# Skaitinė prognozė

# Regresijos modelis



Nagrinėsime duomenis, kurių visi nepriklausomi kintamieji (atributai) $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ ir priklausomas (klasės) kintamasis $Y$ yra skaitiniai. Tokių duomenų $n$ įrašų imtis yra $E = \{(\mathbf{x}_i, y_i)\,,\ i = 1, 2, \ldots, n\}$ . Jos $i$ - tasis įrašas susideda iš $k$ atributų reikšmių $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ ir priklausomo kintamojo $Y$ reikšmės $y_i$. Aptarsime skaitinės prognozės (kitaip dar vadinamus *regresijos* ) modelius. Iš tikrųjų yra konstruojamas kintamojo $Y$ įvertis ( modelis )

$$\hat{Y} = f(\mathbf{X})\,,$$

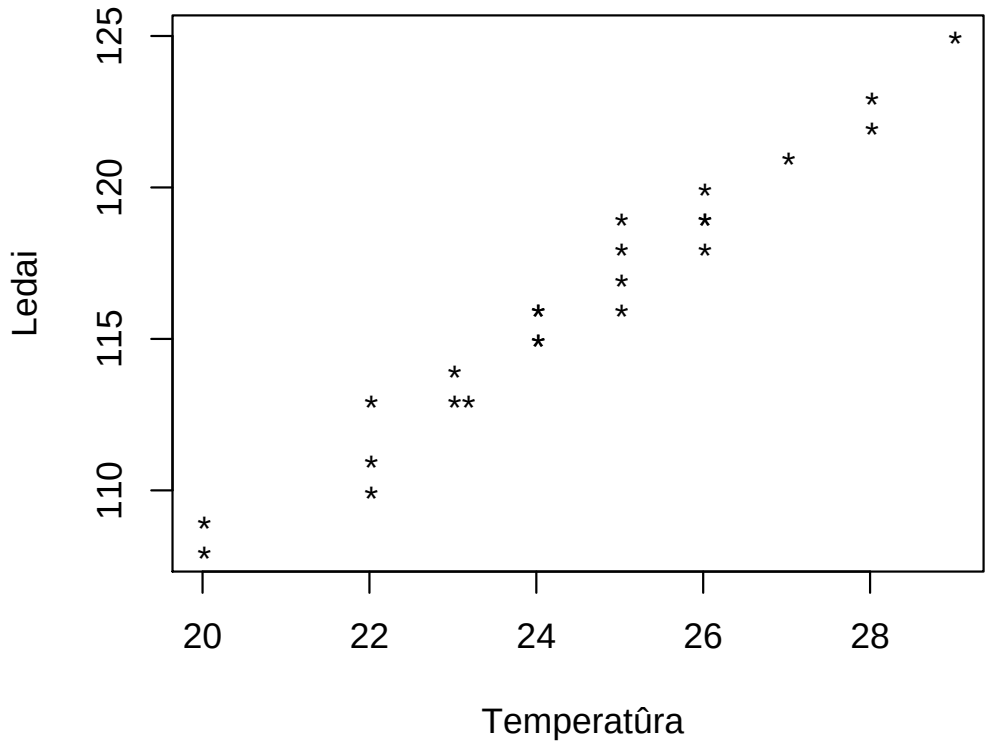
stengiantis, kad jis kuo tiksliau atspindėtų imties duomenis. Kiek įvertis atitinka tikrąjį kintamąjį $Y$ nusako vadinamoji klaidų ( nuostolių ) funkcija $L(Y, f(\mathbf{X}))$ . Konstruojamas toks modelis, kad klaidų funkcijos reikšmių $L(y_i, f(\mathbf{x}_i))$ suma visoje imtyje būtų kuo mažesnė.

# Paprastoji tiesinė regresija



Tarkime, kad imtis $E$ turi tik $k = 1$ atributą $X = X_1$ , t.y. aibė $E$ sudaryta iš $n$ plokštumos taškų $E = \{(x_i, y_i)\,,\ i = 1, 2, \ldots, n\}$ .

**Pavyzdys.** Žinoma kurortinio miestelio keliolikos vasaros dienų vidutinė dienos temperatūra ( $C^\circ$ ) ir vietos restorane suvalgytų ledų kiekis ( kg. ). Visi 24 imties $E$ įrašai pavaizduoti plokštumos taškais.



# Mažiausių kvadratų metodas



Tiesinis modelis

$f(X) = a + bX$ .

Pasirinkę kvadratinę klaidų funkciją rasime koeficientų $a$ ir $b$ įverčius, minimizuojančius suminę klaidą

$$S(a,b) = \sum_{i=1}^{n} L(y_i, f(x_i)) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

$$\hat{b} = \frac{s_y}{s_x} \cdot r , \quad \hat{a} = \bar{y} - \hat{b}\bar{x} .$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i , \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 ,$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 , \quad r = \frac{1}{s_x s_y (n-1)}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

# Regresijos tiesė



**Apibrėžimas.** *Lygtis*

$$\hat{y}(x) = \hat{a} + \hat{b}x\,,$$

*vadinama regresijos tiesės lygtimi.*

$$\hat{e}_i = y_i - \hat{y}(x_i) = y_i - \hat{a} - \hat{b}x_i$$

*vadinama i - tąja liekamąja paklaida,* $i = 1, 2, \ldots, n$ .

$SSE$ (Sum of the Squared Error)

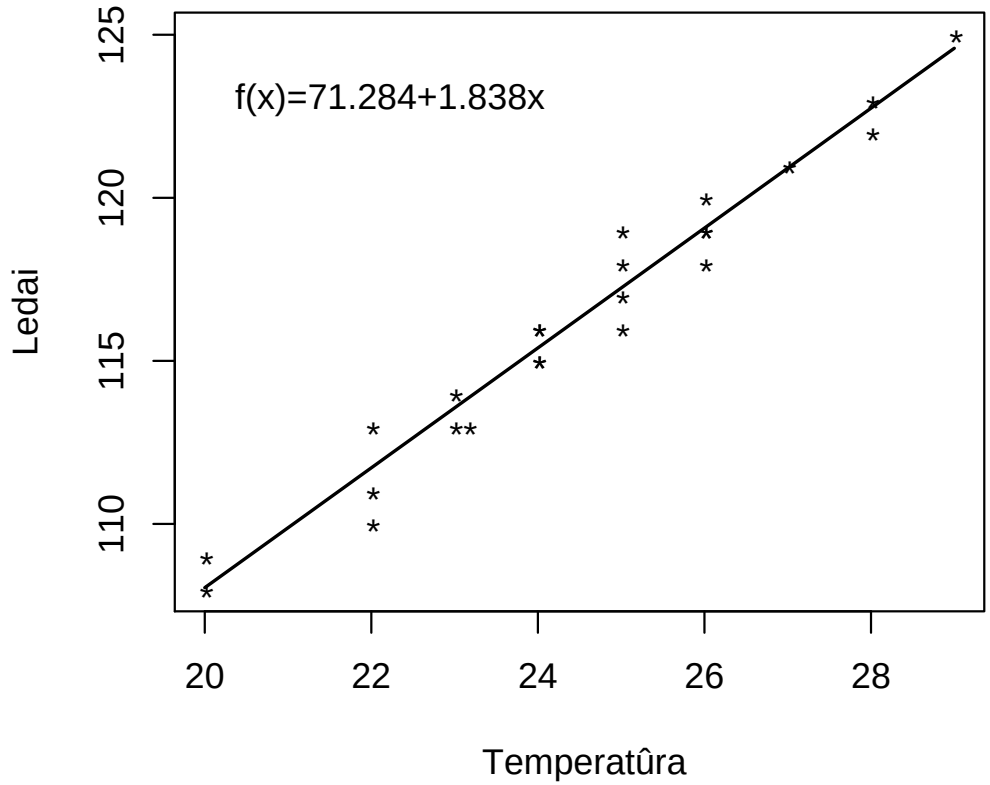
$$SSE = S(\hat{a}, \hat{b}) = \sum_{i=1}^{n} \hat{e}_i^2\,.$$

Įrašę $\hat{a}$ ir $\hat{b}$ išraiškas, gausime

$$\hat{y}(x) = \bar{y} + \frac{s_y}{s_x} \cdot r \cdot (x - \bar{x})\,.$$

# Ledų pardavimo regresijos tiesė

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Temp.($X$) | 25 | 26 | 24 | 26 | 24 | 26 | 22 | 23 | 27 | 20 | 20 | 22 |
| Sv.(kg)($Y$) | 116 | 120 | 115 | 119 | 115 | 118 | 111 | 113 | 121 | 108 | 109 | 110 |
| Temp.($X$) | 28 | 22 | 23 | 23 | 28 | 24 | 26 | 29 | 25 | 25 | 25 | 24 |
| Sv.(kg)($Y$) | 122 | 113 | 113 | 114 | 123 | 116 | 119 | 125 | 118 | 119 | 117 | 116 |



# Atvirkštinės regresijos tiesė



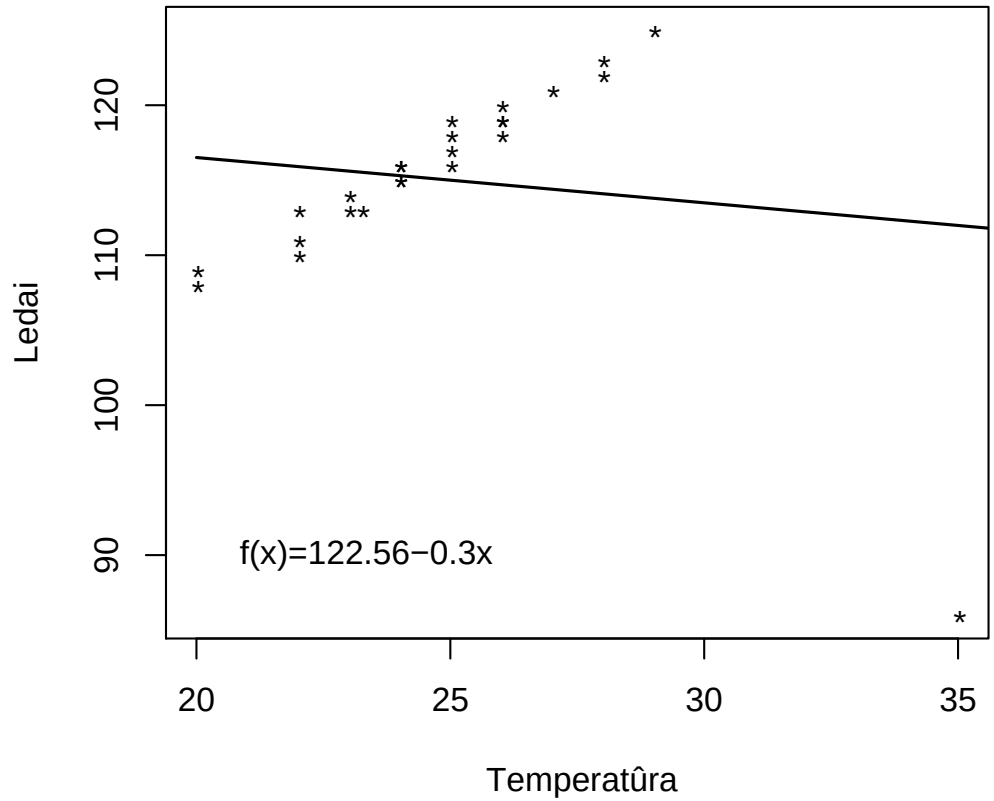
Kai regresijos modelio kintamuosius sieja dvipusė priklausomybė, galima rasti tų pačių duomenų *atvirkštinės regresijos tiesę.* Koreliacijos koeficientas $r$ yra simetriškas kintamųjų atžvilgiu. Todėl gausime tokią atvirkštinės regresijos tiesės lygtį

$$\hat{x}(y) = \bar{x} + \frac{s_x}{s_y} \cdot r \cdot (y - \bar{y}) \,.$$

# Išskirtys



Net ir vienas, labai nuo kitų besiskiriantis įrašas $(x_j, y_j)$ gali radikaliai pakeisti regresijos tiesę. Tarkime, kad ledų pardavimo duomenys yra papildyti: dieną, kai temperatūra siekė 35 lapsnius karščio, suvalgyti 86 kilogramai ledų.



# Įrašo įtakos indeksas



*Įrašo įtakos indeksas* įvertina tik nepriklausomo kintamojo reikšmę ( ar toli nuo $\bar{x}$ yra $x_j$ ). Įrašo $(x_j, y_j)$ įtakos indeksas

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{(n-1)s_x^2} .$$

Kuo regresijos tiesės lygties koeficientai labiau priklauso nuo įrašo, tuo jo įtakos indeksas didesnis. Dažniausiai vadovaujamasi tokia taisykle:

$$\textit{Įrašą } (x_j, y_j) \; laikome \textit{ išskirtimi}, \; jei \; h_j > \frac{4}{n} .$$

# Standartizuotoji liekana



*Standartizuotosios liekanos* yra liekamųjų paklaidų $\hat{e}_j$ $z$-standartizuotos reikšmės

$$SR_j = \hat{e}_j \sqrt{\frac{n-2}{(1-h_j)SSE}}\,.$$

Standartizuotųjų liekanų imties vidurkis lygus nuliui, o imties dispersija - vienetui. Todėl, pagal tikimybių teorijoje žinomą "trijų sigma" taisyklę

$$\textit{Įrašą}\ (x_j, y_j)\ laikome\ \textit{išskirtimi},\ jei\ |SR_j| > 3\,.$$

# Kuko matas



*Kuko matas* atsižvelgia ir į standartizuotąją liekaną ir į įrašo įtakos indeksą. Kuko matas

$$D_j = \frac{(SR_j)^2 h_j}{2(1-h_j)} \, .$$

Kaip matyti iš šios formulės, Kuko matas didelis tada, kai didelis įrašo įtakos indeksas arba didelė standartizuotoji liekana. Supaprastinta, tačiau gana gera taisyklė yra tokia

$$\textit{Įrašą } (x_j, y_j) \; laikome \; \textit{išskirtimi}, \; jei \; D_j > 1 \, .$$

# Paklaidų sumos



Jau anksčiau susidūrėme su paklaidų kvadratų suma $SSE$.

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2 \,.$$

Apibrėšime dar dvi panašias sumas

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \,, \quad SSR = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2 \,.$$

$SST$ (Total Sum of Squares) - visa kvadratų suma,
$SSR$ (Regression Sum of Squares) - regresijos kvadratų suma.
Pastebėsime, kad $SST = (n-1)s_y^2$ ir

$$SSE = SST - SSR \,.$$

# Determinacijos koeficientas



**Apibrėžimas.** *Regresijos modelio determinacijos koeficientu vadinamas santykis*

$$R^2 = \frac{SSR}{SST}.$$

Determinacijos koeficientas neviršija vieneto: $R^2 \leq 1$. Jis naudojamas kaip regresijos modelio tinkamumo indikatorius. Didesnis determinacijos koeficientas reiškia, kad įrašai yra labiau koncentruoti apie regresijos tiesę.
*Dažniausiai reikalaujama, kad būtų tenkinama nelygybė $R^2 \geq 0,25$. Jeigu $R^2 < 0,25$, labai abejotina, ar tiesinės regresijos modelis tinka.*

**Teorema.** *Paprastosios tiesinės regresijos atveju Pirsono koreliacijos koeficiento modulis yra lygus kvadratinei šakniai iš determinacijos koeficiento: $|r| = \sqrt{R^2}$. Jo ženklas sutampa su $\hat{b}$ ženklu.*

# Daugialypės tiesinės regresijos modelis



Daugialypės tiesinės regresijos modelis yra paprastosios regresijos modelio apibendrinimas, kai nepriklausomų kintamųjų yra daugiau nei vienas.

$$f(\mathbf{X}) = \boldsymbol{\beta} \cdot \mathbf{X}^T ;$$

čia

$$\mathbf{X} = (1, X_1, X_2, \ldots, X_k), \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k),$$

kaip įprasta, $\mathbf{X}^T$ žymi transponuotą matricą (šiuo atveju sudarytą iš vieno stulpelio).
Taip modifikavus atributų vektorių $\mathbf{X}$, imties įrašai

$$E = \{(\mathbf{x}_i, y_i), \ i = 1, 2, \ldots, n\}.$$

bus sudaryti iš atributų reikšmių vektoriaus
$\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$ ir priklausomo kintamojo $Y$ reikšmės $y_i$.

# Suminė klaida



Visų atributų reikšmių $n \times (k+1)$ matricą, sudarytą iš vektorių $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ koordinačių, žymėsime

$$\mathfrak{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} .$$

Modelio koeficientus $\boldsymbol{\beta}$ rasime mažiausių kvadratų metodu, t.y. minimizuodami suminę klaidą

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta} \cdot \mathbf{x}_i^T)^2 \\ &= (\mathbf{y} - \mathfrak{X}\boldsymbol{\beta}^T)^T (\mathbf{y} - \mathfrak{X}\boldsymbol{\beta}^T) . \end{aligned}$$

Čia $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ yra priklausomo kintamojo reikšmių stulpelis.

# Tiesinės daugialypės regresijos lygtis



Jei matrica $\mathfrak{X}^T\mathfrak{X}$ neišsigimusi, tai

$$\hat{\boldsymbol{\beta}} = \mathbf{y}^T\mathfrak{X}(\mathfrak{X}^T\mathfrak{X})^{-1}$$

yra modelio koeficientai, minimizuojantys suminę klaidą. Iš čia gauname tiesinės daugialypės regresijos lygtį:

$$\hat{y}(\mathbf{x}) = \hat{\boldsymbol{\beta}} \cdot \mathbf{x}^T \,.$$

Čia $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)\,,\ \ \hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)\,.$

# Daugialypės determinacijos koeficientas



*Daugialypės determinacijos koeficientas* apibrėžiamas taip pat kaip ir vieno nepriklausomo kintamojo atveju.

$$R^2 = \frac{SSR}{SST}\,.$$

Čia

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2\,, \quad SSR = \sum_{i=1}^{n}(\hat{y}(\mathbf{x}_i) - \bar{y})^2\,.$$

Determinacijos koeficientas $R^2 \leq 1$. Galima sakyti, kad kuo $R^2$ reikšmė didesnė, tuo daugiau informacijos apie $Y$ suteikia kintamieji $X_1, X_2, \ldots, X_k$. Taigi tuo geriau tinka ir pasirinktasis regresijos modelis.

# Koreguotasis determinacijos koeficientas



Tačiau, jei nepriklausomų kintamųjų skaičius $k$ nedaug skiriasi nuo įrašų skaičiaus $n$, tai vien todėl determinacijos koeficientas gali būti arti vieneto. Todėl į $R^2$ rekomenduojama atsižvelgti tik tada, kai $k$ daug mažesnis už $n$. Kitais atvejais skaičiuojamas *koreguotasis determinacijos koeficientas*

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1-R^2)\,.$$

Jo reikšmė priklauso ir nuo imties dydžio $n$ ir nuo nepriklausomų kintamųjų skaičiaus $k$. Be to, mažiems $R^2$ koreguotasis determinacijos koeficientas gali įgyti ir neigiamas reikšmes. Koreguotojo determinacijos koeficiento interpretacija lieka ta pati: kuo jis didesnis, tuo geriau $Y$ reikšmes aprašo regresijos modelyje esančių nepriklausomų kintamųjų elgesys.

# Daugialypės koreliacijos koeficientas



*Daugialypės koreliacijos koeficientas R* - tai tiesiog kvadratinė šaknis iš determinacijos koeficiento, t.y

$$R = \sqrt{R^2}\,.$$

Pastebėsime, kad skirtingai nuo Pirsono dviejų kintamųjų koreliacijos koeficiento, jis negali būti neigiamas, nes $0 \leq R \leq 1$.

Daugialypės koreliacijos koeficientas parodo, kaip stipriai prognozuojamas kintamasis priklauso nuo visų nepriklausomų kintamųjų. Žinoma, kalbama apie tiesinę priklausomybę.