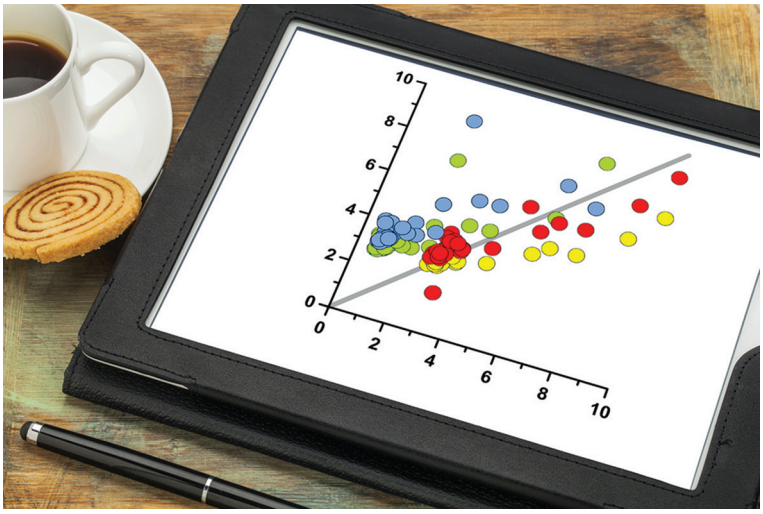
Regresijski pravac



Sanja Varošanec, Zagreb

U Kurikulu za gimnazije i za strukovne škole za treći razred s više od 105 sati nastave matematike pojavljuje se pojam "regresijski pravac" koji u dosadašnjem programu matematike nismo susretali.

U tom se pojmu ogleda korelacija analitičke geometrije i statistike, tj. jednog područja koje na prvi pogled nema veze s geometrijom.

Opišimo situaciju u kojoj se pojavljuje regresijski pravac. U problemu su zadani podatci koji su u obliku uređenih parova. Kad se ti uređeni parovi prikažu kao točke u koordinatnoj ravnini, postavlja se pitanje pronalaženja pravca koji "najbolje aproksimira" zadane točke. Pronalazak takvog pravca omogućava daljnje procjenjivanje i prognoziranje.

Evo jednog zadatka koji se rješava s pomoću regresijskog pravca:

Istraživač želi odrediti postoji li veza između dobi ispitanika i sati koje ispitanik tjedno provede baveći se nekom sportskom aktivnosti. Anketirajući šest ispitanika, dobio je sljedeće podatke:

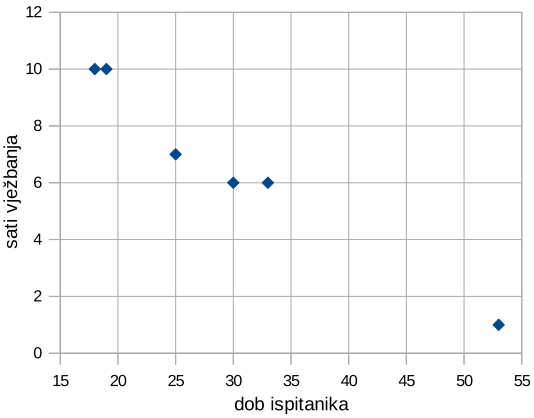
dob (x_i)	sati vježbanja (y_i)
18	10
19	10
25	7
30	6
33	6
53	1

Procijeni koliko bi sati vježbao 40-godišnjak?

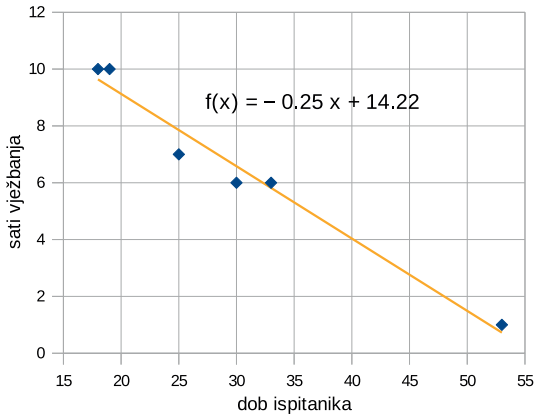
Podatke koji su prikupljeni u anketi prikažimo u obliku uređenih parova koje skicirajmo u koordinatnom sustavu. Za prvu koordinatu uzimamo podatak o dobi ispitanika, a kao drugu koordinatu uzimamo broj sati vježbanja. Grafički prikaz ovih podataka naziva se dijagram rasipanja (engl. *scatter diagram*) i dan je na slici 1.

Oblik dijagrama vodi nas na zaključak da bi se ove točke mogle aproksimirati pravcem. Pravac koji

prof. dr. sc. Sanja Varošanec, Matematički odsjek Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu, varosans@math.hr



Slika 1. Dijagram rasipanja podataka



Slika 2. Dijagram rasipanja podataka i regresijski pravac

najbolje aproksimira dane točke naziva se regresijski pravac. Njegovu jednadžbu obično zapisujemo ovako: $\hat{y} = bx + a$. Oznaka $\hat{y}$ tradicionalno se koristi za zavisnu varijablu regresijskog pravca, koeficijent uz x označava se sa b, a slobodni koeficijent sa a, što odudara od uobičajenih oznaka za koeficijente pravca.

U ovom se trenutku pri rješavanju zadatka obično nude formule za parametre a i b, odnosno upućuje se rješavatelja na upotrebu nekog pomagala kao što je džepni kalkulator koji ima ugrađenu funkciju računanja regresijskog pravca ili na upotrebu proračunskih tablica kao što je Excel, programa dinamičke geometrije kao što je GeoGebra i sl. Drugim riječima, formule za parametre a i b koje glase ovako:

$$b = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$a = \frac{\sum y_i - b\sum x_i}{n}$$

već su ugrađene u razna pomagala. Pri tome je n broj podataka, a $\sum x_i$ jest oznaka za zbroj svih podataka $x_1, \ldots, x_n$ i analogno $\sum y_i = y_1 + \ldots + y_n$.

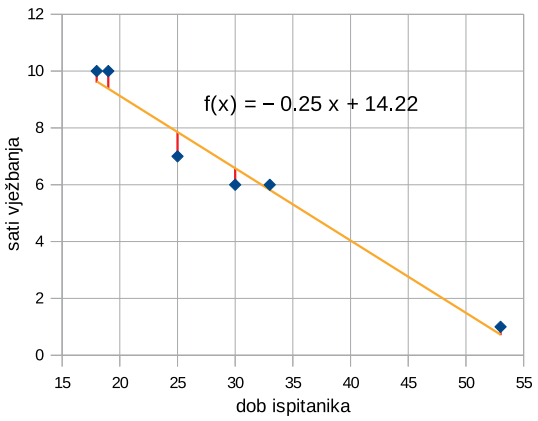
Na slici 2 nalazi se rješenje dobiveno uporabom Excela. Jednadžba regresijskog pravca ispisana je u dijagramu i sad možemo i odgovoriti na pitanje postavljeno u zadatku. Uvrštavajući $x = 40$ u jednadžbu pravca dobivamo $y = 4.04$. Dakle, 40-godišnjak bi vježbao 4.04 sata.

Slične zadatke možete naći u [2], [3], [5] te ostalim novim udžbenicima za treći razred srednje škole. Ovdje se nećemo dalje baviti osmišljavanjem i rješavanjem takvih zadataka, nego ćemo svoju pozornost posvetiti matematici koja stoji iza tih formula.

Cilj ovog članka jest pokazati kako se ove formule izvode, a sam članak može biti osnova nekog projektnog zadatka na ovu temu.

Prikazat ćemo dva dokaza. Jedan se dokaz zasniva na metodi diferencijalnog računa i zbog toga je primjeren onim osobama koje su već upoznate s tim računom. To svakako nisu učenici trećeg razreda. Ali ova je metoda univerzalna i može se koristiti i pri traženju drugih funkcija koje dobro aproksimiraju dane podatke te ju je stoga poželjno poznavati. Drugi se dokaz zasniva na metodi računanja ekstrema kvadratne funkcije koju učenici trećih razreda poznaju jer je ta metoda detaljno obrađena u prethodnom razredu.

Prvo opišimo značenje izraza da neka funkcija "najbolje aproksimira" dane podatke. U problemu su zadane točke (x_i, y_i), $i = 1, \ldots, n$. Ako je $\hat{y}$ funkcija kojom aproksimiramo y, tada su njezine vrijednosti za argumente x_i jednake $\hat{y}_i := \hat{y}(x_i)$ i u koordinatnom sustavu imamo točke $(x_i, \hat{y}_i)$. Razlike ordinata $\hat{y}_i - y_i$ tih točaka (engl. *residual*) na slici 3 označene su crvenom bojom.



Slika 3. Udaljenosti između zadanih točaka i odgovarajućih točaka regresijskog pravca

Ordinate $y_1, \ldots, y_n$ poznatih točaka formiraju jednu n-torku $M = (y_1, y_2, \ldots, y_n)$, dok od ordinata $\hat{y}_i$ formiramo n-torku $\hat{M} = (\hat{y}_1, \ldots, \hat{y}_n)$. Udaljenost tih dviju n-torki (vektora) M i $\hat{M}$ jest veličina na temelju koje za neku funkciju možemo reći da dobro aproksimira dane podatke. Jasno je da kada je $M = \hat{M}$, tada je udaljenost vektora M i $\hat{M}$ jednaka nuli i aproksimativna funkcija $\hat{y}$ upravo prolazi kroz originalno zadane točke. Dakle, problem određivanja funkcije $\hat{y}$ svodi se na minimiziranje udaljenosti vektora M i $\hat{M}$. Za udaljenost dvaju vektora možemo odabrati bilo koju od raznih metrika, a u problemu određivanja regresijskog pravca obično se uzima euklidska metrika koju dobro poznajemo iz dvodimenzionalnog i trodimenzionalnog prostora, a ovdje je proširena na n-dimenzionalni prostor:

$$d(M, \hat{M}) = \sqrt{(\hat{y}_1 - y_1)^2 + \ldots + (\hat{y}_n - y_n)^2}.$$

Budući da nenegativna funkcija i njezin korijen poprimaju ekstreme u istim točkama, problem se svodi na minimizaciju izraza

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

a ova se metoda naziva **metodom najmanjih kvadrata**.

U slučaju kada u familiji linearnih funkcija tražimo onu koja najbolje aproksimira originalne podatke, tada se izraz svodi na sljedeće:

$$F(a, b) = \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

i traži se točka (a_0, b_0) u kojoj funkcija F dviju varijabli a i b postiže svoj minimum.

Dokaz dopunjavanjem do potpunog kvadrata

Riješimo ovaj problem primjereno učenicima trećeg razreda koji su upoznati s postupkom dopunjavanja do potpunog kvadrata. Cilj je funkciju F napisati u obliku

$$F(a, b) = C_1(a - a_0)^2 + C_2(b - b_0)^2 + C_3$$

gdje su C_1 i C_2 pozitivni brojevi. Tada ćemo lako zaključiti da je taj izraz najmanji kad je $a = a_0$ i $b = b_0$.

U prvom nizu koraka stvaramo izraz $C_1(a - a_0)^2$:

$$\begin{aligned}
F(a, b) &= \sum \Big(a - (y_i - bx_i)\Big)^2 \\
&= \sum \Big(a^2 - 2a(y_i - bx_i) + (y_i - bx_i)^2\Big) \\
&= \sum a^2 - 2a\left(\sum y_i - b\sum x_i\right) + \sum (y_i - bx_i)^2 \\
&= na^2 - 2an(\overline{y} - b\overline{x}) + \sum (y_i - bx_i)^2 \\
&= n\Big(a - (\overline{y} - b\overline{x})\Big)^2 - n(\overline{y} - b\overline{x})^2 + \sum (y_i - bx_i)^2 \\
&= n\Big(a - a_0\Big)^2 - n(\overline{y} - b\overline{x})^2 + \sum (y_i - bx_i)^2,
\end{aligned}$$

pri čemu su korištene oznake za a_0 i za aritmetičke sredine $\overline{x}$ i $\overline{y}$:

$$a_0 = \overline{y} - b\overline{x}, \quad n\overline{x} = \sum x_i, \quad n\overline{y} = \sum y_i.$$

U sljedećem nizu koraka stvaramo izraz $C_2(b - b_0)^2$ i pri tome koristimo oznake

$$\sigma_x^2 = \sum x_i^2 - n\overline{x}^2, \quad \sigma_y^2 = \sum y_i^2 - n\overline{y}^2$$

u kojima prepoznajemo formule za disperziju nizova podataka (x_i) i (y_i). Dakle, imamo

$$\begin{aligned}F(a,b) &= n\Big(a-a_0\Big)^2 - n\overline{y}^2 + 2nb\overline{x}\overline{y} - nb^2\overline{x}^2 \\ &\quad + \sum y_i^2 - 2b\sum x_i y_i + b^2\sum x_i^2 \\ &= n\Big(a-a_0\Big)^2 + b^2\Big(\sum x_i^2 - n\overline{x}^2\Big) \\ &\quad - 2b\Big(\sum x_i y_i - n\overline{x}\overline{y}\Big) + \Big(\sum y_i^2 - n\overline{y}\Big) \\ &= n\Big(a-a_0\Big)^2 + \sigma_x^2 b^2 - 2b\Big(\sum x_i y_i - n\overline{x}\overline{y}\Big) + \sigma_y^2 \\ &= n\Big(a-a_0\Big)^2 + \sigma_x^2\left[b - \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sigma_x^2}\right]^2 \\ &\quad - \frac{\left(\sum x_i y_i - n\overline{x}\overline{y}\right)^2}{\sigma_x^2} + \sigma_y^2.\end{aligned}$$

Stavimo li u zadnjoj formuli $b_0 = \dfrac{\sum x_i y_i - n\overline{x}\overline{y}}{\sigma_x^2}$, dobivamo

$$\begin{aligned}F(a,b) &= n\Big(a-a_0\Big)^2 + \sigma_x^2\Big(b-b_0\Big)^2 \\ &\quad - \frac{\left(\sum x_i y_i - n\overline{x}\overline{y}\right)^2}{\sigma_x^2} + \sigma_y^2,\end{aligned}$$

a to je upravo traženi oblik funkcije F iz kojeg očitavamo da se minimum postiže u točki (a_0, b_0) pri čemu vrijedi

$$b_0 = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sigma_x^2} = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

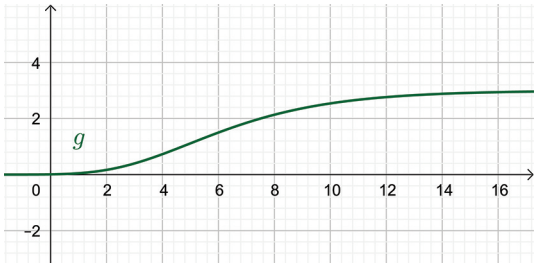
$$a_0 = \overline{y} - b\overline{x} = \frac{\sum y_i - b\sum x_i}{n}.$$

Time smo dokazali formule za koeficijente a i b regresijskog pravca.

Dokaz ispitivanjem ekstrema s pomoću diferencijalnog računa

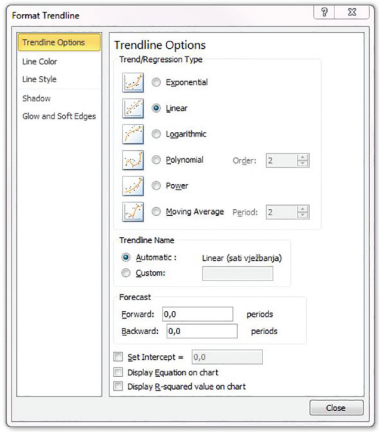
Prikazat ćemo ovdje i drugi dokaz iako se ovaj dokaz ne može provoditi u srednjoj školi jer ispitivanje ekstrema funkcije dviju varijabli nije dio srednjoškolskog gradiva. Ali ovaj je dokaz bitan jer se na taj način traže koeficijenti drugih regresijskih krivulja, ne samo regresijskog pravca.

Naime, u statistici se dani podatci ne aproksimiraju samo s pomoću linearne funkcije. Općenito, u obzir mogu doći sve funkcije. Najčešće se traže funkcije koje su oblika polinoma, eksponencijalne, logaritamske, asimptotske, Gompertzove, logističke funkcije itd. O raznim takvim funkcijama možete pročitati u [4, str. 233–264].



Slika 4. Graf Gompertzove funkcije $f(x) = L \cdot A^{B^{x-C}}$ za $A = 0.5$, $B = 0.7$, $C = 6$, $L = 3$

Neke od tih aproksimacija ugrađene su u proračunske tablice i u džepne kalkulatore. Koristeći se Excelom, osim regresijskog pravca možemo tražiti aproksimativne funkcije koje su eksponencijalne, logaritamske, polinomijalne. U džepnom su kalkulatoru CASIO fx-991ES PLUS primjerice ugrađene, osim regresijskog pravca i regresijske funkcije oblika $y = ab^x$, $y = a + bx + cx^2$, $y = ax^b$.



Slika 5. Izbor regresijskih funkcija u Excelu

S pomoću diferencijalnog računa ekstremi funkcije se traže tako da se odrede stacionarne točke funkcije, tj. točke u kojima je prvi diferencijal jednak nuli. U slučaju kad se radi o funkciji dviju varijabli to se svodi na rješavanje sustava

$$F_a(a,b)=0, \quad F_b(a,b)=0,$$

gdje su F_a i F_b oznake za parcijalne derivacije funkcije F prvog reda po varijablama a, odnosno b.

Vrijedi:

$$F_a(a,b)=\sum_i^n 2(a+bx_i-y_i)=0$$

$$F_b(a,b)=\sum_i^n 2(a+bx_i-y_i)\cdot x_i=0,$$

što se sređivanjem svodi na sustav

$$na+b\sum x_i-\sum y_i=0$$

$$a\sum x_i+b\sum x_i^2-\sum x_iy_i=0$$

u kojemu su nepoznanice a i b. Sustav riješimo metodom supstitucije. Iz prve jednadžbe izrazimo nepoznanicu a i uvrstimo u drugu jednadžbu:

$$a=\frac{\sum y_i-b\sum x_i}{n}$$

$$\frac{\sum y_i-b\sum x_i}{n}\sum x_i+b\sum x_i^2-\sum x_iy_i=0.$$

Kad iz druge jednadžbe izrazimo b, dobivamo

$$b=\frac{n\sum x_iy_i-\left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2-\left(\sum x_i\right)^2}.$$

Time smo dobili stacionarnu točku (a,b). Da bismo pokazali da je to točka u kojoj se postiže minimum, treba proučiti drugi diferencijal u toj točki, odnosno provjeriti dovoljne uvjete koje moraju zadovoljiti parcijalne derivacije drugog reda (vidjeti [1, str. 370]). Ti uvjeti za minimum glase ovako: ako je $F_{aa}F_{bb}-F_{ab}^2>0$ i $F_{aa}>0$, tada funkcija F ima minimum u toj stacionarnoj točki.

Za funkciju F vrijedi

$$F_{aa}=\sum 2=2n$$

$$F_{ab}=\sum 2x_i=2\sum x_i$$

$$F_{bb}=2\sum x_i^2$$

te se lako provjeri da su zadovoljeni dovoljni uvjeti za minimum. Dakle, u točki (a,b) funkcija F postiže svoj minimum i brojevi a i b su koeficijenti pravca koji najbolje aproksimira dane originalne podatke.

Ovime smo formule za koeficijente regresijskog pravca dokazali na dva načina. Jedan način izlazi van okvira srednjoškolskog gradiva, ali nam daje uvid u postupak određivanja raznih drugih regresijskih krivulja. Drugi nam način ilustrira kako metoda obrađena u drugom razredu može biti vrlo efikasna i u rješavanju problema iz područja statistike.

LITERATURA

1/ I. N. Bronštejn i dr. (1975.): *Matematički priručnik za inženjere i studente*, Tehnička knjiga, Zagreb.

2/ B. Dakić, N. Elezović (2020.): *Matematika 3, 2. dio, udžbenik za treći razred gimnazija, 4 ili 5 sati nastave tjedno*, Element, Zagreb.

3/ *Matematika 1, 6. Linearne funkcije, 6.5. Modeliranje linearnom funkcijom*, **edutorij.e-skole.hr**, pristupljeno 21.5.2020.

4/ V. Serdar, I. Šošić (1981.): *Uvod u statistiku*, Školska knjiga, Zagreb.

5/ S. Varošanec (2020.): *Matematika 3, udžbenik za treći razred gimnazija i strukovnih škola, 3 ili 4 sati nastave tjedno*, Element, Zagreb.