KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

Vilija Liutkevičiūtė

Neparametrinių regresinių metodų lyginamasis tyrimas

Magistro darbas

Vadovas
doc. dr. T. Ruzgas

KAUNAS, 2014

KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

TVIRTINU
Katedros vedėjas
doc. dr. N. Listopadskis
2014 06 02

Neparametrinių regresinių metodų lyginamasis tyrimas

Taikomosios matematikos magistro baigiamasis darbas

Vadovas
doc. dr. T. Ruzgas
2014 06 01

Recenzentas
doc. dr. T. Rekašius
2014 06 01

Atliko
FMMM 2 gr. stud.
V. Liutkevičiūtė
2014 05 30

KAUNAS, 2014

KVALIFIKACINĖ KOMISIJA

Pirmininkas: Juozas Augutis, profesorius (VDU)

Sekretorius: Eimutis Valakevičius, profesorius (KTU)

Nariai: Jonas Valantinas, profesorius (KTU)

Vytautas Janilionis, docentas (KTU)

Kristina Šutienė, docentė (KTU)

Zenonas Navickas, profesorius (KTU)

Arūnas Barauskas, dr., direktoriaus pavaduotojas (UAB „Danet Baltic“)

Liutkevičiūtė V. Comparative study of regression methods/ supervisor doc. dr. T. Ruzgas; Department of Applied mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology. – Kaunas, 2014. – 63p.

SUMMARY

In statistics and it‘s application one of the most often solved prediction tasks is valuation of multiple regression. Depending on whether the observed random variable is known to the family of distribution, regression analysis is divided into parametric and non-parametric. In parametric estimation one assumes that the regression f underlying the data $\{y_1, y_2, ..., y_n\}$ belongs to some rather restricted family of functions $f(\cdot;\theta)$ indexed by a small number of parameters $\theta = (\theta_1, \theta_2, ..., \theta_k)$. A regression estimate in the parametric approach is obtained by computing from the data an estimate $\hat{\theta}$ of θ and setting $\hat{f} = f(\cdot;\hat{\theta})$. Such an approach is statistically and computationally very efficient but can lead poor results if none of the family members $f(\cdot;\theta)$ is close to *f*.

In nonparametric regression estimation no parametric assumptions about *f* are made and one assumes instead that *f*, for example, has some smoothness properties (e.g. two continuous derivatives) or that it is square integrable. The shape of the regression function estimate is determined by the data and, in principle, given enough data, arbitrary *f* can be estimated accurately.

In modern data analysis are used many non-parametric methods for multivariate regression dependence of random variables for statistical analysis. Most popular method is the kernel estimator. Also quite popular are histospline, generalized additive models or local regression algorithms. For most popular non-parametric procedures for assessment in practice there is a problem of their choise.

In this work, Monte Carlo approximation was used to perform various non-parametric estimates the accuracy of comparative analysis, in the case, when the valuation model is non-linear and describes the complex surface.

TURINYS

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

ĮVADAS

Statistikoje ir jos taikyme vienas dažniausiai sprendžiamų prognozavimo uždavinių yra daugiamatės regresijos vertinimas. Pagal tai, ar žinoma stebimo atsitiktinio dydžio skirstinio šeima, regresinė analizė skirstoma į parametrinę ir neparametrinę. Parametriniame vertinime daroma prielaida, kad regresijos funkcija *f*, apibūdinanti duomenis, priklauso tam tikrai gan siaurai funkcijų šeimai $f(\cdot;\theta)$, kuri priklauso nuo nedidelio kiekio parametrų $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Regresija, naudojanti parametrinį vertinimą, gaunama iš pradžių apskaičiavus parametro θ įvertį $\hat{\theta}$, o $\hat{f} = f(\cdot;\hat{\theta})$. Toks traktavimas statistiniu požiūriu yra labai efektyvus, tačiau jeigu nei vienas šeimos $f(\cdot;\theta)$ narys nėra artimas funkcijai f, gauti rezultatai gali būti labai netikslūs.

Neparametriniam vertinimui jokios parametrinės prielaidos apie f nėra reikalingos, tačiau vietoj to daromos kitos prielaidos, pavyzdžiui, apie funkcijos f tolydumą (pvz., kad funkcija turi antros eilės tolydžią išvestinę) arba, kad f yra integruojama. Regresijos funkcijos forma yra nustatoma iš turimų duomenų. Turint dideles imtis, funkcija f gali būti apskaičiuota pakankamai tiksliai.

Šiuolaikinėje duomenų analizėje naudojama daugybė neparametrinių metodų, skirtų daugiamačių atsitiktinių dydžių regresinės priklausomybės statistiniam vertinimui. Ypač plačiai paplitę branduoliniai įvertiniai. Gana populiarūs ir splaininiai bei apibendrinti adityvūs ar lokalios regresijos algoritmai. Praktikoje taikant daugumą populiarių neparametrinio vertinimo procedūrų susiduriama su jų pasirinkimo problema.

Šiame darbe Monte Karlo metodu buvo siekiama atlikti įvairių neparametrinių įvertinių tikslumo lyginamąją analizę tuo atveju, kai vertinamas modelis yra netiesinis ir aprašantis sudėtingą paviršių.

Darbo tikslas – ištirti daugiamatės regresijos neparametrinio vertinimo algoritmus sudėtingų, netiesinių modelių atvejais.

Magistro darbo struktūra. Darbą sudaro įvadas, dvi dalys, išvados, literatūros sąrašas ir priedai.

Remiantis literatūros šaltiniais pirmoje dalyje pateikiami magistro darbe sprendžiami uždaviniai ir jų sprendimo metodų analizė. Apžvelgiamas regresinės analizės aktualumas ir statistikos programinės įrangos paketai, pristatomi lyginamojo tyrimo atlikimui naudojami regresijos metodai. Atlikus statistikos programinės įrangos lyginamąją analizę, pateikiamas programinės įrangos pasirinkimo pagrindimas.

Antra dalis skirta eksperimentų ir jų rezultatų analizei. Taip pat joje pateikiami sukurtų algoritmų ir programų taikymą iliustruojantys grafikai.

1. BENDROJI DALIS

1.1. REGRESINĖS ANALIZĖS APŽVALGA

Ankščiausia regresijos forma buvo mažiausių kvadratų metodas, kurį publikavo Ležandras (*Legendre*) 1805 metais ir Gausas (*Gauss*) 1809 metais. Ležandras ir Gausas pritaikė šį metodą astronominiams stebėjimams aiškinti t.y. kūnų skriejimui aplink Saulę (daugiausia kometoms, vėliau ir kitoms, naujai atrastoms, mažesnėms planetoms). 1821 metais Gausas išplatino papildytą mažiausių kvadratų teoriją, kartu su Gauso – Markovo teoremos versija.

Sąvoką „regresija“ XIX – ame amžiuje pirmą kartą įvedė Francis Galtonas (*Francis Galton*), ji buvo skirta apibūdinti neįprastiems biologiniams reiškiniams. Atliktas tyrimas parodė, kad aukštų protėvių palikuonių ūgis linkęs regresuoti žemyn, link populiacijos vidurkio (neįprastas reiškinys taip pat žinomas, kaip regresija link vidurkio). F. Galtonui regresija turėjo tik šią biologinę reikšmę, tačiau, bendresniam statistikos kontekstui, jo darbą vėliau papildė Udnis Julas (*Udny Yule*) ir Karlas Pirsonas (*Karl Pearson*). Julo ir Pirsono darbe daromos prielaidos, kad bendras priklausomo ir nepriklausomų kintamųjų pasiskirstymas yra Gauso (Gaussian). Šią prielaidą savo darbuose 1922 ir 1925 metais sukritikavo R. A. Fišeris (*R. A. Fisher*). Fišeris teigė, kad sąlyginis priklausomo kintamojo pasiskirstymas yra Gauso, tačiau bendras pasiskirstymas nebūtinai turi būti Gauso. Šiuo atžvilgiu, Fišerio prielaida yra artimesnė Gauso formuluotei pateiktai 1821 metais.

XX – to amžiaus šeštajame ir septintajame dešimtmetyje ekonomistai, norėdami apskaičiuoti regresiją, naudojo stalinius elektromechaninius skaičiuotuvus. Iki 1970 metų, rezultatams gauti kartais prireikdavo net 24 valandų.

Regresiniai metodai ir toliau yra aktyvi mokslinių tyrimų sritis. Per pastaruosius dešimtmečius buvo sukurti nauji metodai: robastinė regresija, neparametrinė regresija, Bajeso (*Bayesian*) regresijos metodai, regresija įtraukiant koreliuotus prilausomus kintamuosius, tokius kaip laiko eilutės ir augimo kreivės, regresija, kurioje prognozuojamas priklausomas kintamasis yra kreivės, paveikslėliai, diagramos ar kiti kompleksiniai objektai ir daug kitų. **[1.]**

Pagrindinis regresinės analizės tikslas yra nustatyti stochastinio ryšio tarp priklausomo kintamojo ir nepriklausomų kintamųjų formą bei analitinę išraišką. Nepriklausomi kintamieji taip pat vadinami aiškinančiaisiais kintamaisiais, kovariantėmis, regresoriais. Priklausomas kintamasis dar vadinamas atsaku. Modelis, kai yra vienas nepriklausomas kintamasis, vadinamas vienmate regresine analize, kai yra keli nepriklausomi kintamieji – daugiamate regresine analize.

Regresinės analizės metodai plačiai taikomi įvairiose srityse – medicinoje, gamtos moksluose, istorijoje, ekonomikoje, versle, sociologijoje, psichologijoje, sporte, inžinerijoje, ir kt. Pavyzdžiui, prognozuojama tam tikro vaistinio preparato koncentracija organizme atsižvelgiant į laiką, praėjusį po vaistinio preparato vartojimo [**2.**]; migracija šalies viduje atsižvelgiant į bedarbystės lygį [**3.**];

buto kaina pagal plotą ar vietovę [**4.**]; istorinis objekto amžius remiantis kai kuriomis su objekto amžiumi susijusiomis charakteristikomis [**5.**].

Regresinę analizę galima skirstyti į keletą etapų. Iš pradžių suformuluojamas uždavinys, t.y. klausimai, į kuriuos reikia atsakyti atlikus analizę. Kitas etapas – tinkamų kintamųjų parinkimas, t.y. parinkimas tokių kintamųjų, kurie gerai prognozuotų priklausomą kintamąjį. Pavyzdžiui, prognozuojant buto kainą galimi nepriklausomi kintamieji: buto plotas; kambarių skaičius; namo, kuriame yra butas, tipas, amžius, vietovė; nekilnojamojo turto mokestis ir pan. Dar vienas etapas – modelio parinkimas, t.y. reikia parinkti, kaip susijęs priklausomas kintamasis ir nepriklausomi kintamieji. Atliekama analizė, kurios metu tiriama, ar hipotetinis modelis tinka. Parinktą tinkamą modelį naudojame prognozėms.

1.2. TIESINĖ REGRESINĖ ANALIZĖ

1.2.1. MODELIO LYGTIS

Bendriausias tiesinės priklausomybės modelis, siejantis kintamuosius *Y* ir *X* atrodo taip:

$$Y = a + bX + \varepsilon. \tag{1.1}$$

Šioje lygtyje a, b yra nežinomi parametrai, o ε – atsitiktinė paklaida. Kaip atsiranda atsitiktinė paklaida? Dažnai atsitiktinė paklaida – tai matavimo paklaida. Atsitiktinę paklaidą gali lemti ir individualūs respondentų skirtumai, pavyzdžiui, individuali organizmo reakcija į kraujospūdį mažinančius vaistus.

1.2.2. TIESINĖS REGRESIJOS PRIELAIDOS IR REIKALAVIMAI DUOMENIMS

Kad tiesinės regresijos parametrinis modelis būtų taikoma korektiškai, atsitiktinės paklaidos ε turi tenkinti sąlygas:

1. $\varepsilon_i -$ normaliai pasiskirstę atsitiktiniai dydžiai;
2. visų ε_i vidurkiai lygūs nuliui, t.y.: $E\varepsilon_i = 0$;
3. visų ε_i dispersijos lygios, t.y.: $D\varepsilon_i = \sigma^2$;
4. visi ε_i yra tarpusavyje nepriklausomi.

Jei skaičiuojant parametrus yra tenkinami pirmieji du paminėti reikalavimai, tuomet turime, taip vadinamus „geriausius“ parametrų įverčius, kurie pasižymi trimis savybėmis: yra nepaslinkti, efektyvūs ir suderinti.

Nepaslinktumas reiškia, jog, apskaičiavus *n* regresijos lygčių su skirtingomis duomenų imtimis, gauname įverčius, kurių vidurkis yra lygus tikrajai parametro reikšmei.

Efektyvumas. Įverčiai yra efektyvūs tada, kai jų dispersija yra minimali. Ši savybė reiškia, kad įverčiai išsibarstę įmanomai arti aplink tikrąsias parametro reikšmes.

Suderinti įverčiai reiškia, kad duomenų stebėjimų skaičiui neapibrėžtai augant, parametro įverčio reikšmė artėja prie tikrosios parametro reikšmės.

Trumpai visas aukščiau paminėtas keturias prielaidas galima užrašyti taip: $\varepsilon_i \sim N(0, \sigma^2)$; čia ε_i yra nepriklausomi atsitiktiniai dydžiai. Beje, dispersija σ^2 nežinoma.

Aprašysime iš prielaidų išplaukiančias regresijos modelio savybes:

1. *Paklaidų normavimas.* Kadangi ε_i yra normalieji atsitiktiniai dydžiai, tai su kiekviena fiksuota X_i reikšme kintamieji Y_i irgi yra normalieji atsitiktiniai dydžiai. Regresinės analizės rezultatai nedaug keičiasi ir tuo atveju, kai kintamųjų skirstiniai truputį skiriasi nuo normaliųjų. Praktiškai regresija taikoma ir tuomet, kai kintamieji įgyja ne mažiau kaip 7 skirtingas reikšmes iš tvarkos skalės;
2. *Vidurkių lygybė nuliui.* Atsižvelgiant į tai, jog $Y_i = a + bX_i + \varepsilon_i$ modelyje X_i fiksuoti, apskaičiuokime kintamojo Y_i vidurkį: $EY_i = a + bX_i$. Kintamųjų X ir Y sąryšį aprašo stochastinė lygtis. Tuo tarpu priklausomo kintamojo vidurkis su X_i susietas determinuota tiesine lygtimi. Parametrai a ir b nežinomi. Pakeitus parametrus jų įverčiai, $EY_i = a + bX_i$ lygybės analogas taikomas *Y* reikšmėms prognozuoti;
3. *Dispersijų lygybė.* Ši prielaida vadinama homoskedastiškumo reikalavimu. Kai ji netenkinama, sakoma, kad duomenys heteroskedastiški. Regresinis modelis jautrus šios prielaidos pažeidimams. Skirtingos dispersijos iš esmės gali iškreipti prognozę. Homoskedastiškumas yra reikalavimas, kad su kiekvienu fiksuotu X_i galimų Y_i reikšmių sklaida būtų vienoda. Atkreipiame dėmesį, kad patys Y_i reikšmių vidurkiai kinta tiesiškai, o vienoda tik reikšmių sklaida apie vidurkius.
4. *Paklaidų priklausomybė.* Nepriklausomos paklaidos reiškia, kad visi Y_i nepriklausomi. Iš šio reikalavimo išplaukia, pavyzdžiui, kad Y_i nepriklauso nuo Y_{i-1}. Tarkime, kad X_i yra to paties kintamojo matavimai skirtingais laiko momentais. Tuomet iš paklaidų nepriklausomumo išplaukia, kad priklausomojo kintamojo reikšmė Y_i, įgyta i-uoju laiko momentu, nepriklauso nuo reikšmės Y_{i-1}, įgytos prieš tai, i-1-uoju laiko momentu. Toks reikalavimas ne visada pagrįstas. Tada tiesinės regresijos modelis netaikytinas ir reikia naudotis autoregresija.

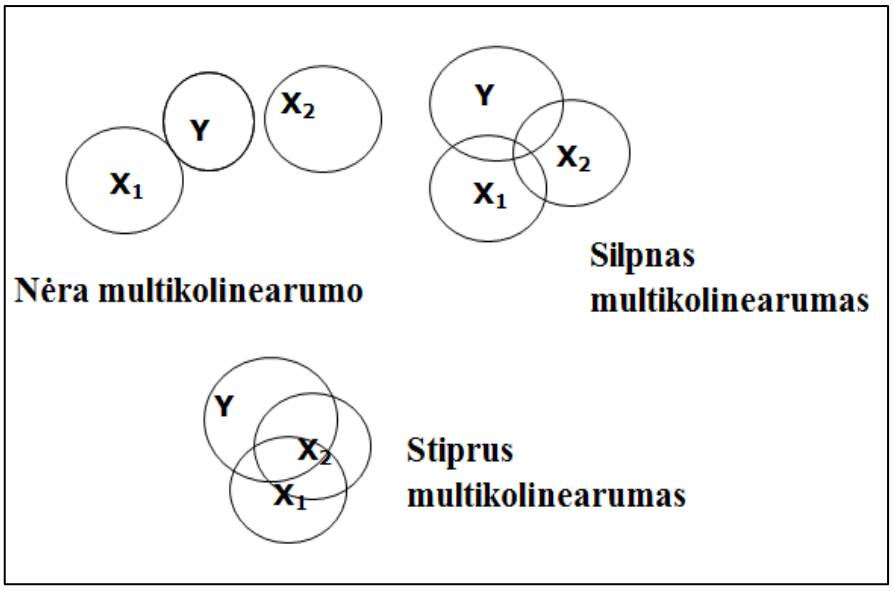
Visi duomenys skaitiniai. Net ir vardų skalės kintamųjų reikšmės užkoduotos skaičiais. Be to: priklausomas kintamasis *Y* yra normaliai pasiskirstęs. Visi kiti kintamieji išmatuoti intervalų skalėje, išskyrus dalį dvireikšmių kintamųjų.

Klasikiniame modelyje tariama, kad nepriklausomi kintamieji matuojami be paklaidų ir yra neatsitiktiniai. Pastarasis reikalavimas praktikoje dažniausiai nėra logiškas, todėl visuotinai priimta nepriklausomus kintamuosius irgi laikyti atsitiktiniais dydžiais. Nepriklausomi kintamieji tuo geriau tinka modeliui, kuo jie panašesni į normaliuosius atsitiktinius dydžius. Faktiškai, tai reikalavimas, kad visų kintamųjų pasiskirstymo tankiai būtų „varpo“ formos. Neretai, siekiant didesnio kintamųjų panašumo į normaliuosius, kintamieji transformuojami. Nepriklausomų kintamųjų normalumo nereikia suabsoliutinti. Socialiniuose moksluose masiškai taikomi regresiniai modeliai, kai joks kintamųjų normalumas netiriamas ir pasitenkinama tuo, kad jie yra išmatuoti intervalų skalėje.

Kartais į modelį įtraukiami ir kategoriniai kintamieji, vadinami pseudokintamieji. Visi jie turi būti perkoduoti taip, kad įgytų tik dvi reikšmes – 0 ir 1. Pavyzdžiui, galima į modelį įtraukti pseudokintamąjį lytis (1 – vyr., 0 – mot.). Vis dėlto pseudokintamieji naudotini tik tada, kai duomenų mažai. Daug tikslesni regresijos modeliai bus gauti, kai atskirai tirsime vyrus ir moteris.

Skirtingų stebėjimų liekamosios paklaidos ε neturi koreliuoti. Praktikoje tai reiškia, kad stebėjimai nesusiję.

Nepriklausomi kintamieji neturi stipriai koreliuoti. Priešingu atveju iškyla vadinamoji multikolinearumo problema. Multikolinearumas yra nepageidaujama situacija, kai koreliacijos tarp nepriklausomų kintamųjų yra stiprios. Kai kuriais atvejais, regresijos rezultatai gali atrodyti paradoksaliai. Pavyzdžiui, modeliui duomenys yra tinkami, nors nė vienas iš nepriklausomų kintamųjų neturi statistiškai reikšmingo poveikio aiškinant priklausomą kintamąjį – *Y*. Tai yra įmanoma tada, kai du nepriklausomi kintamieji yra stipriai koreliuoti ir jie abu pateikia tą pačią informaciją. Kai taip nutinka gauti rezultatai rodo multikolinearumą.



1.1 pav. Multikolinearumo problemos esmė

Multikolinearumas padidina koeficientų standartines paklaidas. Padidėjusios standartinės paklaidos reiškia, kad kai kurių nepriklausomų kintamųjų koeficientai gali būti nereikšmingi, t.y. nesiskirti nuo nulio. Be multikolinearumo ir su mažesnėmis standartinėmis paklaidomis, tie patys koeficientai gali būti gauti su prasminga reikšme ir tyrėjas formuluotų reikšmingas išvadas, priešingai nei pirmuoju atveju, t.y. esant miltikolinearumui.

Dispersijos mažėjimo daugiklio (*VIF*) koeficientas matuoja kiek yra padidėjusi apskaičiuotų koeficientų dispersija, lyginant su atveju, kai tarp kintamųjų *X* nėra koreliacijos. Jei nei vienas iš kintamųjų *X* nėra koreliuotas, tada *VIF* koeficientas bus lygus 1. Jei *VIF* koeficientas vienam iš kintamųjų yra lygus 5 ar didesnis, tai sakome, kad kintamasis yra per daug multikolinearus. Tokiu atveju paprasčiausias sprendimas: jei modelyje yra vienas ar daugiau kintamųjų, kurių *VIF* koeficientas yra 5 ar daugiau, vieną iš šių kintamųjų reikėtų pašalinti iš regresijos modelio. Tam, kad nuspręsti kurį kintamąjį šalinti būtų tikslingiausia, reikėtų paeiliui šalinti kiekvieną atskirai ir parinkti tą regresijos lygtį, kuri paaiškina daugiausia dispersijos (R^2 yra didžiausias).

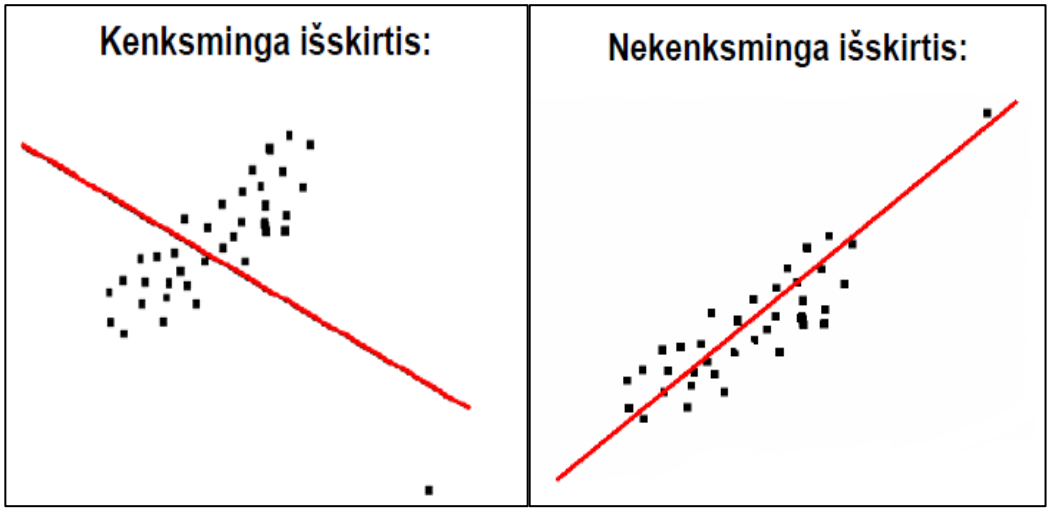
Kiti požymiai, kad egzistuoja multikolinearumas:

- regresijos koeficientai keičiasi radikaliai, kai paliekamas arba pašalinamas nepriklausomas kintamasis;
- regresijos koeficientas gaunamas neigiamas, kai teoriškai *Y* turi didėti didėjant nepriklausomam kintamajam, arba regresijos koeficientas yra teigiamas, kai teoriškai *Y* turi mažėti didėjant nepriklausomam kintamajam;
- egzistuoja stiprios porinės koreliacijos tarp nepriklausomų kintamųjų (tačiau trys ar daugiau nepriklausomų kintamųjų gali būti multikolinearūs kartu, neesant stipriai porinei koreliacijai).

Būdai kaip galima pašalinti multikolinearumą:

- Paprasčiausia išeitis iš analizės pašalinti tarpusavyje koreliuotus kintamuosius.
- Imties tūrio padidinimas yra dažniausiai naudojamas būdas, nes kai imties tūris padidėja, sumažėja standartinė paklaida (visi kiti dydžiai išlieka tokie patys).
- Apjungti koreliuosius kintamuosius į suminį kintamąjį. [**7.**]

Duomenyse neturi būti išskirčių. Stebėjimus, labai išsiskiriančius nuo didžiosios stebėjimų dalies priimta vadinti išskirtimis. Tokie stebėjimai nebūtinai iš esmės keičia parametrų įverčius t.y gali būti nekenksmingos ir kenksmingos išskirtys:



1.2 pav. Kenksminga ir nekenksminga išskirtys.

Galimos atsiradimo priežastys: vedimo klaida, pašalinio kintamojo įtaka. Žinant išskirties priežastis, dažnai tą stebėjimą galima pašalinti, tačiau vien todėl kad stebėjimas yra išskirtis, jo šalinti negalima. Taigi tikslas surasti duomenų išskirtis ir nuspręsti, ką su jomis daryti.

V. Čekanavičius ir G. Murauskas (2002) siūlo išskirtis nustatyti pagal:

a) stebėjimo įtakos indeksą;

b) standartizuotąją liekaną:

c) Kuko matą.

Stebėjimo įtakos indeksas. Stebėjimo (X_i,Y_i) įtakos indeksas h_i apskaičiuojamas pagal formulę:

$$h_i = \frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum_{i=1}^{n}X_i^2-(\sum_{i=1}^{n}X_i)^2/n}, \quad (1.2)$$

čia n – porinės imties didumas, $\bar{X}$ – reikšmių X_1,...,X_n empirinis vidurkis.

Jis įvertina tik pirmosios koordinatės (nepriklausomojo kintamojo) reikšmę (ar toli nuo $\bar{X}$ yra X_i). Kuo regresijos tiesės lygties koeficientai labiau priklauso nuo stebinio, tuo stebinio įtakos indeksas yra didesnis.

Stebinį (X_i,Y_i) laikome išskirtimi, jei

$$h_i > \frac{4}{n}. \quad (1.3)$$

(1.3) sąlyga yra ekvivalenti reikalavimui: $|X_i - \bar{X}| > (3\left(1-\frac{1}{n}\right))^{\frac{1}{2}}s_X$, čia s_X yra imties (X_1,X_2,...,X_n) standartinis nuokrypis, kuris $s_X = \frac{1}{\sqrt{n-1}}\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}$.

Standartizuotosios liekanos (SR_i). Tai yra liekamųjų paklaidų e_i Z – reikšmės (Tarkime, turime duomenų aibę X_1,X_2,...,X_n. Tuomet Z reikšmė skaičiuojama pagal formulę $Z = \frac{X_i-\bar{X}}{s}$, kur X_i – duomenų aibė, $\bar{X}$ yra duomenų aibės vidurkis, o s – standartinis nuokrypis.). Beje, visų e_i kur i=1,…,n aritmetinis vidurkis lygus nuliui. Standartizuotoji liekana SR_i apskaičiuojama pagal formulę:

$$SR_i = \frac{e_i}{\sqrt{\sum_{i=1}^{n}(Y_i - a - bX_i)^2/(n-2)(1-h_i)}}, \quad (1.4)$$

čia e_i yra i – tojo stebinio liekamoji paklaida, h_i – stebinio įtakos indeksas, i=1,…,n, n – imties tūris. Standartizuotųjų liekanų imties vidurkis lygus nuliui, o imties dispersija – vienetui.

Stebinį laikome išskirtimi, jei standartizuotosios liekanos modulis didesnis už 3, t.y. $|SR_i| > 3$.

Kuko matas (D_i) (*atstumo statistika*). Jis atsižvelgia ir į standartizuotąją liekaną ir į stebinio įtakos indeksą. Kuko matas:

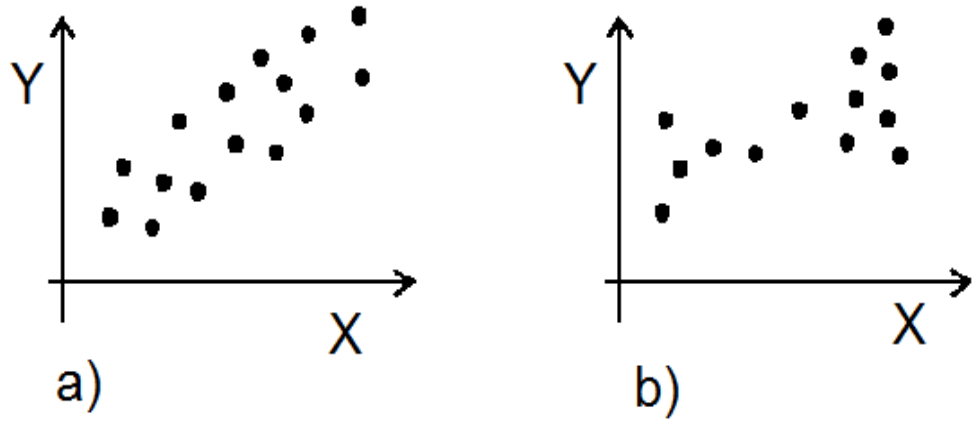
$$D_i = \frac{(SR_i)^2 h_i}{2(1-h_i)}, \quad (1.5)$$

čia SR_i yra i – tojo stebinio standartizuota liekana, o h_i – i – tojo stebinio įtakos indeksas.

Stebinį (X_i,Y_i) laikome išskirtimi, jei $D_i > F_{0.5}$ (2, n-2), čia $F_{0,5}$ (2, n-2) yra Fišerio skirstinio su 2 ir (n-2) laisvės laipsnių 0,5 lygmens kritinė reikšmė. Išskirtimi laikome stebinį, kurio D_i>1.

Radę išskirtis, pirmiausia turime patikrinti ar duomenyse nėra klaidos. Po to turime išsiaiškinti, kaip atsirado duomenų išskirtis. Neišsiaiškinus priežasties jos šalinti negalima. Tokiu atveju rekomenduojama duomenis papildyti naujais stebiniais ir tyrimą kartoti. Dar galima įtarti, kad regresijos modelis nėra tinkamas šiems duomenims.

Duomenys turi būti homoskedastiški. Reikalaujama, kad liekamosios paklaidos dispersija nepriklausytų nuo regresorių reikšmių. Jeigu taip nėra, tai sakome, kad iškilo heteroskedastiškumo problema. Praktiškai heteroskedastiškumas pasireiškia tuo, kad vienoms regresorių reikšmėms priklausomas kintamasis Y įgyja labai skirtingas reikšmes, o kitoms – ne. Nubraižykime Y priklausomybės nuo kiekvieno regresoriaus grafikus ir pažiūrėkime, ar gauti Y reikšmių išsibarstymo „debesėliai“ yra daugmaž vienodo storio visoms X reikšmėms. Jeigu taip, tai duomenys homoskedastiški (a) 1.3 pav.). Jeigu ne, tai – heteroskedastiški (b) 1.3 pav.). Modelis, sudarytas labai heteroskedastiškiems duomenims nėra patikimas.



1.3 pav. Duomenų homoskedastiškumas ir heteroskedastiškumas

1.2.3. TIESINIMO TRANSFORMACIJOS

Neretai tam, kad netiesinę regresiją paversti tiesine ir toliau skaičiuoti tiesinę regresiją atliekamos tiesinimo transformacijos. Ar verta tiesinti paprastai nustatoma atidedant taškus grafiškai paprastose ir transformuotose ašyse ir tada pažiūrime ar gaunasi tiesė. Tai paprastai daroma kompiuteriu ar specialiame grafiniame popieriuje.

Dažniausiai naudojamos transformacijos:

Logistinė (*angl. logistic*) *transformacija*. Ši transformacija konvertuoja atstumus į tikimybinius matavimus esančius intervale nuo nulio iki vieneto. Logistinės transformacijos formulė:

$$Y_i = 1/(1 + e^{-X_i}), \tag{1.6}$$

taip pat, gali būti naudojama tokia logistinės transformacijos formulė:

$$Y_i = 1/(1 + e^{-Z_i}), \tag{1.7}$$

kur $Z_i = (X_i - m)/d$. m yra vidurkis tarp dviejų reikšmių.

Kvadratinės šaknies (*angl. square root*) *transformacija*. Paprastai naudojama, kai duomenys yra kiekis išreikštas vienetais ploto vienete - vabzdžių skaičius ant lapo, eritrocitų skaičius hematocitometre ir pan. Tokie duomenys paprastai būna pasiskirstę ne pagal normalųjį, o pagal Puasono skirstinį. Puasono skirstinyje vidurkis ir dispersija sutampa, tad jie nėra nepriklausomi. Duomenis perskaičiavus į kvadratines jų šaknis dispersija pasidaro nepriklausoma nuo vidurkio. Kai tarp duomenų yra 0, tuomet prie visų duomenų pridedamas 0,5. Tada transformacija - $\sqrt{Y + 0{,}5}$.

Ši transformacija buvo patobulinta:

a) Anskombo (*Anscombe*, 1948): $Y' = \sqrt{Y + \frac{3}{8}}$;

b) ir Frimeno bei Tjukio (*Freeman, Tukey*; 1950): $Y' = \sqrt{Y} + \sqrt{Y + 1}$,

tačiau šie patikslinimai retai naudojami.

Logaritminė (*angl. log*) *transformacija.* Paprastai transformuojamas priklausomas kintamasis. Ištiesina priklausomybę:

$$\hat{Y} = a * e^{bX}\ , \tag{1.8}$$

kuri virsta:

$$\log \hat{Y} = \log a + b * \log_e X. \tag{1.9}$$

Nepriklausomojo kintamojo X logaritminę transformaciją taikyti verta, kai X santykinis kitimas lemia tiesišką Y kitimą.

Bokso-Kokso (*angl. Box-Cox*) *transformacija*. Efektyvi, bet sudėtinga, paprastai taikoma tik transformuojant kompiuteriais, nes reikia iteracinio skaičiavimo. Ji aprašoma taip:

$$f(x) = \begin{cases} (Y^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}. \quad (1.10)$$

Šioje transformacijoje teigiamas priklausomas kintamasis *Y* reguliuojamas parametru λ. Riba, kai λ reikšmė pasiekia nulį yra logistinė transformacija. Kokso – Bokso transformacija gali būti apibrėžiama ir tokiu būdu:

$$f(x) = \begin{cases} -((Y + c)^{\lambda} - 1)/(\lambda g), & \lambda \neq 0 \\ \log(Y + c)/g, & \lambda = 0 \end{cases}, \quad (1.11)$$

čia c = 0. Parametras c gali būti naudojamas perskaičiuojant *Y*, tam kad jis būtų griežtai teigiamas. g = 1. Taip pat, parametras g gali būti lygus $\acute{Y}^{\lambda-1}$, kur $\acute{Y}$ geometrinis *Y* vidurkis. [**8.**]

1.3. NEPARAMETRINĖ REGRESINĖ ANALIZĖ. REGRESIJOS METODŲ APŽVALGA

1.3.1. APIBENDRINTŲ ADITYVIŲ MODELIŲ METODAS

Apibendrintų adityvių modelių metodas, lyginant su kitais neparametrinės regresijos metodais yra universaliausias. Jo metodika yra žymiai lankstesnė, nei tradicinių parametrinių modeliavimo priemonių, tokių kaip tiesinės ar neteisinės regresijos. Šis metodas nereikalauja įprastų parametrinių prielaidų ir parodo struktūrą tarp nepriklausomų ir priklausomų kintamųjų, kuri priešingu atveju gali likti nepastebėta.

Daug neparametrinių metodų yra neefektyvūs, kai modelyje yra didelis nepriklausomų kintamųjų skaičius. Tokiu atveju, duomenų sklaida padidina įverčių poslinkį. Norint to išvengti buvo sukurti adityvūs modeliai. Šie modeliai įvertinami daugiamatės regresijos funkcijos sudėtine aproksimacija. Sudėtinės aproksimacijos nauda yra dviguba. Visų pirma, kiekvienas sudėtinis narys yra įvertinamas atskirai. Antra, kiekvieno nario įvertis parodo, kaip priklausomas kintamasis kinta su atitinkamu nepriklausomu kintamuoju.

Apibendrinti adityvūs modeliai buvo pasiūlyti norint adityviems modeliams naudoti įvairesnius skirstinius. Sakome, kad šiuose modeliuose priklausomojo kintamojo vidurkis priklauso nuo prognozuojamojo kintamojo, naudojant netiesinio ryšio funkciją. Apibendrinti adityvūs modeliai leidžia priklausomo kintamojo tikimybės skirstiniui būti bet kokiu eksponentinio tipo skirstiniu. Daug plačiai taikomų statistinių modelių priklauso šiai klasei, įskaitant adityvius, neparametrinius – logistinius, logistinius – tiesinius modelius. Apibendrintų adityvių modeliui metodui realizuoti skirta SAS (angl. – Statistical Analysis System) paketo procedūra GAM.

Pagrindinės apibendrintų adityvių modelių metodo savybės:

- metodas tinka neparametriniams arba pusiau parametriniams modeliams;
- metodas tinka daugiamačiams duomenims;
- metodas įvertina laisvės laipsnių skaičių arba suglodinimo parametrą.

Apibendrintų adityvių modelių metode gali būti naudojami Gauso, binominiai, Puasono ir Gama skirstiniai. Kiekvienam skirstiniui, bent jau teoriškai, egzistuoja daugiau nei viena forma, tačiau metode visada naudojama kanoninė forma.

Apibendrintas adityvus modelis:

Tarkime, kad Y yra atsitiktinis priklausomas kintamasis, o $X_1,X_2,\ldots,X_p$ – nepriklausomų kintamųjų rinkinys. Regresijos procedūra laikoma metodu rasti Y reikšmei, kai duotas $X_1,X_2,\ldots,X_p$ reikšmių rinkinys. Standartinis tiesinis regresijos modelis turi tiesinę priklausomybę tarp nepriklausomų kintamųjų $X_1,X_2,\ldots,X_p$ ir priklausomo kintamojo Y:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{1.12}$$

čia, ε – regresijos funkcijos paklaida, β_i , $i = 1,\ldots p$ – nežinomi regresijos funkcijos koeficientai. Panaudojus mažiausiųjų kvadratų metodą, randami regresijos lygties koeficientų β_i taškiniai įverčiai b_i. Adityvus modelis apibendrina tiesinį modelį, priklausomybė modeliuojama taip:

$$Y = s_0 + s_1(X_1) + s_2(X_2) + \cdots + s_p(X_p) + \varepsilon, \tag{1.13}$$

čia, $s_i(X)$, i=1,2,…p – glodinančios funkcijos. Tam, kad būtų naudingos, glodinančios funkcijos turi tenkinti tokią sąlygą: $E_{sj}(X_j)$=0. Šios funkcijos nėra parametrinės formos, tačiau naudojamos neparametriniuose modeliuose.

Apibendrinti adityvūs modeliai papildo tradicinius tiesinius modelius įvertindami sąsają tarp $f(X_1,..X_p)$ ir tikėtinos Y reikšmės. Tradiciniai tiesiniai modeliai gali būti naudojami daugelyje statistinių duomenų analizių, tačiau yra tokių atvejų, kuriuose jie yra netinkami. Pavyzdžiui, normalus skirstinys gali būti neadekvatus modeliuojant diskrečiąsias reikšmes. Kadangi, apibendrintuose adityviuose modeliuose be normalaus skirstinio, naudojami kiti įvairūs skirstiniai, šie modeliai gali aprėpti daug platesnę duomenų analizę.

Apibendrinti adityvūs modeliai susideda iš atsitiktinio komponento, adityvaus komponento ir ryšio funkcijos siejančios šiuos du komponentus. Manoma, kad priklausomas kintamasis Y, atsitiktinis komponentas, turi eksponentinio tipo skirstinį:

$$f_Y(y;\theta;\phi) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\}, \tag{1.14}$$

čia, θ – normalusis parametras, ϕ – gama parametras. Normalusis, binominis ir Puasono, kartu su daug kitų skirstinių priklauso eksponentinio tipo skirstiniams. Dydis:

$$\eta = s_0 + \sum_{j=1}^{p} s_j\left(X_j\right), \tag{1.15}$$

čia, $s_i(.),\ldots,s_p(.)$ yra glodinančios funkcijos apibūdinančios adityvųjį komponentą. Ryšys tarp priklausomo kintamojo vidurkio μ ir η apibūdinamas tokia priklausomybe: $g(\mu)=\eta$. Dažniausiai naudojama ryšio funkcija yra kanoninė, kurioje $\eta=\theta$.

Apibendrinti adityvūs modeliai ir apibendrinti tiesiniai modeliai gali būti taikomi panašiais atvejais, tačiau jie atlieka skirtingas analitines paskirtis. Apibendrinti tiesiniai modeliai akcentuoja rezultatus ir išvadas, o apibendrinti adityvūs modeliai tyrinėja duomenų neparametriškumą.

Apibendrinti adityvūs modeliai labiau tinkami duomenų rinkinių analizei ir ryšių tarp priklausomo kintamojo ir nepriklausomų kintamųjų vizualizacijai.

Apibendrintų adityvių modelių taikymas:

Apžvelkime, glodinimo išraiškų s_0, $s_1(.),\ldots,s_p(.)$ įvertinimą adityviąjame modelyje. Yra daug būdų kaip išnagrinėti adityvių modelių formuluotes ir įverčius. *Back-fitting* algoritmas yra pagrindinis algoritmas, kuris tinka adityviems modeliams, naudojant bet kokį regresijos tipą. j – toji dalinių liekanų aibė apibrėžiama taip:

$$R_j = Y - s_0 - \sum_{k \neq j} s_k(X_k), \tag{1.16}$$

Dalinės liekanos panaikina visų kitų kintamųjų poveikį kintamajam Y, todėl jie gali būti naudojami modeliuojant X_j. Tai yra *Back-fitting* algoritmo pagrindas, būdas įvertinti kiekvieną glodinimo funkciją $s_j(.)$. *Back-fitting* algoritmas yra iteracinis, pradedant pradinėmis funkcijomis $s_0,\ldots,s_p$, ir kiekvienoje iteracijoje pritaikant dalinėms funkcijoms tinkančius glodumo parametrus. Iteracijos tęsiamos tol kol apskaičiuoti parametrai nebesikeičia.

Algoritmas kol kas apibūdina tik adityvius modelius. Algoritmas apibendrintiems adityviems modeliams yra šiek tiek sudėtingesnis. Apibendrintus adityvius modelius papildo apibendrinti tiesiniai modeliai, tokiu pačiu būdu kaip ir adityvūs modeliai papildo tiesinės regresijos modelius, t.y. tiesinę formą $\alpha+\sum_j X_j\beta_j$ pakeisdami į adityvią formą $\alpha+\sum_j s_j(X_j)$.

Tokiu pačiu būdu, adityvaus komponento įvertis apibendrintiems adityviems modeliams užbaigiamas pakeičiant pataisyto priklausomo kintamojo svertinę tiesinę regresiją į svertinį *Back-fitting* algoritmą, iš esmės pritaikydamas svertinį adityvų modelį. Šiuo atveju naudojamas algoritmas vadinamas lokaliu taškų apskaičiavimo (*Local scoring*) algoritmu. Jis taip pat yra iteracinis ir pradedamas su pradiniais $s_0,\ldots,s_p$ įverčiais. Kiekvienoje iteracijoje pataisytas priklausomas kintamasis ir svorių rinkinys yra apskaičiuojami ir tada įvertinami glodinimo komponentai, naudojant svertinį *Back-fitting* algoritmą. Taškų apskaičiavimo algoritmas yra stabdomas, kai įverčių nuokrypiai nustoja mažėti.

Apibendrintų adityvių modelių įvertinimo procesas susideda iš dviejų ciklų. Kiekviename lokalaus taškų apskaičiavimo algoritmo žingsnyje (išorinis ciklas), svertinis *Back-fitting* algoritmas (vidinis ciklas) naudojamas iki konvergencijos. Tada, remiantis įverčiais iš šio svertinio *Back-fitting* algoritmo, apskaičiuojamas naujas svorių rinkinys ir toliau pradedama taškų apskaičiavimo algoritmo iteracija.

Bet koks neparametrinis glodinimo metodas gali būti naudojamas $s_i(X)$ reikšmėms gauti. Apibendrintų adityvių modelių metodas realizuoja B – splainų ir lokalios regresijos metodus, vertinant glodumo komponentus pagal vieną požymį ir glodinimą splainais, vertinant glodumo komponentus pagal kelis požymius.

Apibendrintų adityvių modelių metodas turi vienintelį glodumo parametrą. Apibendrinta kryžminio palyginimo (angl. The generalized cross validation) – GCV funkcija plačiai naudojama daugelyje neparametrinės regresijos metodų, kaip glodumo parametrų parinkimo kriterijus. GCV funkcija aproksimuoja tikėtiną prognozės paklaidą. Manoma, kad modelis parinktas pagal šią funkciją yra tinkamiausias. Be automatiškai parenkamo glodumo parametro, apibendrytų adityvių modelių metodas taip pat leidžia kiekvienam glodumo komponentui nurodyti laisvės laipsnį. [**11.**]

1.3.2. LOKALIOS REGRESIJOS METODAS

Vienas iš neparametrinių regresijos kreivės vertinimo metodų – lokali regresija. Lokalios regresijos metodas naudojamas ir daugiamatės regresijos neparametriniam vertinimui, Lokalios regresijos metodui realizuoti skirta programos SAS paketo procedūra LOESS. Lokalios regresijos metodas yra kur kas pranašesnis už kitus tradicinius modeliavimo būdus, nes jis naudojamas ir tuo atveju, kai nežinoma tiksli regresijos paviršiaus parametrinė forma. Be to, šis metodas yra naudojamas ir tada, kai duomenyse yra išskirčių.

Pagrindinės lokalios regresijos metodo savybės:

- metodas tinka neparametriniams modeliams;
- metodas tinka naudojant daugiamačius duomenis;
- metodas tinka naudojant prikausomus daugialypius kintamuosius;
- metodas tinka, kai duomenyse yra išskirčių.

Tarkime, kad i=1,...,n, i – tasis priklausomo kintamojo Y įvertis Y_i ir nepriklausomo kintamojo X įvertis X_i aprašomi tokia funkcine priklausomybe:

$$Y_i = g(X_i) + \varepsilon_i, \tag{1.17}$$

čia, g – regresijos funkcija, ε_i – atsitiktinė paklaida.

Lokalios regresijos metodo mintis yra tokia, kad artimam X=X_0, regresijos funkcija $g(x)$ gali būti aproksimuota, kažkokia apibrėžta parametrinio tipo funkcija. Tokia lokali aproksimacija yra gaunama pritaikant regresijos paviršių duomenų taškams per pasirinktus taško X_0 artinius. Lokalios regresijos metode, svertinis mažiausių kvadratų metodas, naudojamas norint pritaikyti tiesines arba kvadratines funkcijas, prognozuojamiems centriniams artiniams. Artinio plotas parenkamas taip, kad kiekvienas artinys sudarytų vienodą, procentinę duomenų taškų dalį. Duomenų dalis, vadinama glodumo parametru, kiekviename lokaliame artinyje nusako apskaičiuoto paviršiaus glodumą. Lokalios regresijos metodas yra tinkamas, kai ε_i yra atsitiktinis normalus dydis su vidurkiu lygiu nuliui. Šis metodas taip pat gali būti taikomas, kai paklaidų pasiskirstymas yra simetrinis. Jei, taikant šį metodą, imtis yra nedidelė, mažai tikėtina, kad bus rasta visus duomenis atitinkanti funkcija, priešingu atveju – jei imtis yra didelė, tai tikimybė rasti šią funkciją yra labai didelė. Praktikoje lokalios regresijos metodas dažniausiai naudojamas duomenų rinkiniams generuoti, kurie toliau yra analizuojami kitais metodais. [**12.**]

1.3.3. GLODINANČIŲ SPLAINŲ METODAS

Glodindnčių splainų metodas neparametrinės regresijos modeliams naudoja neatitikties mažiausių kvadratų metodą. Jis apskaičiuoja glodinančius splainus, tam, kad jų pagalba būtų galima suglodinti tiriamas daugiamates funkcijas. Tokiu būdu glodinančių splainų metodas regresijos paviršių padengia labai tiksliai, lanksčiai. Šis metodas modeliui nepriskiria jokių parametrinės formos prielaidų. GCV – apibendrinta kryžminio palyginimo funkcija gali būti naudojama parinkti glodinimo kiekiui.

Glodinančių splainų metodas papildo kitus standartinius regresijos metodus. Šie metodai tinkami daugelyje situacijų, kuriose reikia apibrėžti regresijos modelį su fiksuotu parametrų skaičiumi. Glodinančių splainų metodą galime naudoti duomenų modeliavimui, kai apie modelį nėra jokios pradinės informacijos arba, kai žinome, kad modelio duomenys nėra pateikti su fiksuotu parametrų skaičiumi.

Glodinančių splainų metode naudojamas neatitikties mažiausių kvadratų metodas, tam, kad pritaikyti duomenis modeliui, kuriame efektyvių parametrų skaičius gali būti toks didelis, kaip unikalaus modelio taškų skaičius. Pastebėsime, kad kai turimo modelio dydis didėja, modelio plotas taip pat didėja, leisdamas glodinančius splainus naudoti sudėtingose situacijose.

Pagrindinės glodinančių splainų metodo savybės:

- taikant metodą randami neatitikties mažiausių kvadratų įverčiai;
- metodas tinka naudojant daugiamačius duomenis;
- metodas tinka neparametriniams modeliams;
- metodas suteikia galimybes tvarkyti didelius duomenų rinkinius;
- metodas tinka priklausomiems daugialypiams kintamiesiems;
- metodas leidžia pasirinkti tam tikrą modelį, nurodant modelio laisvės laipsnių skaičių ar glodinimo parametrą.

Neatitikties mažiausių kvadratų įverčiai, nurodo būdą, kaip pateikti duomenis sklandžiai ir išvengti per didelio paviršiaus grublėtumo ar staigios variacijos. Neatitikties mažiausių kvadratų įvertis yra paviršius, kuris minimizuoja neatitikties mažiausius kvadratus visų tipų paviršiams tenkinantiems normalumo sąlygas.

Apibrėžkime X_i, kaip d – dimensijos kovariantinį vektorių, o Z_i, kaip p – dimensijos kovariantinį vektorių, o Y_i, kaip stebinį susijusį su (X_i,Z_i). Darant prielaidą, kad ryšys tarp Z_i ir Y_i yra tiesinis, o ryšys tarp X_i ir Y_i yra nežinomas, galima pritaikyti duomenis naudojant pusiau parametrinį modelį:

$$Y_i = f(X_i) + Z_i\beta + \varepsilon_i, \tag{1.18}$$

kur, f – nežinoma funkcija, kuri yra pakankamai glodi, ε_i, i=1,2,...,n yra nepriklausomos, nulinio vidurkio, atsitiktinės paklaidos, o β – p – dimensijos nežinomas parametrinis vektorius. Šis modelis

susideda iš dviejų dalių. $Z_i\beta$ yra parametrinė modelio dalis, o Z_i – regresiniai kintamieji. $f(X_i)$ – neparametrinė modelio dalis, o X_i yra glodininimo kintamieji. Normalūs mažiausių kvadratų metodo įverčiai $f(X_i)$ ir β:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i) - Z_i\beta)^2 \ . \tag{1.19}$$

Tačiau, funkcijos $f(X)$ plotas yra toks didelis, kad visada galime rasti tokią funkciją f, kuri interpoliuotų duomenų taškus. Norint apskaičiuoti įvertį, gerai atitinkantį duomenis ir turintį, tam tikrą glodumo laipsnį, naudojame neatitikties mažiausių kvadratų metodą. Neatitikties mažiausių kvadratų funkcija apibrėžiama taip:

$$S_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i) - Z_i\beta)^2 + \lambda J_2(f)\ , \tag{1.20}$$

kur, $J_2(f)$ – funkcijos f neatitikties nelygumas. Pirmasis narys tikrina tinkamumą, o antrasis narys tikrina funkcijos f glodumą. λ – glodinimo parametras, kuris lemia santykį tarp glodumo ir tinkamumo. Kai parametras λ yra didelis, jis neatitinka įverčių su didelėmis antrosiomis išvestinėmis. Priešingu atveju, maža parametro λ vertė pabrėžia tinkamumą. Įvertis f_λ gali būti pateikiamas, kaip bazinių funkcijų sekų tiesinė kombinacija. Taigi galutiniai f įverčiai gali būti užrašomi taip:

$$f_\lambda(X_i) = \theta_0 + \sum_{j=1}^{d}\theta_j X_{ij} + \sum_{j=1}^{n}\delta_j B_j(X_i)\ , \tag{1.21}$$

kur, B_j – pagrindinė funkcija, kuri priklauso nuo duomenų X_j išsidėstymo, θ_j ir δ_j yra koeficientai, kurie turi būti įvertinami. Fiksuotam parametrui λ, koeficientai (θ, δ, β) apskaičiuojami išsprendus $n \times n$ dydžio sistemą. Glodinimo parametras pasirenkamas minimizuojant GCV funkciją. GCV funkcija žymima – $V(\lambda)$ ir apibrėžiama taip:

$$V(\lambda) = \frac{(\frac{1}{n})||(I - A(\lambda))y||^2}{[(\frac{1}{n})tr(I - A(\lambda))]^2}\ . \tag{1.22}$$

[**13.**]

1.3.4. BRANDUOLINIS REGRESIJOS METODAS

Statistikoje branduolinis regresijos metodas yra neparametrinis metodas skirtas įvertinti sąlygines atsitiktinio kintamojo galimybes. Pagrindinis uždavinys rasti netiesinį ryšį tarp atsitiktinių dydžių X ir Y. Branduolinio regresijos metodo tikslas yra apskaičiuoti ir panaudoti tinkamus svorius $w_{ij(ker)}$:

$$\hat{Y}_{i(\mathrm{ker})} = \sum_{j=1}^{n} w_{ij(ker)}Y_j. \tag{1.23}$$

Pastebėsime, kad lygtis (1.23) yra $\hat{f}(X_i)$, i=1,2,…,n atitikmuo. Kiekvienam duomenų taškui n priskiriamas individualus svoris $w_{ij(ker)}$, j=1,2,...,n, pritaikant jį kiekvienam taškui X_i (arba pritaikant individualius svorius – $w_{0j(ker)}$ taškui X_0). Matriciniame žymėjime, lygtis (1.23) išreiškiama taip:

$$\underline{\hat{Y}}_{(ker)} = W_{(ker)}\underline{Y}\ , \tag{1.24}$$

čia, $W_{(ker)}=(w_{ij(ker)})$ yra žymimas, kaip branduolio svorių matrica. Ši matrica naudojama Y transformuoti į $Y_{(ker)}$.

Vienas iš labiausiai praktikoje paplitusių metodų nustatant svorius apibrėžiamas taip:

$$w_{ij(ker)} = K\left[\frac{X_i - X_j}{h}\right] / \sum_{j=1}^{n} K[\frac{X_i - X_j}{h}], \tag{1.25}$$

čia, $K(u)$ yra mažėjanti u funkcija, o $h>0$ yra vadinamas glodinimo pločiu. $K(u)$, branduolio funkcija, gali būti laikoma tikimybių tankio funkcija (tokia, kaip standartinė Gauso), apibrėžta funkcija lygi nuliui už u intervalo ribų. Branduolio funkcija turi būti simetrinė. Glodinimo plotis, h, yra glodinimo parametras.

Gilesnis lygties (1.25) skaitiklio tyrimas, leidžia šiek tiek numatyti svorių radimo seką, t.y. daugiau svorių siejami su stebiniais išsidėsčiusiais arčiau X_i (tinkama vieta) ir mažiau stebimi svoriai, kurie yra nutolę. Vardiklio pagalba nustatomos $W_{(ker)}$ eilutės, tam kad suvienodinti sumą.

Apskaičiuota regresijos kreivė užbaigiama gaunant prognozes. Visame duomenų intervale funkcija f dar yra nežinoma. Turint neparametrinės regresijos branduolinę formą, reikia pažymėti, kad neegzistuoja jokia tiksli funkcijos f išraiška.

Galime pasirinkti glodinimo plotį (h), tačiau negalime pasirinkti branduolio funkcijos, kuri turėtų turėti neparametrines savybes. Taigi, supaprastinta, normalioji branduolio funkcija yra naudojama šioje išraiškoje:

$$K(u) = \exp(-u^2). \tag{1.26}$$

Glodinimo plotis h, nustato kaip greit svoriai mažėja, kai atstumas nuo taško X_0 didėja. Intervalas, kuriame svoriai mažėja atitinkamose X_i vietose, savo ruožtu reguliuoja f rezultatų įverčių glodumą.

Tarkime, kad h yra mažas (arti nulio). Prognozės taškas pats turi didžiąją svorio dalį ir tik su artimiausiais stebėjimais šiame taške gaunama likusi svorio dalis (prisiminkime, kad svoriai suvienodina sumą). Pagal tokią seką, rezultatas būtų sujungti taškai, suformuoti stebint duomenų taškus, o jie turi būti suglodinti ir pasižymėti didele variacija. Kitaip tariant, šiame procese vietoj to, kad gauti vieną vienintelį pagrindinį sprendinį, imami skirtingi mėginiai, kad dėl mėginių kintamumo būtų gauta daug skirtingų sprendinių.

Dabar tarkime, kad h yra didelis (lygus arba beveik lygus visam X reikšmių intervalui). Užuot sutelkus svorius ant vieno taško ar kelių taškų, svoris yra tolygiai paskirstomas visiems stebiniams. Tokia priklausomybė yra laikoma suglodinta (su didele paklaida), nes ji tikrai atitinka kiekvieno duomenų taško $\overline{Y}$ vertę.

Yra viena problema būdinga branduoliniui regresijos metodui. Apsvarstykime, kas atsitinka, kai X artėja prie duomenų intervalo krašto (dešnio arba kairio). Tada branduolio svoriai jau gali nebūti simetriniai. Kad būtų aiškiau, tarkime turime dešnįjį duomenų intervalo kraštą. Tiksliau, tarkim prognozuojamo kintamojo $\hat{Y}_0$ prie X_0 gavimo procesas, kur X_0 yra ant arba šalia šio dešniojo

intervalo krašto. Tik taškai, kurie yra taško X_0 kairėje naudojami gauti branduoliniams svoriams (išskyrus patį X_0). Tiesiog nėra taškų X_0 dešnėje, kad būtų galima apskaičiuoti svorius.

Jei duomenys (ir tikroji funkcija *f*) mažėja einant link dešniojo intervalo krašto, tada visos Y vertės svertinės sumos, kurios naudojamos $\hat{Y}_0$ reikšmei gauti, labiausiai tikėtina, yra didesnės arba lygios Y_0 reikšmei prie X_0. Intervalo krašto paklaidos lems prognozę, $\hat{Y}_0$, kuri bus per didelė. [**10.**]

1.4. STATISTINĖS PROGRAMINĖS ĮRANGOS APŽVALGA IR LYGINAMOJI ANALIZĖ

Šiuolaikinė statistika yra neatsiejama nuo kompiuterinės duomenų analizės, padedančios greitai ir efektyviai spręsti įvairius statistikos uždavinius. Taikomosios statistikos metodai plačiai taikomi priimant svarbius sprendimus įvairiose srityse.

Egzistuojančią duomenų analizės programinę įrangą galima suskirstyti į kelias grupes:

- duomenų analizės uždavinių programų bibliotekos universaliose programavimo kalbose (Pascal, C ir kt.);
- matematinių uždavinių sprendimo universalios sistemos (MathCad, Maple, MatLab, Mathematica ir kt.);
- duomenų analizės universalios sistemos (SAS, SPSS, Statistica ir kt.);
- ekspertinės duomenų analizės sistemos, skirtos konkrečiai analizei (TABLE CURVE – vieno kintamojo regresinė analizė, ABP – laiko eilučių analizė ir kt.);
- kitos paskirties sistemos (Excel ir kt.).

Ne visos duomenų analizei taikomos programinės įrangos yra pakankamai efektyvios ir patikimos, ypač kai duomenų yra labai daug, pavyzdžiui MS Excel, SPSS. SPSS leidžia analizuoti duomenis ir vaizduoti analizės rezultatus. Pagrindinis SPSS programinio paketo privalumas – didelė šiuolaikinių statistinių analizės metodų pasirinktis ir duomenų analizės rezultatų vizualizavimo priemonių (duomenų pateikimo lentelių, diagramų, skirstinių kreivių) įvairovė, lengvai įvaldoma dialoginė sąsaja. SPSS programinis paketas dažniausiai taikomas sociologijos, psichologijos, biologijos, medicinos, rinkodaros, kokybės valdymo procese.

Šiame darbe pateikiama statistinės analizės vartotojo sąsaja realizuota panaudojus programų paketo SAS instrumentus.

Programinės įrangos kūrimui pasirinkta sistema SAS dėl to, kad tai yra viena iš populiariausių pasaulyje universalių duomenų analizės sistemų, galinti atlikti įvairias funkcijas ir netgi turinti labai geras taikomųjų programų kūrimo priemones. Kitose populiariose sistemose taikomųjų programų kūrimo priemonių arba nėra, arba jos yra labai elementarios. Be to, kitos sistemos atlieka mažiau funkcijų bei yra prastesnė vartotojo sąsaja.

SAS sistema buvo sukurta 1960 m. Eksperimentinės Statistikos skyriuje valstybiniame Šiaurės Karolinos universitete. 1976 m. buvo įkurtas SAS Institutas, nuo to laiko sistemos SAS programinę įrangą gamina JAV kompanija SAS Institute INC. Iš pradžių SAS sistema buvo sukurta ūkinės informacijos apdorojimui IBM skaičiavimo mašinų pagalba. Po to SAS pradėjo sparčiai vystytis, ir laikui bėgant tapo viena iš galingiausių pasaulyje sistemų statistinės duomenų analizės srityje. 1980 m. SAS sistema buvo integruota į skaičiuojamąsias mašinas, mini kompiuterius ir personalinius kompiuterius, kas pritraukė daugiau vartotojų. Šiuo metų SAS sistema turi apie 3.5 milijonų vartotojų ir yra naudojama virš 120 šalių.

Šiuo metu SAS sistema yra viena iš galingiausių duomenų analizės sistemų, ji naudojama dideliam duomenų kiekiui apdoroti. SAS sistema turi privalumų prieš kitus paketus. SAS kalboje yra įmanomas duomenų importas ir eksportas iš tokių paketų kaip: *Oracle, DB2, Lotus, Sybase, Access* ir kitų, ko negalima būtų pasakyti apie kai kuriuos kitus paketus. SAS kalba turi gerai išvystytą makro kalbą, matricų kalbą, kurių net neturi kai kurie kiti paketai.

SAS sistema gali būti valdoma komandų ir meniu pagalba. SAS programavimo kalba turi daug įvairių komandų, funkcijų, operatorių ir kitokių programavimo priemonių, kurios užtikrina norimo rezultato pasiekimą. Sistemos SAS programavimo kalba nėra sudėtinga, jos sintaksė yra labai panaši į kitų procedūrinių kalbų sintaksę.

1.1 lentelė.

Paketų palyginimas

Galimybės	Mathematica	MiniTab	SAS	S-Plus	SPSS	Statistica
MANIPULIAVIMAS DUOMENIMIS:						
Jungimas	Taip	Taip	**Taip**	Taip	Taip	Taip
Skaidymas	Taip	Taip	**Taip**	Taip	Taip	Taip
Pjūvių formavimai	Taip	Taip	**Taip**	Taip	Taip	Taip
Įterpimas į duomenų failus	Taip	Taip	**Taip**	Taip	Taip	Taip
Išsaugojimas HTML formatu	Taip	Ne	**Taip**	Ne	Taip	Ne
Metodinis pastabų pridėjimas	Taip	Taip	**Taip**	Taip	Taip	Ne
Disponavimas vartotojo suprogramuotomis procedūromis	Taip	Taip	**Taip**	Taip	Taip	Taip
Makro kalba	Ne	Taip	**Taip**	Taip	Taip	Taip
Matricų kalba	Taip	Ne	**Taip**	Taip	Taip	Taip
Trumpas išvedimas	Taip	Taip	**Taip**	Taip	Taip	Taip
Ataskaitų pateikimas	Taip	Ne	**Taip**	Ne	Taip	Taip
Ataskaitų formavimas	Taip	Ne	**Taip**	Ne	Taip	Taip
Apribojimai duomenų įrašų skaičiui	Ne	Taip	**Ne**	Ne	Taip	Ne
Apribojimai kintamųjų skaičiui	Ne	Ne	**Ne**	Ne	Taip	Taip

1.5. SPRENDŽIAMI UŽDAVINIAI

Magistro baigiamojo darbo užduotis: Monte Karlo metodu palyginti populiarius bei praktikoje dažnai sutinkamus neparametrinius regresinius metodus. Atlikti Lietuvos ir užsienio literatūros, skirtos tiriamos problemos nagrinėjimui, apžvalgą. Apžvelgus literatūros šaltinius, susisteminti informaciją, išanalizuoti tiriamos problemos sprendimui taikomų matematinių metodų ir programinių priemonių privalumus ir trūkumus sprendžiant nagrinėjamą uždavinį. Tuomet, suformuluoti reikalavimus pasirinktiems modeliams ir kuriamai programinei įrangai bei atliekamiems tyrimams. Sukurti programinę įrangą, kuri palengvintų sprendimų priėmimą ir pagrindimą atliekat regresinių metodų lyginamąjį tyrimą. Gautus tyrimo rezultatus apibendrinti, magistro baigiamojo darbo ataskaitoje juos pateikti lentelių bei grafikų pavidalu, taip pat pateikti išvadas bei rekomendacijas.

Sumodeliuotų duomenų tiesioginei analizei reikalinga SAS arba SPSS (angl. – Statistical Package for the Social Science) programos, kurios galėtų iš turimų duomenų išgauti naudingą informaciją, padedančią priimti sprendimus. Programai realizuoti pasirinkta SAS programinė įranga, kadangi šioje sistemoje plačiai išvystytos statistinės duomenų analizės galimybės ir ji turi nesudėtingą programavimo kalbą, taip pat galima kurti vartotojui patogią interaktyvią sąsają, pateikti duomenis grafikų, diagramų, lentelių pavidalu. SAS programavimo kalba yra paprasta ir lengvai įsimenama – daug gerų pavyzdžių ir patarimų galima rasti SAS pagalboje arba internete.

Atliekant regresinių metodų lyginamąjį tyrimą priklausomumas tarp kintamųjų yra išreiškiamas matematine lygtimi, kuri vadinama regresijos lygtimi arba regresiniu modeliu. Šiame darbe bus analizuojami kitų mokslininkų tirti neparametrinės regresijos modeliai

Magistro darbe bus atliekamas keturių regresinių metodų lyginamasis tyrimas. Pasirinkti metodai:

- apibendrintų adityvių modelių metodas;
- lokalios regresijos metodas;
- glodinančių splainų metodas;
- branduolinis regresijos metodas.

Pagrindiniai regresinių metodų palyginimo kriterijai: apibrėžtumo koeficientas, vidutinės kvadratinės paklaidos šaknis ir vidutinė procentinė absoliutinė paklaida. Šių statistinių metrikų pagalba išrinksime geriausią regresijos metodą.

2.TIRIAMOJI DALIS

2.1. TYRIMO SCHEMA

Ankstesnėje dalyje – 1.4 skyriuje – aprašytų regresinių metodų lyginamasis tyrimas atliktas Monte Karlo metodu. Toks palyginimo būdas sudarė galimybes apskaičiuoti tikrąsias priklausomo kintamojo Y_i reikšmes, bei šių reikšmių įverčius. Tyrimą sudaro keturi etapai:

1) Generuojamos nepriklausomų kintamųjų vektoriaus X_i reikšmės.
2) Apskaičiuojamos priklausomo kintamojo Y_i tikrosios reikšmės taškuose X_i.
3) Apskaičiuojami priklausomo kintamojo Y_i reikšmių įverčiai $\hat{Y}_i$ taškuose X_i.
4) Pirmame ir antrame etape gautos reikšmės panaudojamos paklaidų apskaičiavimui. Rezultatai naudojami apskaičiuoti apibrėžtumo koeficientą (žr. (2.5) formulę), vidutinės kvadratinės paklaidos šaknį (žr. (2.6) formulę) ir vidutinę procentinę absoliutinę paklaidą (žr. (2.7) formulę) (ši apskaičiavimo procedūra atliekama įvairiems imčių tūriams).

2.1.1. DUOMENŲ MODELIAVIMO SKIRSTINIAI

Lyginamasis apibendrintų adityvių modelių, lokalios regresijos, glodinančių splainų, branduolinių regresijos metodų tyrimas buvo atliktas pasirinkus kitų autorių darbuose nagrinėtus regresinius modelius [**15.**], [**16.**] ir [**17.**]. Duomenims modeliuoti naudojami normalusis ir tolygusis skirstiniai. Duomenų tankio funkcijos apibrėžiamos taip:

Normalusis skirstinys.

Šio skirstinio pasiskirstymo tankis:

$$w(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-m)^2}{2\sigma^2}], \tag{2.1}$$

čia m atitinka vidurkį, o σ^2 – dispersiją. Normaliojo atsitiktinio dydžio tikimybių pasiskirstymo funkcija išreiškiama taip:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} \exp[-\frac{(u-m)^2}{2\sigma^2}]\, du. \tag{2.2}$$

Tolygusis skirstinys.

Šio skirstinio pasiskirstymo tankis:

$$w(x) = f(x) = \begin{cases} \frac{1}{b-a}, & kai\ a \le x \le b; \\ 0\ , & kai\ x < a\ ir\ x > b. \end{cases} \tag{2.3}$$

Tolygaus atsitiktinio dydžio tikimybių pasiskirstymo funkcija išreiškiama taip:

$$F(x) = \begin{cases} 0, & kai\ x \le a; \\ \frac{x-a}{b-a}, & kai\ a < x < b; \\ 1, & kai\ x \ge b. \end{cases} \tag{2.4}$$

2.1.2. MODELIO TINKAMUMO MATAI

Regresinio metodo tikslumą nusako tokie matai:

- *Apibrėžtumo koeficientas (R^2).* Tai svarbiausia metodo tikimo duomenims charakteristika, kuri privaloma visuose regresijos modelių aprašymuose. Apibrėžtumo koeficientas lygina skirtumus tarp *Y* reikšmių, kai atsižvelgiama į regresijos modelį, su skirtumais tarp *Y* reikšmių, kai į modelį neatsižvelgiama. Labai apytikslė R^2 interpretacija, padedanti geriau suvokti jo prasmę, yra tokia – kiek procentų *Y* elgesio paaiškina kintamųjų *X, Z, W* elgesys. Apibrėžtumo koeficientas įgyja reikšmes iš intervalo [0, 1]. Kuo koeficiento reikšmė didesnė, tuo metodas geriau tinka duomenims. Blogai, kai $R^2 < 0{,}20$. Apskritai nėra ko labai džiaugtis ir tada, kai $R^2 = 0{,}25$ (modelis tinka tik iš dalies). O štai, jei $R^2 = 0{,}89$, tai modelis labai gerai aprašo duomenis. Tiesa, tai dar neužtikrina, kad visi kintamieji jame būtini, o pats modelis yra prasmingas.

$$R^2 = 1 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 / \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2. \tag{2.5}$$

- Vidutinės kvadratinės paklaidos šaknis (*RMSE*):

$$RMSE = \sqrt{MSE}.$$

$$MSE = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{2.6}$$

čia, *n* – stebinių skaičius, *k* – metodo parametrų skaičius. Vidutinės kvadratinės paklaidos dydis nusako paklaidos dispersiją, ir, ja remiantis, parenkami optimalūs prognozavimo metodo parametrai. Pasirenkamas metodas kurio vidutinės kvadratinės paklaidos šaknies reikšmė yra mažiausia, tokiu atveju metodas yra tiksliausias.

- Vidutinė procentinė absoliutinė paklaida (*MAPE*):

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|/Y_i. \tag{2.7}$$

Reikia atkreipti dėmesį, kad šis dydis neskaičiuojamas, kai $Y_i = 0$. Vidutinė procentinė absoliutinė paklaida nusako santykinį prognozavimo tikslumą ir remiantis juo galima palyginti skirtingas prognozes.

2.1 lentelė

Prognozavimo tikslumo nustatymas

MAPE %	Prognozavimo tikslumas
<10	Labai tikslus
10÷20	Tikslus
20÷50	Pakankamas
>50	Nepakankamas

2.2. MODELIAVIMO PLANAS

Priklausomumas tarp kintamųjų išreiškiamas matematine lygtimi, kuri vadinama regresijos lygtimi arba regresiniu modeliu. Tyrimui atlikti buvo pasirinkti šeši, kitų mokslininkų nagrinėti, regresiniai modeliai. Taip pat, jie pasirinkti siekiant palyginti neparametrinės regresijos metodus naudojant skirtingų priklausomybių modelius, t.y. eksponentinius, trigonometrinius, multiplikatyvinius.

Pirmas regresinis modelis:

$$Y_i = 1 - X_i + e^{-200(X_i - 1/2)^2} + \varepsilon_i, X_i \sim U(0,1), \varepsilon_i \sim N(0,1), \quad (2.8)$$

čia, Y_i – priklausomasis (atsako) kintamasis, X_i – nepriklausomasis (aiškinamasis) kintamasis, ε_i – atsitiktinė paklaida. Šiame modelyje nepriklausomo kintamojo X_i duomenys modeliuojami Monte Karlo metodu pasiskirstę pagal tolygųjį skirstinį intervale [0;1], o atsitiktinės paklaidos ε_i duomenys – pagal normalųjį.

Antras regresinis modelis:

$$Y_i = 2\sin(2\pi X_i) + \varepsilon_i, X_i \sim U(0,1), \varepsilon_i \sim N(0,1), \quad (2.9)$$

čia, Y_i – priklausomasis (atsako) kintamasis, X_i – nepriklausomasis (aiškinamasis) kintamasis, ε_i – atsitiktinė paklaida. Šiame modelyje nepriklausomo kintamojo X_i ir atsitiktinės paklaidos ε_i duomenys modeliuojami Monte Karlo metodu pasiskirstę pagal normalųjį skirstinį.

Trečias regresinis modelis:

$$Y_i = 1 + 0{,}25(X_i)^2 + 0{,}1(X_i)^3 + \varepsilon_i, X_i \sim N(0,1), \varepsilon_i \sim N(0,1), \quad (2.10)$$

čia, Y_i – priklausomasis (atsako) kintamasis, X_i – nepriklausomasis (aiškinamasis) kintamasis, ε_i – atsitiktinė paklaida. Šiame modelyje nepriklausomo kintamojo X_i ir atsitiktinės paklaidos ε_i duomenys modeliuojami Monte Karlo metodu pasiskirstę pagal normalųjį skirstinį.

Ketvirtas regresinis modelis:

$$Y_i = X_i + \sqrt{0{,}1 + 0{,}4X_i^2}\varepsilon_i, X_i \sim N(0,1), \varepsilon_i \sim N(0,1), \quad (2.11)$$

čia, Y_i – priklausomasis (atsako) kintamasis, X_i – nepriklausomasis (aiškinamasis) kintamasis, ε_i – atsitiktinė paklaida. Šiame modelyje nepriklausomo kintamojo X_i ir atsitiktinės paklaidos ε_i duomenys modeliuojami Monte Karlo metodu pasiskirstę pagal normalųjį skirstinį.

Penktas regresinis modelis:

$$Y_i = -\sin\big(2\pi(X_i - 0{,}5)\big)\cos\big(2\pi X_j\big) + \varepsilon_i, X_i \sim N(0,1), X_j \sim N(0,1), \varepsilon_i \sim N(0,1),$$

čia, Y_i – priklausomasis (atsako) kintamasis, X_i – nepriklausomasis (aiškinamasis) kintamasis, X_j – nepriklausomasis (aiškinamasis) kintamasis, ε_i – atsitiktinė paklaida. Šiame modelyje nepriklausomų kintamųjų X_i ir X_j ir atsitiktinės paklaidos ε_i duomenys modeliuojami Monte Karlo metodu pasiskirstę pagal normalųjį skirstinį.

Šeštas regresinis modelis:

$$Y_i = sin\sqrt{{X_i}^2 + {X_j}^2} + \varepsilon_i, X_i \sim N(0,1), X_j \sim N(0,1), \varepsilon_i \sim N(0{,}0{,}05),$$

čia, Y_i – priklausomasis (atsako) kintamasis, X_i – nepriklausomasis (aiškinamasis) kintamasis, X_j – nepriklausomasis (aiškinamasis) kintamasis, ε_i – atsitiktinė paklaida. Šiame modelyje nepriklausomų kintamųjų X_i ir X_j ir atsitiktinės paklaidos ε_i duomenys modeliuojami Monte Karlo metodu pasiskirstę pagal normalųjį skirstinį.

Norint atlikti regresinių metodų lyginamąjį tyrimą, šiems modeliams įvertinti naudojami keturi skirtingi regresinės analizės metodai:

- apibendrintų adityvių modelių metodas (AAMM);
- lokalios regresijos metodas (LRM);
- glodinančių splainų metodas (GSM);
- branduolinis regresijos metodas (BRM).

Taikant metodą, atskirai kiekvienam modeliui yra keičiamas imties tūris *n*:
n = 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. Kiekvienu atveju generuota po 100 imčių.

Regresinės analizės metodų efektyvumui nustatyti naudoti šie jų vertinimo tikslumą nusakantys kriterijai:

- apibrėžtumo koeficientas (R^2);
- vidutinės kvadratinės paklaidos šaknis (RMSE);
- vidutinė procentinė absoliutinė paklaida (MAPE).

Minėti tikslumo vertinimo kriterijai, kiekvienam metodui, su kiekvienu imties dydžiu pateikiami grafikuose ir lentelėse (žr. 2. PRIEDAS).

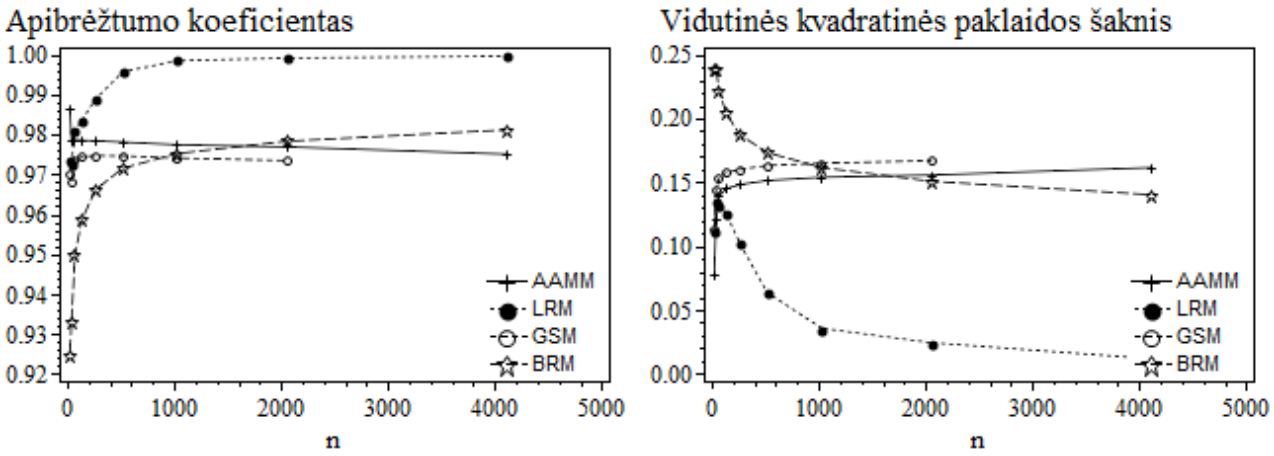
2.3. MODELIAVIMO REZULTATAI

2.3.1. PIRMO MODELIO VERTINIMO REZULTATAI

2.1 paveiksle atvaizduojami pirmo regresinio modelio apibrėžtumo koeficiento ir vidutinės kvadratinės paklaidų šaknies vidutinės reikšmės priklausomumas nuo imties tūrio.

Apibrėžtumo koeficientai pateikiami visiems naudotiems metodams. Visų metodų rezultatai pateikiami skirtingo tūrio imtims. Jei apibrėžtumo koeficiento vertė lygi vienetui, tai sakome, kad metodas yra labai tikslus, priešingu atveju, jei vertė lygi nuliui, tai rodo, kad metodas negali tiksliai aprašyti duomenų. Paveiksle matome, kad didinant imties tūrį tiksliausias modelis gaunamas naudojant lokalios regresijos metodą (žym. LRM). Didinant imties tūrį, kai taikomas branduolinis regresijos metodas (žym. BRM) modelio tikslumas sparčiai didėja, tačiau tokio tikslumo, kaip taikant lokalios regresijos metodą nepasiekia. Naudojant apibendrintų adityvių modelių metodą (žym. AAMM) ir glodinančių spainų metodą (žym. GSM) ir didinant imties tūrį apibrėžtumo koeficiento vertė nežymiai mažėja, modelio tikslumas mažėja.

Vidutinių kvadratinių paklaidų šaknys pateikiamos visiems panaudotiems metodams. Visų metodų rezultatai pateikiami įvairaus tūrio imtims. Pasirenkamas modelis kurio vidutinės kvadratinės paklaidos šaknies reikšmė yra mažiausia, tokiu atveju modelis yra patikimiausias. Paveiksle matome, kad taikant lokalios regresijos metodą (žym. LRM) ir didinant imties tūrį gauname mažėjančias paklaidas, šio modelio atveju kai imtis yra didžiausia gauname mažiausią paklaidą. Branduolinio regresijos metodo (žym. BRM) atveju didinant imties tūrį paklaidos mažėja, tačiau negaunamos tokios mažos, kaip taikant ankščiau minėtą metodą. Taikant glodinančių splainų (žym. GSM) ir branduolinės regresijos (žym. BRM) metodus ir didinant imtį paklaidos didėja.



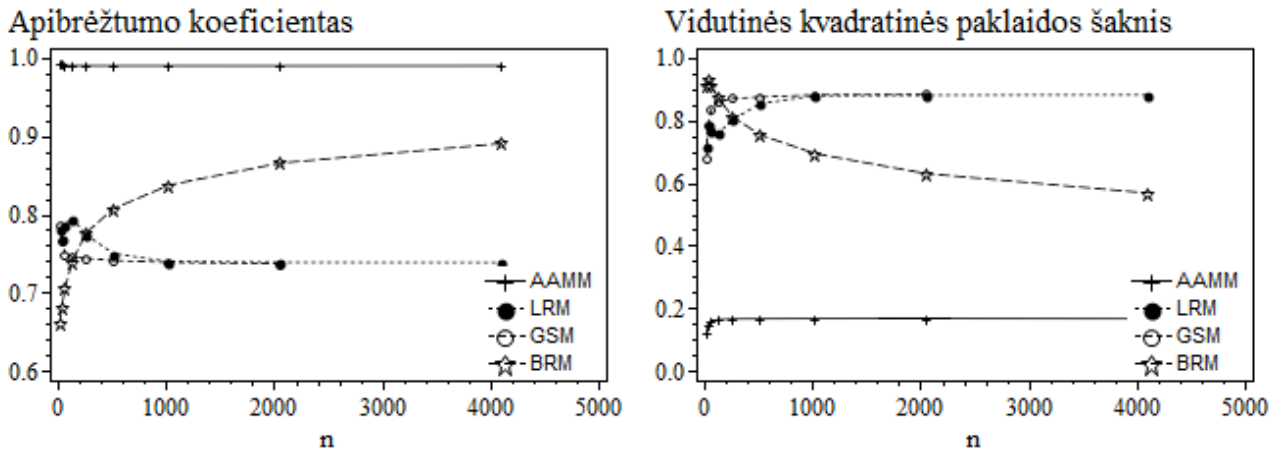
2.1 pav. Pirmo regresijos modelio vertinimo rezultatai

2.3.2. ANTRO MODELIO VERTINIMO REZULTATAI

2.2 paveiksle atvaizduojami antro regresinio modelio apibrėžtumo koeficiento ir vidutinės kvadratinės paklaidų šaknies vidutinės reikšmės priklausomumas nuo imties tūrio.

Apibrėžtumo koeficientai pateikiami visiems naudotiems metodams. Visų metodų rezultatai pateikiami skirtingo tūrio imtims. Paveiksle matome, kad modelis yra patikimiausias naudojant apibendrintų adityvių modelių metodą (žym. AAMM). Branduolinės regresijos metodo (žym. BRM) atveju didinant imtį pastebėsime, kad modelio tikslumas didėja, tačiau tokio patikimumo, kaip apibendritų adityvių modelių metodo atveju, nepasiekia. Praščiausi rezultatai gaunami taikant lokalios regresijos (žym. LRM) ir glodinančių splainų (žym. GSM) metodus.

Vidutinių kvadratinių paklaidų šaknys pateikiamos visiems panaudotiems metodams. Visų metodų rezultatai pateikiami įvairaus tūrio imtims. Paveiksle matome, kad modelis su mažiausiomis paklaidomis gaunamas, kai naudojamas apibendrintų adityvių modelių metodas (žym. AAMM). Taikant branduolinės regresijos metodą (žym. BRM) ir didinant imties tūrį paklaidos mažėja, tačiau paklaidos vis tiek gaunamos didesnės, nei apibendrintų adityvių modelių (žym. AAMM) atveju.



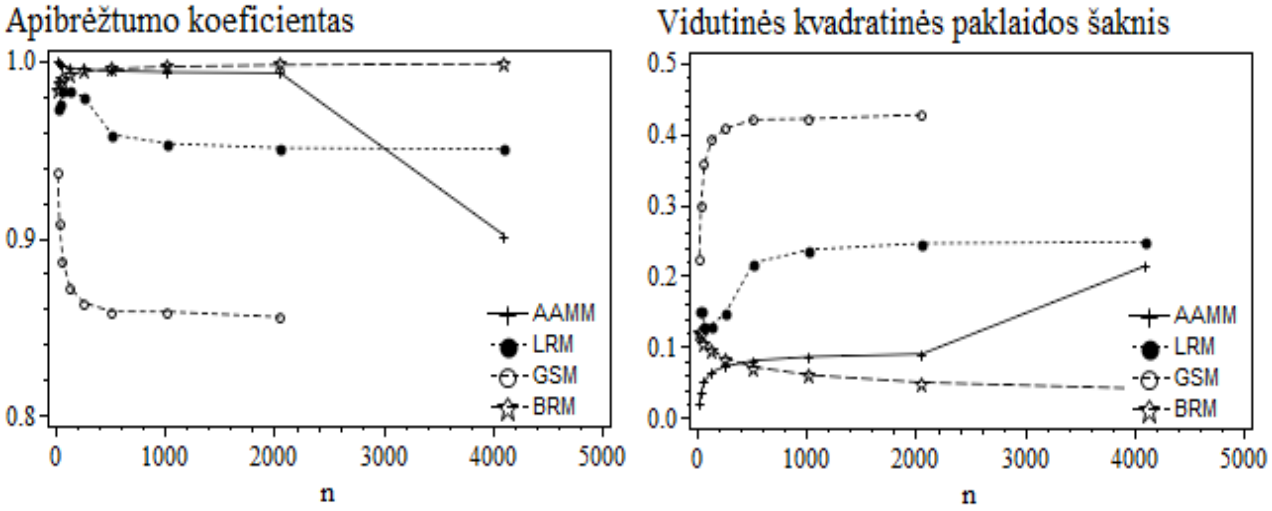
2.2 pav. Antro regresijos modelio vertinimo rezultatai

2.3.3. TREČIO MODELIO VERTINIMO REZULTATAI

2.3 paveiksle atvaizduojami antro regresinio modelio apibrėžtumo koeficiento ir vidutinės kvadratinės paklaidų šaknies vidutinės reikšmės priklausomumas nuo imties tūrio.

Apibrėžtumo koeficientai pateikiami visiems naudotiems metodams. Visų metodų rezultatai pateikiami skirtingo tūrio imtims. Paveiksle matome, kad pradžioje patikimiausias modelis gaunamas taikant apibendrintų adityvių modelių metodą (žym. AAMM), bet didinant imties tūrį apibrežtumo koeficiento vertė tolsta nuo 1, o branduolinės regresijos metodas (žym. BRM), didinant imties tūrį susilygina su apibendrintų adityvvių modelių metodu (žym. AAMM) ir kai n=1024 jį aplenkia.

Vidutinių kvadratinių paklaidų šaknys pateikiamos visiems panaudotiems metodams. Visų metodų rezultatai pateikiami įvairaus tūrio imtims. Paveiksle matome, kad mažiausios paklaidos gaunamos taikant apibendrintų adityvių modelių metodą (žym. AAMM), tačiau tik tada kai imties dydis yra mažas, didinant imties tūrį paklaidos smarkiai padidėja. Naudojant branduolinį regresijos metodą (žym. BRM), didinant imties tūrį paklaidos mažėja ir gaunamas patikimiausias modelis. Lokalios regresijos metodo atveju paklaidos su maža imtimi mažėja, tačiau vėliau, didinat imtį, pradeda didėti. Šiam modeliui pats nepatikimiausias metodas yra glodinančių splainų (žym. GSM).



2.3 pav. Trečio regresijos modelio vertinimo rezultatai

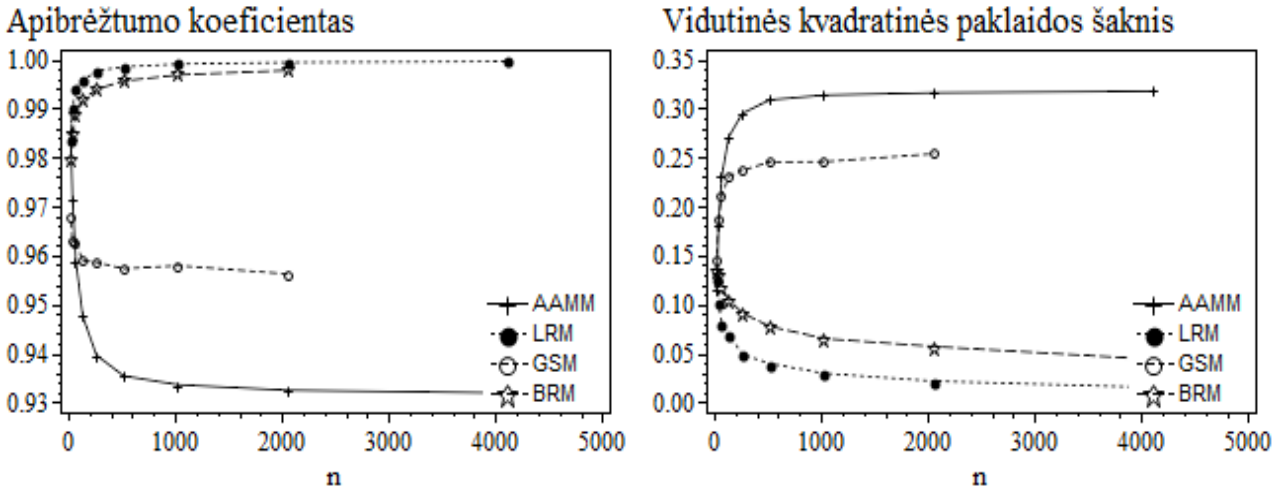
2.3.4. KETVIRTO MODELIO VERTINIMO REZULTATAI

2.4 paveiksle atvaizduojami trečio regresinio modelio apibrėžtumo koeficiento ir vidutinės kvadratinės paklaidų šaknies vidutinės reikšmės priklausomumas nuo imties tūrio.

Apibrėžtumo koeficientai pateikiami visiems naudotiems metodams. Visų metodų rezultatai pateikiami skirtingo tūrio imtims. Paveiksle matome, kad patikimiausias modelis gaunamas, taikant lokalios regresijos metodą (žym. LRM) . Didinant imtį, modelio patikimumas didėja, apibrėžtumo koeficiento reikšmė artėja prie vieneto. Taip pat didinant imtį, modelio patikimumas didėja naudojant ir branduolinį regresijos metodą (žym. BRM), tačiau modelio tikslumas gaunamas prastesnis, nei lokalios regresijos metodo atveju. Priešingai, nei su ankščiau minėtais metodais, didinant imties tūrį, kai taikomi glodinančių splainų metodas (žym. GSM) ir apibendrintų adityvių modelių metodas, modelio tinkamumas mažėja.

Vidutinių kvadratinių paklaidų šaknys pateikiamos visiems panaudotiems metodams. Visų metodų rezultatai pateikiami įvairaus tūrio imtims. Paveiksle matome, kad paklaidos yra mažiausios, kai naudojamas lokalios regresijos metodas (žym. LRM). Didinant imtį, paklaidos mažėja naudojant ir branduolinį regresijos metodą (žym. BRM), visgi jos gaunamos didesnės, nei

lokalios regresijos metodo atveju. Su kitais, t.y. apibendrintų adityvių modelių metodu (žym. AAMM) ir glodinančių splainų metodu (žym. GSM), didinant imties tūrį paklaidos didėja.



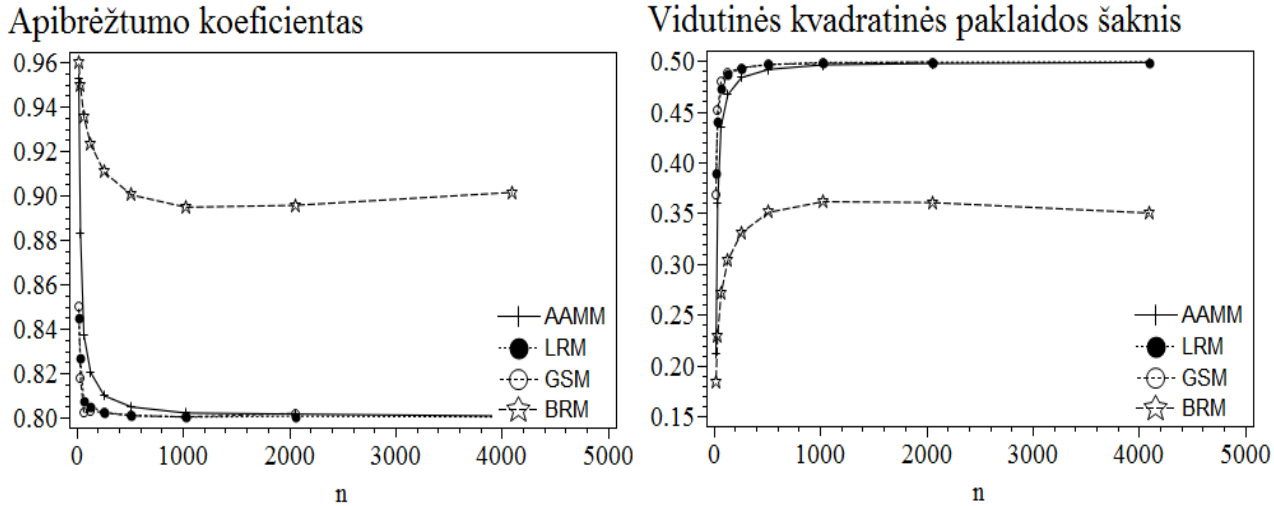
2.4 pav. Ketvirto regresijos modelio vertinimo rezultatai

2.3.5. PENKTO MODELIO VERTINIMO REZULTATAI

2.5 paveiksle atvaizduojami trečio regresinio modelio apibrėžtumo koeficiento ir vidutinės kvadratinės paklaidų šaknies vidutinės reikšmės priklausomumas nuo imties tūrio.

Apibrėžtumo koeficientai pateikiami visiems naudotiems metodams. Visų metodų rezultatai pateikiami skirtingo tūrio imtims. Paveiksle matome, kad taikant apibendrintų adityvių modelių metodą (žym. AAMM), glodinančių splainų metodą (žym. GSM) ir lokalios regresijos metodą (žym. LRM) gautos apibrėžtumo koeficientų vertės yra panašios, tačiau patikimiausias modelis gaunamas, taikant branduolinį regresijos metodą (žym. BRM).

Vidutinių kvadratinių paklaidų šaknys pateikiamos visiems panaudotiems metodams. Visų metodų rezultatai pateikiami įvairaus tūrio imtims. Paveiksle matome, kad paklaidos yra mažiausios, kai naudojamas branduolinės regresijos metodas (žym. BRM).



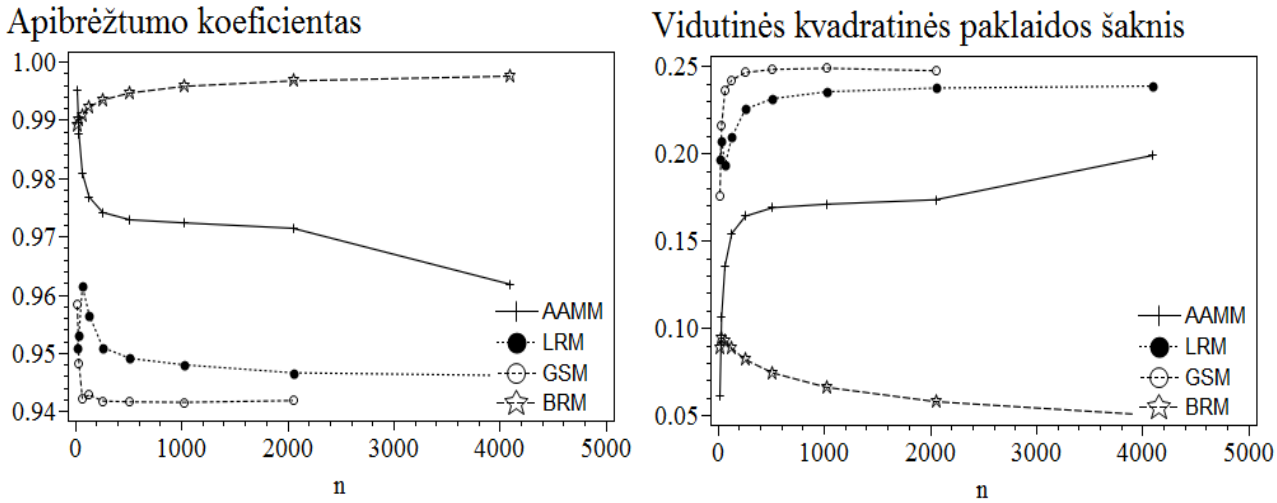
2.5 pav. Penkto regresijos modelio vertinimo rezultatai

2.3.6. ŠEŠTO MODELIO VERTINIMO REZULTATAI

2.6 paveiksle atvaizduojami trečio regresinio modelio apibrėžtumo koeficiento ir vidutinės kvadratinės paklaidų šaknies vidutinės reikšmės priklausomumas nuo imties tūrio.

Apibrėžtumo koeficientai pateikiami visiems naudotiems metodams. Visų metodų rezultatai pateikiami skirtingo tūrio imtims. Paveiksle matome, kad kai imties tūris yra 16 patikimiausias modelis gaunamas taikant apibendrintų adityvių modelių metodą (žym. AAMM), tačiau didinant imties tūrį, pastebėsime, kad modelio patikimumas mažėja. O patikimiausias modelis gaunamas naudojant branduolinį regresijos metodą (žym. BRM).

Vidutinių kvadratinių paklaidų šaknys pateikiamos visiems panaudotiems metodams. Visų metodų rezultatai pateikiami įvairaus tūrio imtims. Paveiksle matome, kad su mažiausiu imties tūriu (16), mažiausia paklaida gaunama taikant apibendrintų adityvių modelių metodą (žym. AAMM), tačiau didinant imties tūrį mažiausios paklaidos gaunamos taikant branduolinį regresijos metodą (žym. BRM).



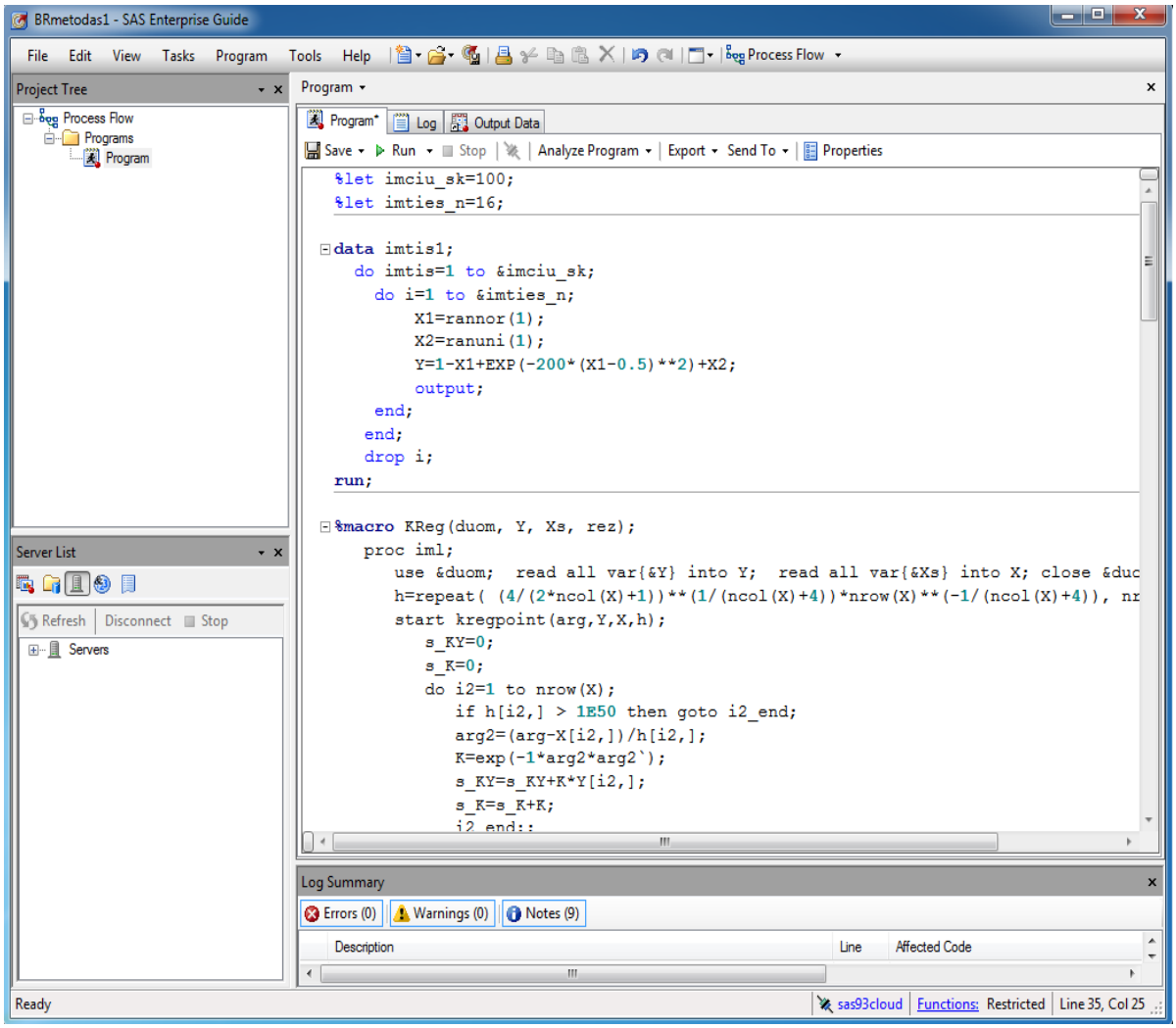
2.6 pav. Šešto regresijos modelio vertinimo rezultatai

2.4 PROGRAMINĖ REALIZACIJA IR INSTRUKCIJA VARTOTOJUI

Programinės įrangos kūrimui pasirinkta sistema SAS dėl to, kad tai yra viena iš universalių duomenų analizės sistemų, galinti atlikti įvairias funkcijas ir turinti labai geras taikomųjų programų kūrimo priemones. SAS sistema gali būti valdoma komandų ir meniu pagalba, jos programavimo kalba nėra sudėtinga, o sintaksė yra labai panaši į kitų procedūrinių kalbų sintaksę.

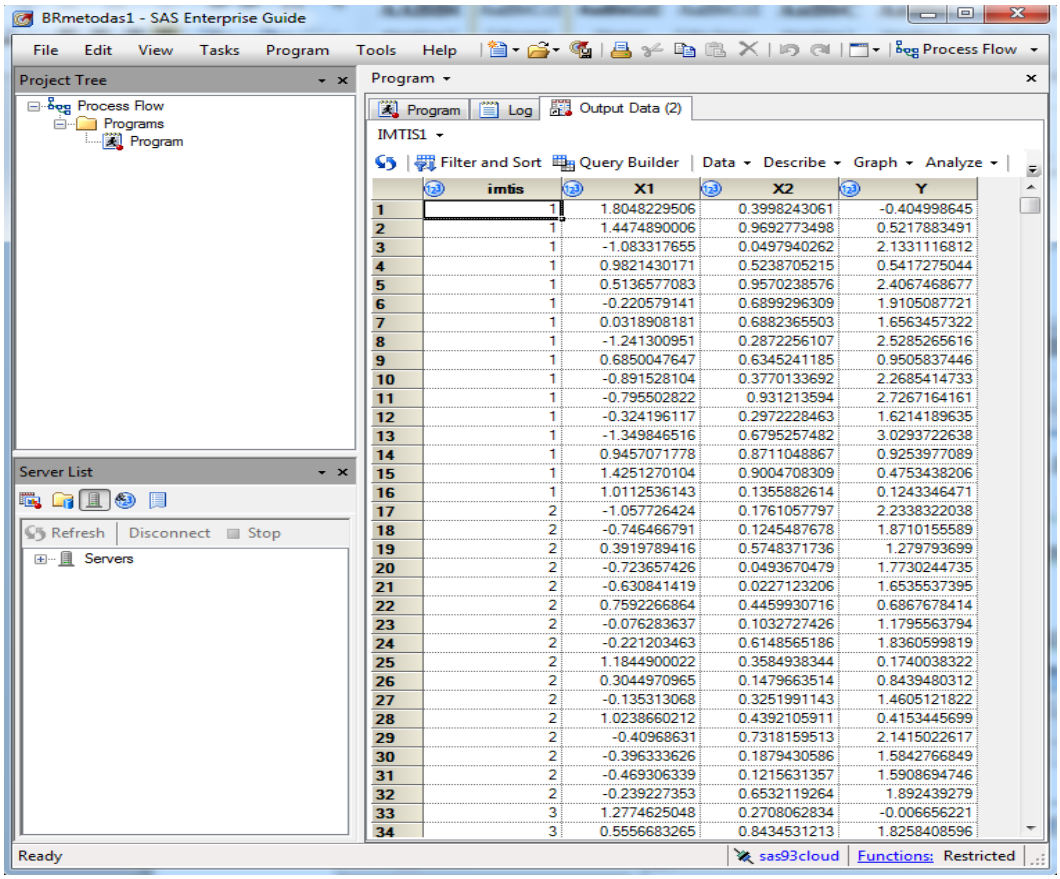
Sukurta programa, kurios instrukcija aprašoma šiame skyriuje.

2.7 paveiksle pavaizduota branduolinio regresijos metodo realizacija pirmąjam regresijos modeliui, pirmoje programos kodo eilutėje esantis skaičius žymi generuotų imčių skaičių (atliktame tyrime kiekvienu atveju buvo generuota po 100 imčių). Antroje eilutėje apibrėžiamas imties tūris – *n* (*n* = 16, 32, 64, 128, 256, 512, 1024, 2048, 4096.). Toliau programoje aprašomi nepriklausomo kintamojo ir atsitiktinės paklaidos duomenų pasiskirstymai ir regresijos modelio funkcija.



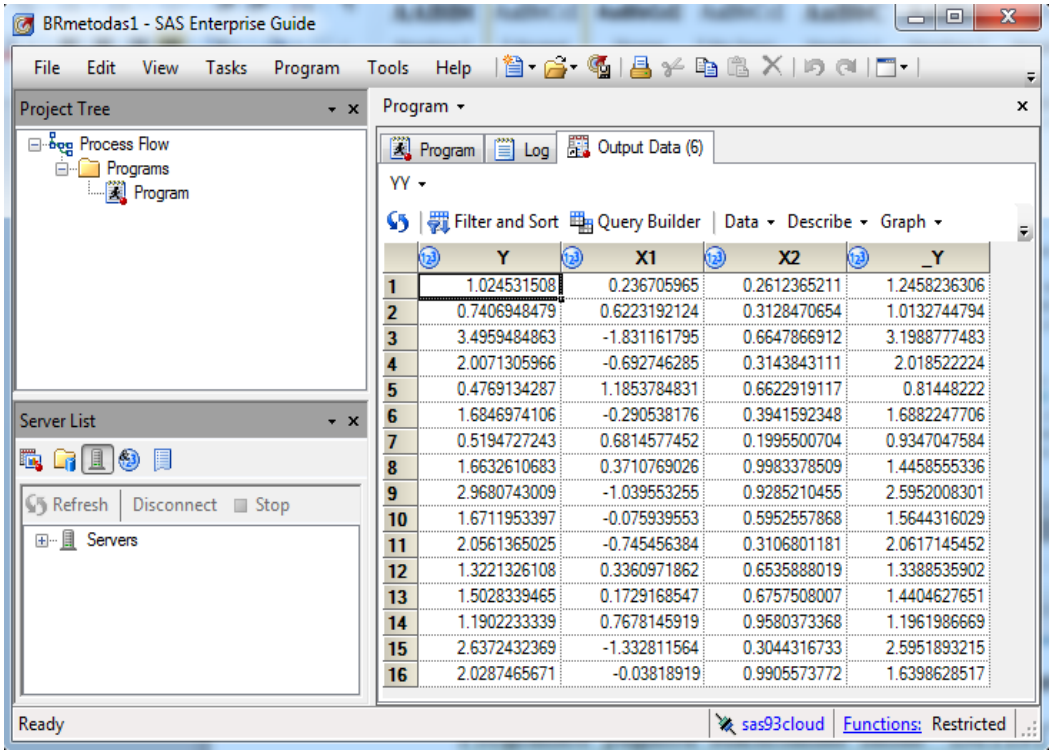
2.7 pav. Programos langas

Programos pagalba sukuriamas failas “IMTIS1” (2.8 pav.), kuriame pateikiamos 100 – ui imčių su norimu imties tūriu sugeneruotos nepriklausomo kintamojo (*X1*), atsitiktinės paklaidos (*X2*) ir priklausomo kintamojo (*Y*) tikrosios reikšmės.



2.8 pav. Failo “IMTIS1” langas

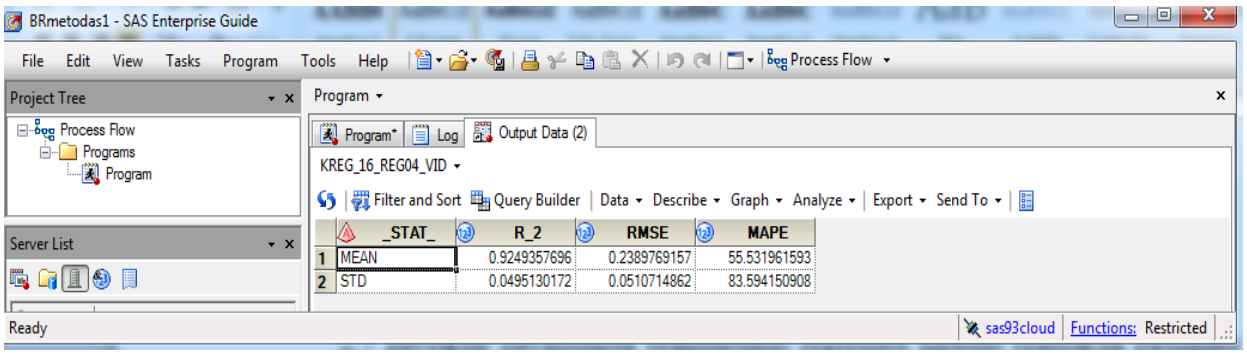
Realizavus branduolinį regresijos metodą faile “YY” (2.9 pav.) pateikiami apskaičiuoti priklausomo kintamojo reikšmių įverčiai (*_Y*).



2.9 pav. Failo “YY” langas

Galiausiai, apskaičiuojami neparametrinių regresijos metodų lyginamajam tyrimui naudojami vertinimo kriterijai, t.y. apibrėžtumo koeficientas (R_2), vidutinės kvadratinės paklaidos šaknis

(RMSE) ir vidutinė procentinė absoliutinė paklaida (MAPE). Paveiksle 2.10, esančioje lentelėje pirmoje eilutėje pateikiamos minėtų vertinimo kriterijų reikšmės, o antroje eilutėje – šių reikšmių standartiniai nuokrypiai.



2.10pav. Failo “KREG_16_REG04_VID” langas

IŠVADOS

Magistro darbe atlikus regresinių metodų tyrimą pasirinktiems neparametrinės regresijos modeliams ir juos palyginus galima nustatyti kiekvieną modelį geriausiai įvertinantį metodą.

1. Adytivios eksponentinės formos neparametrinės regresijos modelį, geriausiai įvertina lokalios regresijos metodas. Apibrėžtumo koeficiento reikšmės gaunamos artimiausios vienetui t.y. modelis yra patikimiausias. Vidutinių kvadratinių paklaidų šaknys ir vidutinių procentinių absoliutinų paklaidų reikšmės gaunamos mažiausios.
2. Sinusoidinės formos neparametrinės regresijos modelį, geriausiai įvertina apibendrintų adityvių modelių metodas. Apibrėžtumo koeficiento reikšmės gaunamos artimiausios vienetui t.y. modelis yra patikimiausias. Vidutinių kvadratinių paklaidų šaknys ir vidutinių procentinių absoliutinų paklaidų reikšmės gaunamos mažiausios.
3. Polinominės formos neparametrinės regresijos modelį, geriausiai įvertina branduolinis regresijos metodas. Apibrėžtumo koeficiento reikšmės gaunamos artimiausios vienetui t.y. modelis yra patikimiausias. Vidutinių kvadratinių paklaidų šaknys ir vidutinių procentinių absoliutinų paklaidų reikšmės gaunamos mažiausios.
4. Semiadityvios formos neparametrinės regresijos modelį, geriausiai įvertina lokalios regresijos metodas. Apibrėžtumo koeficiento reikšmės gaunamos artimiausios vienetui t.y. modelis yra patikimiausias. Vidutinių kvadratinių paklaidų šaknys ir vidutinių procentinių absoliutinų paklaidų reikšmės gaunamos mažiausios.
5. Multiplikatyvios sinusoidinės formos neparametrinės regresijos modelį, geriausiai įvertina branduolinis regresijos metodas. Apibrėžtumo koeficiento reikšmės gaunamos artimiausios vienetui t.y. modelis yra patikimiausias. Vidutinių kvadratinių paklaidų šaknys ir vidutinių procentinių absoliutinų paklaidų reikšmės gaunamos mažiausios.
6. Sinusoidinio paraboloido formos neparametrinės regresijos modelį, geriausiai įvertina branduolinis regresijos metodas. Apibrėžtumo koeficiento reikšmės gaunamos artimiausios vienetui t.y. modelis yra patikimiausias. Vidutinių kvadratinių paklaidų šaknys ir vidutinių procentinių absoliutinų paklaidų reikšmės gaunamos mažiausios.

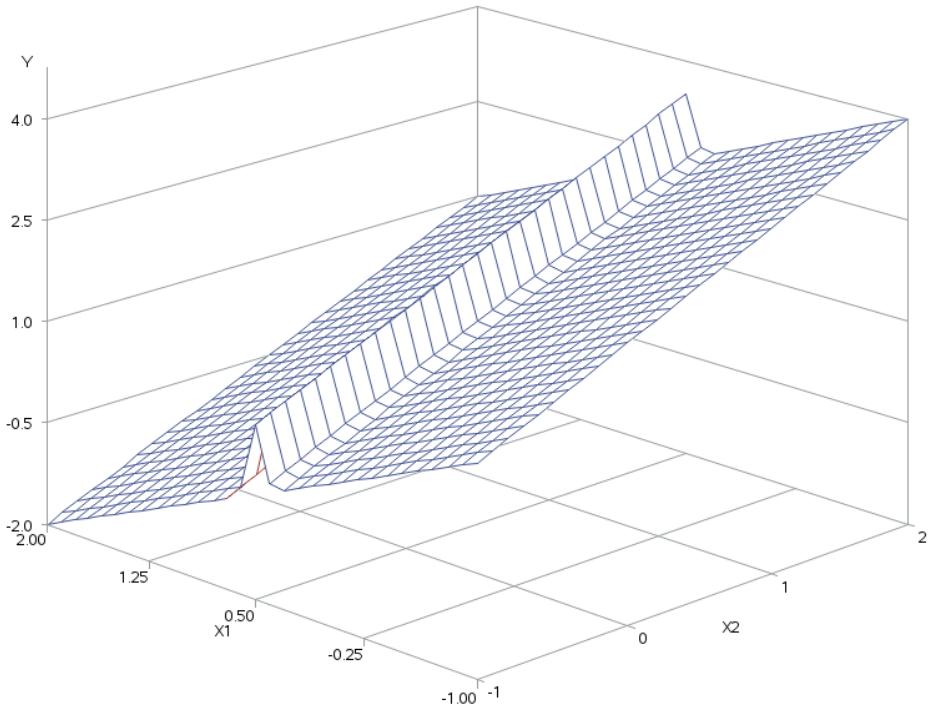
Netiesinių, sudėtingų paviršių atvejais, geriausia taikyti branduolinį regresijos metodą.

LITERATŪRA

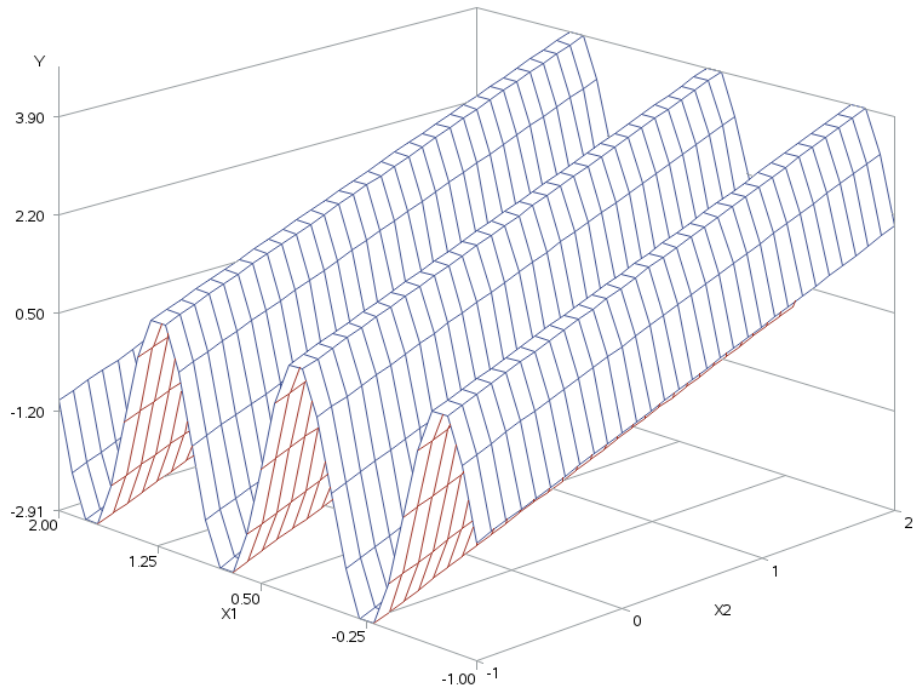
1. http://en.wikipedia.org/wiki/Regression_analysis
2. Yushi U. Adachi, Kazuhiko Watanabe, Hideyuki Higuchi, Tetsuo Satoh. The Determinants of Propofol Induction of Anasthesia Dose. http://www.anesthesia-analgesia.org/content/92/3/656.full, 2000.
3. Angus CF Kwok, Lubanski Lam. Effects of Economic Performance and Immigration on Youth Unemployment: The Hong Kong Experience. http://www.eurojournals.com/IRJFE_69_07.pdf, 2011.
4. Butų vertę Vilniuje įtakojančių veiksnių statistinė analizė. http://www.butastau.lt/gallery/paveiksleliai/statistine_analize/statistine_analize.pdf
5. N. Jagannathan, P. Neelakantan, C. Thiruvengadam, P. Ramani, P. Premkumar, A. Natesan, J.S. Herald, H.U. Luder. Age Estimation in an Indian Population Using Pulp/Tooth Volume Ratio of Mandibular Canines Obtained from Cone Beam Computed Tomography. http://www.iofos.eu/Journals/JFOS%20Jun11/1_AGE%20ESTIMATION%20IN%20AN%20INDIAN%20POPULATION.pdf, 2011.
6. Čekanavičius V., Murauskas G. Statistika ir jos taikymai II. Vilnius, 2002. 123 – 168.
7. http://www.chsbs.cmich.edu/fattah/courses/empirical/multicollinearity.html
8. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_transreg_sect015.htm
9. http://www.scritub.com/limba/lituaniana/EKONOMETRIJA-Paskait-konspekta921019172.php
10. http://www.sas.com
11. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#gam_toc.htm
12. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#loess_toc.htm
13. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#tpspline_toc.htm
14. Čekanavičius V., Murauskas G., Statistika ir jos taikymai I. Vilnius, 2003. 240 – 274.
15. http://support.sas.com/kb/25/602.html
16. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&ved=0CIwBEBYwCQ&url=http%3A%2F%2Fwww.ms.uky.edu%2F~mai%2Fbiostat277%2FLN.ppt&ei=dLT7Upj9AqeO7AbbhoHYCA&usg=AFQjCNFwww09mcZumgstAMrYKHo2l2-Wtw&sig2=nrHKXJZij8I2v4S3o6UuCw&bvm=bv.61190604,d.ZGU
17. http://www-personal.ksu.edu/~wxyao/material/localcov.pdf.

1. PRIEDAS. MODELIŲ ILIUSTRACIJOS

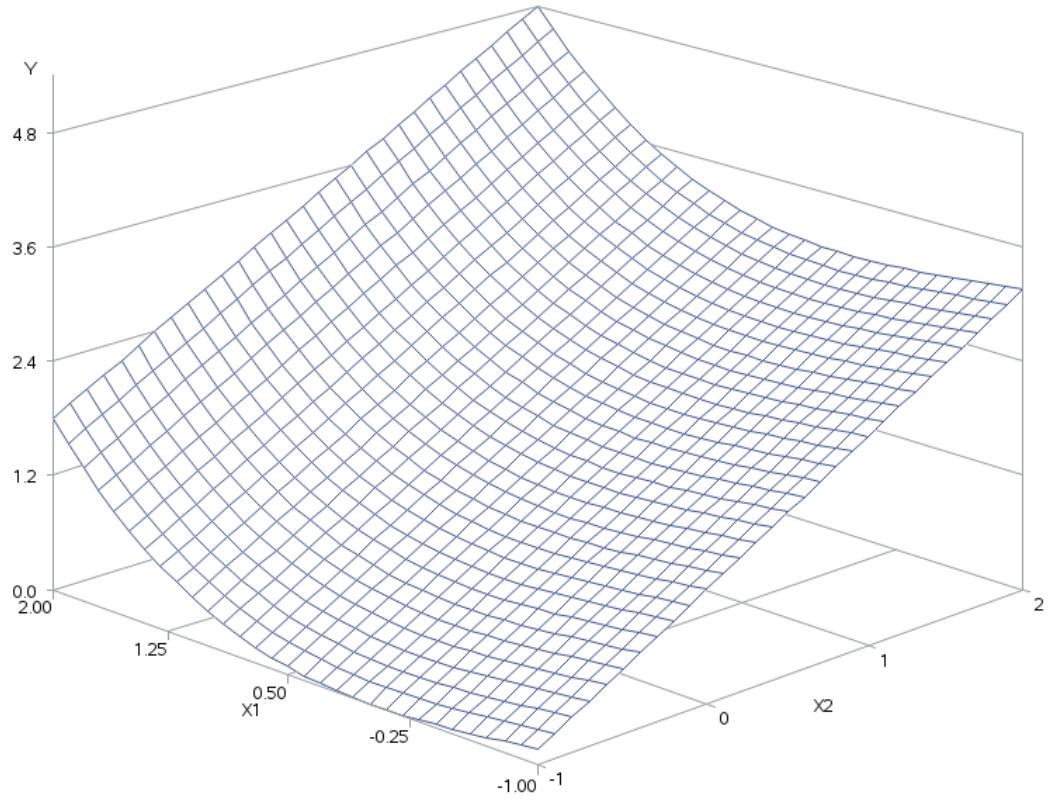
Pirmo modelio iliustracija



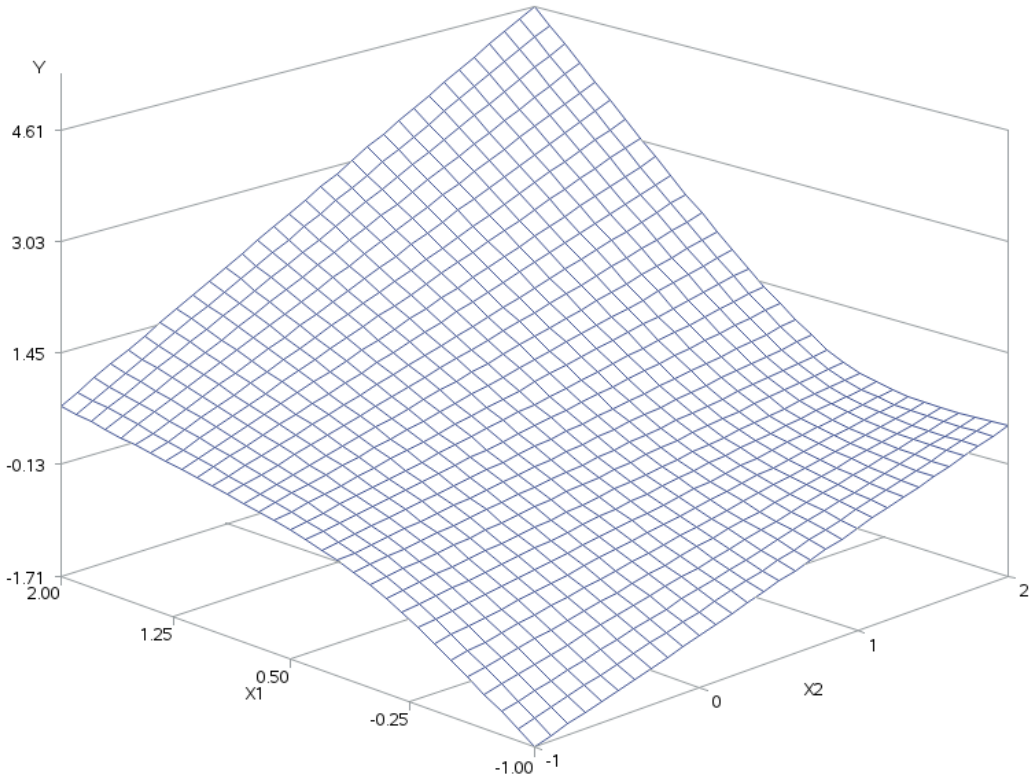
Antro modelio iliustracija



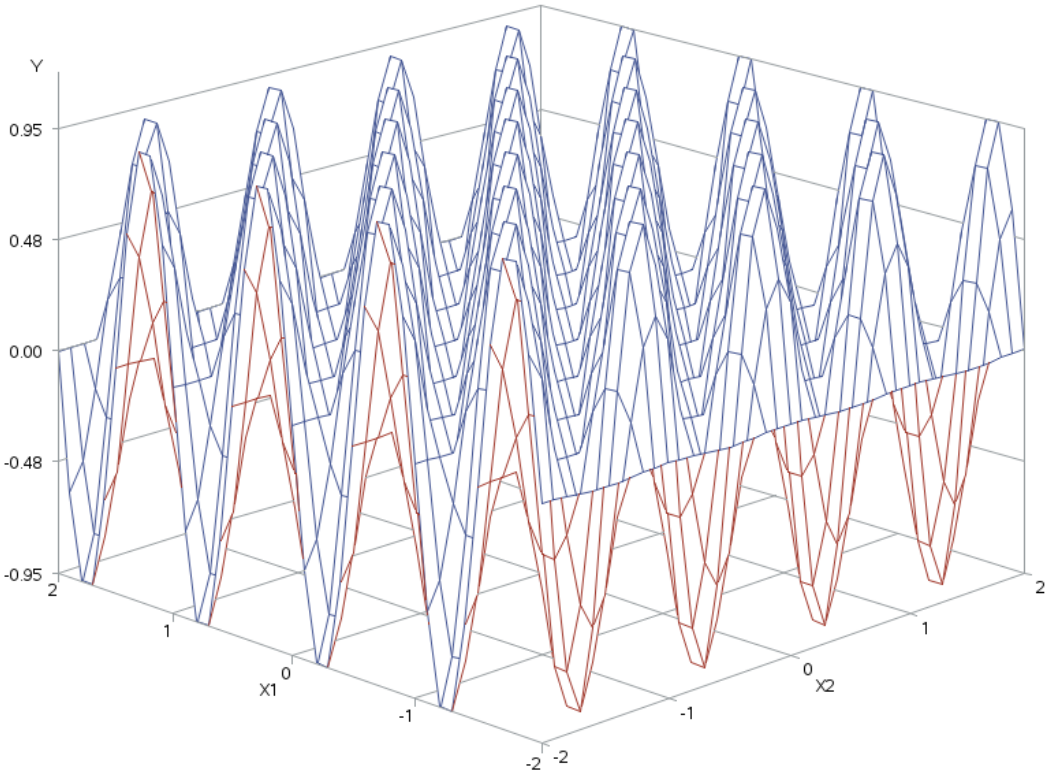
Trečio modelio iliustracija



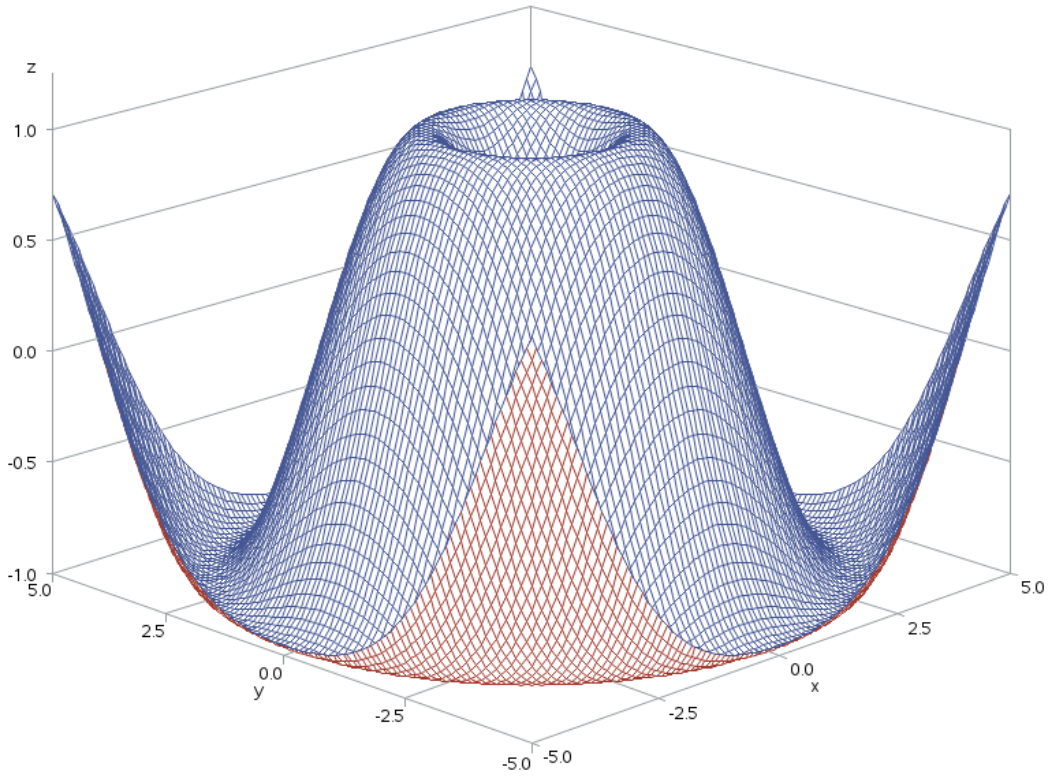
Ketvirto modelio iliustracija



Penkto modelio iliustracija



Šešto modelio iliustracija



2. PRIEDAS. LENTELĖS

1.lentelė

Pirmo regresijos modelio R^2 vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.9868250593	0.9737740726	0.9704367883	0.9249357696
	0.0174559967	0.032822131	0.0412107998	0.0495130172
32	0.9787581109	0.9729566308	0.9684202864	0.9335564655
	0.0177319322	0.0245868751	0.0291307565	0.0328465857
64	0.978862953	0.981139384	0.9736848577	0.950027518
	0.0111147933	0.0108406127	0.0160569572	0.0169166596
128	0.9787277588	0.9837669952	0.9745861605	0.9587763205
	0.0078108545	0.0075351918	0.0107301479	0.0105157316
256	0.9786119628	0.9890436436	0.9749073872	0.9663707611
	0.0065462775	0.0066628261	0.0086077118	0.0069649202
512	0.9782892252	0.9959342287	0.9748313776	0.9717574892
	0.0047424755	0.0022817964	0.00584396	0.0044324674
1024	0.9776358629	0.9987089729	0.9743008094	0.9754500853
	0.0033576749	0.0006523549	0.0040821569	0.0027030398
2048	0.9770923004	0.9993736513	0.9736643665	0.9785216574
	0.0021961396	0.0003257621	0.0025550576	0.0015310895
4096	0.9752781706	0.9998456855	-	0.9813685173
	0.0033690462	0.0000960828	-	0.0010783946

2. lentelė

Pirmo regresijos modelio RMSE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.0790012822	0.1120556787	0.1138014536	0.2389769157
	0.0683762127	0.0980531264	0.1048114548	0.0510714862
32	0.1218699113	0.1354656525	0.1442955366	0.2395399734
	0.0662249468	0.0775911446	0.0861773371	0.0385284447
64	0.140412309	0.1321171524	0.1542387967	0.2216894297
	0.0387289926	0.0403706205	0.0504768402	0.0257708079
128	0.1457869458	0.1260266772	0.1582137156	0.2045047413
	0.0239251194	0.0266073214	0.0312918058	0.0169417824
256	0.1489192632	0.1026788598	0.1607922954	0.1878309883
	0.0201162165	0.0326900785	0.0247135608	0.0135055342
512	0.1521152645	0.0641393176	0.1638762508	0.173953612
	0.0149956793	0.015788991	0.0172800141	0.0106569601
1024	0.154633939	0.0362728173	0.1656268678	0.1622166911
	0.0104714328	0.0085767025	0.0119794972	0.007164269
2048	0.1563835694	0.0250498745	0.1676538829	0.1514895903
	0.0069130278	0.0064700577	0.0080223597	0.0042912115

4096	0.1621800263	0.012508801	-	0.1410375265
	0.0101626688	0.0029681571	-	0.0032911359

3. lentelė

Pirmo regresijos modelio MAPE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	9,1709268328	13,5380015800	16,1539856460	55,5319615930
	18,4111912280	24,4074874640	38,3024107380	83,5941509080
32	14,1993501260	21,6845577450	17,5890826970	85,2263497730
	22,3936704880	89,9822015100	23,0972881240	287,7211662100
64	12,9813605250	11,6629072160	15,5785452900	56,2353413450
	12,8056374730	21,1674685100	14,6002059320	101,1555174000
128	54,5618523070	15,4279844870	115,2233279900	115,5017549900
	412,3643621900	80,8141604210	985,8027877500	685,5772069900
256	20,3613731070	3,6895477191	42,0358434650	43,2882959670
	66,1191027650	2,2281652979	241,5487192600	48,2246201040
512	19,3271324110	1,5312664855	48,1168458810	58,7094093770
	35,8164767880	0,4410103972	249,9898792200	147,8151240500
1024	17,9471302600	0.7318117973	33,6097724170	38,9868622430
	21,1624901710	0.2121391548	112,4146317200	41,4867728700
2048	34,2898842440	0.4520280375	79,6162626110	61,4652217770
	119,6565612000	0.1291361293	184,8118584500	208,7681303900
4096	38,0320336250	0.1920459636	-	47,3179197490
	102,7631866000	0.0414138846	-	150,4387882900

4.lentelė

Antro regresijos modelio R^2 vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.9927468253	0.7817730331	0.786913261	0.6632698295
	0.0050813734	0.1101086868	0.1418111715	0.1136224064
32	0.99114162	0.7696828106	0.7666909866	0.6825550257
	0.0029312038	0.0779213312	0.0829761579	0.0749583011
64	0.9905749525	0.7884615277	0.7510179987	0.7077628144
	0.0016731519	0.0670762842	0.0556388577	0.0490405643
128	0.9904578104	0.7961352794	0.746802976	0.7406965732
	0.0010039392	0.0712842077	0.0336025153	0.0302431006
256	0.9903705848	0.7749456587	0.7449763177	0.7773352649
	0.000739968	0.0678627576	0.0252857084	0.0206181655
512	0.9903037121	0.751075556	0.7421777011	0.8083627245
	0.0005495261	0.0430179402	0.016410257	0.012780309
1024	0.9903013139	0.7409415209	0.739846484	0.8383416402
	0.0004138666	0.0122880617	0.0125638117	0.0082779547

2048	0.9902739189	0.7398738968	0.738406251	0.866393321
	0.0002887409	0.008770851	0.0051299599	0.0050908125
4096	0.9903514286	0.739651829		0.8914121906
	0.0007171341	0.0063205436		0.0029994623

5. lentelė

Antro regresijos modelio RMSE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.1263044361	0.7171944129	0.6786168476	0.9113357162
	0.0292069792	0.1350265868	0.2341324189	0.1311996125
32	0.1535303805	0.7842410434	0.7867550916	0.928570993
	0.0150033742	0.0982620105	0.1128881542	0.0777976076
64	0.1637757811	0.7680742356	0.8394529458	0.9123790295
	0.0085696588	0.1102250059	0.0702922746	0.0432510962
128	0.1673694719	0.7602726691	0.8610894871	0.8720249687
	0.0048763208	0.1454653247	0.0385884286	0.0285638974
256	0.1694133761	0.8043664418	0.8712630001	0.8142317512
	0.0029973311	0.1566368951	0.0262836066	0.0203767505
512	0.1699830534	0.8549075393	0.8764491166	0.7555125942
	0.0020695027	0.1064572771	0.0178497359	0.0134614952
1024	0.1704258103	0.8807763097	0.8832136712	0.6956809178
	0.0013603414	0.0129220254	0.0135212323	0.0084867339
2048	0.170609342	0.8822962822	0.8837908227	0.6322700153
	0.0011102539	0.0096256787	0.0074934351	0.0070232125
4096	0.1699246138	0.8831340354		0.5703265177
	0.0063766295	0.0061122031		0.0042794638

6. lentelė

Antro regresijos modelio MAPE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	29,8627252310	171,9858790000	141,8068708700	118,4482325100
	63,0839390740	340,2015388800	180,1049171700	75,0520824340
32	36,8714448880	183,8424523900	179,9783021000	157,7415973400
	43,0531456290	196,5132211100	187,2680351000	164,5944958700
64	51,7749202580	250,2141088000	311,4007293200	240,7209527300
	109,8801341300	605,8738376300	666,5202156400	592,9504730000
128	50,0047115570	266,1606581200	300,0553669000	227,8039845300
	54,6203545670	471,3574196600	441,7880517600	311,0241131200
256	50,2870544240	253,8034794600	274,5163234700	208,8155073000
	47,1539393910	282,4467064300	275,5003128200	197,3521276200
512	50,6994840680	258,2340566300	265,5313839100	202,6223583600
	29,7655710480	165,4585814200	158,5951017700	124,7580669400

1024	49,6405170090	263,6029620200	275,9621047300	187,5914008200
	20,9725939350	119,2873714100	136,5908164000	81,3318767180
2048	63,0660075590	341,0610261100	339,7971343400	230,8447471700
	60,4685849560	370,9925775900	390,6417254200	281,6063259800
4096	65,2583888200	356,2470624700		214,4459156500
	44,2892896080	264,0801616400		157,2071690900

7.lentelė

Trečio regresijos modelio R^2 vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.9994315469	0.9739169589	0.9375863639	0.9837222796
	0.0007405125	0.0265070567	0.0773481538	0.0092257483
32	0.9986736229	0.9761362447	0.9084025014	0.9871751351
	0.0011349732	0.0217590356	0.0809702598	0.0053461057
64	0.9976423	0.9838632751	0.8872673319	0.989758189
	0.0015142152	0.0154737824	0.0802943433	0.0045595629
128	0.9963664635	0.9837989946	0.8728290853	0.9922784366
	0.0019387318	0.0158847734	0.0641955258	0.0019723921
256	0.9954729358	0.9795504517	0.8640649366	0.9941611623
	0.0020339373	0.0167040911	0.0466478742	0.0009963411
512	0.9945371234	0.9587174511	0.859125134	0.9957593229
	0.0020759246	0.0203574651	0.0314029452	0.0005259711
1024	0.9939045189	0.9536840221	0.8590416452	0.9970126935
	0.0018520113	0.0181540433	0.0232789976	0.0002722112
2048	0.9934004424	0.9512138718	0.8560291058	0.9979379274
	0.0015446054	0.0147265892	0.0141018173	0.0001314042
4096	0.9018335035	0.9505421628		0.9985887085
	0.4082210148	0.0144899643		0.0000670897

8. lentelė

Trečio regresijos modelio RMSE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.0208235505	0.1517462603	0.2239015594	0.1229497896
	0.0132173262	0.0898288589	0.1686591403	0.0186850605
32	0.0366510425	0.1543009399	0.2998788509	0.1174994696
	0.0169494003	0.0830119452	0.1583158009	0.0158231389
64	0.0522976153	0.1306178314	0.3578662259	0.10764518
	0.0217318904	0.0717372817	0.1707246361	0.0113759385
128	0.0659227515	0.1320277526	0.39261805	0.0968038299
	0.0212658207	0.0690164119	0.1283420729	0.0071053262
256	0.0739313482	0.1487444532	0.4087695772	0.0848349698
	0.0193186124	0.0666757375	0.0884321655	0.0051062848

512	0.0820616229	0.2193934412	0.4209897754	0.0730167954
	0.0164651047	0.070675849	0.0581950872	0.0031422367
1024	0.0872943437	0.2381604888	0.4223026779	0.0615642968
	0.0140744205	0.0518541201	0.0425386316	0.0020849951
2048	0.0911774569	0.2468760305	0.4278448219	0.0512098704
	0.0114177004	0.0383985585	0.0293778604	0.001192963
4096	0.2155949626	0.2488235955		0.0423995741
	0.2810180502	0.0349299582		0.0007202117

9. lentelė

Trečio regresijos modelio MAPE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	4,206794205	39,332763865	46,549706995	46,590246181
	9,175402266	63,338325462	62,240418498	125,206924300
32	6,952147523	30,440825426	75,704589013	36,880134461
	12,621079623	38,152939952	164,469625160	59,170923936
64	8,017070835	26,458667609	73,057608371	29,682696353
	9,671296336	56,171188090	88,098201621	41,803414111
128	15,043268606	22,443676983	99,025643681	37,966545613
	68,694976918	21,102254492	228,140750210	116,291032760
256	7,781413219	50,857820799	90,860120976	33,254474587
	5,718981731	258,635139290	105,542364440	90,095045279
512	8,856072092	52,965504922	102,196743560	28,700594429
	11,212444322	80,296207768	132,743892740	64,688772453
1024	9,798748919	84,228338713	110,439525740	28,937084882
	12,731809571	220,400000980	181,245988260	60,429412905
2048	8,961902905	71,912829993	80,762867395	19,996739229
	6,563320897	121,713686740	18,123301661	26,627688295
4096	50,189214538	69,861567831		17,385434571
	114,085942190	78,162212083		30,014250582

10.lentelė

Ketvirto regresijos modelio R^2 vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0,985556669	0,983700463	0,967951324	0,979978765
	0,014764228	0,012749118	0,041345668	0,01327229
32	0,971596147	0,990264202	0,963216657	0,984981824
	0,016932188	0,007735359	0,033481646	0,006545162
64	0,959080001	0,994385346	0,962819687	0,988966297
	0,016946373	0,004550621	0,023113169	0,004054673
128	0,947787268	0,995871249	0,959139034	0,991933883
	0,011651317	0,003816363	0,021601632	0,002142309

256	0,939825489	0,997914863	0,958801696	0,994114363
	0,010269794	0,002486353	0,018486695	0,001130725
512	0,935582196	0,99871219	0,957484556	0,99583547
	0,00748077	0,001196214	0,01701932	0,000575037
1024	0,933771329	0,999292621	0,95811296	0,997074175
	0,00592568	0,000627438	0,014220306	0,000285953
2048	0,932661206	0,999616894	0,956415725	0,99798746
	0,004767481	0,000335214	0,010757131	0,000137869
4096	0,932098419	0,99981002		0,998650235
	0,003396606	0,000124056		0,0000729837

11. lentelė

Ketvirto regresijos modelio RMSE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0,115872236	0,126069699	0,147331389	0,134611728
	0,059982835	0,05078035	0,11889816	0,024050558
32	0,181208968	0,102524317	0,188742853	0,131096776
	0,053757435	0,041433492	0,099894235	0,014809862
64	0,23229361	0,081456367	0,211964746	0,118651772
	0,05300343	0,03388645	0,081078464	0,010299485
128	0,271927067	0,070178478	0,231837251	0,105485141
	0,038661607	0,028321194	0,073609519	0,006832781
256	0,296041693	0,050010259	0,237878629	0,092060249
	0,029977572	0,021931777	0,057498825	0,004351306
512	0,309829259	0,040613237	0,246406893	0,078519937
	0,023516864	0,01621568	0,050695284	0,00289028
1024	0,314697351	0,030518295	0,246696406	0,066060372
	0,016457718	0,010933812	0,041190195	0,001878377
2048	0,317113837	0,022643568	0,254749411	0,057961835
	0,012218226	0,007541014	0,031495347	0,001082941
4096	0,318606464	0,016172146		0,044909437
	0,007915707	0,00475997		0,000815769

12. lentelė

Ketvirto regresijos modelio MAPE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	141,0283962	112,5156889	269,2894289	131,396936
	902,853272	297,6497919	1918,584335	753,7927335
32	142,2599147	111,988476	196,6276559	61,80013469
	593,5971428	720,6911307	886,0732318	146,4285267
64	185,7830432	46,79309299	167,1146669	59,17447071
	395,6893037	88,35592798	347,9388892	82,22566499

128	157,2739549	26,03370643	159,7287701	45,00938965
	237,2294709	48,36866399	349,9906518	52,50999396
256	146,7693157	12,06364238	131,2489187	32,46372646
	118,7276001	12,63078567	141,5240578	32,36059312
512	185,6175274	7,494279901	169,5313764	22,81513638
	359,1845871	15,11644969	374,4358666	13,50781047
1024	238,3853928	6,070932201	154,3815582	27,2853256
	590,194993	22,38179633	169,9729142	71,36096428
2048	255,4468781	3,074283094	183,2987858	19,29407662
	423,9322419	8,50403607	134,131237	29,93624978
4096	286,710924	1,27114672		23,82038673
	491,4333326	2,168673912		88,62033552

13.lentelė

Penkto regresijos modelio R^2 vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.9529983178	0.8452579301	0.850571007	0.9598797825
	0.0324900226	0.0933392033	0.0940199602	0.0367741592
32	0.8836462851	0.8270959166	0.8179279995	0.9495122719
	0.042555418	0.0557698657	0.059074984	0.0296026548
64	0.8376006157	0.8078893146	0.8025435682	0.93581346
	0.0359432914	0.0429175517	0.0443858408	0.0212323648
128	0.8206578329	0.8054278282	0.8035455072	0.9233366978
	0.028652184	0.029500684	0.030228571	0.0169968625
256	0.8101769194	0.8026327154	0.8025322758	0.9107728682
	0.0214289958	0.0222291675	0.0222624022	0.0139108415
512	0.8051459471	0.8014065722	0.8012223324	0.9005642779
	0.0141927503	0.01466545	0.0142194675	0.0090065306
1024	0.8024482522	0.800578318	0.8006847832	0.8950269869
	0.0106408371	0.0108634825	0.0111625865	0.0065806952
2048	0.8019353183	0.8010157211	0.802274205	0.8958738255
	0.007397673	0.0073697752	0.0095316285	0.0039714986
4096	0.8010455735	0.8006951147		0.9016995823
	0.0053621695	0.0053699633		0.0024865351

14. lentelė

Penkto regresijos modelio RMSE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.2120384193	0.3897972519	0.3696408179	0.184724845
	0.0637860849	0.078108402	0.1259958589	0.0712357303
32	0.3605229487	0.4406744891	0.4521096567	0.2297009803
	0.0537568015	0.0519350444	0.0598004797	0.0562551156

64	0.4358037288	0.4738902159	0.4801942597	0.2711887163
	0.0392398901	0.0416966091	0.0439700871	0.0353222777
128	0.4672204245	0.4868769384	0.489201328	0.3042263799
	0.0265245272	0.0262856595	0.0269799258	0.0265879811
256	0.4840663943	0.4935979982	0.4937526489	0.3311957665
	0.0148894918	0.0153436753	0.0163458923	0.016446252
512	0.4921450481	0.4968290159	0.4970974129	0.3513743333
	0.0105665915	0.0106584662	0.0107385227	0.0099756391
1024	0.4966148939	0.498952878	0.0111625865	0.3619288046
	0.0092347408	0.0093497621	0.0094791178	0.0073704796
2048	0.4979307057	0.4990878333	0.4994774397	0.3610154108
	0.0070973236	0.0070926883	0.0059935658	0.0042011367
4096	0.4989347395	0.4993742261		0.3507043369
	0.0050385758	0.005058176		0.0022856897

15. lentelė

Penkto regresijos modelio MAPE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	5371,4468513	5587,5777163	6642,7302976	2045,0470534
	52909,289849	54435,605568	65020,986279	19740,469774
32	5560,0360808	5706,4134023	7115,5618207	366,38261200
	54172,794216	54739,944685	69066,957374	2424,54817950
64	2949,8019052	2687,9764143	2193,8089225	358,01328231
	27112,919841	23930,359663	19090,089767	2383,62103860
128	1055,2063455	1099,9719421	1150,8606006	921,55276757
	7223,0198516	7497,3132877	8061,7733092	7318,29624710
256	666,98287497	688,91951551	726,50269379	567,94723215
	2620,561376	2751,1124573	3247,77102	2792,21003070
512	488,82602949	541,36598429	531,51998884	528,29609165
	1416,1136158	1816,3471489	1790,5473235	2800,17600500
1024	499,6992059	496,6181633	487,07800489	369,19092484
	1244,7745195	1230,718651	1247,6856467	1076,8804748
2048	406,35506336	406,13665494	1444,1435409	299,37149862
	606,322948	604,41419734	1936,5699677	601,97244429
4096	328,50778588	328,751509		224,39724247
	330,68555207	332,90549778		276,94337352

16.lentelė

Šešto regresijos modelio R^2 vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.9952226974	0.9509131215	0.9585458299	0.9892169858
	0.0037207977	0.0379402999	0.031468203	0.0102661912

32	0.9876102434	0.9532284352	0.9482620442	0.9900187847
	0.0052886064	0.0204638635	0.0231824141	0.004916843
64	0.9809577464	0.9615730619	0.9422022308	0.9908747163
	0.0059314323	0.0134799644	0.0194993029	0.0032617187
128	0.9768483332	0.9566081226	0.9428691431	0.9923331298
	0.0046512253	0.0145832508	0.0131146726	0.0017863516
256	0.9742061726	0.9510181386	0.9418253158	0.9935233995
	0.003613387	0.0106967007	0.0090481339	0.0010057183
512	0.9729486836	0.949143975	0.9416772695	0.9947282163
	0.0025018012	0.0075501843	0.0062183898	0.0005703131
1024	0.9724176538	0.9480375509	0.9415939692	0.9958619101
	0.0018093415	0.0055443231	0.00428644	0.0002756072
2048	0.9714885376	0.9466715976	0.9419043083	0.9968092278
	0.002619626	0.0035676881	0.0037498662	0.000161339
4096	0.9618582569	0.9461950794		0.9976105119
	0.0107667057	0.0024406756		0.000084744

17. lentelė

Šešto regresijos modelio RMSE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas
16	0.0614070863	0.1971231441	0.1761564902	0.0885915105
	0.0189303288	0.0525770212	0.0693195283	0.0233725501
32	0.1068840085	0.2076550462	0.2163117425	0.0946432229
	0.0197024218	0.0385973857	0.0461744998	0.0140478638
64	0.1359124296	0.1934207397	0.2362472323	0.0932691763
	0.0164464269	0.0370957938	0.0312294383	0.0095656433
128	0.1544285199	0.2095731447	0.2422392925	0.0885384494
	0.0114898192	0.0323167026	0.0219188978	0.0057566944
256	0.1642905765	0.2255884041	0.2465278756	0.0821808397
	0.0097105772	0.0231165004	0.0146924295	0.0036047046
512	0.1691651035	0.2315796164	0.2482980131	0.0745998029
	0.0065755769	0.0162820695	0.0113940095	0.002404644
1024	0.1711858885	0.0055443231	0.2489760484	0.0662964784
	0.0045334201	0.0115321598	0.0075577494	0.0014499324
2048	0.173701891	0.2376420986	0.2475566334	0.0581380511
	0.0068964524	0.0071345897	0.0053518672	0.0009532496
4096	0.1991266021	0.2387821335		0.0503239391
	0.0284659114	0.0048226787		0.0006216308

18. lentelė

Šešto regresijos modelio MAPE vertinimo rezultatai

imties dydis	Apibendrintų adityvių modelių metodas	Lokalios regresijos metodas	Glodinančių splainų metodas	Branduolinis regresijos metodas

16	27,903069986	107,256242310	72,090944153	43,783648678
	52,348344452	256,976894230	151,646217760	157,572057500
32	59,751348496	104,799466620	107,374648330	39,175243562
	102,154150540	160,424211100	150,018622820	67,826994550
64	69,066422719	93,235918860	117,315243490	37,075012822
	85,616564835	124,442614890	155,161244240	41,405670673
128	132,761058950	159,973304520	191,676991400	66,510259851
	529,408262330	584,897002100	659,675997100	239,756031150
256	106,641069170	137,957094420	147,442166970	52,589664990
	285,213118220	354,885004920	303,088066950	140,328861760
512	107,856134600	142,714994530	162,839364980	38,118786010
	216,841861670	270,417904370	320,244594720	57,311581337
1024	100,330161610	142,246334640	156,383978370	28,378841488
	123,218835320	187,096188950	192,058659670	20,838017894
2048	114,340233430	151,832639920	141,235970230	27,201385562
	132,603477140	154,497020010	97,001438421	22,151725305
4096	136,347216230	167,980714030		27,544636291
	139,822361940	175,724022330		28,901041530

3 PRIEDAS. PROGRAMŲ TEKSTAI

```
%let imciu_sk=100;
%let imties_n=16;
```

Pirmo modelio realizacija:

```
 data imtis1;
     do imtis=1 to &imciu_sk;
       do i=1 to &imties_n;
              X1=rannor(1);
                   X2=ranuni(1);
                   Y=1-X1+EXP(-200*(X1-0.5)**2)+X2;
                   output;
          end;
      end;
      drop i;
   run;
```

Antro modelio realizacija:

```
data imtis1;
  do imtis=1 to &imciu_sk;
    do i=1 to &imties_n;
              X1=ranuni(1);
                   X2=rannor(1);
                   Y=2*SIN(2*3.14*X1)+X2;
                   output;
          end;
   end;
   drop i;
run;
```

Trečio modelio realizacija:

```
data imtis1;
  do imtis=1 to &imciu_sk;
    do i=1 to &imties_n;
              X1=rannor(1);
                   X2=rannor(1);
                   Y=1+0.25*X1**2+0.1*X1**3+X2;
                   output;
          end;
   end;
   drop i;
run;
```

Ketvirto modelio realizacija:

```
data imtis1;
  do imtis=1 to &imciu_sk;
     do i=1 to &imties_n;
        X1=rannor(1);
        X2=rannor(1);
        Y=X1+sqrt(0.1+0.4*X1**2) * X2;
        output;
     end;
```

```
  end;
  drop i;
run;
```

Penkto modelio realizacija:

```
data imtis1;
  do imtis=1 to &imciu_sk;
    do i=1 to &imties_n;
              X1=rannor(1);
                   X2=rannor(1);
                   X3=rannor(1);
                   Y=-SIN(2*3.14*(X1-0.5))*COS(2*3.14*X2)+X3;
                   output;
         end;
   end;
   drop i;
run;
```

Šešto modelio realizacija:

```
data imtis1;
  do imtis=1 to &imciu_sk;
    do i=1 to &imties_n;
              X1=rannor(1);
                   X2=rannor(1);
                   X3=rannor(0.05);
                   Y=SIN(SQRT(X1*X1+X2*X2))+X3;
                   output;
         end;
   end;
   drop i;
run;
```

Apibendrintų adityvių modelių metodo realizacija:

```
  %macro reg_gam(duom,modelis,rez);
    ods _all_ close;
    proc gam data=&duom.;
       model &modelis.;
       output out=&rez. PRED=_;
    run;
    ods html;
    ods listing;
  %mend reg_gam;

  %macro regresijos_vertinimas(imtis,d,metodas,rezultatai);
    %let starttime=%sysfunc(datetime());

    proc sql noprint;
       select max(imtis)
          into :imciu_sk
          from &imtis
       ;
    quit;
    %if %sysfunc(exist(&rezultatai.)) %then
```

```
  %do;
     proc sql noprint;
        select max(imtis)+1
           into :eil_imtis_nuo
           from &rezultatai.
        ;
     quit;
  %end;
  %else %let eil_imtis_nuo=1;

  %do eil_imtis=&eil_imtis_nuo %to &imciu_sk;
     %put Imtis: &eil_imtis is %sysfunc(left(&imciu_sk));
     %put Taikomas metodas: &metodas%str(;) duomenys: &imtis;
     data XY (keep=X1-X&d Y);
        set &imtis;
        where imtis=&eil_imtis;
     run;

     %&metodas;

     proc sql noprint;
        create table _tmp as
           select *,
                  (Y-_Y)**2 as SK1,
                  (Y-mean(Y))**2 as SK2,
                  abs((Y-_Y)/Y) as APE
           from YY
        ;
     quit;
     proc summary data=_tmp;
        var SK1 SK2 APE;
        output out=_tmp(drop=_TYPE_) sum()=;
     run;
     data _tmp(keep=IMTIS R_2 RMSE MAPE);
        set _tmp;
        IMTIS = &eil_imtis.;
        R_2 = 1 - SK1 / SK2;
        RMSE = sqrt(SK1 / _FREQ_);
        MAPE = APE / _FREQ_ * 100;
     run;

     data &rezultatai.;
        set
           %if &eil_imtis.>1 %then %str(&rezultatai.);
           _tmp;
     run;
  %end;
  proc means data=&rezultatai. noprint;
     var R_2 RMSE MAPE;
     output out=&rezultatai._vid(where=(_STAT_
in('MEAN','STD'))drop=_FREQ_ _TYPE_) ;
  run;

  data _null_;
```

```
      exectime=datetime()-&starttime;
      put;
      put 'Vykdymo trukme ' exectime time.;
    run;
  %mend regresijos_vertinimas;

  %regresijos_vertinimas(imtis1,2,reg_gam(XY,%str(Y=spline (X1)
  loess (X2)/dist=gaussian),YY),work.reg_gam_n16_reg01);
```

Branduolinio regresijos metodo realizacija:

```
%macro KReg(duom, Y, Xs, rez);
   proc iml;
      use &duom;  read all var{&Y} into Y;  read all var{&Xs} into
X; close &duom;
      h=repeat( (4/(2*ncol(X)+1))**(1/(ncol(X)+4))*nrow(X)**(-
1/(ncol(X)+4)), nrow(X), 1);
      start kregpoint(arg,Y,X,h);
         s_KY=0;
         s_K=0;
         do i2=1 to nrow(X);
            if h[i2,] > 1E50 then goto i2_end;
            arg2=(arg-X[i2,])/h[i2,];
            K=exp(-1*arg2*arg2`);
            s_KY=s_KY+K*Y[i2,];
            s_K=s_K+K;
            i2_end:;
         end;
         m=s_KY/s_K;
         return(m);
      finish kregpoint;

     do i1=1 to nrow(X);
         arg=X[i1,];
         m=kregpoint(arg,Y,X,h);
         if i1=1 then Y_P=m;
         else Y_P=Y_P//m;
      end;
      REZ_MATRICA = Y||X||Y_P;
      create &rez. (rename=(Y_P=_Y)) from REZ_MATRICA [colname={
&Y. &Xs. Y_P}];
      append from REZ_MATRICA;
   quit;
 %mend KReg;

%macro regresijos_vertinimas(imtis,d,metodas,rezultatai);
  %let starttime=%sysfunc(datetime());

  proc sql noprint;
     select max(imtis)
        into :imciu_sk
        from &imtis
```

```
    ;
  quit;
  %if %sysfunc(exist(&rezultatai.)) %then
  %do;
     proc sql noprint;
        select max(imtis)+1
           into :eil_imtis_nuo
           from &rezultatai.
        ;
     quit;
  %end;
  %else %let eil_imtis_nuo=1;

  %do eil_imtis=&eil_imtis_nuo %to &imciu_sk;
     %put Imtis: &eil_imtis is %sysfunc(left(&imciu_sk));
     %put Taikomas metodas: &metodas%str(;) duomenys: &imtis;
     data XY (keep=X1-X&d Y);
        set &imtis;
        where imtis=&eil_imtis;
     run;

     %&metodas;

     proc sql noprint;
        create table _tmp as
           select *,
                  (Y-_Y)**2 as SK1,
                  (Y-mean(Y))**2 as SK2,
                  abs((Y-_Y)/Y) as APE
           from YY
        ;
     quit;
     proc summary data=_tmp;
        var SK1 SK2 APE;
        output out=_tmp(drop=_TYPE_) sum()=;
     run;
     data _tmp(keep=IMTIS R_2 RMSE MAPE);
        set _tmp;
        IMTIS = &eil_imtis.;
        R_2 = 1 - SK1 / SK2;
        RMSE = sqrt(SK1 / _FREQ_);
        MAPE = APE / _FREQ_ * 100;
     run;

     data &rezultatai.;
        set
           %if &eil_imtis.>1 %then %str(&rezultatai.);
           _tmp;
     run;
  %end;
  proc means data=&rezultatai. noprint;
     var R_2 RMSE MAPE;
     output out=&rezultatai._vid(where=(_STAT_ in('MEAN','STD'))
drop=_FREQ_ _TYPE_) ;
```

```
  run;

  data _null_;
     exectime=datetime()-&starttime;
     put;
     put 'Vykdymo trukme ' exectime time.;
  run;
%mend regresijos_vertinimas;

   %regresijos_vertinimas(imtis1,2,Kreg(XY,Y, X1
   X2,YY),work.Kreg_16_reg04);
```

Glodinančių splainų metodo realizacija:

```
%macro reg_tpspline(duom,modelis,rez);
 ods _all_ close;
 proc tpspline data=&duom.;
    model &modelis.;
    output out=&rez.(rename=(P_Y=_Y));
 run;
 ods html;
 ods listing;
%mend reg_tpspline;

%macro regresijos_vertinimas(imtis,d,metodas,rezultatai);
  %let starttime=%sysfunc(datetime());

  proc sql noprint;
     select max(imtis)
        into :imciu_sk
        from &imtis
     ;
  quit;
  %if %sysfunc(exist(&rezultatai.)) %then
  %do;
     proc sql noprint;
        select max(imtis)+1
           into :eil_imtis_nuo
           from &rezultatai.;
     quit;
  %end;
  %else %let eil_imtis_nuo=1;

  %do eil_imtis=&eil_imtis_nuo %to &imciu_sk;
     %put Imtis: &eil_imtis is %sysfunc(left(&imciu_sk));
     %put Taikomas metodas: &metodas%str(;) duomenys: &imtis;
     data XY (keep=X1-X&d Y);
        set &imtis;
        where imtis=&eil_imtis;
     run;

     %&metodas;
```

```
      proc sql noprint;
         create table _tmp as
            select *,
                   (Y-_Y)**2 as SK1,
                   (Y-mean(Y))**2 as SK2,
                   abs((Y-_Y)/Y) as APE
            from YY
         ;
      quit;
      proc summary data=_tmp;
         var SK1 SK2 APE;
         output out=_tmp(drop=_TYPE_) sum()=;
      run;
      data _tmp(keep=IMTIS R_2 RMSE MAPE);
         set _tmp;
         IMTIS = &eil_imtis.;
         R_2 = 1 - SK1 / SK2;
         RMSE = sqrt(SK1 / _FREQ_);
         MAPE = APE / _FREQ_ * 100;
      run;

      data &rezultatai.;
         set
            %if &eil_imtis.>1 %then %str(&rezultatai.);
            _tmp;
      run;
   %end;
   proc means data=&rezultatai. noprint;
      var R_2 RMSE MAPE;
      output out=&rezultatai._vid(where=(_STAT_ in('MEAN','STD'))
drop=_FREQ_ _TYPE_) ;
   run;

   data _null_;
      exectime=datetime()-&starttime;
      put;
      put 'Vykdymo trukme ' exectime time.;
   run;
%mend regresijos_vertinimas;

   %regresijos_vertinimas(imtis1,2,reg_tpspline(XY,%str(Y=X1
   (X2)),YY),work.reg_tpspline_n16_reg03);
```

Lokalios regresijos metodo realizacija:

```
%macro reg_loess(duom,modelis,rez);
   ods _all_ close;
   proc loess data=&duom.;
      model &modelis.;
      ods output OutputStatistics=&rez.;
   run;
   ods html;
   ods listing;
%mend reg_loess;
```

```
%macro regresijos_vertinimas(imtis,d,metodas,rezultatai);
  %let starttime=%sysfunc(datetime());

  proc sql noprint;
     select max(imtis)
        into :imciu_sk
        from &imtis
     ;
  quit;
  %if %sysfunc(exist(&rezultatai.)) %then
  %do;
     proc sql noprint;
        select max(imtis)+1
           into :eil_imtis_nuo
           from &rezultatai.
        ;
     quit;
  %end;
  %else %let eil_imtis_nuo=1;

  %do eil_imtis=&eil_imtis_nuo %to &imciu_sk;
     %put Imtis: &eil_imtis is %sysfunc(left(&imciu_sk));
     %put Taikomas metodas: &metodas%str(;) duomenys: &imtis;
     data XY (keep=X1-X&d Y);
        set &imtis;
        where imtis=&eil_imtis;
     run;

     %&metodas;

     proc sql noprint;
        create table _tmp as
           select *,
                  (DepVar-Pred)**2 as SK1,
                  (DepVar-mean(DepVar))**2 as SK2,
                  abs((DepVar-Pred)/DepVar) as APE
           from YY
        ;
     quit;
     proc summary data=_tmp;
        var SK1 SK2 APE;
        output out=_tmp(drop=_TYPE_) sum()=;
     run;
     data _tmp(keep=IMTIS R_2 RMSE MAPE);
        set _tmp;
        IMTIS = &eil_imtis.;
        R_2 = 1 - SK1 / SK2;
        RMSE = sqrt(SK1 / _FREQ_);
        MAPE = APE / _FREQ_ * 100;
     run;

     data &rezultatai.;
        set
```

```
            %if &eil_imtis.>1 %then %str(&rezultatai.);
            _tmp;
      run;
   %end;
   proc means data=&rezultatai. noprint;
      var R_2 RMSE MAPE;
      output out=&rezultatai._vid(where=(_STAT_ in('MEAN','STD'))
drop=_FREQ_ _TYPE_) ;
   run;

   data _null_;
      exectime=datetime()-&starttime;
      put;
      put 'Vykdymo trukme ' exectime time.;
   run;
%mend regresijos_vertinimas;

%regresijos_vertinimas(imtis1,2,reg_loess(XY,%str(Y=X1
X2),YY),work.reg_loess_n16_reg02);
```