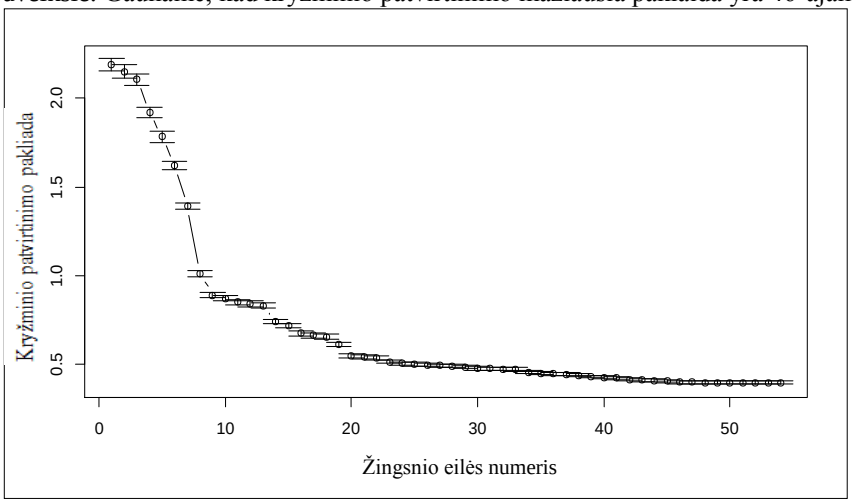
Kintamasis	$\lambda_{\min}$	$\lambda_{\min+1se}$	Kintamasis	$\lambda_{\min}$	$\lambda_{\min+1se}$
Laisvasis narys	3,671	3,671	12 valanda	0,700	0,700
Žiema	-0,332	-0,332	13 valanda	0,895	0,895
Pavasaris			14 valanda	0,869	0,869
Vasara			15 valanda	0,790	0,790
Ruduo	0,178	0,178	16 valanda	0,826	0,826
Metai	0,434	0,434	17 valanda	1,085	1,085
Vasaris			18 valanda	1,455	1,455
Kovas	0,028	0,028	19 valanda	1,374	1,374
Balandis			20 valanda	1,142	1,142
Gegužė	0,078	0,078	21 valanda	0,853	0,853
Birželis			22 valanda	0,619	0,619
Liepa	-0,091	-0,091	23 valanda	0,386	0,386
Rugpjūtis	-0,008	-0,008	24 valanda		
Rugsėjis	0,041	0,041	Šventinė diena	-0,097	-0,097
Spalis	0,035	0,035	Antradienis	-0,024	-0,024
Lapkritis			Trečiadienis	-0,022	-0,022
Gruodis			Ketvirtadienis	-0,016	-0,016
2 valanda	-1,140	-1,140	Penktadienis		
3 valanda	-1,667	-1,667	Šeštadienis	0,092	0,092
4 valanda	-2,273	-2,273	Sekmadienis	0,085	0,085
5 valanda	-2,516	-2,516	Darbo diena / savaitgalis		
6 valanda	-1,431	-1,431	Rūkas, debesuota	-0,0002	0,000
7 valanda	-0,238	-0,238	Lengvas sniegas, smarkus debesuotumas, lietus	-0,509	-0,509
8 valanda	0,605	0,605	Sniegas, stiprus lietus, kruša, stiprus rūkas		
9 valanda	1,228	1,228	Temperatūra	1,576	1,576
10 valanda	0,959	0,959	Drėgmė	-0,426	-0,426
11 valanda	0,613	0,613	Vėjo stiprumas	-0,192	-0,192
RMSLE	0,6038	0,6038	RMSLE	0,6038	0,6038

Lentelėje matyti, kad dėl parinktų λ reikšmių labai mažo skirtumo kintamųjų koeficientai yra identiški. Taip pat nors naudota λ reikšmė yra labai artima nuliui, o tai reiškia, kad gauti rezultatai turėtų būti labai panašūs į paprasto mažiausių kvadratų metodo su visais kintamaisiais rezultatus, vis dėlto LASSO regresija į modelį neįtraukė net vienuolikos kintamųjų.

4.4. LARS regresijos modeliavimo rezultatai

Tai pačiai apmokymo aibei taikomas LARS algoritmas. Tam, kad būtų galima nustatyti, kuriame žingsnyje LARS modelis yra optimaliausias, panaudojama kryžminio patvirtinimo procedūra (angl. *cross-validation*), kai grupių skaičius $K = 10$. Optimalus žingsnis parenkamas atsižvelgiant į kryžminio patvirtinimo paklaidas. Tačiau taip pat dažnai optimaliu žingsniu pasirenkamas tas, kuriame statistikos *Cp* reikšmė, kuri paprastoje tiesinėje regresijoje atitiktų Akaike informacinį kriterijų (AIC), yra mažiausia. *Cp* parodo vidutinės kvadratinės paklaidos įvertį kiekviename modelyje, kuris priklauso nuo atrinktų kintamųjų poaibio ir kas kart yra pakoreguojamas priklausomai nuo kintamųjų skaičiaus

tame poaibyje. Šiame darbe buvo išbandyti abu šie variantai. Kryžminio patvirtinimo paklaidos kiekviename žingsnyje kitimas pateiktas 7 paveiksle. Gauname, kad kryžminio patvirtinimo mažiausia paklaida yra 46-ajame žingsnyje.



7 pav. Kryžminio patvirtinimo paklaidos kitimas

Toliau randami LARS modelių parametrų įverčiai ir modelių paklaidos, gautos atsižvelgiant į prognozės rezultatus testavimo aibėje. LARS modelio parametrų įverčiai, atrinkti kintamieji ir modelių paklaidos pateiktos 6 lentelėje.

6 lentelė. LARS regresijos rezultatai

Kintamasis	*Min* Cp	Minimali kryžminio patikrinimo paklaida	Kintamasis	*Min* Cp	Minimali kryžminio patikrinimo paklaida
Žiema	-0,341	-0,317	13 valanda	1,506	0,688
Pavasaris			14 valanda	1,481	0,661
Vasara	0,047		15 valanda	1,406	0,579
Ruduo	0,256	0,175	16 valanda	1,445	0,614
Metai	0,462	0,420	17 valanda	1,699	0,874
Vasaris	0,111		18 valanda	2,070	1,242
Kovas	0,137		19 valanda	1,983	1,164
Balandis	0,085		20 valanda	1,738	0,939
Gegužė	0,170	0,056	21 valanda	1,441	0,652
Birželis	0,060		22 valanda	1,199	0,420
Liepa	-0,102	-0,073	23 valanda	0,965	0,188
Rugpjūtis	-0,019		24 valanda	0,572	-0,039
Rugsėjis	0,081	0,021	Šventinė diena	-0,137	-0,067
Spalis	0,054	0,015	Antradienis	-0,046	-0,005
Lapkritis	-0,016		Trečiadienis	-0,052	
Gruodis	-0,015		Ketvirtadienis	-0,043	
2 valanda	-0,665	-1,278	Penktadienis	0,014	
3 valanda	-1,194	-1,799	Šeštadienis	0,113	0,083
4 valanda	-1,806	-2,402	Sekmadienis	0,106	0,075
5 valanda	-2,048	-2,646	Darbo diena /savaitgalis		
6 valanda	-0,963	-1,560	Rūkas, debesuota	-0,051	
7 valanda	0,231	-0,368	Lengvas sniegas, smarkus debesuotumas, lietus	-0,594	-0,478

Kintamasis	*Min* Cp	Minimali kryžminio patikrinimo paklaida	Kintamasis	*Min* Cp	Minimali kryžminio patikrinimo paklaida
8 valanda	1,174	0,413	Sniegas, stiprus lietus, kruša, stiprus rūkas	-0,043	
9 valanda	1,804	1,033	Temperatūra	1,523	1,590
10 valanda	1,546	0,760	Drėgmė	-0,301	-0,458
11 valanda	1,206	0,411	Vėjo stiprumas	-0,385	-0,061
12 valanda	1,300	0,497	RMSLE	0,588	0,617

6 lentelės rezultatai rodo, kad pirmuoju atveju gaunama šiek tiek mažesnė paklaida. Palyginus LASSO ir LARS rezultatus, nesunku pastebėti, kad LASSO atveju atrinktų kintamųjų skaičius yra mažesnis, tačiau didesnis tikslumas pasiekiamas naudojant LARS regresiją.

Išvados ir rezultatai

1. Išskaidytos regresinės analizės metu nustatyta, kad parametrų įverčių pogrupiuose tiek paprastas, tiek svorinis vidurkis yra labai artimas arba yra lygus visos duomenų aibės modelio parametrų įverčiams. Nors literatūroje nurodoma, kad svorinis vidurkis yra efektyvesnis už paprastą vidurkį, tačiau naudotiems duomenims svorinio vidurkio skaičiavimas nebuvo tikslingas, nes skaičiavimai yra sudėtingesni, užima daugiau laiko, o gauti rezultatai yra analogiški paprasto vidurkio rezultatams. Taip pat nustatyta, kad skirtumai kiekvienos iteracijos metu yra minimalūs, o pogrupių skaičius tikslumui įtakos neturi.
2. Atlikta stebinių įtakos indeksu pagrįsta regresinė analizė parodė, kad tiksliausias modelis gaunamas, kai yra naudojamas stebinių įtakos indeksu paremtas tikimybių skaičiavimas. Lyginant mažiausių kvadratų metodo skirtingas svorių realizacijas, galima daryti išvadą, kad svorių įtraukimas tikslesnių rezultatų nedavė. Taip pat nustatyta, kad tiksliausi modeliai gaunami, kai imtis yra kiek įmanoma didesnė.
3. Sudarant LASSO regresijos modelius, nustatyta, kad nors naudota λ reikšmė yra labai artima nuliui, LASSO regresija į modelį neįtraukė net vienuolikos kintamųjų. Kai λ reikšmė yra labai artima nuliui, gauti rezultatai turėtų būti labai panašūs į paprasto mažiausių kvadratų metodo su visais kintamaisiais rezultatus, todėl tikėtina, kad panašius rezultatus pavyko gauti su žymiai mažesniu kintamųjų skaičiumi modelyje.
4. Palyginus LASSO ir LARS rezultatus, pastebėta, kad LASSO atveju atrinktų kintamųjų skaičius yra mažesnis, tačiau didesnis tikslumas pasiekiamas, naudojant LARS regresiją.
5. Lyginant LASSO ir LARS regresijų paklaidas su išskaidytos arba stebinių įtakos indeksu pagrįstos regresijos paklaidomis, nustatyta, kad LASSO ir LARS regresijos padėjo jas sumažinti beveik dvigubai. Tai patvirtina kintamųjų atrinkimo svarbą ir efektyvumą regresijos modelių sudaryme.

Literatūra

1. *Bike Sharing Dataset* [interaktyvus], https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset (žiūrėta 2018-03-20).
2. Fonti V. 2017: *Feature Selection using LASSO*, Research Paper in Business Analytics.
3. Hastie T.; Tibshirani R., Friedman J. 2009: *The Elements of Statistical Learning*, Springer.
4. Iturbide E., Cerda J., Graff M. 2013: *A Comparison between LARS and LASSO for Initialising the Time-Series Forecasting Auto-Regressive Equations*, Iberoamerican Conference on Electronics Engineering and Computer Science.
5. Jianqing F., Fang H., Han L. 2014: *Challenges of Big Data Analysis.* NIH Public Access Author Manuscript.
6. Ma P., Sun X. 2015: Leveraging for big data regression, *WIREs Computational Statistics*, 7, p. 70–76
7. Sunghae J., Seung-Joo L., Jea-Bok R. 2015: *A Divided Regression Analysis for Big Data*, International Journal of Software Engineering and Its Applications.
8. Tsai-Hung F., Dennis L., Kuang-Fu C. 2006: *Regression analysis for massive datasets*, Data & Knowledge Engineering.
9. Xiang P., Wentao W., Jis X. 2001: *Will You Be in Hospital Next Year: Leveraging Machine Learning in Improving Healthcare*, University of Wisconsin-Madison.
10. Buuren S., Groothuis-Oudshoorn C. 2011: MICE: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*, 45. 10.18637/jss.v045.i03.

APPLICATION OF REGRESSION ANALYSIS TO BIG DATA

Indrė Baltušninkaitė, Nomeda Bratčikovienė

Abstract. Opportunities and challenges of regression analysis for big data are investigated in the present article. Firstly, the main characteristics describing big data are identified and explained, and then potential challenges that arise in big data analytics are identified. According to the identified challenges, some methods used in the regression analysis for big data are proposed. These methods reduce the calculation burden and select variables that best describe the response variable, thus achieving sufficient statistical accuracy and reducing costs and time of calculations. One of the main purposes of this article is to apply the methods for real data set. Simulation and real data regression models are formed and parameters are estimated using divided regression and regression based on leverage techniques. The LASSO and LARS regressions are used to select the best subset of variables. Finally, model diagnostics, accuracy estimation and comparisons of results are performed.

Keywords: big data, regression analysis, leveraging, LASSO, LARS, RMSLE