2 Skyrius. Tiesinė regresija

Bendra samprata

Šaltinis: Zhou, H. (2023). *Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods*. Berkeley, CA: Apress

Tiesinė regresija yra prognostinis modelis, kuriame mokymo duomenys naudojami tiesiniam modeliui sukurti, kad būtų galima prognozuoti kitus duomenis. Kai kalbame apie tiesinį modelį, turime omenyje, kad ryšys tarp tikslo (priklausomo kintamojo) ir atributo (-ų) (nepriklausomų kintamųjų) yra tiesinis. Regresinėje analizėje įprasta vartoti terminus „nepriklausomas kintamasis“ ir „priklausomas kintamasis“. Todėl šiame skyriuje atributą pakeisime nepriklausomu kintamuoju ir pakeisime priklausomą tikslinį kintamąjį.

Tiesinės regresijos analizėje gali būti vienas ar daugiau nepriklausomų kintamųjų. Kai yra tik vienas nepriklausomas kintamasis, tiesinis modelis išreiškiamas dažniausiai naudojam tiesine funkcija $y = mx + b$, kur y yra priklausomasis kintamasis, m yra tiesės krypties koeficientas, o b nurodo, kuriame taške tiesė kerta y ašį. Daugeliu atvejų yra daugiau nei vienas nepriklausomas kintamasis; todėl tiesinis modelis vaizduojamas kaip $y = m_1x_1 + m_2x_2 + ... + m_nx_n + b$, kur yra n nepriklausomų kintamųjų skaičius, o m_i yra koeficientas, susietas su konkrečiu nepriklausomu kintamuoju x_i, kai i = 1 ... n. Norėdami sukurti tokį tiesinį modelį, turime rasti m_i ir b reikšmes, remiantis žinomomis y ir x_i reikšmėmis mokymo duomenų rinkinyje.

https://mokymai.github.io/biostatistika/tiesine-regresija.html

Regresijos modelis – statistinis modelis, leidžiantis vieno kintamojo reikšmes prognozuoti pagal kito kintamojo reikšmes. Regresija – statistinė vieno atsitiktinio dydžio reikšmių priklausomybė nuo kito – neatsitiktinio – dydžio (arba kelių kitų dydžių), turinti griežtą funkcinį ryšį (t.y., aprašoma lygtimi). Pavyzdžiai:

- pirmas pavyzdys, kuris ateina į galvą – yra kalibravimas, kai paruošiame žinomos koncentracijos tirpalus, supilstome į kiuvetes bei išmatuojame šviesos sugertį. Susidarome koncentracijos ir šviesos sugerties priklausomybės (t.y., kalibracinę) kreivę bei lygtį. Vėliau į mėgintuvėlį įpylę nežinomos koncentracijos tirpalą pagal šviesos sugerties intensyvumą galime nustatyti koncentraciją. Tai dažnas uždavinys per biochemijos laboratorinius darbus;
- regresija gali būti naudojama ir tada, kai, tarkime, norime nustatyti, koks įprastai būna pulsas padarius tam tikrą kiekį pritūpimų;
- koks būna vabzdžio šarvo storis, jei žinome vabzdžio svorį ir ilgį;
- kaip svoris priklauso nuo amžiaus;
- kaip plaučių tūris priklauso nuo per savaitę surūkomų cigarečių skaičiaus.

Daugiau tiesinės regresijos taikymo pavyzdžių įvairiose srityse:

Ekonomika ir finansai:

- ***Akcijų rinkos prognozės.*** Tiesinė regresija naudojama siekiant prognozuoti akcijų kainas remiantis įvairiais ekonominiais rodikliais, tokiais kaip infliacija, palūkanų normos ar ekonomikos augimas.
- ***Kainų nustatymas.*** Modeliai gali padėti prognozuoti prekių ar paslaugų kainas, atsižvelgiant į gamybos sąnaudas, paklausą ir pasiūlą.

Sveikatos apsauga:

- ***Ligų prognozavimas.*** Naudojama analizuoti sveikatos rodiklius, kaip kraujospūdį, cukraus kiekį kraujyje, ir numatyti ligų riziką, pavyzdžiui, širdies ligas ar diabetą.
- ***Medikamentų efektyvumas.*** Gali būti naudojama vertinant, kokią įtaką tam tikri veiksniai (pvz., dozė, paciento amžius) turi gydymo rezultatams.

Rinkodara ir pardavimai:

- ***Reklamos efektyvumas.*** Naudojama analizuoti reklamos išlaidų poveikį pardavimams ir optimizuoti marketingo kampanijas.
- ***Klientų elgsenos analizė.*** Gali būti taikoma siekiant suprasti, kokią įtaką skirtingi veiksniai (pvz., pajamos, amžius) turi klientų pasirinkimams.

Socialiniai mokslai:

- ***Demografiniai tyrimai.*** Tiesinė regresija naudojama siekiant nustatyti, kaip įvairūs demografiniai veiksniai (pvz., amžius, pajamos, išsilavinimas) daro įtaką tokiems rodikliams kaip nedarbo lygis ar šeimos dydis.
- ***Psichologiniai tyrimai.*** Naudojama analizuoti sąsajas tarp psichologinių kintamųjų, pvz., kaip stresas veikia darbo našumą.

Inžinerija ir gamyba:

- ***Kokybės kontrolė.*** Gali būti naudojama prognozuoti gaminių kokybę ar defektus atsižvelgiant į gamybos parametrus.
- ***Išteklių optimizavimas.*** Inžinieriai naudoja regresijos modelius optimizuojant gamybos procesus ir sumažinant sąnaudas.

Aplinkosauga ir klimato kaita:

- ***Klimato pokyčiai.*** Tiesinė regresija naudojama modeliuojant klimato duomenis, pvz., analizuojant temperatūros pokyčių įtaką ledynų tirpimui ar jūros lygio kilimui.
- ***Teršalų analizė.*** Gali būti naudojama analizuoti, kokia tam tikrų veiksnių įtaka (pvz., gamyklų skaičius, transporto priemonių kiekis) oro taršai.

Sportas:

- ***Sporto rezultatų prognozavimas.*** Tiesinė regresija gali būti naudojama numatyti komandos ar individualaus žaidėjo pasirodymus, remiantis tokiais duomenimis kaip ankstesni rezultatai, sužaidžiamų rungtynių skaičius, komandos sudėtis.
- ***Žaidėjų vertinimas.*** Naudojama siekiant nustatyti, kaip tam tikri veiksniai, pvz., treniruočių intensyvumas ar žaidėjo amžius, gali paveikti žaidėjų meistriškumą.

Nekilnojamojo turto vertinimas:

- ***Nekilnojamojo turto kainos.*** Tiesinė regresija dažnai naudojama siekiant nustatyti, kaip veiksniai, pvz., vieta, būsto dydis, metų skaičius nuo statybos, veikia nekilnojamojo turto kainas.

Transportas ir logistika:

- ***Kuro suvartojimo prognozavimas***. Analizuoja, kaip įvairūs veiksniai, pvz., kelio būklė, transporto priemonės amžius, paveikia degalų suvartojimą.
- ***Eismo prognozės.*** Naudojama numatant eismo intensyvumą ir optimizuojant transporto sistemas.

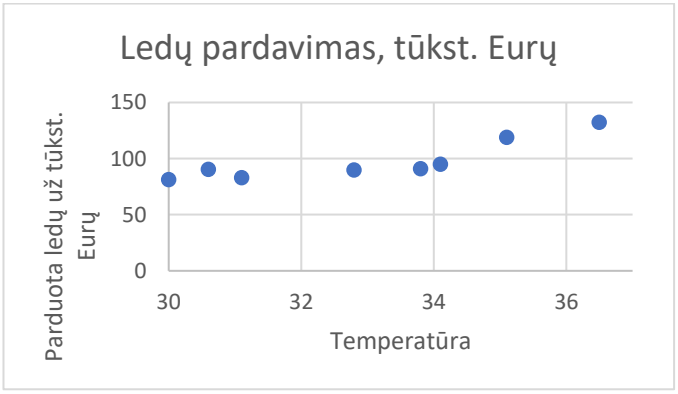
Tiesinė regresija yra efektyvi priemonė, naudojama įvairiose srityse nuo finansų iki sveikatos apsaugos, siekiant spręsti realius uždavinius, prognozuoti ateities įvykius ir priimti argumentuotus sprendimus.

Išnagrinėkime pavyzdį, kaip ledų pardavimas priklauso nuo oro temperatūros (žr. 2-1 lentelės duomenis, 2Skyrius_1.xlsx).

2-1 lentelė. Ledų pardavimų priklausomybė nuo oro temperatūros

	A	B
1	**Temperatūra, ^{0}C**	**Ledų pardavimas, tūkst. Eurų**
2	32,8	89,8
3	30,6	90,2
4	30	81,1
5	31,1	83
6	33,8	90,9
7	35,1	119
8	34,1	94,9
9	36,5	132,4

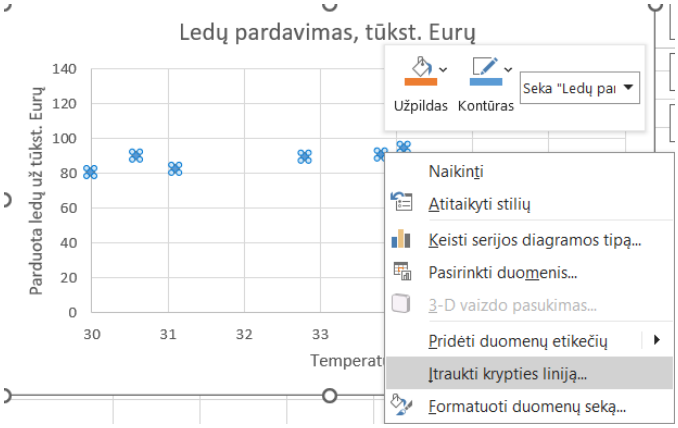
2-1 pav. pateikta pardavimų priklausomybė nuo temperatūros.



2-1 pav. Oro temperatūros ir ledų pardavimų priklausomybė

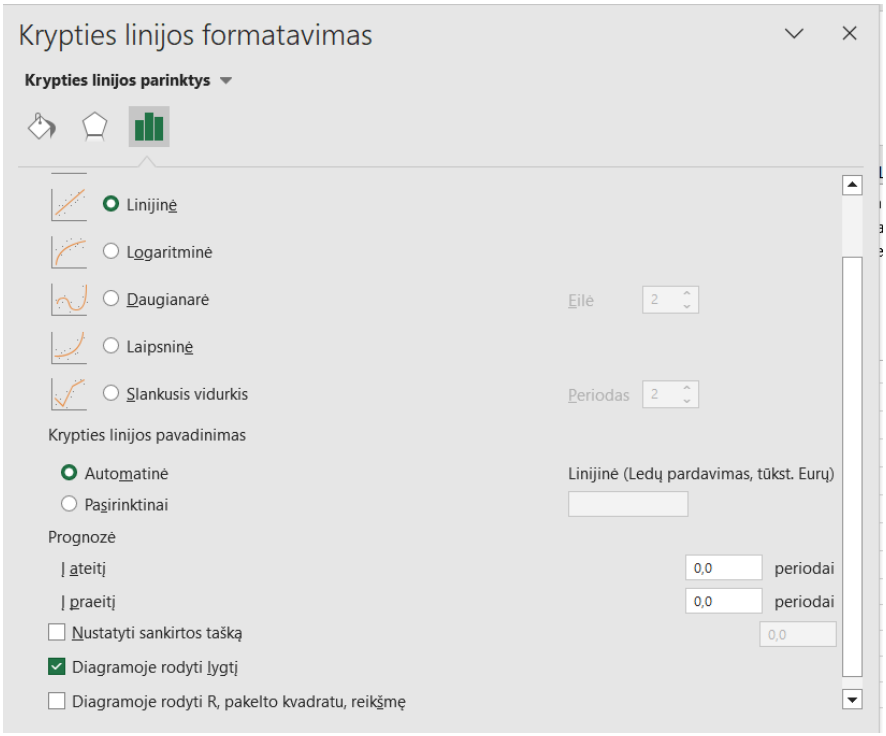
Jei norėtume prognozuoti ledų pardavimus, kai oro temperatūra 32,0 ^{0}C, iš pateiktos diagramos būtų sunku tiksliai nustatyti.

Kairiuoju pelės mygtuku pažymėkite taškus ir spustelėję dešinįjį pelės klavišą kontekstiniame meniu pasirinkite **Įtraukti krypties liniją...** (žr. 2-2 pav.)

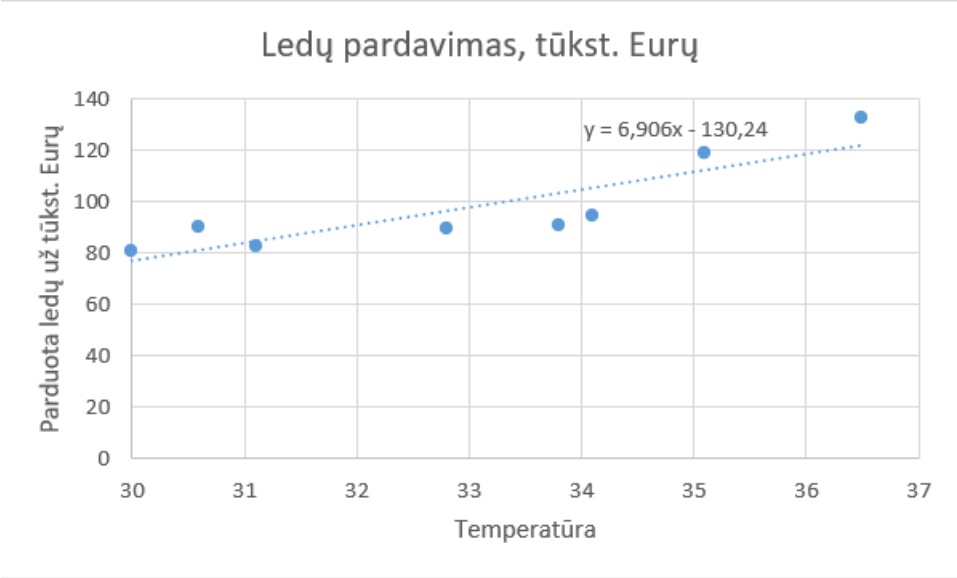


2-2 pav. Tendencijų (krypties) linijos pridėjimas

Pasirinkite formatavimo nustatymus, kaip pateikta 2-3 pav. ir patvirtinę pasirinkimą turėtumėte matyti vaizdą, panašų į pateiktą 2-4 pav.



2-3 pav. Krypties linijos formatavimo pasirinkimas



2-4 pav. Sklaidos diagrama su regresijos lygtimi

Panaudodami regresijos lygtį y = 6,906x – 130,24, galime apskaičiuoti, kokie bus ledų pardavimai y, kai oro temperatūra x = 31 ^{0}C: y = 6,906 * 31 – 130,24 = 83,8 tūkst. Eurų.

Tiesinė lygtis apibrėžia tiesinį duomenų tyrybos modelį šiam specifiniam tiesinės regresijos tyrimui. Suprantame, kad tokio tipo tiesinei lygčiai y = mx + b, m ir b yra koeficientai. Radus m ir b, modelis sukonstruotas ir užbaigiamas. Konstruojant modelį randamos m ir b reikšmės.

2-4 pav. matome, kad tendencijos (krypties) linija eina ne per kiekvieną tašką, ji tik seka tendenciją. Gali būti kelios skirtingos tendencijų linijos, jei keli žmonės ją braižo rankiniu būdu. Kaip „Excel“ suranda šią konkrečią tendencijų liniją?

Norint suformuoti konkrečią tiesinę lygtį y = mx + b forma, naudojamas **mažiausių kvadratų metodas**[1]**.**

Mažiausių kvadratų metodas pagrįstas tuo, kad skaičiuojamos paklaidos tarp numatomos vertės (tendencijos, krypties linijos) ir faktinių duomenų. Taikant šį metodą skaičiuojama paklaidų kvadratų suma:

$$E = \sum_{i=1}^{n} \left(y_i - \left(mx_i + b \right) \right)^2$$

(2-1 formulė)

čia y_i yra ledų pardavimai, kai oro temperatūra yra x_i apmokymo duomenų rinkinyje.

Tikslas yra minizuoti E. Tokiu atveju sprendžiamos lygtys su dalinėmis išvestinėmis:

[1] Mažiausiųjų kvadratų metodą 1794–95 sukūrė C. F. Gaussas ir 1805–06 A.-M. Legendre'as ir pritaikė geodeziniams bei astronominiams matavimams apdoroti. Matematiškai pagrindė A. Markovas, A. Kolmogorovas. (https://www.vle.lt/straipsnis/maziausiuju-kvadratu-metodas/)

$$\frac{\partial E}{\partial x} = 0 \text{ and } \frac{\partial E}{\partial y} = 0$$

(2-2 formulė)

Sprendiniai:

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

(2-3 formulė)

$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

(2-4 formulė)

Tiesinės regresijos realizavimas Excel

1 būdas. Išmoksime apskaičiuoti koeficientus m ir b pagal (2-3) ir (2-4) formules pritaikant Excel funkcijas SUM, SUMPRODUCT ir COUNT. Mūsų pavyzdyje (žr. 2-1 lentelę) oro temperatūra atitinka x, ledų pardavimai – y kintamąjį. Papildykime 2-1 lentelę papildomais skaičiavimais (2-5 pav.).

	A	B	C	D
1	x	y		
2	91	89.8		
3	87	90.2	SUM(X)	=SUM(A:A)
4	86	81.1	SUM(Y)	=SUM(B:B)
5	88	83	SUM(XY)	=SUMPRODUCT(A:A, B:B)
6	92.8	90.9	SUM(X^2)	=SUMPRODUCT(A:A,A:A)
7	95.2	119	n	=COUNT(A:A)
8	93.3	94.9	m	=(D7*D5-D3*D4)/(D7*D6-D3^2)
9	97.7	132.4	b	=(D6*D4-D3*D5)/(D7*D6-D3^2)

2-5 pav. Koeficientų m ir b skaičiavimas taikant mažiausių kvadratų metodą

Skaičiuojant panaudotos Excel funkcijos: SUM – sumos skaičiavimui (čia A;A nurodo, kad sumuojami stulpelio A duomenys, analogiškai galima nurodyti ir stulpelio B duomenis), SUMPRODUCT – sandaugų suma, COUNT – skaitinėmis vertėmis užpildytų langelių skaičius.

Apskaičiavę turėtumėte gauti (2-6 pav.):

	A	B	C	D
1	**Temperatūra, ^{0}C**	**Ledų pardavimas, tūkst. Eurų**		
2	32,8	89,8		
3	30,6	90,2	**SUM(X)**	264
4	30	81,1	**SUM(Y)**	781,3
5	31,1	83	**SUM(XY)**	26037,87
6	33,8	90,9	**SUM(X^2)**	8748,92
7	35,1	119	**n**	8
8	34,1	94,9	**m**	6,906013001
9	36,5	132,4	**b**	-130,235929

2-6 pav. Koeficientai m ir b apskaičiuoti taikant Excel funkcijas SUM, SUMPRODUCT ir COUNT

2 būdas. Koeficientų m ir b skaičiavimas panaudojant Excel funkcijas SLOPE ir INTERCEPT. Langelyje D2 įrašykite formulę =SLOPE(B2:B9;A2:A9), langelyje D3 =INTERCEPT(B2:B9;A2:A9). Turėtumėte matyti rezultatus, kaip pateikta 2-7 pav.

	A	B	C	D	E
1	**Temperatūra, ^{0}C**	**Ledų pardavimas, tūkst. Eurų**			
2	32,8	89,8	**SLOPE(Y,X)**	6,906013001	m
3	30,6	90,2	**NTERCEPT(Y,X**	-130,235929	b
4	30	81,1			
5	31,1	83			
6	33,8	90,9			
7	35,1	119			
8	34,1	94,9			
9	36,5	132,4			

2-7 pav. Koeficientai m ir b apskaičiuoti taikant Excel funkcijas SLOPE ir INTERCEPT.

Antrasis būdas yra lengvesnis, nes nereikia atlikti papildomų skaičiavimų. Tačiau labai svarbu tinkamai nurodyti kintamųjų x ir y blokų koordinates (negalima nurodyti tik stulpelio kaip skaičiuojant pirmuoju būdu).

m ir b koeficientų reikšmes galima apskaičiuoti naudojant Excel įrankį „Data Analysis Tools“ (neišversta į lietuvių kalbą), tačiau trūkumas yra tas, kad pakeitus apmokymo duomenis, m ir b reikšmės automatiškai neatsinaujina.

Daugialypės tiesinės regresijos realizavimas Excel

Papildykime lentelę 2-1 papildomais duomenimis (2Skyrius-2.xlsx).

2-2 lentelė. Ledų pardavimų priklausomybė nuo oro temperatūros, turistų ir saulėtų dienų per savaitę skaičiaus

	A	B	C	D
1	**Temperatūra, ^{0}C**	**Turistų skaičius**	**Saulėtų dienų per savaitę skaičius**	**Ledų pardavimas, tūkst. Eurų**
2	32,8	998	4	89,8
3	30,6	1256	7	90,2
4	30	791	6	81,1
5	31,1	705	5	83
6	33,8	1089	3	90,9
7	35,1	1135	6	119
8	34,1	1076	4	94,9
9	36,5	1198	7	132,4

Įveskite langeliuose A11:A14 tekstą Temperatūra, Turistų skaičius, Saulėtų dienų per savaitę skaičius, y-intercept ir langeliuose B11:B14 skaičius 1, 2, 3, 4 (žr. 2-8 pav.).

	A	B	C	D
1	**Temperatūra, ^{0}C**	**Turistų skaičius**	**Saulėtų dienų per savaitę skaičius**	**Ledų pardavimas, tūkst. Eurų**
2	32,8	998	4	89,8
3	30,6	1256	7	90,2
4	30	791	6	81,1
5	31,1	705	5	83
6	33,8	1089	3	90,9
7	35,1	1135	6	119
8	34,1	1076	4	94,9
9	36,5	1198	7	132,4
10				
11	Temperatūra	1		
12	Turistų skaičius	2		
13	Saulėtų dienų per savaitę skaičius	3		
14	y-intercept	4		

2-8 pav. Duomenys, kurie bus reikalingi tiesinės daugialypės regresijos koeficientų nustatymui

Langelyje C11 įveskite formulę

=INDEX(LINEST(D$2:D$9;A$2:C$9;TRUE;TRUE);1;B11)

Šioje formulėje pirmasis funkcijos LINEST įvesties argumentas yra D2:D9, priklausomos kintamųjų reikšmės. Antrasis įvesties argumentas yra A2:C9, tai blokas, kuriame yra x1, x2 ir x3 reikšmės. Grąžintas masyvas iš LINEST įvedamas į funkciją INDEX kaip įvestis. Kadangi B11 = 1, funkcija INDEX[2] paima pirmąjį elementą grąžintame masyve, kuris yra m_3. LINEST grąžina masyvą, kuris išdėsto koeficientus tokia tvarka: m_n, m_{n-1}, m_{n-2}, …, m_1, b. Šiame konkrečiame pavyzdyje duomenys

[2] https://support.microsoft.com/lt-lt/office/funkcija-index-a5dcf0dd-996d-40a4-a822-b56b061328bd

grąžintame masyve yra m_3, m_2, m_1 ir b nepriklausomiems kintamiesiems Saulėtų dienų skaičius per savaitę, Turistų skaičius, Temperatūra ir y-intercept.

Nukopijuokite formulę iš langelio C11 į langelį C14. Atkreipkite dėmesį į B11, B12, B13 ir B14 naudojimą.

Langelyje C12 ankstesnė formulė automatiškai tampa

=INDEX(LINEST(D$2:D$9;A$2:C$9;TRUE;TRUE),1,B12).

Kadangi B12 = 2, C12 formulė teisingai nustato turistų skaičiaus koeficientą (m_2). Ta pati logika taikoma C13 ir C14 formulėms. Iš anksto įvedę skaičius į langelius B11:B14, teisingai įvedę pradinę formulę, galime gauti visas reikalingas reikšmes automatiškai užpildydami.

Funkcija LINEST[3] pakartotinai iškviečiama kiekvienam koeficientui.

Apskaičiuokime, kokios reikšmės gaunamos prognozuojant. Langelyje E1 įrašykime „Prognozė“. Langelyje E2 įveskite formulę

=A2*C$13+B2*C$12+C2*C$11+C$14

ir ją nukopijuokite žemyn į langelius E3:E9.

Norėdami ištirti tiesinės regresijos modelio kokybę, galime toliau apskaičiuoti paklaidas pagal (2-1) lygtį. F1 langelyje įveskite tekstą „Paklaidos“, o langelyje F2 – formulę =POWER(D2-E2;2), tada automatiškai užpildykite formulėmis bloką F3:F9.

Langelyje D11 įrašykite tekstą „Paklaidų suma“, langelyje E11 – formulę =SUM(F2:F9). Turėtumėte matyti vaizdą, pateiktą 2-9 pav.

[3] https://support.microsoft.com/lt-lt/office/linest-funkcija-linest-84d7d0d9-6e50-4101-977a-fa7abf772b6d

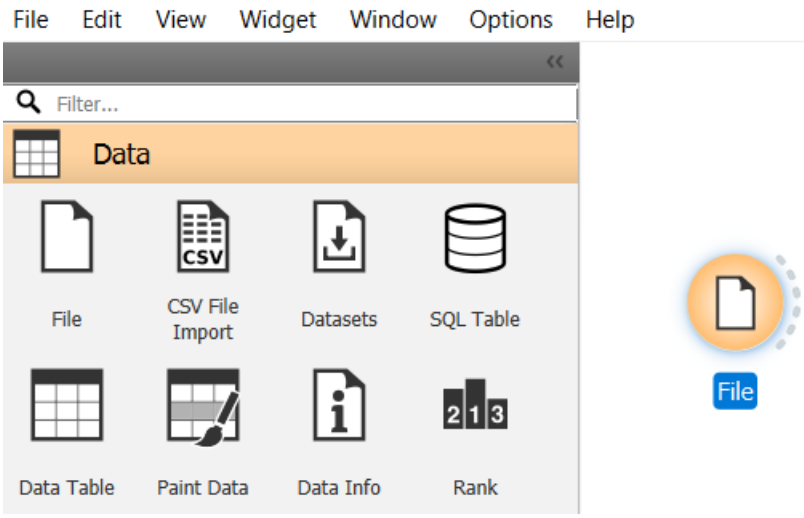
	A	B	C	D	E	F
1	**Temperatūra, ^{0}C**	**Turistų skaičius**	**Saulėtų dienų per savaitę skaičius**	**Ledų pardavimas, tūkst. Eurų**	**Prognozė**	**Paklaidos**
2	32,8	998	4	89,8	88,773508	1,0536855
3	30,6	1256	7	90,2	90,412367	0,0450996
4	30	791	6	81,1	81,092432	5,728E-05
5	31,1	705	5	83	83,187129	0,0350173
6	33,8	1089	3	90,9	89,760346	1,298812
7	35,1	1135	6	119	117,07283	3,7139981
8	34,1	1076	4	94,9	97,966192	9,401534
9	36,5	1198	7	132,4	133,0352	0,4034789
10						
11	Temperatūra	1	6,016574579	Paklaidų suma	15,951683	
12	Turistų skaičius	2	-0,002189783			
13	Saulėtų dienų per savaitę skaičius	3	7,202682329			
14	y-intercept	4	-169,3553671			
15						

2-9 pav. Daugialypė tiesinė regresija

Šiame pavyzdyje turistų skaičiaus įtaka ledų pardavimams yra neigiama, t.y., didėjant turistų skaičiui ledų parduodama mažiau. Tačiau ši reikšmė yra artima nuliui, todėl galima teigti, kad turistų skaičius įtakos ledų pardavimams neturi.

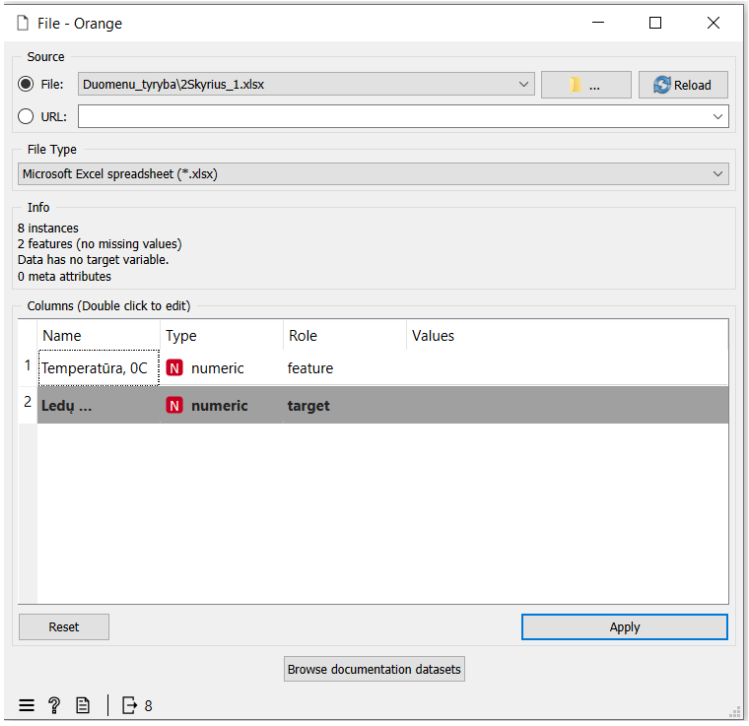
Tiesinės regresijos realizavimas Orange

1. Atvėrę Orange programą, valdiklių lange pasirinkite File ir paspaudę kairįjį pelės klavišą nutempkite File į darbo lauką (žr. 2-10 pav.).



2-10 pav. Komponento File pasirinkimas iš valdiklių Data grupės

2. Kairiuoju pelės klavišu dukart spragtelkite File darbo lauke ir atsivėrusiame lange pasirinkite failą 2Skyrius_1.xlsx ir nustatę tinkamas parametrų reikšmes paspauskite mygtuką Apply (žr. 2-11 pav.).

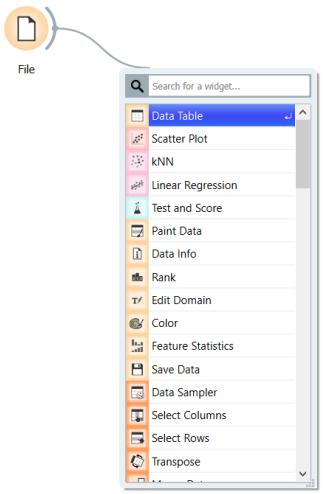


2-11 pav. Failo ir jame esančių duomenų parametrų nustatymas

Paspaudę mygtuką Apply, langą pašalinkite paspausdami lango pašalinimo mygtuką dešinėje viršuje.

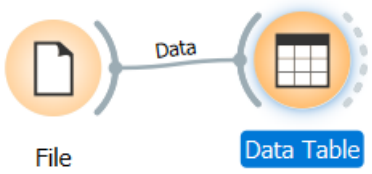
3. Paspaudę kairįjį pelės klavišą ties punktyrine linija 

vilkite ir atleidę turėtumėte matyti vaizdą, panašų į pateiktą 2-12 pav.



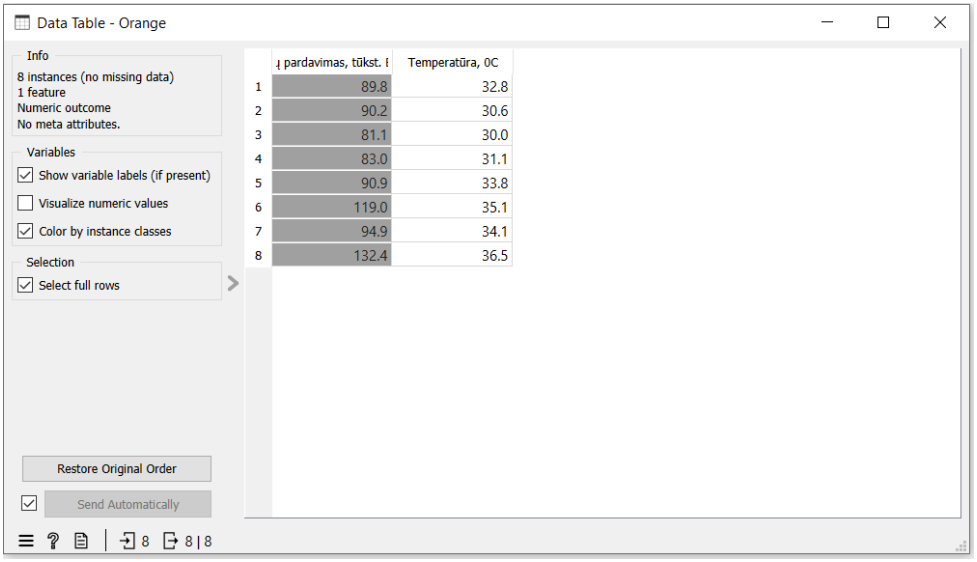
2-12 pav. Išsiskleidžiantis meniu su pasirinkimais

4. Pasirinkite komponentą Data Table (jei komponento nesimato meniu, jį galite greitai surasti įvedę Data Table į meniu paieškos laukelį). Jei viską atlikote tinkamai, turėtumėte matyti vaizdą, panašų į pateiktą 2-13 pav.



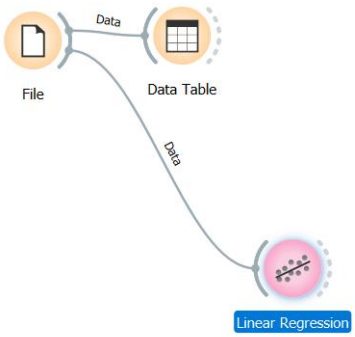
2-13 pav. Sujungti File ir Data Table

5. Kairiuoju pelės klavišu dukart spragtelkite ant Data Table ir turėtumėte matyti vaizdą, panašų į 2-14 pav.



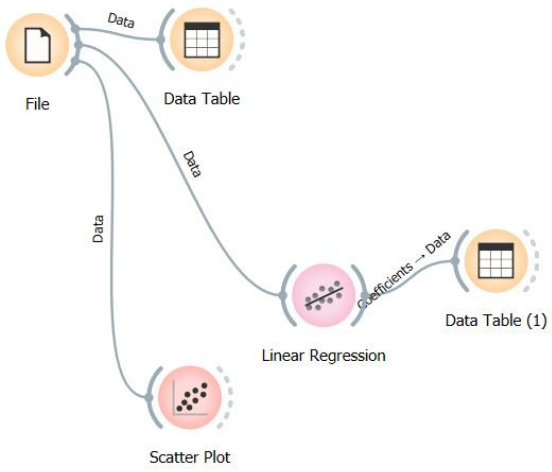
2-14 pav. Pradiniai duomenys

6. Komponentą File pakartodami 3-4 žingsnius sujunkite su tiesinės regresijos komponentu Linear Regression (žr. 2-15 pav.).



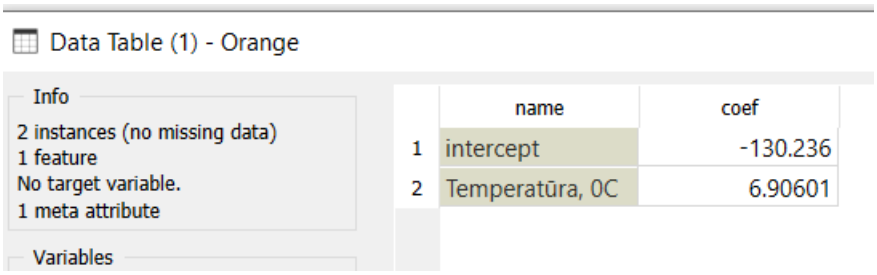
2-15 pav. Tiesinės regresijos komponento prijungimas

Kad matytume tiesinės regresijos rezultatus, prie tiesinės regresijos komponento reikia prijungti Data Table, o prie File Scatter Plot (žr. 2-16 pav.). Pastaba: jungiant Data Table komponentą pasirodys informacinis langas. Pasirinkite OK ir langas užsivers.



2-16 pav. Rezultatus pateikiančių komponentų prijungimas

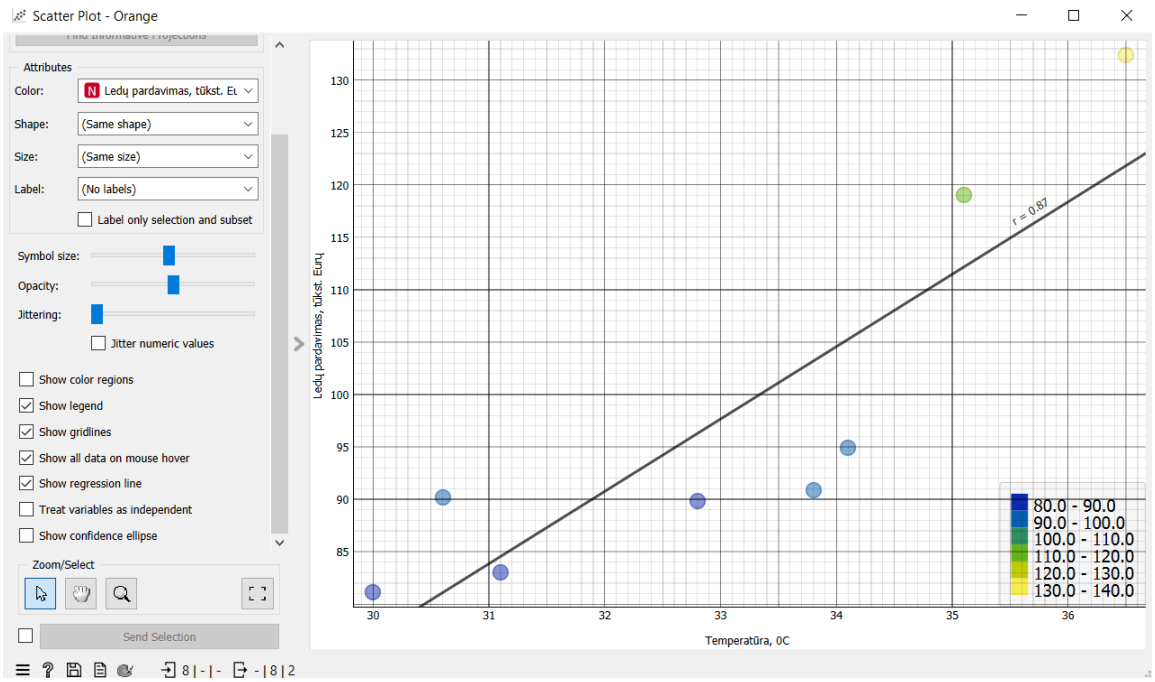
7. Kairiuoju pelės klavišu dukart spragtelkite ant Data Table (1) ir pamatysite tiesinės regresijos rezultatus (žr. 2-17 pav.).



2-17 pav. Tiesinės regresijos koeficientai

Palyginkite rezultatus su rezultatais, kuriuos gavote atlikdami užduotį Excel. Rezultatai sutampa.

Grafinę ledų pardavimo priklausomybę nuo temperatūros matysime dukart kairiuoju pelės klavišu spragtelėję Scatter Plot (žr. 2-18 pav.).

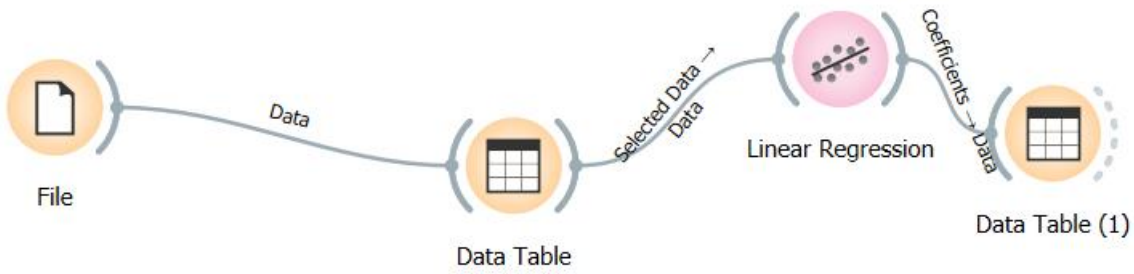


2-18 pav. Tiesinės regresijos grafinis vaizdavimas

Pilnas tiesinės regresijos modelis yra faile 2Skyrius_1.ows.

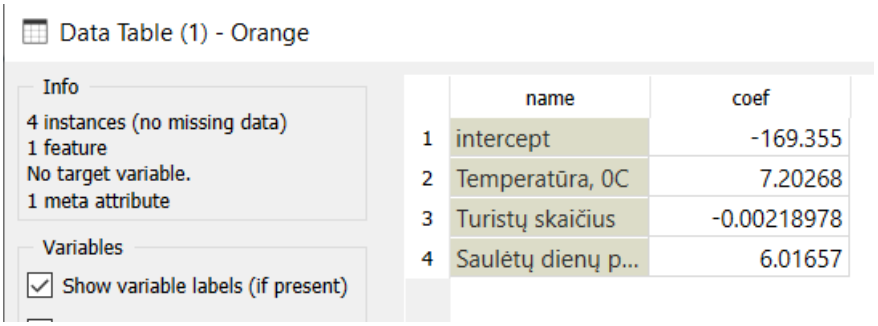
Daugialypės tiesinės regresijos realizavimas Orange

Daugialypė tiesinė regresija Orange realizuojama labai panašiai kaip ir tiesinė regresija. Atlikite 1-6 veiksmus, tik su naujais duomenimis, kurie saugomi faile 2Skyrius_2.xlsx. Nepamirškite, kad Ledų pardavimai yra tikslo kintamasis (jis turi būti target). Pilnai parengtas daugialypės regresijos modelis pateiktas 2-19 pav. ir 2Skyrius_2.ows faile.



2-19 pav. Daugialypės tiesinės regresijos modelis

Daugialypės regresijos rezultatai pateikiami 2-20 pav. Kaip matote, jie sutampa su rezultatais, gautais atliekant užduotį Excel.



2-20 pav. Daugialypės regresijos rezultatai

Paklaidų priežastys atliekant tiesinę regresiją

Atliekant tiesinę regresiją, paklaidos yra neišvengiamos ir atsiranda dėl įvairių priežasčių. Paklaidos rodo skirtumą tarp realių duomenų ir modelio prognozių. Štai kelios pagrindinės priežastys, dėl kurių gali atsirasti paklaidos atliekant tiesinę regresiją:

- Kartais regresijos modelis neapima visų reikšmingų kintamųjų, kurie gali turėti įtakos priklausomam kintamajam. Pvz., jei prognozuojame būsto kainą, bet neįtraukiame kintamųjų, tokių kaip kriminogeninė padėtis ar mokyklų kokybė, modelio tikslumas gali sumažėti, o prognozavimo paklaidos padidėti. Jei reikšmingi veiksniai neįtraukti, modelis negali tiksliai prognozuoti rezultatų.
- Jei realus duomenų ryšys nėra tiesinis, bet pasirenkamas tiesinis modelis, prognozės bus netikslios. Tokiais atvejais reikalingi sudėtingesni modeliai (pvz., polinominė regresija), kurie geriau atspindėtų realų duomenų ryšį.
- Kai kurie kintamieji gali būti neteisingai išmatuoti ar užfiksuoti. Pvz., klaidos gali kilti dėl prietaisų netikslumų, subjektyvių vertinimų ar duomenų įvedimo klaidų. Šios klaidos tiesiogiai padidina paklaidas ir sumažina modelio tikslumą.
- Paklaidos gali atsirasti dėl veiksnių, kurie negali būti tiksliai numatyti arba išmatuoti, tačiau daro įtaką priklausomam kintamajam. Pvz., ekonominiai sukrėtimai, gamtos nelaimės ar netikėti visuomeniniai pokyčiai gali sukelti nenumatytas pokyčių tendencijas, kurios neatitinka modelio.
- Duomenys natūraliai kinta, todėl neišvengiama, kad atsiranda tam tikra atsitiktinė paklaida, net jei modelis gerai atspindi pagrindinę tendenciją.
- Tiesinės regresijos modelis daro prielaidą, kad paklaidų dispersija yra pastovi visoms nepriklausomų kintamųjų reikšmėms. Jei paklaidos didėja arba mažėja priklausomai nuo kintamųjų reikšmių, modelio tikslumas gali sumažėti.
- Jei nepriklausomi kintamieji yra stipriai susiję tarpusavyje, regresijos koeficientai tampa nepatikimi. Tai gali sukelti didesnes paklaidas prognozuojant rezultatus ir sunkinti modelio interpretaciją.
- Jei duomenų rinkinyje yra ekstremalių arba išskirtinių taškų (anomalijų), jie gali neproporcingai paveikti tiesinę regresiją, todėl bendros tendencijos tampa klaidingos. Šiuos taškus būtina identifikuoti ir įvertinti, ar tai klaidingi duomenys, ar anomalija.

2 skyriaus užduotys

Užduotys, skirtos įgūdžiams įtvirtinti, pateiktos faile ***2Skyrius-SD.xlsx*** *(užduotys adaptuotos iš* Čekanavičius, V., & Murauskas, G. (2006). Statistika ir jos taikymai, II dalis. *TEV, Vilnius.).*

Pastaba: atlikdami užduotis Orange kiekvienai užduočiai pasidarykite atskirus failus, nusikopijuodami pradinius duomenis iš failo 2Skyrius-SD.xlsx.

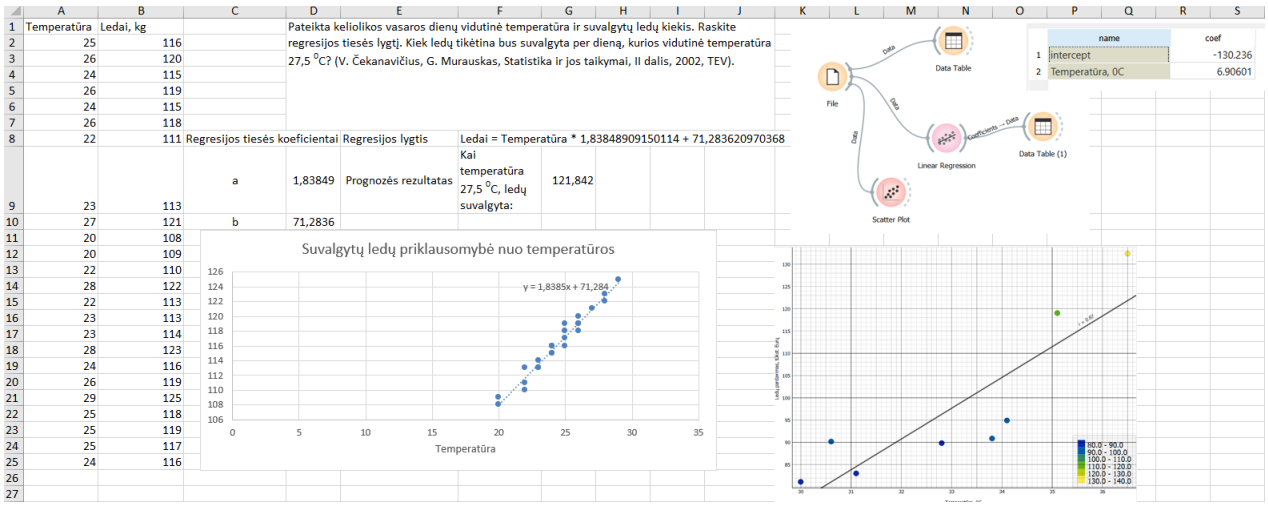
Antro skyriaus 1-7 užduotis (tiesinė regresija) atlikite Excel ir Orange, rezultatus pateikite apskaičiuodami tiesinės regresijos lygties koeficientus ir grafiškai (tiesinės regresijos lygtis).

8-10 užduotis (daugialypė tiesinė regresija) atlikite pasirinktu įrankiu (Excel arba Orange).

1. Pateikta keliolikos vasaros dienų vidutinė temperatūra ir suvalgytų ledų kiekis. Raskite regresijos tiesės lygtį. Kiek ledų tikėtina bus suvalgyta per dieną, kurios vidutinė temperatūra 27,5 ^{0}C?

	A	B
1	Temperatūra	Ledai, kg
2	25	116
3	26	120
4	24	115
5	26	119
6	24	115
7	26	118
8	22	111
9	23	113
10	27	121
11	20	108
12	20	109
13	22	110
14	28	122
15	22	113
16	23	113
17	23	114
18	28	123
19	24	116
20	26	119
21	29	125
22	25	118
23	25	119
24	25	117
25	24	116

Atliktos užduoties pavyzdys:



	A	B	C	D	E	F	G	H	I	J
1	Temperatūra	Ledai, kg		Pateikta keliolikos vasaros dienų vidutinė temperatūra ir suvalgytų ledų kiekis. Raskite regresijos tiesės lygtį. Kiek ledų tikėtina bus suvalgyta per dieną, kurios vidutinė temperatūra 27,5 °C? (V. Čekanavičius, G. Murauskas, Statistika ir jos taikymai, II dalis, 2002, TEV).						
2	25	116								
3	26	120								
4	24	115								
5	26	119								
6	24	115								
7	26	118								
8	22	111	Regresijos tiesės koeficientai		Regresijos lygtis		Ledai = Temperatūra * 1,83848909150114 + 71,283620970368			
9	23	113	a	1,83849	Prognozės rezultatas	Kai temperatūra 27,5 °C, ledų suvalgyta:	121,842			
10	27	121	b	71,2836						
11	20	108								
12	20	109								
13	22	110								
14	28	122								
15	22	113								
16	23	113								
17	23	114								
18	28	123								
19	24	116								
20	26	119								
21	29	125								
22	25	118								
23	25	119								
24	25	117								
25	24	116								
26										
27										

2. Pateikta keliolikos metų avarijų ir jose sužeistų žmonių skaičius. Raskite sužeistųjų skaičiaus priklausomybę nuo avarijų skaičiaus.

	A	B	C
1	Metai	Avarijų skaičius	Sužeistųjų skaičius
2	1990	5135	5491
3	1991	6067	6638
4	1992	4049	4251
5	1993	4319	4555
6	1994	3902	4146
7	1995	4144	4508
8	1996	4579	5223
9	1997	5319	6198
10	1998	6445	6356
11	1999	7669	7696

3. Visuomeninis centras „Madam“ kreipėsi į vyriausybę su prašymu visus valdininkus lankyti jų centro organizuojamus džentelmeniškumo kursus. Sudarykite tiesinės regresijos modelį ir nustatykite, kiek komplimentų pasakys vyras treniravęsis 267 kartus. Lentelėje pateikti 32 vyrų duomenys.

	A	B
1	Treniruotės	Komplimentai
2	254	124
3	230	108
4	254	85
5	300	152
6	320	140
7	364	198
8	312	182
9	264	125
10	274	130

4. Pateiktas automobilių amžius ir mechaninių gedimų taisymo trukmė. Kokią taisymo trukmę galima prognozuoti 17 metų amžiaus automobiliui?

	A	B
1	Automobilio amžius	Taisymo trukmė
2	1	2,5
3	2	3
4	6	9,5
5	8	11
6	3	4
7	2	2,5
8	11	15
9	15	20,5
10	8	10,5
11	4	5,5
12	6	9
13	12	16,5
14	9	12,5
15	10	13,5
16	3	4,5
17	4	5,5

5. Pateikta vandens temperatūra skirtingomis mėnesio dienomis ir maksimalus besimaudančiųjų skaičius. Parenkite regresijos modelį. Lentelėje yra 23 įrašai.

	A	B
1	Vandens temperatūra	Besimaudančiųjų skaičius
2	17	79
3	18	83
4	16	78
5	18	82
6	16	78
7	18	81
8	14	74
9	15	76
10	19	84

6. Medicinos studentas tyrė kraujospūdžio ir insulto rizikos balo priklausomybę. Kokį rizikos balą galima prognozuoti pacientui, kurio kraujospūdis yra 180? Lentelėje yra 22 įrašai.

	A	B
1	Kraujospūdis	Insulto rizika
2	150	27
3	162	40
4	155	28
5	173	60
6	195	43
7	188	55
8	155	33
9	120	43
10	136	49

7. Buvo tiriama, kaip greitai žmonės suranda tekste nurodytą specialų simbolį. Prognozuokite, kokia bus paieškos trukmė, kai simbolis yra 20 eilutėje. Lentelėje yra 16 įrašų.

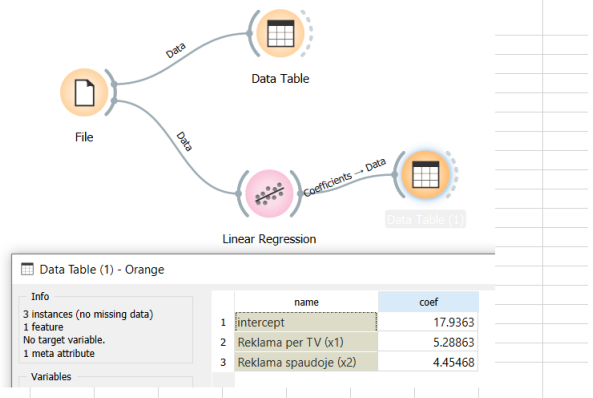
	A	B
1	Eilutė	Paieškos trukmė
2	8	7
3	11	8
4	15	8
5	18	6
6	19	10
7	21	15
8	23	13
9	24	14
10	25	16

8. Lentelėje pateikti duomenys, kiek tūkst. eurų išleista produkto X reklamai TV bei spaudoje. Raskite regresijos funkciją. Kokį pelną tikėtina gauti, išleidus reklamai TV 7, o reklamai spaudoje 5 tūkst. eurų?

	A	B	C
1	Reklama per TV (x_1)	Reklama spaudoje (x_2)	Pelnas (y)
2	8,5	6,0	93,0
3	8,0	1,0	61,0
4	8,5	5,0	80,0
5	7,5	1,9	66,5
6	6,7	4,9	72,0
7	7,7	3,0	77,5
8	5,8	4,0	66,5
9	7,5	4,1	72,0
10	8,4	7,0	95,1
11	7,9	1,0	62,1
12	8,4	4,2	84,1
13	7,4	2,2	67,7
14	6,6	4,2	73,1
15	7,6	3,1	78,6
16	5,7	4,0	67,6
17	7,4	5,1	73,1

Atliktos užduoties pavyzdys:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Reklama per TV (x_1)	Reklama spaudoje (x_2)	Pelnas (y)	Lentelėje pateikti duomenys, kiek tūkst. eurų išleista produkto X reklamai TV bei spaudoje. Raskite regresijos funkciją.														
2	8,5	6,0	93,0	Kokį pelną tikėtina gauti, išleidus reklamai TV 7, o														
3	8,0	1,0	61,0	reklamai spaudoje 5 tūkst. eurų? (V. Čekanavičius, G.														
4	8,5	5,0	80,0	Murauskas, Statistika ir jos taikymai, II dalis, 2002,														
5	7,5	1,9	66,5	TEV).														
6	6,7	4,9	72,0															
7	7,7	3,0	77,5															
8	5,8	4,0	66,5															
9	7,5	4,1	72,0															
10	8,4	7,0	95,1															
11	7,9	1,0	62,1															
12	8,4	4,2	84,1															
13	7,4	2,2	67,7															
14	6,6	4,2	73,1															
15	7,6	3,1	78,6															
16	5,7	4,0	67,6															
17	7,4	5,1	73,1															
18																		
19	Reklama spaudoje	1	4,4546837															
20	Reklama per TV	2	5,2886344				Regresijos lygtis		Reklama per TV (x1)*4,45468371967052+Reklama spaudoje (x2)*5,28863435684246+17,9362518211025									
21	b	3	17,936252				Tikėtinas pelnas		77,23011									



9. Lentelėje pateikti duomenys apie vadybininkų IQ, patirtį mėnesiais ir gaunamą atlyginimą. Raskite regresijos funkciją. Koks atlyginimas prognozuojamas vadybininkui, kurio IQ yra 120, o patirtis 18 mėnesių? Lentelėje yra 22 įrašai. Aptarkite gautą rezultatą. Atlikite analizę, kaip gaunamas atlyginimas priklauso nuo kiekvieno iš parametrų atskirai.

	A	B	C
1	IQ (x_1)	Patirtis (x_2)	Atlyginimas (y)
2	149	49	5000,0
3	159	21	2000,0
4	154	39	4000,0
5	113	26	2500,0
6	144	29	3000,0
7	112	36	3500,0
8	114	25	2500,0
9	113	31	3000,0
10	111	50	5100,0

10. Lentelėje pateikti duomenys apie parengiamųjų biologijos žinių testų ir egzamino rezultatus. Raskite regresijos funkciją, prognozuojančią egzamino rezultatus. Koks tikėtinas egzamino rezultatas, kai studentas už pirmąjį testą gavo 15, o už antrąjį 13,5 taško įvertinimą? Lentelėje yra 39 įrašai.

	A	B	C
1	I testas (x_1)	II testas (x_2)	Egzaminas (y)
2	13,3	13,8	120
3	13,3	14,1	123
4	13,4	14,6	124
5	13,7	14,9	130
6	13,9	15,3	134
7	14,2	15,6	136
8	14,4	15,6	136
9	14,5	15,6	134
10	14,7	15,4	133

Daugiau šaltinių ir duomenų rinkinių, skirtų tiesinės regresijos užduotims

https://smartone.ai/blog/10-best-open-source-datasets-for-linear-regression/
https://www.telusinternational.com/insights/ai-data/article/10-open-datasets-for-linear-regression
https://www.interviewquery.com/p/regression-datasets-and-projects
https://hackernoon.com/top-10-regression-datasets-for-machine-learning-projects-ce4i3wuu
https://imerit.net/blog/25-open-datasets-for-data-science-projects-all-pbm/
https://www.justintodata.com/linear-regression-machine-learning-python-tutorial/