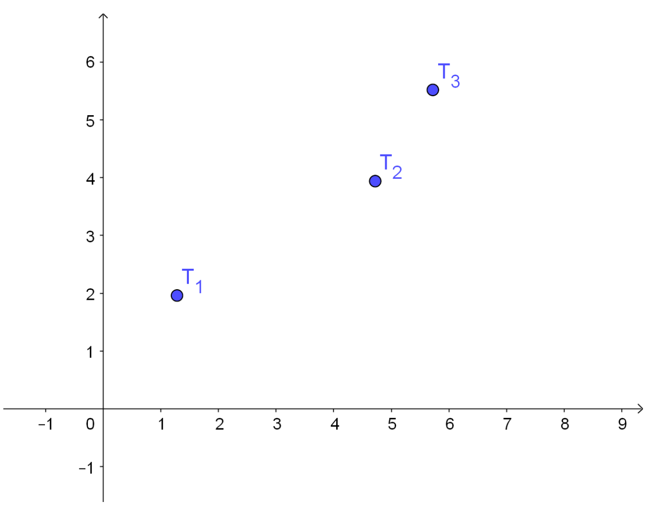
Regresijska premica

Peter Legiša

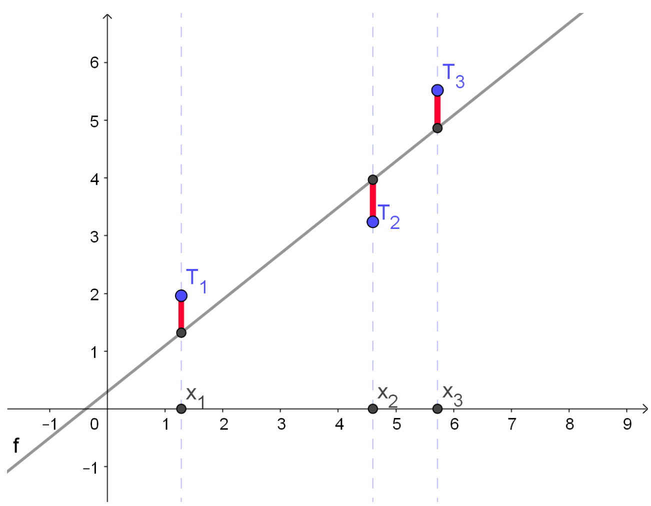
Denimo, da imamo v pravokotnem koordinatnem sistemu na ravnini tri točke, $T_1(x_1, y_1)$, $T_2(x_2, y_2)$, $T_3(x_3, y_3)$, ki sicer ne ležijo na isti premici. Ampak v mislih si vseeno lahko narišemo premico, od katere bodo vse le malo oddaljene. Ilustracijo imamo na sliki 1.



SLIKA 1.
Tri točke, ki ne ležijo na isti premici.

Denimo, da so te tri točke rezultati treh meritev para količin (x, y) in vemo, da je zveza med količinama teoretično linearna v obliki $y = f(x) = px + q$. Tu sta p in q neznana in ju moramo še določiti. Zaradi slučajnih napak pri meritvah prihaja zmeraj do rahlih odstopanj od pravih vrednosti in zato točke niso kolinearne.

Kako bi kar najbolje izkoristili rezultate vseh treh meritev in določili premico, ki se »najbolje« prilega vsem trem meritvam? Radi bi, da premica kar najbolje podaja rezultate meritev, se pravi, da je $f(x_i)$ kolikor je mogoče blizu y_i za $i = 1, 2, 3$. Odstopanja, to so vrednosti $|f(x_i) - y_i|$, naj bi bila kolikor je mogoče majhna. Na sliki 2 smo narisali premico, ki se na oko kar dobro prilega danim podatkom. Dolžine debelo narisanih navpičnih rdečih daljic so velikosti odstopanj meritvenih vrednosti od ustreznih vrednosti na premici. Nekatera odstopanja so navzgor, druga navzdol. Nekako moramo minimizirati ta odstopanja v celoti. Obstajajo močni razlogi (ki jih pa tu ne moremo razlagati), da uporabimo **metodo najmanjših kvadratov**. To pomeni, da skušamo minimizirati vsoto kvadratov odstopanj, torej vsoto kvadratov dolžin rdečih daljic na sliki 2.



SLIKA 2.
Rdeče so narisana odstopanja.

Števili p in q moramo torej izbrati tako, da bo vsota

$$W=(y_1-f(x_1))^2+(y_2-f(x_2))^2+(y_3-f(x_3))^2 \quad (1)$$

minimalna. Premico, ki zadošča tem pogojem, imenujemo *regresijska premica*.

To lahko posplošimo. Če imamo točke $T_1(x_1,y_1)$, $T_2(x_2,y_2),\ldots,T_n(x_n,y_n)$, je **regresijska premica** $y=f(x)=px+q$ tista, za katero je vsota

$$W=(y_1-f(x_1))^2+(y_2-f(x_2))^2+\ldots+(y_n-f(x_n))^2 \quad (2)$$

najmanjša.

V programu Geogebra je regresijska premica imenovana *trendna črta*. Napišemo seznam točk in ukaz *Trendna črta* nam nariše regresijsko premico in da njeno enačbo. Uporabimo lahko tudi program za preglednice Excel. Grafična računala znajo potegniti tako premico.

Z vektorji bomo izpeljali enačbo regresijske premice v primeru treh točk. V ta namen se bomo preselili v trirazsežni prostor. Če vas zanima le rezultat, lahko nadaljujete pri enačbah (3) in (4).

Število W v enačbi (1) je kvadrat razdalje med znano točko $Y(y_1,y_2,y_3)$ s krajevnim vektorjem $\vec{y}=(y_1,y_2,y_3)$ in neznano točko $Z(f(x_1),f(x_2),f(x_3))$ s krajevnim vektorjem $\vec{z}=(f(x_1),f(x_2),f(x_3))$. Ta W mora biti minimalen. Zato mora biti **razdalja med** Y in Z **minimalna**!

Vektor $\vec{z}$ lahko zapišemo, kot lahko takoj preverite, zaradi $f(x_i)=px_i+q$ (za i=1,2,3) takole:

$$\vec{z}=p(x_1,x_2,x_3)+q(1,1,1).$$

Označimo

$$(x_1,x_2,x_3)=\vec{x},$$
$$(1,1,1)=\vec{c}.$$

Torej je

$$\vec{z}=p\vec{x}+q\vec{c}.$$

Tako točka Z leži na ravnini, napeti na znana vektorja $\vec{x}$ in $\vec{c}$. Z drugimi besedami, točka Z leži na ravnini Σ skozi izhodišče ter točki $X(x_1,x_2,x_3)$ in $C(1,1,1)$. Iščemo torej točko Z na ravnini Σ, ki je najmanj oddaljena od točke Y. To je točki Y najbližja točka na ravnini Σ. Geometrija nam pove, da je iskana točka Z **pravokotna projekcija** točke Y na ravnino Σ.

Vektor $\overrightarrow{YZ}=\vec{z}-\vec{y}=p\vec{x}+q\vec{c}-\vec{y}$ je torej pravokoten na vektorja $\vec{x}$ in $\vec{c}$, ki razpenjata ravnino Σ. Tako je skalarni produkt

$$(p\vec{x}+q\vec{c}-\vec{y})\cdot\vec{x}=p\vec{x}\cdot\vec{x}+q\vec{c}\cdot\vec{x}-\vec{y}\cdot\vec{x}=0.$$

Še

$$(p\vec{x}+q\vec{c}-\vec{y})\cdot\vec{c}=p\vec{x}\cdot\vec{c}+q\vec{c}\cdot\vec{c}-\vec{y}\cdot\vec{c}=0.$$

Dobili smo sistem dveh linearnih enačb za p in q:

$$p\vec{x}\cdot\vec{x}+q\vec{c}\cdot\vec{x}=\vec{y}\cdot\vec{x}$$
$$p\vec{x}\cdot\vec{c}+q\vec{c}\cdot\vec{c}=\vec{y}\cdot\vec{c}$$

$$\vdots$$

Izračunamo skalarne produkte v zadnjih dveh enačbah:

$$\begin{aligned}
\vec{y}\cdot\vec{x}&=y_1x_1+y_2x_2+y_3x_3,\\
\vec{x}\cdot\vec{x}&={x_1}^2+{x_2}^2+{x_2}^2,\\
\vec{y}\cdot\vec{c}&=(y_1,y_2,y_3).(1,1,1)=y_1+y_2+y_3,\\
\vec{x}\cdot\vec{c}&=x_1+x_2+x_3,\\
\vec{c}\cdot\vec{c}&=3.
\end{aligned}$$

Sistem enačb za p in q lahko tako prepišemo v enačbi

$$p({x_1}^2+{x_2}^2+{x_3}^2)+q(x_1+x_2+x_3)=y_1x_1+y_2x_2+y_3x_3 \quad (3)$$

$$p(x_1+x_2+x_3)+3q=y_1+y_2+y_3. \quad (4)$$

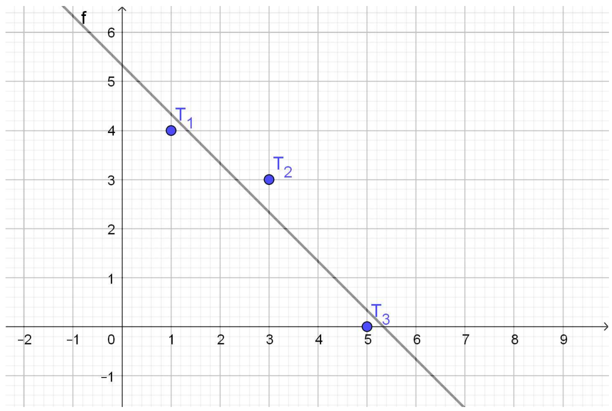
Primer

Vzemimo točke (1,4), (3,3), (5,0) v ravnini. Tako je $x_1=1$, $y_1=4$ itd. Če te podatke vstavimo v (3) in (4), dobimo

$$35p+9q=13,$$
$$9p+3q=7.$$

Pomnožimo drugo enačbo z -3 in prištejmo prvi, pa dobimo $8p = -8$ in tako je $p = -1$. Izračunamo še $q = \frac{16}{3}$. Premica, ki se najbolje prilega danim trem točkam, ima torej enačbo $y = -x + \frac{16}{3}$.

Narisana je na sliki 3.



SLIKA 3.
Regresijska premica

Brez težav boste verjeli, da formuli (3) in (4) lahko posplošimo na primer n točk in dobimo za parametra p in q regresijske premice enačbi:

$$p({x_1}^2 + {x_2}^2 + \ldots + {x_n}^2) + q(x_1 + x_2 + \ldots + x_n) = y_1x_1 + y_2x_2 + \ldots + y_nx_n,$$
$$p(x_1 + x_2 + \ldots + x_n) + nq = y_1 + y_2 + \ldots + y_n.$$

Na koncu povejmo še nekaj o zgodovini metode najmanjših kvadratov. Izvira iz potreb astronomov in geodetov po povečanju natančnosti z večkratnimi meritvami. Dubrovčan Rudjer Bošković je leta 1757 predlagal kriterij za najboljšo aproksimacijo, ki bi na sliki 2 pomenil najti premico, pri kateri je največje odstopanje najmanjše, se pravi, največja med rdečimi daljicami naj bo kolikor je mogoče kratka.

Slavni francoski matematik Pierre Simon Laplace (1754–1827) je avtor izredno vplivne monografije (v petih knjigah) o nebesni mehaniki *Traité de mécanique céleste*. Laplace je predlagal minimiziranje vsote dolžin rdečih daljic na sliki 2.

Francoz Adrien-Marie Legendre je leta 1805 prvi jasno predstavil metodo najmanjših kvadratov. Naredil je tudi primerjavo z Laplaceovo metodo pri računanju oblike Zemlje. Uporabil je kar Laplaceove podatke. Astronomi in geodeti so hitro razumeli prednosti nove metode.

Matematik Carl Friedrich Gauss (1777–1855), ki je kasneje dobil naziv *knez matematikov*, je že leta 1801 uporabil metodo najmanjših kvadratov pri določitvi orbite pritlikavega planeta Cerere. To je odkril italijanski astronom Giuseppe Piazzi. V dobrem mesecu je opravil 24 opazovanj. Potem je zbolel, med njegovo boleznijo pa je Cerera izginila v siju Sonca. Podatke je Piazzi kasneje poslal trem drugim astronomom. Ko naj bi bilo opazovanje spet mogoče, planeta niso več mogli najti. Štiriindvajsetletni Gauss je izdelal novo metodo za približno določevanje orbit in po intenzivnem računanju s Piazzijevemi podatki eno leto po odkritju sporočil, kje naj iščejo Cerero (Ceres). Dejansko so jo našli pol stopinje od predvidenega mesta.

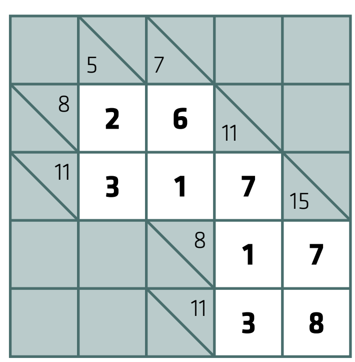
Takratni astronomi si verjetno niti v sanjah niso mogli predstavljati, da bo dobri dve stoletji kasneje (leta 2015) Cerero obiskala in jo od blizu proučila vesoljska sonda Dawn. Vsi prej omenjeni so bili del znanstvene skupnosti, ki je postavila temelje, da je bilo to sploh mogoče.

Gaussova trditev, da je metodo najmanjših kvadratov razvil že leta 1795 (a tega ni objavil), je vodila v spor z Legendrom. Kasneje pa je Gauss veliko uporabnost metode utemeljil z verjetnostnim računom in to tudi objavil.

× × ×

Križne vsote

Rešitev s strani 8



× × ×