12. Linearna regresija (MAST)

Profesor Milan Merkle
emerkle@etf.rs milanmerkle.etf.rs

Matematička Statistika-jesen 2019

Zavisnost između dve slučajne promenljive

$$Y = f(X) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \text{nezavisno od } X$$

- $f(X)$ - objašnjena zavisnost
- ε - neobjašnjena zavisnost

Regresiona funkcija:

$$f(x) = \mathrm{E}\,(Y|X = x)$$

Svođenje na problem ocenjivanja parametara:

- Pretpostavke o zajedničkoj funkciji raspodele za (X, Y).
- Pretpostavke o obliku funkcije f.

Zavisnost između dve slučajne promenljive

$$Y = f(X) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \text{nezavisno od } X$$

- $f(X)$ - objašnjena zavisnost
- ε - neobjašnjena zavisnost

Regresiona funkcija:

$$f(x) = \mathrm{E}\,(Y|X = x)$$

Svođenje na problem ocenjivanja parametara:

- Pretpostavke o zajedničkoj funkciji raspodele za (X, Y).
- Pretpostavke o obliku funkcije f.

Zavisnost između dve slučajne promenljive

$$Y = f(X) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \text{nezavisno od } X$$

- $f(X)$ - objašnjena zavisnost
- ε - neobjašnjena zavisnost

Regresiona funkcija:

$$f(x) = \mathrm{E}\,(Y|X = x)$$

Svođenje na problem ocenjivanja parametara:

- Pretpostavke o zajedničkoj funkciji raspodele za (X, Y).
- Pretpostavke o obliku funkcije f.

Zavisnost između dve slučajne promenljive

$$Y = f(X) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \text{nezavisno od } X$$

- $f(X)$ - objašnjena zavisnost
- ε - neobjašnjena zavisnost

Regresiona funkcija:

$$f(x) = \mathrm{E}\,(Y|X = x)$$

Svođenje na problem ocenjivanja parametara:

- Pretpostavke o zajedničkoj funkciji raspodele za (X, Y).
- Pretpostavke o obliku funkcije f.

Pretpostavka o zajedničkoj funkciji raspodele

Primer 188. Pretpostavimo da slučajni vektor (X, Y) ima dvodimenzionalnu normalnu raspodelu sa parametrima μ_X, μ_Y, σ_X^2, σ_Y^2 i ρ. Prema rezultatu primera 180, imamo da je

$$f(x) = \mathrm{E}(Y|X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

Dakle, da bi se ocenila regresiona funkcija (u ovom slučaju regresiona prava), potrebno je oceniti pet nepoznatih parametara.

Pretpostavka o obliku funkcije f

Pretpostavimo da je

$$f(x) = a\cos x + b\sin x.$$

Koeficijenti a i b određuju se kao oni realni brojevi za koje funkcija

$$R(a,b) = \mathrm{E}\,(Y - a\cos X - b\sin X)^2$$

dostiže minimum.

$$\begin{aligned} \frac{\partial R(a,b)}{\partial a} &= -2\mathrm{E}\left(\cos X \cdot (Y - a\cos X - b\sin X)\right) = 0 \\ \frac{\partial R(a,b)}{\partial b} &= -2\mathrm{E}\left(\sin X \cdot (Y - a\cos X - b\sin X)\right) = 0 \end{aligned}$$

Najjednostavniji model: Regresiona prava

$$f(x) = ax + b\,, \quad \text{tj.} \quad Y = aX + b + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

gde se pretpostavlja da su X i ε nezavisne.
Parametre a i b nalazimo iz uslova $E(Y - aX - b)^2 \longrightarrow \min$.

$$a = \frac{\mathrm{E}\,XY - \mathrm{E}\,X\mathrm{E}\,Y}{\mathrm{Var}\,X}, \quad b = \mathrm{E}\,Y - a\mathrm{E}\,X\,, \tag{1}$$

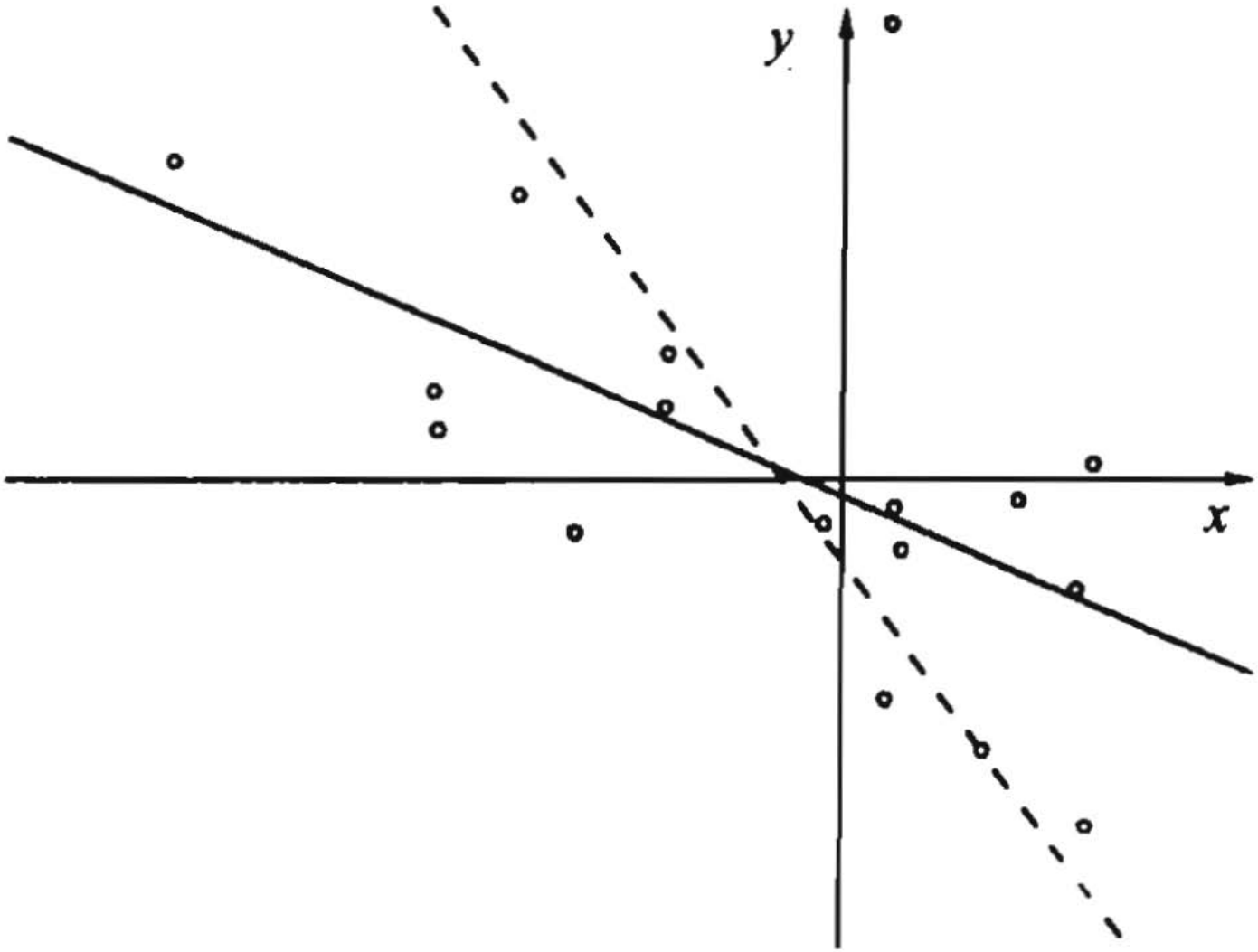
Ocene parametara a i b dobijaju se iz uzorka

$$(X_1, Y_1),\ (X_2, Y_2),\ \ldots\ , (X_n, Y_n)$$

na sledeći način:

$$\hat{a} = \frac{n\sum_{k=1}^{n} X_k Y_k - \sum_{k=1}^{n} X_k \sum_{k=1}^{n} Y_k}{n\sum_{k=1}^{n} X_k^2 - \left(\sum_{k=1}^{n} X_k\right)^2}, \quad \hat{b} = \frac{1}{n}\sum_{k=1}^{n} Y_k - \frac{\hat{a}}{n}\sum_{k=1}^{n} X_k \tag{2}$$

Primer 190 - nije svejedno!



Slika 39. Regresione prave $y = y(x)$ (puna linija) i $x = x(y)$.

Opšta linearna regresija

Opšti problem linearne regresije jeste da se, na osnovu uzorka, ocene koeficijenti $a_0, a_1, \ldots, a_m$ u pretpostavljenoj jednakosti

$$Y = a_0 f_0(X) + a_1 f_1(X) + \cdots + a_m f_m(X) + \varepsilon\,,$$

gde su X i Y slučajne promenljive, $f_0, \ldots, f_m$ su date funkcije, a ε je slučajna promenljiva nezavisna od X sa $\mathcal{N}(0, \sigma^2)$ raspodelom.
Zavisnost ovog oblika je linearna po nepoznatim koeficijentima $a_0, \ldots, a_m$.
Zavisnost oblika, na primer,

$$Y = a \exp(\sin X) + b \exp(\cos X) + cX^3 + \varepsilon$$

je takođe primer linearne regresije, iako se pojavljuju funkcije koje nisu linearne.
Ako je navedeni model tačan, onda je

$$a_0 f_0(X) + a_1 f_1(X) + \cdots + a_m f_m(X) = \mathrm{E}\,(Y|X)\,,$$

pa se parametri $a_0, \ldots, a_m$ nalaze iz uslova

$$R(a_0, \ldots, a_m) = \mathrm{E}\,(Y - a_0 f_0(X) - a_1 f_1(X) - \cdots - a_m f_m(X))^2 \longrightarrow \min\,.$$

Kontrolisana x-promenljiva

Vrednostima $x_1, \ldots, x_n$ kontrolisane promenljive odgovaraju vrednosti slučajne promenljive Y:

$$Y_i = a_0 f_0(x_i) + a_1 f_1(x_i) + \cdots + a_m f_m(x_i) + \varepsilon_i \,,$$

pri čemu važi:

- $x_1, \ldots, x_n$ imaju fiksirane (unapred određene) brojne vrednosti, pri čemu neke od njih mogu biti međusobno jednake.
- $\varepsilon_1, \ldots, \varepsilon_n$ su nezavisne slučajne promenljive sa istom raspodelom $\mathcal{N}(0, \sigma^2)$, gde je σ^2 nepoznato.
- Parametri $a_0, \ldots, a_m$ su nepoznati, dok su funkcije $f_0, \ldots, f_m$ date.

Terminologija je ista kao kada je X slučajna promenljiva. Funkcija

$$f(x) = a_0 f_0(x) + a_1 f_1(x) + \cdots + a_m f_m(x)$$

naziva se regresionom funkcijom ili regresionom krivom.

Kontrolisana x-promenljiva

Vrednostima $x_1, \ldots, x_n$ kontrolisane promenljive odgovaraju vrednosti slučajne promenljive Y:

$$Y_i = a_0 f_0(x_i) + a_1 f_1(x_i) + \cdots + a_m f_m(x_i) + \varepsilon_i \,,$$

pri čemu važi:

- $x_1, \ldots, x_n$ imaju fiksirane (unapred određene) brojne vrednosti, pri čemu neke od njih mogu biti međusobno jednake.
- $\varepsilon_1, \ldots, \varepsilon_n$ su nezavisne slučajne promenljive sa istom raspodelom $\mathcal{N}(0, \sigma^2)$, gde je σ^2 nepoznato.
- Parametri $a_0, \ldots, a_m$ su nepoznati, dok su funkcije $f_0, \ldots, f_m$ date.

Terminologija je ista kao kada je X slučajna promenljiva. Funkcija

$$f(x) = a_0 f_0(x) + a_1 f_1(x) + \cdots + a_m f_m(x)$$

naziva se regresionom funkcijom ili regresionom krivom.

Kontrolisana x-promenljiva

Vrednostima $x_1, \ldots, x_n$ kontrolisane promenljive odgovaraju vrednosti slučajne promenljive Y:

$$Y_i = a_0 f_0(x_i) + a_1 f_1(x_i) + \cdots + a_m f_m(x_i) + \varepsilon_i \,,$$

pri čemu važi:

- $x_1, \ldots, x_n$ imaju fiksirane (unapred određene) brojne vrednosti, pri čemu neke od njih mogu biti međusobno jednake.
- $\varepsilon_1, \ldots, \varepsilon_n$ su nezavisne slučajne promenljive sa istom raspodelom $\mathcal{N}(0, \sigma^2)$, gde je σ^2 nepoznato.
- Parametri $a_0, \ldots, a_m$ su nepoznati, dok su funkcije $f_0, \ldots, f_m$ date.

Terminologija je ista kao kada je X slučajna promenljiva. Funkcija

$$f(x) = a_0 f_0(x) + a_1 f_1(x) + \cdots + a_m f_m(x)$$

naziva se regresionom funkcijom ili regresionom krivom.

Za ocenu parametara koristimo princip minimiziranja srednjeg kvadratnog odstupanja. Neka su $x_1, \ldots, x_n$ vrednosti kontrolisane promenljive kojima odgovaraju realizovane (izmerene) vrednosti $y_1, \ldots, y_n$ slučajnih promenljivih Y_i.

Metod najmanjih kvadrata. Koeficijenti a_i se određuju iz uslova

$$\sum_{i=1}^{n} (y_i - a_0 f_0(x_i) - a_1 f_1(x_i) - \cdots - a_m f_m(x_i))^2 \longrightarrow \min$$

Čest slučaj je tzv. linearna regresija reda m ili polinomska regresija, gde se pretpostavlja da je

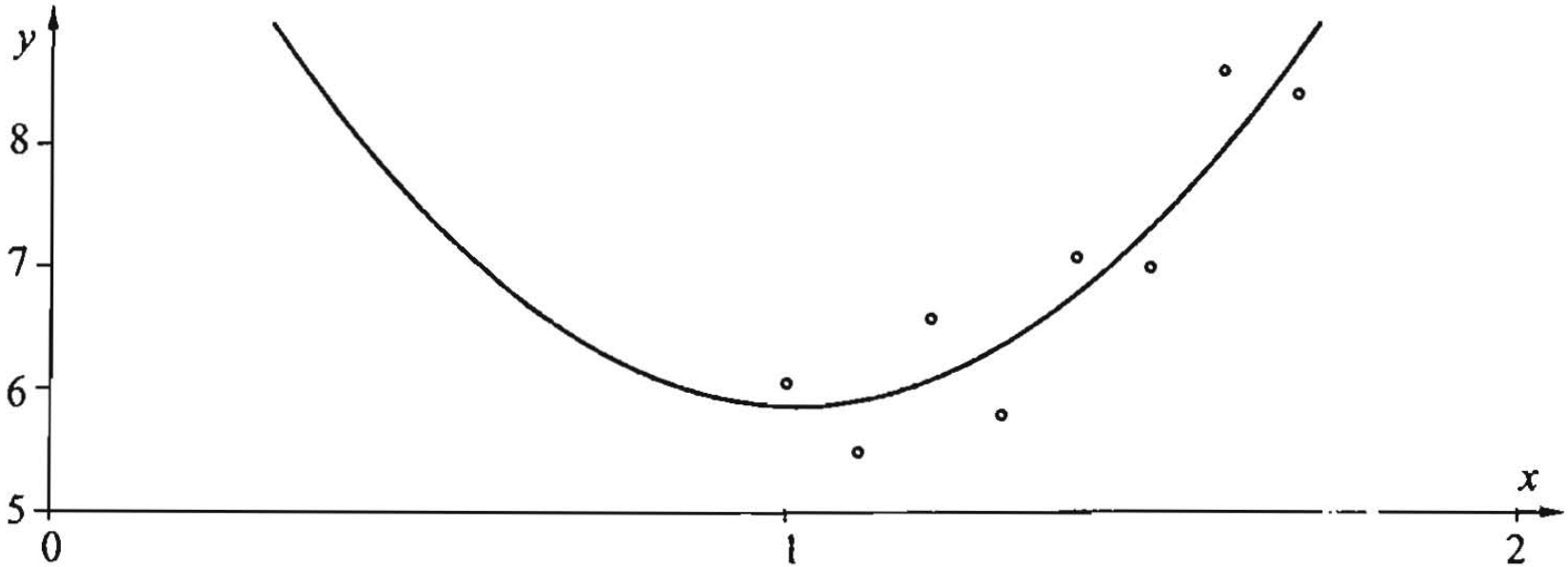
$$Y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + \varepsilon$$

Regresija ili interpolacija

Metod regresije se suštinski razlikuje od metoda interpolacije, mada oba metoda imaju za cilj „fitovanje". Naime, za svaki dati skup od n tačaka (x_i, y_i) postoji jedinstven interpolacioni polinom stepena $n-1$ koji prolazi kroz te tačke. Regresija nema jedinstveno rešenje ako je broj tačaka veći od broja parametara. Ako bismo, umesto principa najmanjih kvadrata usvojili neki drugi kriterijum, recimo najmanji zbir apsolutnih vrednosti odstupanja, dobili bismo drugu krivu. Regresiju koristimo kada nam treba oblik polinoma određenog stepena, dok nam interpolacija služi da bismo kroz merene tačke provukli glatku krivu.

Primer 191

x	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7
y	6.05	5.49	6.58	5.79	7.07	6.99	8.60	8.41



Slika 41. Uz primer 191.

Pretpostavićemo zavisnost oblika

$$Y = ax^2 + bx + c + \varepsilon\,.$$

Koeficijente a, b, c nalazimo iz uslova

$$R(a,b,c) = \sum_{k=1}^{8}(y_i - ax_i^2 - bx_i - c)^2 \;\longrightarrow\; \min\,.$$

Kako je

$$\begin{aligned}
\frac{\partial R(a,b,c)}{\partial a} &= -2\sum_{k=1}^{8} x_i^2(y_i - ax_i^2 - bx_i - c)\\
\frac{\partial R(a,b,c)}{\partial b} &= -2\sum_{k=1}^{8} x_i(y_i - ax_i^2 - bx_i - c)\\
\frac{\partial R(a,b,c)}{\partial c} &= -2\sum_{k=1}^{8}(y_i - ax_i^2 - bx_i - c)\,,
\end{aligned}$$

Linearni sistem jednačina:

$$
\begin{aligned}
a\sum_{k=1}^{8} x_i^4 - b\sum_{k=1}^{8} x_i^3 - c\sum_{k=1}^{8} x_i^2 &= \sum_{k=1}^{8} x_i^2 y_i \\
a\sum_{k=1}^{8} x_i^3 - b\sum_{k=1}^{8} x_i^2 - c\sum_{k=1}^{8} x_i &= \sum_{k=1}^{8} x_i y_i \\
a\sum_{k=1}^{8} x_i^2 - b\sum_{k=1}^{8} x_i - 8c &= \sum_{k=1}^{8} y_i
\end{aligned}
$$

Izračunavanjem koeficijenata i rešavanjem sistema dobijamo da je $a = 6.13$, $b = -12.4$, $c = 12.2$. Parabola $y = 6.13x^2 - 12.4x + 12.2$,

Regresiona prava: Ocene parametara

Uzorak: $(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$

$$\bar{x} = \frac{1}{n}\sum x_k\,, \quad \bar{y} = \frac{1}{n}\sum y_k\,,$$

$$S_x^2 = \sum (x_k - \bar{x})^2 = \sum x_k^2 - n\bar{x}^2\,, \quad S_y^2 = \sum (y_k - \bar{y})^2 = \sum y_k^2 - n\bar{y}^2\,,$$

$$S_{xy} = \sum (x_k - \bar{x})(y_k - \bar{y}) = \sum x_k y_k - n\bar{x}\bar{y}\,.$$

Ovde su $\bar{x}$ i $\bar{y}$ srednje vrednosti uzorka po x i po y, a S_x^2 i S_y^2 su zbirovi kvadrata odstupanja od srednjih vrednosti.
Sa uvedenim oznakama, ocene parametara regresione prave mogu se napisati u obliku

$$\hat{a} = \frac{S_{xy}}{S_x^2}\,, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}\,.$$

$$S^2 = \sum (y_k - \hat{a}x_k - \hat{b})^2 = S_y^2 - \hat{a}^2 S_x^2$$