Teorija pogrešaka

Linearna Regresija

Procjena pogrešaka

Pretpostavka: mjerimo neku veličinu relativno nepreciznim instrumentom, te uvijek dobivamo istu vrijednost $\bar{x} = x_1 = x_2 = ... = x_n$.

Znači li to da nema pogreške?

Odgovor: nikako!

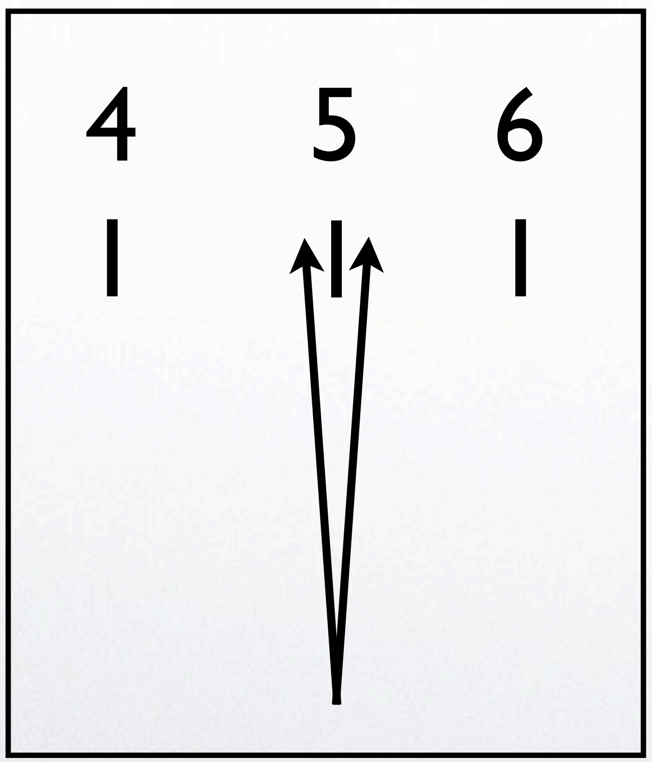
To samo znači da instrument nije dovoljno osjetljiv.

U tom slučaju procijenjujemo pogrešku i pišemo:

$$X = \bar{x} \pm \Delta x, \quad R = \frac{\Delta x}{\bar{x}} \cdot 100(\%)$$

Procjena pogrešaka

Kako procijeniti Δx?



Δx je najmanji razmak u kojem na određenom instrumentu možemo uočiti promjenu

Δx ima značenje *maksimalne pogreške*, koju možemo definirati i na skupu mjerenih rezultata $x_1, x_2, ..., x_n$

$$\Delta x = \left|\bar{x} - x_i\right|_{\max}$$

Pogreška ovisnih veličina

Pretpostavimo da je fizikalna veličina F funkcija neposredno mjerenih veličina $y_1, y_2, ..., y_n$

Primjeri:

1. mjerimo duljinu, širinu i visinu kvadra, a zanima nas njegov obujam ($V = a \cdot b \cdot c$)
2. mjerimo mase tri predmeta, m_1, m_2 i m_3, a zanima nas ukupna masa ($m = m_1+m_2+m_3$)
3. mjerimo vrijeme t padanja kuglice s vrha nebodera, a želimo odrediti visinu nebodera ($h = gt^2/2$)

Najvjerojatnija vrijednost veličine $\overline{F}$ je:

$$\boxed{\overline{F} = F\left(\overline{y}_1, \overline{y}_2, ..., \overline{y}_n\right)}$$

Pogreška ovisnih veličina

Primjer 1. $\bar{a} = 2.31\,\text{cm}, \bar{b} = 1.43\,\text{cm}, \bar{c} = 4.11\,\text{cm}$

$$\bar{V} = \bar{a} \cdot \bar{b} \cdot \bar{c} = 13.58\,\text{cm}^3$$

No, što s pogreškom? Znamo pogreške osnovnih veličina, ali u kakvoj su one vezi s pogreškom ovisne veličine?

Iz zakona rasprostranjenja pogreške znamo da je uz $F = y_1 + y_2 + \ldots + y_k$:

$$m^2(F) = m^2(y_1) + m^2(y_2) \ldots + m^2(y_k)$$

Što možemo pisati i kao:

$$m^2(F) = \left[\frac{\partial F}{\partial y_1} m(y_1)\right]^2 + \left[\frac{\partial F}{\partial y_2} m(y_2)\right]^2 + \ldots + \left[\frac{\partial F}{\partial y_k} m(y_k)\right]^2 = \sum_{j=1}^{k} \left[\frac{\partial F}{\partial y_i} m(y_i)\right]^2$$

Pogreška ovisnih veličina

Pokazuje se da se ovaj pristup može poopćiti i na nelinearne funkcije F(y_1, y_2, ..., y_k), te da vrijedi:

$$m^2(F) = \sum_{j=1}^{k} \left[\frac{\partial F}{\partial y_j} \right]^2 m^2(y_j)$$ srednja pogreška

$$\sigma^2(F) = \sum_{j=1}^{k} \left[\frac{\partial F}{\partial y_j} \right]^2 \sigma^2(y_j)$$ standardna devijacija

$$M^2(F) = \sum_{j=1}^{k} \left[\frac{\partial F}{\partial y_j} \right]^2 M^2(y_j)$$ nepouzdanost

Pogreška ovisnih veličina

Primjer: $F = a \cdot b + c = F(a,b,c)$

$$M^2(F) = \left\{ \left(\frac{\partial F}{\partial a} M_a \right)^2 + \left(\frac{\partial F}{\partial b} M_b \right)^2 + \left(\frac{\partial F}{\partial c} M_c \right)^2 \right\}_{\substack{a=\overline{a} \\ b=\overline{b} \\ c=\overline{c}}} =$$

$$= \left(\overline{b} \cdot M_a \right)^2 + \left(\overline{a} \cdot M_b \right)^2 + \left(1 \cdot M_c \right)^2$$

Pogreška ovisnih veličina

Primjer iz praktikuma (vježba 8):
Modul elastičnosti čelika određuje se mjerenjem dimenzija čelične šipke, savijenosti čelične šipke po jedinici sile i udaljenosti potpornja. Teorijska formula za savijenost šipke je:

$$\lambda = \frac{1}{4E}\frac{L^3}{ab^3}F,$$

gdje je E modul elastičnosti čelika, L je udaljenost potporanja, a je širina šipke, b je debljina šipke, a F je sila teže utega.

Izmjerene su sljedeće veličine:

$$A = \frac{\lambda}{F} = (0{,}76 \pm 0{,}01)\,\frac{\mathrm{mm}}{\mathrm{N}} = (7{,}6 \pm 0{,}1)\cdot 10^{-4}\,\frac{\mathrm{m}}{\mathrm{N}} = \bar{A} \pm M_A$$

$$a = (10{,}26 \pm 0{,}05)\,\mathrm{mm} = (1{,}026 \pm 0{,}005)\cdot 10^{-2}\,\mathrm{m} = \bar{a} \pm M_a$$

$$b = (1{,}53 \pm 0{,}03)\,\mathrm{mm} = (1{,}53 \pm 0{,}03)\cdot 10^{-3}\,\mathrm{m} = \bar{b} \pm M_b$$

$$L = (29{,}0 \pm 0{,}1)\,\mathrm{cm} = (2{,}90 \pm 0{,}01)\cdot 10^{-1}\,\mathrm{m} = \bar{L} \pm M_L$$

Najvjerojatnija vrijednost za modul elastičnosti čelika je:

$$\bar{E} = \frac{1}{4}\frac{\bar{L}^3}{\bar{a}\bar{b}^3}\frac{1}{\bar{A}} = 2{,}183226\ \mathrm{N/m^2}$$

Varijanca modula elastičnosti je:

$$M_E^2 = \left(\frac{\partial E}{\partial A}\right)^2 M_A^2 + \left(\frac{\partial E}{\partial a}\right)^2 M_a^2 + \left(\frac{\partial E}{\partial b}\right)^2 M_b^2 + \left(\frac{\partial E}{\partial L}\right)^2 M_L^2 =$$

Pogreška ovisnih veličina

$$= \left(\frac{\overline{L}^3}{4\overline{a}\,\overline{b}^3 \overline{A}^2} \right)^2 M_A^2 + \left(\frac{\overline{L}^3}{4\overline{a}^2 \overline{b}^3 \overline{A}} \right)^2 M_a^2 + \left(\frac{3\overline{L}^3}{4\overline{a}\,\overline{b}^4 \overline{A}} \right)^2 M_b^2 + \left(\frac{3\overline{L}^2}{4\overline{a}\,\overline{b}^3 \overline{A}} \right)^2 M_L^2 =$$

$$= \overline{E}^2 \left[\left(\frac{M_A}{\overline{A}} \right)^2 + \left(\frac{M_a}{\overline{a}} \right)^2 + \left(3\frac{M_b}{\overline{b}} \right)^2 + \left(3\frac{M_L}{\overline{L}} \right)^2 \right] =$$

$$= \overline{E}^2 \left[1{,}73 \cdot 10^{-4} + 2{,}37 \cdot 10^{-5} + 3{,}46 \cdot 10^{-3} + 1{,}07 \cdot 10^{-4} \right]$$

Standardna pogreška je:

$$M_E = \overline{E}\sqrt{3{,}76 \cdot 10^{-3}} = 0{,}13 \cdot 10^{11}\ \mathrm{N/m^2}$$

Rezultat za modul elastičnosti pišemo:

$$E = (2{,}2 \pm 0{,}1) \cdot 10^{11}\ \mathrm{N/m^2}$$

Pogreška ovisnih veličina

Što ako imamo samo procijenjene pogreške Δy_i?

Tada ne možemo govoriti o rasprostranjenju pogreške, što je zakon definiran na statističkom skupu.

Umjesto toga razvijamo F u red oko $\bar{F}$ i računamo *maksimalnu pogrešku*.

$$F = F\left(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k\right) \pm \sum_{j=1}^{k} \left(\frac{\partial F}{\partial y_j}\right)_{y_j=\bar{y}_j} \times \Delta y_j = \bar{F} \pm \Delta F$$

Budući da je $\Delta F \geq 0$, uzimamo apsolutne vrijednosti derivacija, pa je:

$$\boxed{\Delta F = \sum_{j=1}^{k} \left|\frac{\partial F}{\partial y_j}\right|_{y_j=\bar{y}_j} \times \Delta y_j}$$

Pogreška ovisnih veličina

Primjer:

$$F = a \cdot b + c$$

$$\Delta F = \left\{ \left| \frac{\partial F}{\partial a} \right| \Delta a + \left| \frac{\partial F}{\partial b} \right| \Delta b + \left| \frac{\partial F}{\partial c} \right| \Delta c \right\}_{\substack{a=\bar{a} \\ b=\bar{b} \\ c=\bar{c}}} = \bar{b}\,\Delta a + \bar{a}\,\Delta b + \Delta c$$

Mjerenja s različitim statističkim težinama

Pretpostavimo da je u više različitih laboratorija neovisno mjerena ista veličina *X*. Ostvareni su rezultati:

$$X_1 = \bar{x}_1 \pm M_1, X_2 = \bar{x}_2 \pm M_2, \ldots, X_l = \bar{x}_l \pm M_l$$

Kako će netko tko dobije samo listu rezultata, a ne zna ništa o pojedinostima eksperimenta, izraziti rezultat ovih mjerenja u obliku:

$$\boxed{X = \bar{\bar{x}} + \bar{M}}$$

gdje je $\bar{\bar{x}}$ *opća aritmetička sredina*, a $\bar{M}$ *opća nepouzdanost*.

Mjerenja se obično izvode na različite načine i s različitim pouzdanostima pa ih ne smijemo tretirati ravnopravno. Kažemo da mjerenja imaju različite statističke težine.

Mjerenja s različitim statističkim težinama

Pretpostavimo da svakom rezultatu pridijelimo *statističku težinu* w_j, tako da je:

$$\bar{\bar{x}} = \sum_{j=1}^{l} w_j \bar{x}_j; \quad \sum_{j=1}^{l} w_j$$

Dakle, w_j ima slično značenje kao i relativna frekvencija pojavljivanja događaja, a $\bar{x}_j$ kao pojedinačno mjerenje.

U praksi se najčešće uzima (veća težina za manji M_j!):

$$\boxed{w_j = \frac{\frac{1}{M_j}}{\sum_{k=1}^{l} \frac{1}{M_k^2}}} \Rightarrow \boxed{\text{zadovoljava} \quad \sum_{j=1}^{l} w_j = 1}$$

Mjerenja s različitim statističkim težinama

Prema tome, opća aritmetička sredina je:

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{l} \frac{\bar{x}_j}{M_j^2}}{\sum_{j=1}^{l} \frac{1}{M_j^2}}$$

Mjerenja s različitim statističkim težinama

Koliki je *M*?

Zamislimo jedno mjerenje koje ima statističku težinu $w = 1$ i nepouzdanost $\bar{M}$, a u potpunosti zamjenjuje *l* rezultata.

Znamo otprije:

$$\frac{w_i}{w_j} = \frac{M_j^2}{M_i^2} \quad \text{i stavimo} \quad \begin{aligned} w_i &= 1 \\ M_i &= \bar{M} \end{aligned}$$

$$\frac{1}{w_j} = \frac{M_j^2}{\bar{M}^2} \quad \Rightarrow \quad \bar{M}^2 = w_j M_j^2 = \frac{1}{\sum_{k=1}^{l} \frac{1}{M_k^2}}$$

Mjerenja s različitim statističkim težinama

Prema tome, opća nepouzdanost je:

$$\bar{M} = \sqrt{\frac{1}{\sum_{j=1}^{l} \frac{1}{M_j^2}}}$$

Mjerenja s različitim statističkim težinama

Kako računati?

(i) $\bar{x}_i \approx \bar{x}_j \quad \mathrm{i} \quad M_i \approx M_j, \quad \forall i, j$

koristiti gornje formule za $\bar{\bar{x}}$ i $\bar{M}$

(ii) M_k << M_i, za neki k, tj. rezultat $X_k = \bar{x}_k \pm M_k$ puno je precizniji od svih ostalih.

w_k >> w_i za sve i, pa uzimamo w_k = 1:

$$\bar{\bar{x}} = \bar{x}_k; \quad \bar{M} = M_k$$

(iii) $M_i \approx M_j$ za sve i, j, ali $\bar{x}_i$ i $\bar{x}_j$ se znatno razlikuju za sve i, j

zanemarujemo M_i, te smatramo $\bar{x}_i$
neposredno mjerenim veličinama:

$$\bar{\bar{x}} = \frac{1}{l}\sum_{j=1}^{l} \bar{x}_j; \quad \bar{M} = \sqrt{\frac{\sum_{j=1}^{l}\left(\bar{\bar{x}} - \bar{x}_j\right)}{l(l-1)}}$$

Korelacija u linearnoj regresiji

Zamislimo neki događaj na kojem su definirana dva obilježja *x* i *y*.

Svako promatranje tog događaja daje par brojeva (*x*, *y*).

Npr. elastična sila, obilježja: masa utega i duljina opruge
slobodni pad, obilježja: vrijeme i prijeđeni put
studenti, obilježja: uspjeh na prijemnom i uspjeh na studiju

Pitamo se: jesu li obilježja *x* i *y* na bilo koji način povezana, tj. *korelirana*?

Drugim riječima, postoji li funkcionalna veza *y(x)* koja može dobro opisati parove *(x, y)*?

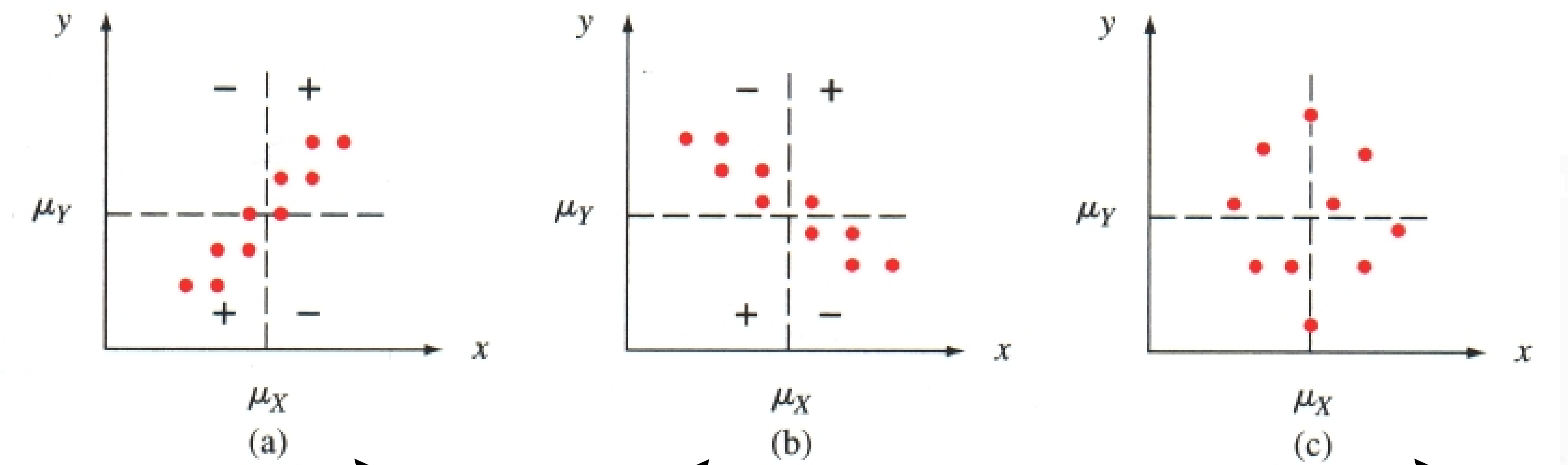
Mi ćemo se zadržati na najjednostavnijem slučaju, tj. istraživanjem postoji li zadovoljavajući opis u okvirima *linearne regresije*.

$$\boxed{y = ax + b}$$

Korelacija u linearnoj regresiji

Prva ocjena može se donijeti crtanjem grafa, tj. točaka u *x*-*y* ravnini:

Primjer:



- možemo "od oka" povući pravac; korelacija vjerojatno postoji:
1. koja je jednadžba pravca?
2. koja je kvanitativna ocjena postojanja korelacije

- ne možemo "od oka" povući pravac; korelacija vjerojatno ne postoji:
1. koja je kvanitativna ocjena nepostojanja korelacije

Korelacija u linearnoj regresiji

Događaj možemo tretirati kao 2D slučajnu varijablu, pri čemu frekvencija pojavljivanja nekog događaja ima značenje (nenormirane) vjerojatnosti:

$$P\left(x_i, y_j\right) \Leftrightarrow \frac{f_{i,j}}{N}, \quad N = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{i,j}$$

$$f_i = \sum_{j=1}^{m} f_{ij};\; f_j = \sum_{i=1}^{n} f_{ij}$$

Npr. {(1,1), (1,1), (1,2), (2,2), (2,2), (2,2)}

$f_{11} = 2$, $f_{12}=1$, $f_{22}=2$; $N=2+1+2=5$

Korelacija u linearnoj regresiji

Također vrijedi:

$$\bar{x} = \frac{1}{N}\sum_{i,j} f_{i,j} x_i = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

$$\bar{y} = \frac{1}{N}\sum_{i,j} f_{i,j} y_y = \frac{1}{N}\sum_{j=1}^{m} f_j y_j$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i,j} f_{i,j}\left(x_i - \bar{x}\right)^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i,j} f_{i,j}\left(y_j - \bar{y}\right)^2 = \frac{1}{N}\sum_{j=1}^{n} f_j\left(y_j - \bar{y}\right)^2$$

$$\sigma_{xy} = \frac{1}{N}\sum_{i,j} f_{i,j}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \frac{1}{N}\sum_{i,j} f_{i,j} x_i y_j - \bar{x}\cdot\bar{y}$$

Korelacija u linearnoj regresiji

Sjetimo se da je za nezavisna svojstva *x* i *y*, $\sigma_{xy} = 0$

Definiramo *koeficijent korelacije r*:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$\sigma_{xy} \to 0;$ *r* je mali, korelacija je slaba

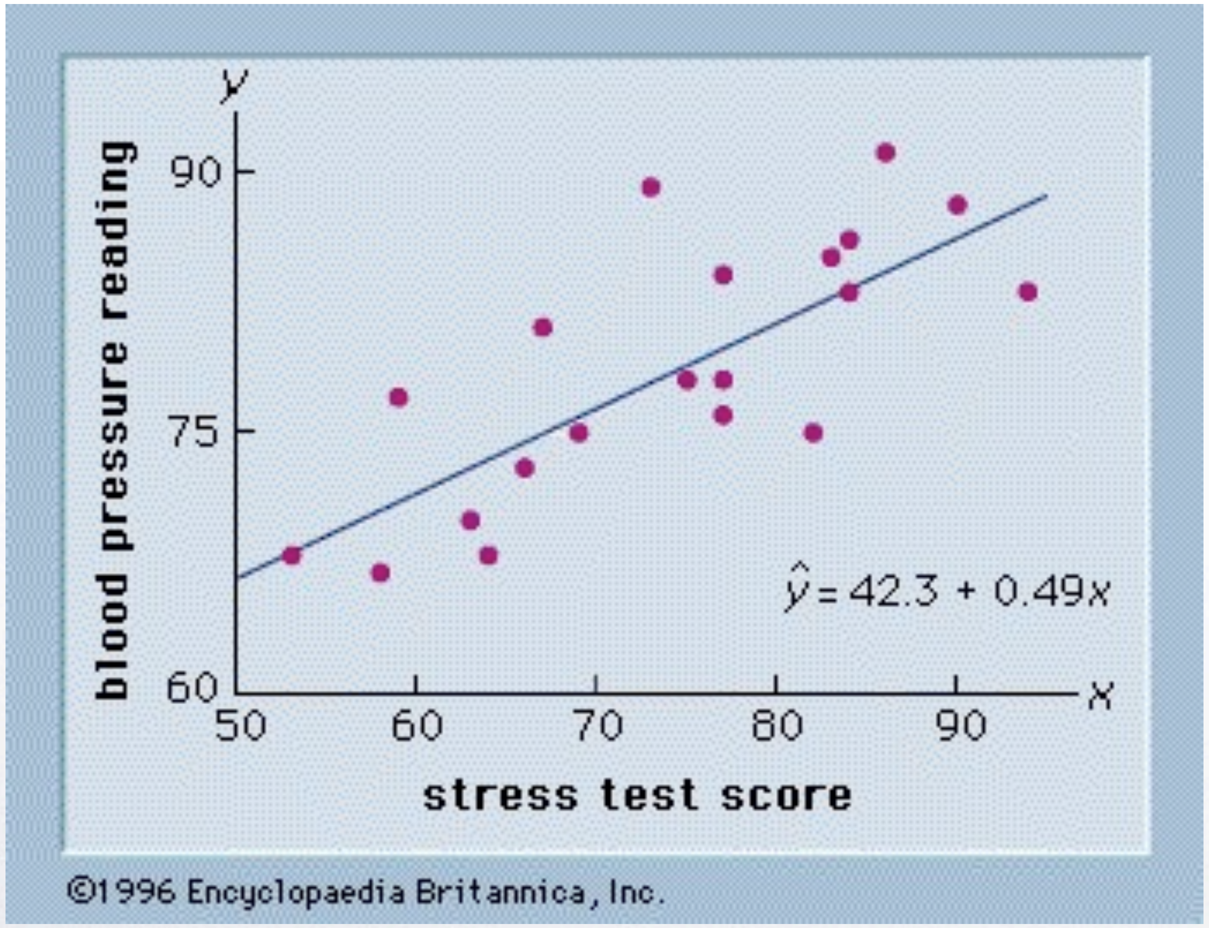
Najčešći kriterij dobrote korelacije:
- korelacija je dobra ukoliko je $|r| \geq 0.5$

U praksi je kriterij često stroži, npr $|r| \geq 0.7$

Dakle, *r* je kvantitativni pokazatelj prisustva ili odsustva korelacije veličina *x* i *y*

Korelacija u linearnoj regresiji

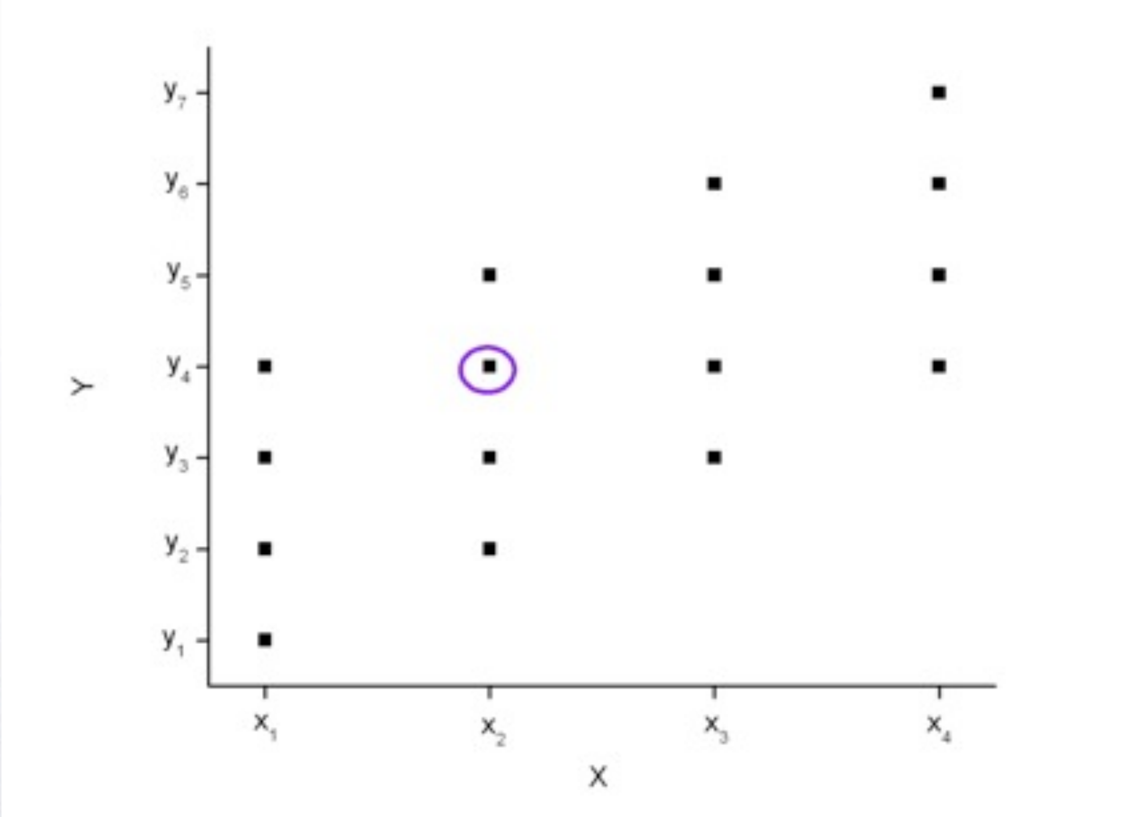
No, kako povući najbolji pravac *y(x)* ili *x(y)*?



Od oka?
Previše proizvoljno

Korelacija u linearnoj regresiji

Pretpostavimo da postoji funkcionalna veza *y(x)* = *ax*+*b*, gdje su *a* i *b* koeficijenti koji nabolje opisuju eksperimentalne rezultate



Preuredimo malo notaciju:

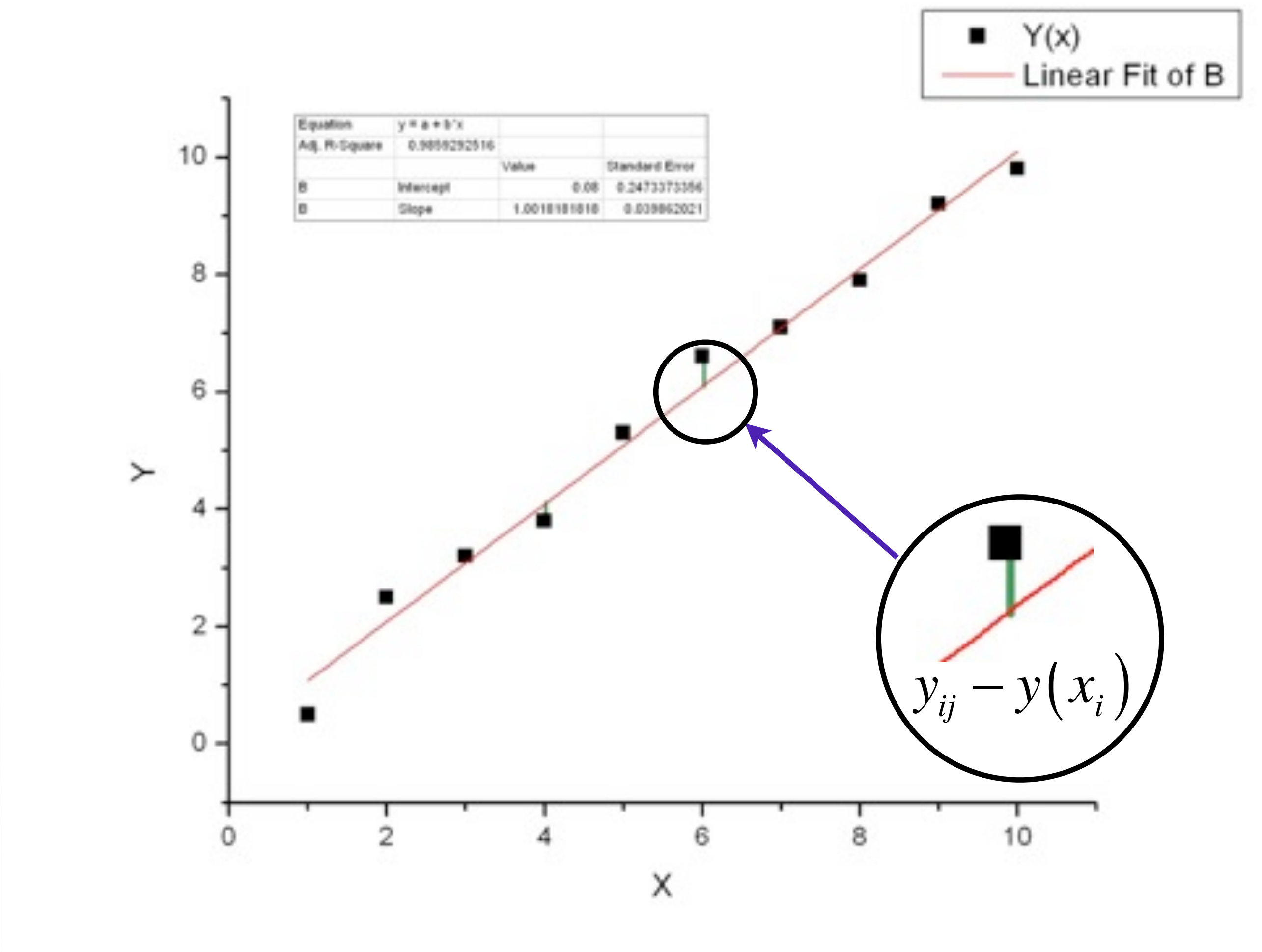
$$\left(x_i, y_j\right) \leftrightarrow \left(x_i, y_{ij}\right)$$

npr. zaokruženu točku možemo pisati

$$\left(x_2, y_4\right) \text{ ili } \left(x_2, y_{24}\right)$$

Primijetimo da vrijedi: $\sum_{i,j} f_{ij} x_i y_j = \sum_{i,j} f_{ij} x_i y_{ij}$

Korelacija u linearnoj regresiji



Uvodimo sume kvadrata odstupanja:

$$S^2 = \frac{1}{N} \sum_{i,j} f_{ij} \left(y_{ij} - y(x_i) \right)^2$$

Moramo kvadrirati razlike, jer dobar opis *y(x)* = *ax*+*b* postiže se samo ako je podjednako pozitivnih i negativnih razlika, pa je:

$$\sum_{i,j} f_{ij} \left(y_{ij} - y(x_i) \right) \approx 0$$

Korelacija u linearnoj regresiji

$$S^2 = \frac{1}{N}\sum_{i,j} f_{ij}\left(y_{ij} - ax_i - b\right)^2$$ $\Rightarrow$ za najbolji odabir, ovaj izraz je minimalan

$$\boxed{\frac{\partial S^2}{\partial a} = 0; \quad \frac{\partial S^2}{\partial b} = 0}$$

$$\frac{\partial S^2}{\partial a} = -\frac{2}{N}\sum_{i,j} f_{ij}x_i\left(y_{ij} - ax_i - b\right) = 0$$

$$\frac{\partial S^2}{\partial b} = -\frac{2}{N}\sum_{i,j} f_{ij}\left(y_{ij} - ax_i - b\right) = 0$$

Korelacija u linearnoj regresiji

To nam daje sustav jednadžbi za *a* i *b*:

$$(1)\quad \sum_{i,j} f_{ij} y_{ij} = a \sum_{i,j} f_{ij} x_i + b \sum_{i,j} f_{ij}$$

$$(2)\quad \sum_{i,j} f_{ij} y_{ij} x_i = a \sum_{i,j} f_{ij} x_i^2 + b \sum_{i,j} f_{ij} x_i$$

Prepoznajmo sljedeće:

$$\sum_{i,j} f_{ij} y_{ij} = \sum_{j} f_j y_j = N\overline{y}; \quad \sum_{i,j} f_{ij} x_i = N\overline{x}$$

$$\sum_{i,j} f_{ij} y_{ij} x_i = N\left(\sigma_{xy} + \overline{x} \cdot \overline{y}\right); \quad \sum_{i,j} f_{ij} = N$$

$$\sum_{i,j} f_{ij} x_i^2 = \sum_{i} f_i x_i^2 = N\vec{x}^2$$

Korelacija u linearnoj regresiji

$$(1)\quad N\overline{y} = aN\overline{x} + bN \Rightarrow \overline{y} = a\overline{x} + b$$

pravac ide kroz točku $(\overline{x}, \overline{y})$

$$(2)\quad N\left(\sigma_{xy} + \overline{xy}\right) = \left(a\overline{x^2} + b\overline{x}\right)N$$

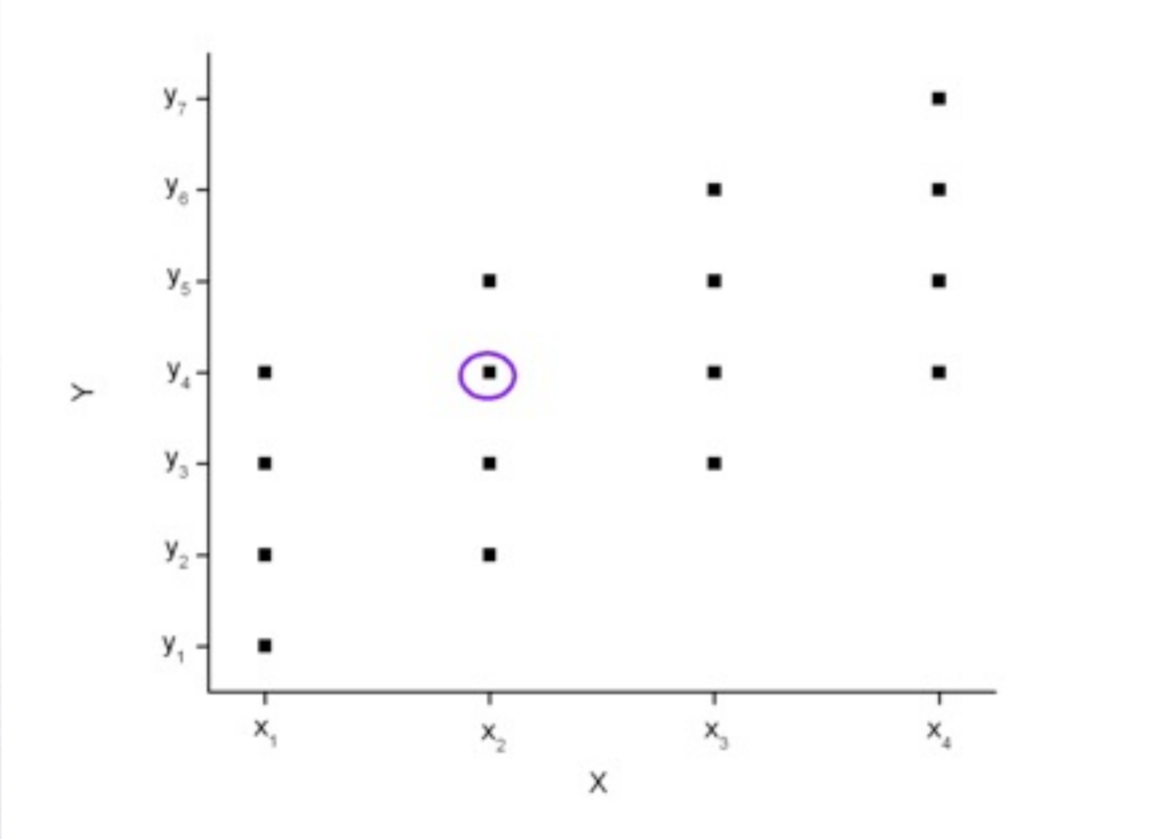
$$\sigma_{xy} + \overline{xy} = a\overline{x^2} + b\overline{x}$$

To sve skupa daje:

$$\boxed{a = \frac{\sigma_{xy}}{\overline{x^2} - \overline{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2}, \quad b = \overline{y} - \frac{\sigma_{xy}}{\sigma_x^2}\overline{x}}$$

Korelacija u linearnoj regresiji

Pravac *y(x)=ax+b* zovemo 'pravac *y* s obzirom na *x*', i on pokazuje kako aritmetičke sredine točaka y_{ij} ovise o *x*



- zaokružena točka je $\overline{y}_2$, tj. aritmetička sredina svih točaka y_{2j}

Korelacija u linearnoj regresiji

Ekvivalentno, možemo promatrati kako $\overline{x}_j$ ovisi o *y*, i tad dobivamo pravac '*x* s obzirom na *y*'

$$x = Ay + B; \quad A = \frac{\sigma_{xy}}{\sigma_y^2} \quad B = \overline{x} - \frac{\sigma_{xy}}{\sigma_y^2}\overline{y}$$

Korelacija u linearnoj regresiji

Pogledajmo još kolika je suma kvadrata odstupanja za *y(x)*:

$$S^2 = \frac{1}{N}\sum_{i,j} f_{ij}\left(y_{ij} - ax_i - b\right)^2 = \frac{1}{N}\sum_{i,j} f_{ij}\left(y_{ij} - \frac{\sigma_{xy}}{\sigma_x^2} x_i - \overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}\overline{x} \right)^2 =$$

$$= \frac{1}{N}\sum_{i,j} f_{ij}\left(\left(y_{ij} - \overline{y}\right) - \frac{\sigma_{xy}}{\sigma_x^2}\left(x_i - \overline{x}\right) \right)^2 =$$

$$= \frac{1}{N}\sum_{i,j} f_{ij}\left(y_{ij} - \overline{y}\right)^2 + \frac{\sigma_{xy}^2}{\sigma_x^4}\frac{1}{N}\sum_{i,j} f_{ij}\left(x_i - \overline{x}\right)^2 - 2\sum_{i,j} f_{ij}\left(y_{ij} - \overline{y}\right)\left(x_i - \overline{x}\right)\frac{\sigma_{xy}}{\sigma_x^2} =$$

$$= \sigma_y^2 + \frac{\sigma_{xy}^2}{\sigma_x^4}\sigma_x^2 - 2\frac{\sigma_{xy}}{\sigma_x^2}\sigma_{xy} = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_y^2\left(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\right) = \sigma_y^2\left(1 - r^2\right)$$

$$\boxed{S^2 = \sigma_y^2\left(1 - r^2\right)}$$

Korelacija u linearnoj regresiji

Očito je sljedeće:

(1) $S^2 > 0$, jer $|r| \leq 1$

(2) Za $r = 1$, $S^2 = 0$, tj. korelacija je potpuna i sve točke leže na pravcu $y = ax+b$