L5. LINEARNA REGRESIJA

PREDOLOČEN SISTEM LINEARNIH ENAČB

Težko dostopen objekt v ravnini xy so opazovali s štirih različnih lokacij:

Prvi opazovalec je izmeril, da objekt leži na premici $a_{11}x + a_{12}y = b_1$.

Drugi —"— $a_{21}x + a_{22}y = b_2$.

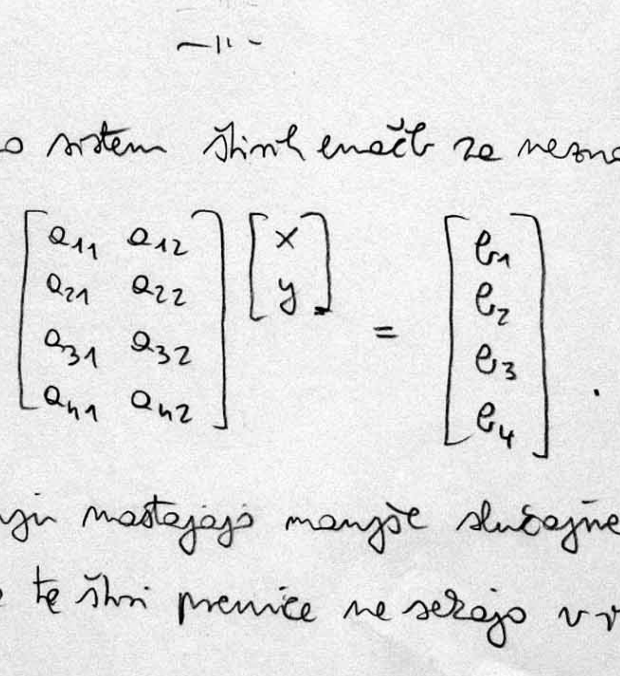
Tretji —"— $a_{31}x + a_{32}y = b_3$

Četrti —"— $a_{41}x + a_{42}y = b_4$.

Dobili smo sistem štirih enačb za neznanki x, y:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}.$$

Pri merjenju nastajajo manjše slučajne napake (šum ...), tako da se te štiri premice ne sekajo v isti točki (slika 1).



(OVERDETERMINED)

Sistem torej formalno ni rešljiv. Pravimo, da je PREDOLOČEN.

Načeloma je vsaka meritev nekaj vredna. Kako izkoristiti vse štiri meritve za kar se da dobro določitev lege objekta?

Oglejmo si splošno situacijo.

Imamo predoločen sistem m linearnih enačb za n neznank:

$$\boxed{A\vec{x} = \vec{y}} \qquad (A \in \mathbb{R}^{m\times n},\ m > n);$$

z več enačbami, kot je neznank. Če že ne moremo najti tak $\vec{x}$, da bi bil $A\vec{x} - \vec{y} = 0$, se bomo zadovoljili s takim $\vec{x}$, da bo

$$\boxed{\|A\vec{x} - \vec{y}\| \text{ minimalna.}}$$

Torej iščemo najboljšo aproksimacijo za $\vec{y}$ v $\mathrm{im}\,A = A\mathbb{R}^n$.
To pa je pravokotna projekcija vektorja $\vec{y}$ na $\mathrm{im}\,A$. Zato je $(A\vec{x} - \vec{y}) \perp \mathrm{im}\,A$, se pravi $\langle A\vec{x} - \vec{y}, A\vec{z}\rangle = 0$ za vse $\vec{z} \in \mathbb{R}^n$.
Od tod je $\langle A^T(A\vec{x} - \vec{y}), \vec{z}\rangle = 0$ za vse $\vec{z} \in \mathbb{R}^n$, torej $(A^TA\vec{x} - A^T\vec{y}) \perp \mathbb{R}^n$ in od tod seveda $A^TA\vec{x} - A^T\vec{y} = 0$:

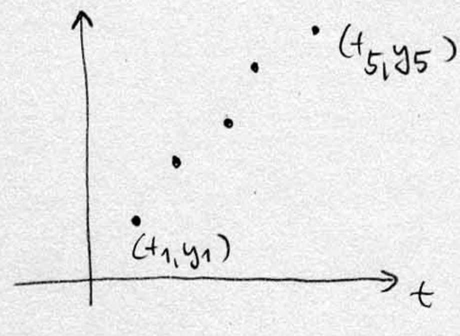
$$\boxed{A^TA\vec{x} = A^T\vec{y}.}$$ (NORMALNA ENAČBA

Tu je A^TA kvadratna matrika velikosti $n \times n$.

*[Če je $\ker A = \{0\}$, je A^TA nesingularna, zato ima pridruženi sistem natanko eno rešitev $\vec{x}$. (Res, če je $A^TA\vec{z} = 0$, je $\langle A^TA\vec{z}, \vec{z}\rangle = \langle A\vec{z}, A\vec{z}\rangle = 0$, zato $A\vec{z} = 0$ in tako $\vec{z} = 0$. Tako je $\ker(A^TA) = \{0\}$.)]

REGRESIJSKA PREMICA

Imamo vrsto meritev skalarne količine y. V času t_i smo zanjo izmerili vrednost y_i $(i=1,2,\ldots,m)$. Narišimo rezultate meritev v koordinatnem sistemu (slika 1).



Slika 1

Iz slike 1 vidimo, da ne manjka dosti, da bi te točke ležale na premici. Pri meritvah nastajajo slučajne napake... Kako narisati premico, ki bi se najbolje prilagajala tem podatkom, tako da bi maksimalno izkoristili rezultate meritev?

Če bi te točke ležale na premici z enačbo $y = pt + q$, bi veljalo

$$t_i p + q = y_i \quad (i=1,2,\ldots,m).$$

Spet imamo predoločen sistem enačb za p, q:

$$\begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \vec{y}.$$

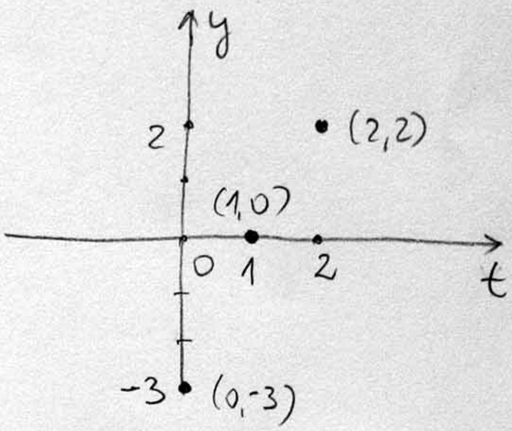
Če pišemo $A = \begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ \cdots & \cdots \\ t_m & 1 \end{bmatrix}$, je $A \begin{bmatrix} p \\ q \end{bmatrix} = \vec{y}$. Rešiti moramo torej normalni sistem:

$$A^T A \begin{bmatrix} p \\ q \end{bmatrix} = \vec{y}.$$

Če je $\begin{bmatrix} p \\ q \end{bmatrix}$ rešitev normalnega sistema, je premica $t \mapsto pt + q$
REGRESIJSKA PREMICA.

Primer 1: Imamo tabelo

t	0	1	2
y	-3	0	2



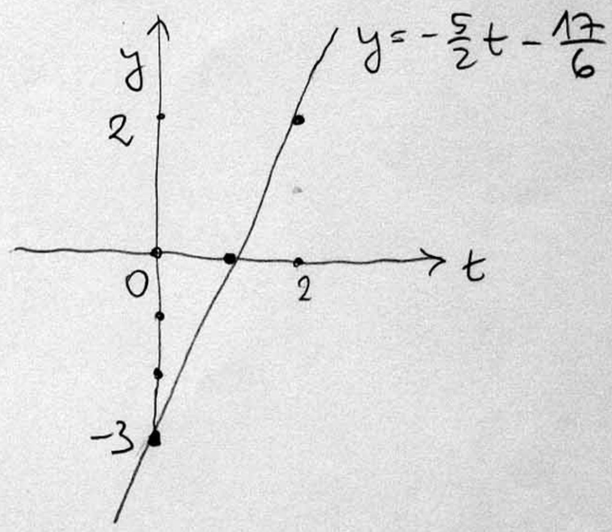
Te tri točke ne ležijo na isti premici (slika 2).

Če bi ležale na premici $y = pt + q$, bi veljalo:

$$p \cdot 0 + q = -3$$
$$p \cdot 1 + q = 0$$
$$p \cdot 2 + q = 2$$

Imamo predoločen sistem:

$$\underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}}_{A} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} -3 \\ 0 \\ 2 \end{bmatrix}$$



Pomnožimo z A^T na obeh straneh, da dobimo normalni sistem:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ 0 \\ 2 \end{bmatrix}$$

ali

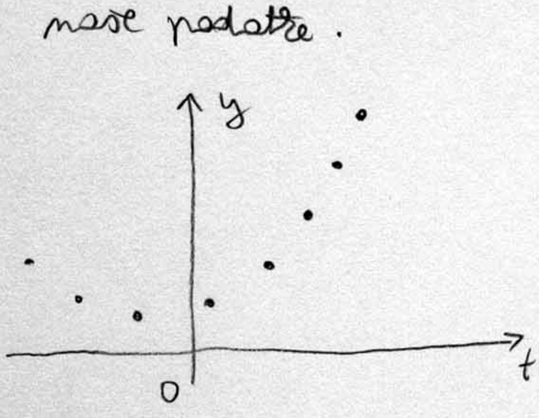
$$\begin{bmatrix} 5 & 3 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} 4 \\ -1 \end{bmatrix}.$$

Od tod je $p = \frac{5}{2}$, $q = -\frac{17}{6}$. Enačba regresijske premice je (slika 3)

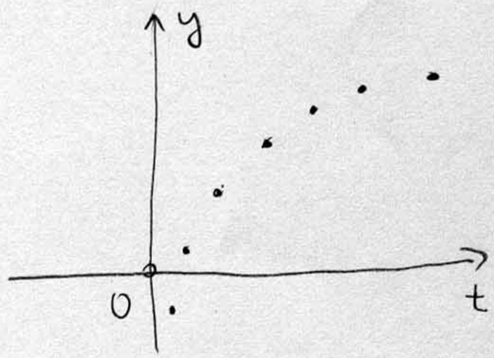
$$\underline{\underline{y = \frac{5}{2}t - \frac{17}{6}}}.$$

LINEARNA REGRESIJA

Včasih je očitno, da premica ni dobra aproksimacija za naše podatke.



Slika 3



Slika 4

Na sliki 3 bi morda poskusili z aproksimacijo $y = p_0 1 + p_1 t + p_2 t^2$, na sliki 4 morda z $y = p_1 1 + p_2 \ln t$.

Denimo, da želimo naše podatke aproksimirati s funkcijo

$$f = p_1 f_1 + p_2 f_2 + \dots + p_n f_n,$$

kjer so $f_1, \dots, f_n$ znane funkcije. Idealno bi bilo najti vektor $\vec{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}$, da bi veljalo

$$f(t_i) = f_1(t_i) p_1 + f_2(t_i) p_2 + \dots + f_n(t_i) p_n = y_i$$

za $i = 1, \dots, m$ ali

$$A \vec{p} = \vec{y},$$

kjer je

$$A = \begin{bmatrix} f_1(t_1), & f_2(t_1), & \dots, & f_n(t_1) \\ f_1(t_2), & f_2(t_2), & \dots, & f_n(t_2) \\ \text{---} & \text{---} & \text{---} & \text{---} \\ f_1(t_m), & f_2(t_m), & \dots, & f_n(t_m) \end{bmatrix}$$

Ker to navadno ne gre, iščemo tak $\vec{p}$, da bo

$$\|A\vec{p} - \vec{y}\| \text{ minimalna.}$$

Od tod sledi zat prej normalna enačba

$$A^T A \vec{p} = A^T \vec{y}. \qquad (1)$$

Vemo, da bo potem $\|A\vec{p} - \vec{y}\|^2$ minimalna, se pravi

$$\sum_{i=1}^{m} \left[f_1(t_i)p_1 + \dots + f_n(t_i)p_n - y_i \right]^2$$

je minimalna. Če je $\vec{p}$ rešitev enačbe (1) in

$$f = \sum_{i=1}^{n} p_i f_i,$$

je torej $\sum_{i=1}^{m} \left[f(t_i) - y_i \right]^2$ minimalna. Vsota kvadratov odstopanj od REGRESIJSKE FUNKCIJE f je torej minimalna. Zato naši metodi rečemo METODA NAJMANJŠIH KVADRATOV.

V letih 1800 - 1809 sta to metodo razvila francoski matematik LEGENDRE in nemški matematik GAUSS. Oba sta preučevala orbite planetov in asteroidov, Gauss pa se je ukvarjal tudi z geodezijo, kjer je bilo treba pogosto ponavljati merjenja.

Primer 2. Iščemo najboljšo aproksimacijo v obliki $y = p_1 + p_2 \ln t$ za tabelo

t	1	2	3	4
y	2	3	3,5	4

Tu je $f_1(t) = 1$ za vsak t in $f_2(t) = \ln t$. Pišimo $\vec{p} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$,

$$A = \begin{bmatrix} 1 & 0 \\ 1 & \ln 2 \\ 1 & \ln 3 \\ 1 & \ln 4 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} 2 \\ 3 \\ 3{,}5 \\ 4 \end{bmatrix}$$

Potem je
$$A^TA = \begin{bmatrix}1 & 1 & 1 & 1\\ 0 & \ln 2 & \ln 3 & \ln 4\end{bmatrix}\begin{bmatrix}1 & 0\\ 1 & \ln 2\\ 1 & \ln 3\\ 1 & \ln 4\end{bmatrix} = \begin{bmatrix}4, & \ln 24\\ \ln 24, & c_{22}\end{bmatrix},$$

kjer je $c_{22} = (\ln 2)^2 + (\ln 3)^2 + (\ln 4)^2 \approx 3,609$.

$$A^Ty = \begin{bmatrix}1 & 1 & 1 & 1\\ 0 & \ln 2 & \ln 3 & \ln 4\end{bmatrix}\begin{bmatrix}2\\ 3\\ 3,5\\ 4\end{bmatrix} = \begin{bmatrix}12,5\\ z_2\end{bmatrix},$$

kjer je $z_2 = 3\ln 2 + 3,5\ln 3 + 4\ln 4 \approx 11,467$.

Normalni sistem je:

$$4p_1 + (\ln 24)p_2 = 12,5$$
$$(\ln 24)p_1 + c_{22}p_2 = z_2$$

Od tod je $p_1 \approx 1,998$, $p_2 \approx 1,419$.

Za aproksimacijo vzamemo $y = 1,998 + 1,419 \ln t$.

Izračunamo lahko:
$$y(1) \approx 1,998 \qquad y(3) \approx 3,557$$
$$y(2) \approx 2,982 \qquad y(4) \approx 3,965,$$

tako da je aproksimacija zelo dobra.

Linearna regresija z grafičnim računalom SHARP

Gremo v STATISTIČNI MENI.

Vnesemo podatke: Prva kartica:

$(t_1) = x_1 = \ldots$

$y_1 = \ldots$

Druga kartica ...

Nato gremo v meni za GRAFIČNE PRIKAZE V STATISTIKI:

Izberemo podmeni REG (regresija).

Izberemo, v kateri množici funkcij bomo iskali aproksimacijo.
Moje računalo ima za linearno regresijo naslednje možnosti:

$a + bx \qquad a + b\log x$

$a + b\ln x \qquad a + \frac{b}{x}$.

Program naredi krivuljo.
Pritisnemo tipko MENU, izberemo REG in dobimo vrednosti za a, b.

NALOGE

1. Poišči regresijsko premico za tabelo

t	-1	0	1	2
y	-7	-3	3	8

2. Za podatke iz primera 2 poišči najboljšo aproksimacijo v obliki $y = p_1 + \frac{p_2}{x}$.