8 Paskaita. Koreliacinė analizė. Porinė tiesinė regresija.

8.1 Koreliacijos koeficiento taškinis įvertis.

Tarkime, stebime kiekybinius kintamuosius X ir Y. Turime atsitiktinę imtį $(X_1, Y_1), \ldots, (X_n, Y_n)$. Norime įvertinti atsitiktinių dydžių X ir Y tiesinės priklausomybės matą – Pirsono koreliacijos koeficientą ρ_{XY}. Koreliacijos koeficiento ρ_{XY} taškiniam įverčiui (arba empiriniam koreliacijos koeficientui) skaičiuoti naudojama formulė

$$R = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{(n\sum X_i^2 - (\sum X_i)^2)\,(n\sum Y_i^2 - (\sum Y_i)^2)}}. \tag{1}$$

Koreliacijos koeficientas, apskaičiuotas iš imties duomenų, matuoja dviejų kiekybinių kintamųjų tiesinio ryšio stiprumą ir šio ryšio kryptį. Imties koreliacijos koeficientas žymimas simboliu yra R (jo realizacija – simboliu r). Koreliacijos koeficiento įgyjamos reikšmės yra nuo -1 iki +1. Jei tarp kintamųjų yra stiprus teigiamas tiesinis ryšys, r reikšmė bus artima +1. Jei yra stiprus neigiamas tiesinis ryšys tarp kintamųjų, r reikšmė bus artima -1. Kai nėra kintamųjų tiesinio ryšio arba yra tik silpnas tiesinis ryšys, r reikšmė bus artima 0. Koreliacijos koeficiento R realizacija r turi šias savybes:

1. Jei koreliacijos koeficientas $r > 0$, tai didėjant X, Y taip pat didėja. Jei $r < 0$, tai didėjant X, Y mažėja.
2. r nepriklauso nuo X ir Y matavimo skalių. Pavyzdžiui, visus X padauginus iš 1000, r reikšmė nepasikeis.
3. $r(X, Y) = r(Y, X)$.
4. r neparodo netiesinės priklausomybės.
5. r priklauso nuo duomenų homogeniškumo, t.y., kuo X arba Y reikšmės vienodesnės, tuo r mažesnis. Jei visi x_i vienodi, tai $r = 0$.
6. r yra suderintasis ρ įvertis, t.y. kuo didesnė imtis, tuo r yra artimesnis tikrajai koreliacijos koeficiento reikšmei ρ.
7. Koreliacijos koeficientas r nenusako priežastingumo.

Pavyzdys. Apskaičiuokite turtingiausių JAV žmonių amžiaus ir jų turto empirinio koreliacijos koeficiento reikšmę pagal pateiktus 1 lentelėje duomenis: Norėdami išspręsti uždavinį užpildykime pagalbinę 2 lentelę. Įstatę reikšmes iš paskutinės eilutės į formulę (1), gausime:

$$R = \frac{8 \cdot 13363.6 - (519)(208.1)}{\sqrt{(8 \cdot 34227 - (519)^2)\,(8 \cdot 6495.41 - (208.1)^2)}} = -0.176.$$

1 lentelė: Turtingiausių JAV žmonių amžiaus ir jų turto duomenys

Person	Age x	Net wealth y
A	73	16
B	65	26
C	53	50
D	54	21.5
E	79	40
F	69	16
G	61	19.6
H	65	19

2 lentelė: Pagalbinė lentelė

Person	Age x	Net wealth y	xy	x^2	y^2
A	73	16	1168	5329	256
B	65	26	1690	4225	676
C	53	50	2650	2809	2500
D	54	21.5	1161	2916	462.25
E	79	40	3160	6241	1600
F	69	16	1104	4761	256
G	61	19.6	1195.6	3721	384.16
H	65	19	1235	4225	361
Iš viso:	$\sum x = 519$	$\sum y = 208.1$	$\sum xy = 13363.6$	$\sum x^2 = 34227$	$\sum y^2 = 6495.41$

r reikšmė rodo labai silpną neigiamą ryšį tarp kintamųjų. Ar šis ryšys yra statistiškai reikšmingas, išsiaiškinsime kitame skyrelyje.

8.2 Hipotezė apie koreliacijos koeficiento lygybę nuliui.

Norime patikrinti, ar dviejų atsitiktinių dydžių X ir Y koreliacijos koeficientas $\rho_{X,Y}$ reikšmingai skiriasi nuo nulio. Tikrinama hipotezė:

$$\begin{cases} H_0 : \rho_{X,Y} = 0, \\ H_1 : \rho_{X,Y} \neq 0. \end{cases}$$

Žinoma, kad statistika

$$T = R\sqrt{\frac{n-2}{1-R^2}}$$

turi Stjudento skirstinį su $n-2$ laisvės laipsniais, jei H_0 teisinga. Tarkime, kad α yra reikšmingumo lygmuo. Kritinės sritys nurodytos 3-je lentelėje.

3 lentelė: Kritinės sritys

H_1	H_0 atmetama
$\rho_{X,Y} \neq 0$	$\lvert t\rvert > t_{\alpha/2}(n-2)$
$\rho_{X,Y} < 0$	$t < -t_{\alpha}(n-2)$
$\rho_{X,Y} > 0$	$t > t_{\alpha}(n-2)$

Patikrinsime hipotezę apie koreliacijos tarp JAV žmonių amžiaus ir jų turto koeficiento lygybės nuliui su dvipuse alternatyva ir $\alpha = 0.05$.

$$t = -0.176\sqrt{\frac{8-2}{1-(-0.176)^2}} \approx -0.44, t_{0.025}(6) = 2.447.$$

Kadangi $t = -0.44 \notin W = (-\infty; -2.447) \cup (2.447; +\infty)$, H_0 neatmetama, t.y. galime laikyti, kad koreliacijos koeficientas tarp JAV žmonių amžiaus ir jų turto lygus nuliui.

8.3 Porinė tiesinė regresija. Mažiausių kvadratų metodas.

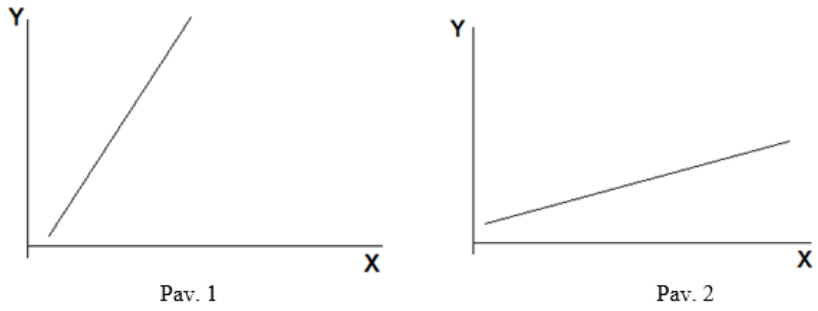
Imties koreliacijos koeficientas rodo tiesinio ryšio tarp dviejų stebimų kintamųjų stiprumą. Tačiau koreliacijos koeficiento žinojimo nepakanka, norint prognozuoti, daryti statistines išvadas.

Pavyzdys. Tarkime, žinome, kad pajamos iš produkcijos pardavimo ir išlaidos reklamai labai stipriai koreliuoja tarpusavyje. Tačiau, negalime pasakyti, kiek padidės pardavimų apimtis, išlaidas reklamai padidinus 1000 eurų. Žinant tik koreliacijos koeficiento dydį, negalime atsakyti į klausimą, ar apsimoka skirti pinigų produkcijos reklamai. 1 pav. (apsimoka) 2 pav.(neapsimoka), nors koreliacijos koeficiento dydžiai abiem atvejais nesiskiria.

Regresinė analizė turi tikslą nustatyti kelių dydžių funkcinį ryšį. Kitas regresinės analizės tikslas yra priklausomo kintamojo Y reikšmių prognozavimas pagal nepriklausomo kintamojo X reikšmes. Imkime pavyzdį, kai norima nustatyti ryšį tarp namų ūkio pajamų ir išlaidų maistui. Nustatę funkcinės priklausomybės formą ir įvertinę modelio parametrus galėsime atsakyti į klausimus:

- kiek vidutiniškai pasikeistų išlaidos maistui namų ūkio pajamoms padidėjus 100 eurų/mėn.?
- ar gali vidutinės išlaidos maistui sumažėti padidėjus pajamoms?
- kokias prognozuotume vidutiniškas išlaidas maistui namų ūkiams, kurių pajamos per mėnesį sieks 800 eurų?

Šiuo atveju kintamasis Y – namų ūkio išlaidos maistui, vadinamas endogeniniu (paaiškinamuoju) kintamuoju, kintamasis X – namų ūkio pajamos, yra



1 pav.: X – išlaidos reklamai, Y – produkcijos pardavimai

egzogeninis (paaiškinananantysis). *Paprastas tiesinis regresinis modelis* aprašomas lygtimi:

$$Y = \beta_1 + \beta_2 X + \varepsilon, \tag{2}$$

Abejose šios lygties pusėse yra atsitiktiniai dydžiai. Fiksuokime namų ūkio pajamas $X = 700$ eurų/mėn. ir apklauskime n atrinktų namų ūkių, turinčių tokias mėnesines pajamas, gausime imties realizaciją $y_1, y_2, ..., y_n$. Galime suskaičiuoti imties sąlyginį vidurkį su sąlyga $X = 700$, tai yra teorinio sąlyginio vidurkio $\mu_{Y|700} = E(Y|X = 700)$ įvertį $\hat{\mu}_{Y|700} = \frac{1}{n}\sum_i y_i$. Panašiai galime fiksuoti pajamas ties tam tikra konkrečia reikšme ir skaičiuoti išlaidų maistui sąlyginio vidurkio įvertį kitai namų ūkio pajamų X reikšmei, pvz., $\hat{\mu}_{Y|800}$. Bendru atveju vidutinių išlaidų maistui $\mu_{Y|x} = E(Y|X = x)$, kai pajamos fiksuotos $X = x$, priklausomybė nuo pajamų aprašoma lygtimi:

$$E(Y|X = x) = \beta_1 + \beta_2 x. \tag{3}$$

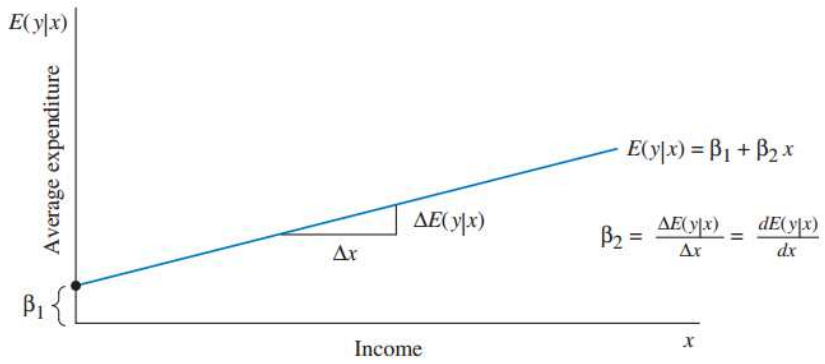
Funkcinė priklausomybė gali būti ir kitokia. Dažnai funkcinio ryšio formą nustatyti padeda duomenų grafinis atvaizdavimas. Funkcija (3) vadinama *paprastos tiesinės regresijos funkcija*, o nežinomi parametrai β_1, β_2 – *regresijos parametrais*. β_1 vadinamas *laisvuoju nariu (angl.* intercept). Jis reiškia namų ūkio, kurio mėnesio pajamos yra nulinės, vidutines išlaidas maistui. β_2 vadinamas *marginaliu polinkiu išleisti maistui (angl.* slope). Jis reiškia vidutinišką išlaidų maistui pasikeitimą, pajamoms pakitus 1 eurų. β_2 lygus sąlyginio vidurkio $E(Y|X = x)$ išvestinę x atžvilgiu:

$$\beta_2 = \frac{\Delta E(Y|x)}{\Delta x} = \frac{dE(Y|x)}{dx}.$$

Geometrinė regresijos koeficientų interpretacija pavaizduota pav. 2

8.4 Paprasto regresinio modelio prielaidos

Norėdami įvertinti regresijos parametrus, pirmiausia renkame duomenis apie namų ūkių išlaidas maistui ir mėnesines pajamas, t.y. $(Y_t, X_t), t = 1, 2, ..., T$.



2 pav.: Paprastas tiesinis regresinis modelis

(Y_i, X_i) yra i-jo namų ūkio duomenys. Šios imties generavimo modelis yra

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t, t = 1, 2, ..., T. \tag{4}$$

Regresinės analizės esmė yra tai, kad priklausomas kintamasis Y skaidomas į 2 komponentes: sisteminę ir atsitiktinę. Sisteminė Y komponentė yra jo vidurkis $E(Y|x) = \beta_1 + \beta_2 x$, kuris nėra atsitiktinis dydis. Atsitiktinė Y komponentė yra skirtumas tarp Y ir jo sąlyginio vidurkio $E(Y|x)$. Ji vadinama *atsitiktiniu nariu* arba *paklaida* ir apibrėžiama taip:

$$\varepsilon = y - \beta_1 - \beta_2 x. \tag{5}$$

Paklaida ε apima visus kitus faktorius, įtakojančius Y, išskyrus faktorių X. Suformuluokime paprasto regresinio modelio

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$$

prielaidas paklaidoms ε_t:

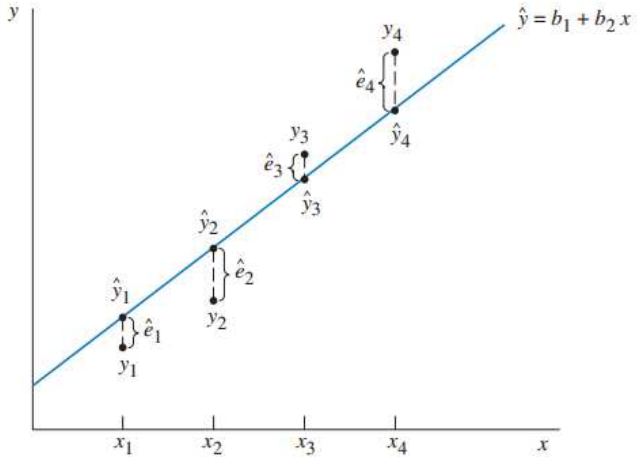
1. Nulinių vidurkių prielaida: $E\varepsilon_1 = ... = E\varepsilon_T = 0$. Ši prielaida ekvivalenti $E(Y_t|X_t) = \beta_1 + \beta_2 X_t, t = 1, 2, ..., T$.

2. *Homoskedastiškumo* prielaida: $var(\varepsilon_t) = \sigma^2, t = 1, 2, ..., T$. Ši prielaida ekvivalenti $var(Y_t|X_t) = \sigma^2, t = 1, 2, ..., T$, nes šie dydžiai skiriasi tik konstanta. Jei dispersija priklauso nuo t ir $var(\varepsilon_t) \neq var(\varepsilon_s)$ kuriems nors $t \neq s$, tai sakome, kad modelis yra *heteroskedastinis*.

3. Nekoreliuotų paklaidų prielaida: $cov(\varepsilon_t, \varepsilon_s) = cov(Y_t, Y_s) = 0$, kai $t \neq s$.

4. Kintamasis X nėra atsitiktinis dydis ir yra bent 2 jo skirtingos reikšmės.

5. *Gausinių paklaidų* prielaida: atsitiktiniai dydžiai $\varepsilon_1, ..., \varepsilon_T$ yra nepriklausomi ir $\varepsilon_t \sim N(0, \sigma^2), t = 1, 2, ..., T$. Ši prielaida ekvivalenti prielaidai, kad duomenys turi sąlyginį Gausinį skirstinį $(Y_t|X_t) \sim N(\beta_1 + \beta_2 X_t, \sigma^2), t = 1, 2, ..., T$.



3 pav.: Mažiausių kvadratų principas

8.5 Regresijos parametrų įvertinimas

Norėdami įvertinti regresijos parametrus β_1, β_2, pirmiausia turime surinkti duomenis. Tarkime, kad turime 40 respondentų imties realizaciją $(Y_t, X_t), t = 1, 2, ..., 40$. Tarkime, kad šią imtį generuojantis procesas yra

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$$

ir galioja prielaidos paklaidoms 1.– 4.
Koeficientų β_1, β_2 radimui pritaikysime *mažiausių kvadratų principą*, kurio esmė – vertikalių atstumų nuo taškų iki regresijos tiesės kvadratų suma turi būti minimizuojama. Iš regresijos lygties apskaičiuotos reikšmės yra

$$\hat{y}_i = \beta_1 + \beta_2 x_i.$$

Vertikalūs atstumai nuo kiekvieno taško iki regresijos tiesės

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \beta_1 - \beta_2 x_i$$

pavaizduoti pav. 3. Ieškome tokius β_1, β_2, kad suma

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} \hat{e}_i^2 = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$$

būtų mažiausia. Apskaičiuojame funkcijos $S(\beta_1, \beta_2)$ dalines išvestines pagal β_1, β_2 ir prilyginame jas nuliui:

$$\begin{aligned} \frac{\partial S}{\partial \beta_1} &= 2N\beta_1 - 2\sum y_i + 2\beta_2 \left(\sum x_i\right) = 0 \\ \frac{\partial S}{\partial \beta_2} &= -2\sum x_i y_i + 2\beta_1 \left(\sum x_i\right) + 2\beta_2 \left(\sum x_i^2\right) = 0 \end{aligned} \tag{6}$$

Perrašome (6) lygčių sistemą tokiu būdu:

$$\begin{array}{l} N\beta_1 + \beta_2 \left(\sum x_i\right) = \sum y_i \\ \beta_1 \left(\sum x_i\right) + \beta_2 \left(\sum x_i^2\right) = \sum x_i y_i \end{array} \tag{7}$$

Padauginę pirmąją lygtį iš $-\sum x_i$, o antrąją iš N ir abi lygtis sudėję, gausime

$$-\beta_2 \left(\sum x_i\right)^2 + \beta_2 N \sum x_i^2 = N \sum x_i y_i - \sum x_i \sum y_i$$

Iš čia gauname koeficiento β_2 mažiausių kvadratų įvertį:

$$\hat{\beta}_2 = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - \left(\sum x_i\right)^2}. \tag{8}$$

Lygčių sistemos (7) pirmoje lygtyje išsireiškę β_1 gausime šio koeficiento mažiausių kvadratų įvertį:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}. \tag{9}$$

Čia $\bar{y} = \sum y_i/N$ ir $\bar{x} = \sum x_i/N$. Jei nebūtų išpildyta prielaida 4. ir visos x_i reikšmės būtų vienodos, lygties (8) vardiklis būtų lygus nuliui. Pasinaudodami formulėmis (8) ir (9) gauname koeficientų įverčius pajamų – išlaidų uždaviniui:

$$\hat{\beta}_2 = 10.21, \hat{\beta}_1 = 283.5735 - (10.21)(19.6048) = 83.416$$

Gautoji regresijos tiesės lygtis yra

$$\hat{y}_i = 83.416 + 10.21\hat{x}_i. \tag{10}$$

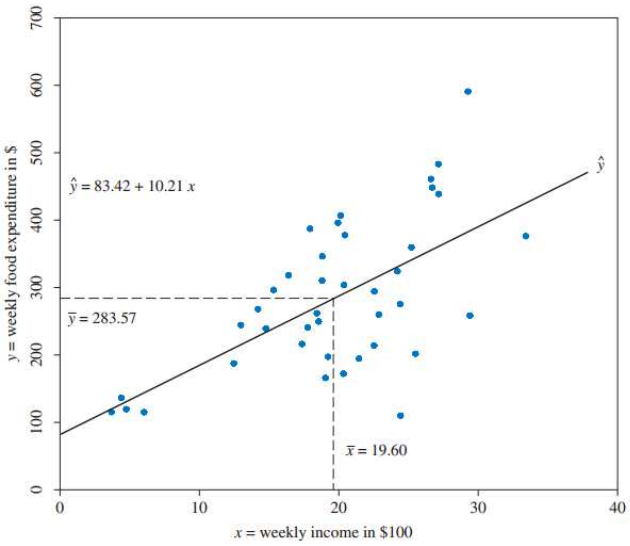
Duomenys ir pajamų ir išlaidų modelį atitinkančioji regresijos tiesė pavaizduoti pav. 4. Atkreipkime dėmesį į tai, kad visada regresijos tiesė eina per tašką, kuris nusakomas empirinių vidurkių $(\bar{x}, \bar{y})$, šiuo atveju per tašką $(19.6048, 283.5735)$.

8.6 Gauso-Markovo teorema.

Nagrinėkime klasikinį tiesinį regresinį modelį

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t, t = 1, 2, ..., T.$$

- Mažiausių kvadratų (MK) metodu gauti β_1 ir β_2 įverčiai yra *tiesiniai įverčiai*, nes jie yra tiesiniai y_i dariniai.
- Šie įverčiai yra *nepaslinkti*, t.y. $E\hat{\beta}_1 = \beta_1, E\hat{\beta}_2 = \beta_2$, kai galioja prielaidos liekanoms 1.– 4.
- Iš visų nepaslinktų įverčių *geriausias* yra tas, kurio *dispersija yra mažiausia*, nes tada tikimybė, kad parametro įvertis yra arčiau tikrosios parametro reikšmės, yra didesnė.



4 pav.: Įvertintas pajamų ir išlaidų modelis

GAUSO-MARKOVO TEOREMA. Klasikinio tiesinio regresinio modelio su prielaidomis liekanoms 1.– 4. parametrų β_1 ir β_2 įverčiai, gauti mažiausių kvadratų metodu, turi mažiausią dispersiją tarp visų tiesinių nepaslinktų įverčių.

Taigi MK metodu gauti įverčiai yra geriausi tarp visų tiesinių nepaslinktų įverčių arba **best linear unbiased estimators (BLUE)** įverčiai.

- Teorema neteigia, kad MK įverčiai yra geriausi iš visų galimų.
- Jei negalioja bent viena iš prielaidų 1.– 4., tada $\hat{\beta}_1$ ir $\hat{\beta}_2$ nebus BLUE (arba geriausiais) įverčiais.
- Gauso-Markovo teorema **nepriklauso nuo gausiškumo prielaidos** 5.