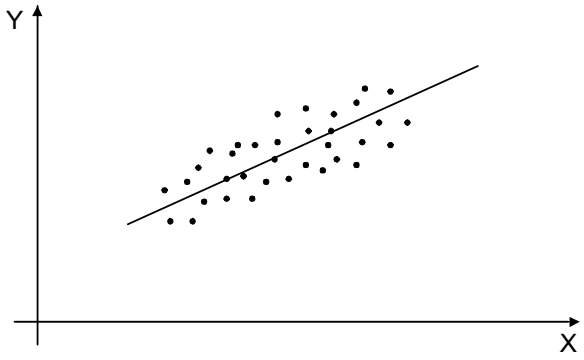
2.9. Regresiona analiza

U prethodnom tekstu je navedeno da se u ekonomskim analizama mogu koristiti različite matematičke funkcije za opisivanje zavisnosti između posmatranih veličina. Za funkciju ukupnih troškova, na primer, navedeno je pet različitih oblika funkcija koje su u upotrebi. Do odgovora koju od mogućih funkcija treba koristiti u konkretnom slučaju dolazi se empirijskim istraživanjima.

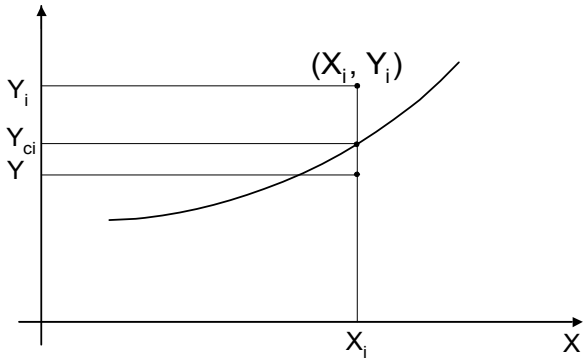
Istraživanja se zasnivaju na sakupljanju i obradi relevantnih statističkih podataka. Hipotezu o statističkoj zavisnosti posmatranih slučajnih promenljivih potrebno je zatim jasno formulisati i testirati.

Ovde razmatramo najjednostavniji slučaj kada postoje samo dve promenljive X i Y, npr. obim proizvodnje i troškovi. Promenljiva X se po pravilu naziva nezavisno promenljivom, a promenljiva Y zavisno promenljivom. Prvi korak je sakupiti podatke $X_1,X_2,...,X_n$ i odgovarajuće podatke $Y_1,Y_2,...,Y_n$, npr. podatke o obimu proizvodnje i odgovarajućim troškovima. Podaci se obično prikazuju u tabelama. Sledeći uobičajeni korak je nacrtati tačke (X_1,Y_1), (X_2,Y_2), ... , (X_n,Y_n) u koordinatnoj ravni X0Y. Rezultujući skup tačaka naziva se dijagram rasejanja, skater dijagram ili skatergram.

Sa dijagrama rasejanja često se lako uočava da postoji neka glatka kriva koja aproksimira podatke. Takva kriva naziva se aproksimativna kriva. Na slikama 2.12. i 2.13. prikazana su dva dijagrama rasejanja i dve moguće aproksimativne krive.



Slika 2.12. Dijagram rasejanja



Slika 2.13. Dijagram rasejanja

Podaci na slici 2.12. dobro se aproksimiraju pravom linijom pa se zaključuje da između posmatranih promenljivih postoji linearna relacija. Podatke na slici 2.13. prava linija ne bi dobro aproksimirala što znači da između posmatranih promenljivih postoji neka nelinearna relacija.

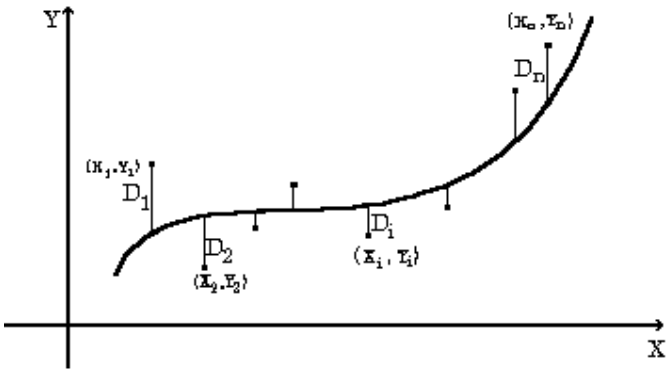
Problem određivanja glatke krive koja dobro aproksimira razmatrane podatke naziva se problem fitovanja krive**.** Ako je analiza usmerena na proveravanje hipoteze da se radi o međusobno zavisnim veličinama pa treba utvrditi zakon koji opisuje tu zavisnost, onda se problem fitovanja krive naziva problemom regresije. (Sama reč regresija je izvorno značila nazadovanje, a u statistici i vraćanje na srednju vrednost.) Odgovarajuća kriva se naziva regresionom krivom ili regresionom linijom. Kada se čitava analiza radi da bi se procenila vrednost zavisno promenljive za neku (proizvoljnu) vrednost nezavisne promenljive, onda se kaže da se rešava problem estimacije.

Do krive koja dobro fituje podatke može se doći slobodnom rukom. Mnoge linearne zavisnosti klasično se utvrđuju upravo na taj način: student ili analitičar najpre na dijagramu rasejanja pomoću lenjira nacrta pravu liniju a potom određuje njene parametre. Ovaj način ima ograničene mogućnosti i zavisi od individualnih procena. To nije u skladu sa opštom namerom naučne metode da se rezultat, u ovom slučaju regresiona linija, odredi na što je moguće objektivniji način. Da bi se taj cilj postigao, obično se za određivanje regresione linije koristi metoda najmanjih kvadrata.

Kada se umesto originalnih podataka o dve promenljive koristi regresiona linija, onda se za neko određeno X_i dobija na regresionoj liniji vrednost Y_{ci} koja se u opštem slučaju razlikuje od originalnog podatka Y_i . Razlika

$$e_i = Y_i - Y_{ci} \qquad (2.32)$$

naziva se greška, devijacija ili rezidual. Ona može biti pozitivna, negativna ili jednaka nuli, slika 2.14.



Slika 2.14. Greška fitovanja

Kao mera koja pokazuje koliko dobro neka kriva aproksimira date podatke koristi se zbir kvadrata greške

$$F = e_1^2 + ... + e_n^2 . \qquad (2.33)$$

Od svih aproksimativnih krivih ona kriva koja za dati skup podataka ima osobinu da je zbir kvadrata greške minimalan naziva se najbolja kriva fitovanja.

Ako bi se problem modifikovao tako da se kao zavisno promenljiva posmatra X a kao nezavisno promenljiva Y, onda bi u račun trebalo uzeti horizontalne umesto vertikalnih devijacija, odnosno greške po X, a ne po Y. Tako dobijena kriva najmanjih kvadrata ne bi se poklopila sa prethodnom. Moguće je kao krivu najmanjih kvadrata koristiti i krivu računatu sa normalnim odstojanjem tačke od linije, ali se takav pristup u praksi primenjuje mnogo ređe.

2.9.1. Linearna regresija

Pokažimo kako se određuju parametri linearne regresije

$$Y = a_0 + a_1 X . \qquad (2.34)$$

Neka je dato n parova podataka (X_i,Y_i). Radi jednostavnosti pisanja u ovom delu teksta izostavljaćemo indekse promenljivih X i Y kada se ove nalaze ispod znaka za sumiranje

$$\sum X = \sum_{i=1}^{n} X_i$$

$$\sum Y = \sum_{i=1}^{n} Y_i \qquad (2.35)$$

$$\sum XY = \sum_{i=1}^{n} X_i Y_i \ .$$

Zbir kvadrata grešaka je funkcija parametara a_0 i a_1

$$F(a_0, a_1) = \Sigma (Y - a_0 - a_1X)^2 \qquad (2.36)$$

Potrebni uslovi za minimum funkcije $F(a_0, a_1)$ su

$$\frac{dF}{da_0} = 0$$

$$\frac{dF}{da_1} = 0$$

koji daju sistem jednačina

$$n\, a_0 + a_1 \Sigma X = \Sigma Y$$

$$a_0 \Sigma X + a_1 \Sigma X^2 = \Sigma XY. \qquad (2.37)$$

Dobijeni sistem jednačina čije rešavanje daje parametre regresione linije naziva se sistem normalnih jednačina.

Parametri a_0 i a_1 mogu se odrediti sledećim formulama

$$a_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$a_0 = \frac{(\sum X^2)(\sum Y) - (\sum XY)(\sum X)}{n\sum X^2 - (\sum X)^2}, \text{ ili}$$

$$a_0 = \frac{\sum Y - a_1 \sum X}{n} \qquad (2.38)$$

Označimo srednje vrednosti podataka sa X_{sr} i Y_{sr}

$$X_{sr} = \Sigma X/n \qquad Y_{sr} = \Sigma Y/n \qquad (2.39)$$

i uvedimo smene $x_i = X_i - X_{sr}$ $y_i = Y_i - Y_{sr}$. Jednačina linije najmanjih kvadrata može se tada napisati

$$y = \frac{\sum xy}{\sum x^2} x \qquad \text{i l i} \qquad y = \frac{\sum xY}{\sum x^2} x. \qquad (2.40)$$

U posebnom slučaju kada su podaci takvi da je $\Sigma X = 0$, tj. $X_{sr} = 0$, jednačina regresione linije postaje

$$Y = Y_{sr} + \frac{\sum XY}{\sum X^2} X. \qquad (2.41)$$

Iz ovih jednačina je jasno da linija najmanjih kvadrata prolazi kroz tačku (X_{sr}, Y_{sr}) koja se zove centroid ili centar gravitacije podataka.

Opisani postupak odgovara pretpostavki da je Y zavisna, a X nezavisna promenljiva. Ako se X posmatra kao zavisna, a Y kao nezavisna

promenljiva, onda treba odrediti parametre b_0 i b_1 jednačine $X = b_0 + b_1Y$. Rezultujuća linija, kao što je već rečeno, ne poklapa se u opštem slučaju sa prethodnom.

Primer 2.14. Sledeća tabela daje podatke o obimu proizvodnje (u tonama) i troškovima (u hiljadama dinara) za jedno preduzeće.

Q /tone/	30	31	34	36	37	38	39
C /hiljade din/	58	60	61	62	65	72	75

Odrediti zavisnost troškova od obima proizvodnje.

Rešenje: Mada danas i xepni kalkulatori imaju ugrađene programe za računanje parametara linearne regresije, ovde se daje klasičan postupak koji se oslanja na podatke i međurezultate date u sledećoj tabeli

i	Q	C	QC	Q^2	C_c
1	30	58	1740	900	56,6
2	31	60	1860	961	58,2
3	34	61	2074	1156	63,1
4	36	62	2232	1296	66,3
5	37	65	2405	1369	68,0
6	38	72	2736	1444	69,6
7	39	75	2925	1521	71,2
Σ	245	453	15972	8674	

Sistem normalnih jednačina je

$$7\,a_0 + 245\,a_1 = 453$$

$$245\,a_0 + 8674 a_1 = 15972$$

a njegovo rešenje $a_0 = 7,84$ i $a_1 = 1,625$ tj.

$$C = 7,84 + 1,625\,Q.$$

U poslednjoj koloni prethodne tabele date su vrednosti troškova izračunate na osnovu dobijene regresione linije.

Da je zadatak bio utvrditi zavisnost obima proizvodnje Q od ukupnih troškova $Q = b_0 + b_1C$, dobila bi se regresiona prava $Q = 4,4 + 0,47\,C$ koja se razlikuje od prethodne. ♦

2.9.2. Nelinearna regresija

Za opisivanje regresione linije često se koriste polinomne jednačine

$Y = a_0 + a_1X + a_2X^2$ Parabola ili kvadratna kriva

$Y = a_0 + a_1X + a_2X^2 + a_3X^3$ Kubna kriva

$Y = a_0 + a_1X + ... + a_kX^k$ Kriva k-tog stepena

Desne strane gornjih jednačina zovu se polinomi drugog, trećeg i k-tog stepena, respektivno. Sledeći sistem normalnih jednačina za krivu k-tog stepena $Y = a_0 + a_1X + ... + a_kX^k$ dobijen je izjednačavanjem sa nulom parcijalnih izvoda fukcije zbira kvadrata greške po parametrima

$$na_0 + a_1\Sigma X + \ldots + a_k\Sigma X^k = \Sigma Y$$
$$a_0\Sigma X + a_1\Sigma X^2 + \ldots + a_k\Sigma X^{k+1} = \Sigma XY$$
$$\ldots$$
$$a_0\Sigma X^k + a_1\Sigma X^{k+1} + \ldots + a_k\Sigma X^{2k} = \Sigma X^kY \qquad (2.42)$$

Od velikog broja drugih mogućih jednačina za regresione linije u praksi se sledeće koriste veoma često jer se određivanje njihovih parametara svodi na postupak računanja parametara linearne regresije.

$Y = \frac{1}{a_0 + a_1X}$ ili $\frac{1}{Y} = a_0 + a_1X$ Hiperbola

$Y = ab^X$ ili $\log Y = \log a + (\log b)X = a_0 + a_1X$ Eksponencijalna kriva

$Y = aX^b$ ili $\log Y = \log a + (\log b)X = a_0 + a_1X$ Geometrijska kriva

$Y = ab^X + g$ Modifikovana eksponencijalna kriva

$Y = aX^b + g$ Modifikovana geometrijska kriva

$Y = pq^{b^X}$ ili $\log Y = \log p + b^X \log q = ab^X + g$ Gomprecovakriva

$Y = pq^{b^X} + h$ Modifikovana Gomprecova kriva

$Y = \frac{1}{ab^X + g}$ ili $\frac{1}{Y} = ab^X + g$ Logisti ~ka kriva

$Y = a_0 + a_1(\log X) + a_2(\log X)^2$

Pokažimo kako se problem određivanja parametara modifikovane geometrijske krive $Y = aX^b + g$ svodi na problem linearne regresije.

Prebacivanjem g na levu stranu i logaritmovanjem jednačine dobija se $\log (Y-g) = \log a + b \log X$.

Smenom $y = \log(Y-g)$, $a_o = \log a$, $a_1 = b$, i $x = \log X$ dobija se linearna jednačina $y = a_0 + a_1X$.

Primer 2.15. U fabrici kablova privode se kraju pripreme za proizvodnju novog kabla čija će prodajna cena iznositi 3,78, a troškovi

proizvodnje 2,93 novčanih jedinica (n.j) po dužnom metru. Zavisnost tražnje x od cene p sličnih proizvoda data je tabelom.

Cena /n.j. po m/	2,00	2,50	3,00	4,00	5,00
Tražnja / km/	880	810	750	680	670

Koliku godišnju dobit treba očekivati od prodaje ovog kabla ako se pretpostavi sledeća teorijska zavisnost tražnje od cene $x = a\,p^{b}$?

Rešenje: Najpre treba predvideti tražnju koja odgovara ceni p=3,78n.j. U tu svrhu odredićemo parametre a i b koristeći znanje o linearnoj regresiji.

Logaritmovanjem se dobija

$$\log x = \log a + b \log p$$

Smenama $X = \log x$, $A = \log a$ i $P = \log p$ problem se prevodi u određivanje parametara A i b linearne regresije

$$X = A + bP.$$

Rešavanjem se dobija $A \approx 3{,}03$, $a = 1075$ i $b \approx -0{,}311$, tj.

$$x = 1075\,p^{-0{,}311}\ /km/$$

Tražnja za cenu p = 3,78 je x(3,78) = 710,9km. Dobit po dužnom metru je d = p - c = 3,78 - 2,93 = 0,85, a ukupna dobit D = 604 265n.j. ♦

Dijagram rasejanja pomaže pri odlučivanju koju od navedenih krivih treba koristiti. Ukoliko dijagram pokazuje nelinearnu zavisnost između promenljivih, korisnim se može pokazati dijagram transformisanih promenljivih. Naprimer, za geometrijsku krivu treba koristiti logaritme originalnih podataka ili specijalni, logaritamski kalibrisan papir.

2.9.3. Korelacija

Regresiona linija izražava prirodu odnosa između dve promenljive. Regresiona jednačina pokazuje kako se zavisno promenljiva menja kao rezultat promene nezavisno promenljive. Parametri regresije se izračunavaju na osnovu statističkih podataka korišćenjem opisanih algoritama. Sami parametri ne sadrže informaciju koliko dobro regresiona linija reprezentuje originalne podatke. U svrhu utvrđivanja stepena zavisnosti između posmatranih veličina treba uraditi dodatnu statističku analizu. Stepen zavisnosti naziva se korelacijom. Ukoliko su odstojanja realnih podataka od regresione linije manja, korelacija je veća, odnosno, regresija bolje opisuje posmatrani skup podataka.

Da bismo uveli meru korelacije između posmatranih promenljivih, podsetimo se definicija standardne devijacije i varijanse.

Kada se posmatra samo skup šY¡ć, definiše se standardna devijacija S_Y promenljive Y

$$S_Y = \sqrt{\frac{\sum (Y - Y_{sr})^2}{n}} = \sqrt{\frac{\sum y^2}{n}}$$

Analogno se definiše standardna devijacija S_X promenljive X

$$S_X = \sqrt{\frac{\sum (X - X_{sr})^2}{n}} = \sqrt{\frac{\sum x^2}{n}}$$

Standardna devijacija promenljive Y za dato X je

$$S_{Y.X} = \sqrt{\frac{\sum (Y - Y_c)^2}{n}}$$

a standardna devijacija X za dato Y

$$S_{X.Y} = \sqrt{\frac{\sum (X - X_c)^2}{n}}$$

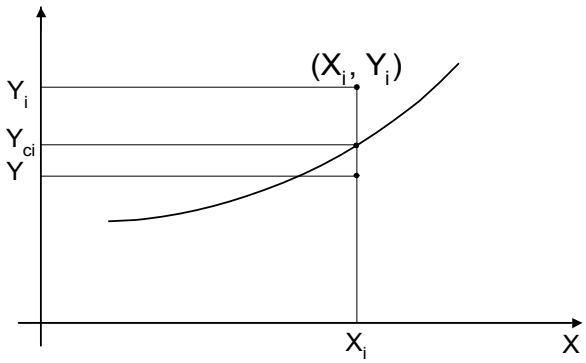
Za male uzorke koriste se modifikovane standardne devijacije, npr.

$$S_{Y.X} = \sqrt{\frac{\sum (Y - Y_c)^2}{n-2}}$$

Varijansa je kvadrat odgovarajuće standardne devijacije.

Ukupna greška $e_i = Y_i - Y_{sr}$ razlaže se na dva dela: objašnjenu grešku $Y_{ci} - Y_{sr}$ i neobjašnjenu grešku $Y_i - Y_{ci}$, slika 2.15.

$e_i = (Y_i - Y_{ci}) + (Y_{ci} - Y_{sr}) = e_{in} + e_{i0}$.



Slika 2.15. Ukupna, objašnjena i neobjašnjena greška

Greška e_{i0} se zove objašnjenom jer je potpuno određena regresionom linijom. Neobjašnjena greška je potpuno slučajna. Jasno je da regresiona linija

bolje opisuje date podatke ukoliko su neobjašnjene greške manje. Kada neobjašnjenih grešaka ne bi bilo, regresija bi bila perfektna i svi podaci bi ležali na regresionoj liniji.

Varijacija je zbir kvadrata greške. Ukupna varijacija je zbir kvadrata ukupnih grešaka i jednaka je zbiru objašnjene i neobjašnjene varijacije

$$\sum (Y - Y_{sr})^2 = \sum (Y - Y_c)^2 + \sum (Y_c - Y_{sr})^2$$

Koeficijent determinacije je odnos objašnjene i ukupne varijacije greške

$$d = \frac{\text{obja\{wena varijacija}}{\text{ukupna varijacija}} = \frac{\sum (Y_c - Y_{sr})^2}{\sum (Y - Y_c)^2}$$

Jasno je da regresiona linija bolje aproksimira statističke podatke ako je koeficijent determinacije bliži jedinici.

Za određivanje stepena korelacije dve promenljive uobičajeno se koristi koeficijent korelacije r

$$r = \pm \sqrt{d}\,.$$

Koeficijent korelacije r se odnosi na procenat varijacije u Y koji se može objasniti regresionom linijom.

Koeficijent korelacije je između -1 i 1. Znak koeficijenta korelacije ukazuje da li sa rastom X promenljiva Y opada, r negativno, ili raste, r pozitivno.

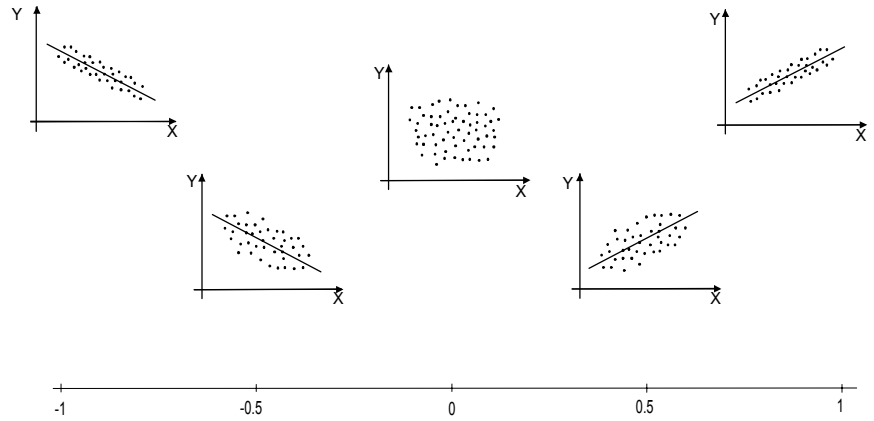
Često korišćeni obrasci za koeficijent korelacije su

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}}$$

$$r = \sqrt{1 - \frac{S^2_{Y.X}}{S^2_Y}}$$

Ne postoji jedno čvrsto pravilo kako da se tumači konkretna vrednost koeficijenta korelacije već to zavisi od konkretnog istraživanja. Jedno opšte pravilo je sledeće. Ako je r > 0,7, kaže se da postoji jaka korelacija između posmatranih promenljivih. Ukoliko je $0{,}5 < r < 0{,}7$, kaže se da podaci ukazuju da između posmatranih promenljivih postoji korelacija. Za r < 0,4 kaže se da podaci ne ukazuju na međusobnu korelisanost posmatranih promenljivih, slika 2.16.



Slika 2.16. Ilustracija koeficijenta korelacije

Primer 2.16. Praćenjem troškova i obima proizvodnje u toku poslednjih dvanaest meseci dobijeni su podaci prikazani u tabeli.

Mes	mar	apr	maj	jun	jul	avg	sep	okt	nov	dec	jan	feb
C	34,4	34,9	34,6	32,2	31,8	33,8	34,3	34,2	36,3	38,2	37,1	36,2
Q	21	19	23	18	17	19	22	24	21	26	24	25

Primenom metode najmanjih kvadrata odrediti funkciju aproksimativne prave linije koja pokazuje kako se ukupni troškovi menjaju u zavisnosti od obima proizvodnje. Izračunati koeficijent korelacije. Ako se planom za sledeći mesec predviđa obim proizvodnje 25, kolike troškove treba očekivati.

Rešenje: Ranije opisanim postupkom dobija se a_0=23,64 i a_1=0,52, tj. C=23,64+0,52Q. Koeficijent korelacije je 80,4% što ukazuje na jaku korelaciju između prikazanih troškova i obima proizvodnje. Za Q=25 dobija se C_c = 36,6. ♦

Jaka korelacija između dve posmatrane stohastičke veličine ne mora da znači da između njih postoji uzročno posledična veza. Zbog toga treba biti obazriv u tumačenju i korišćenju regresionih linija i koeficijenta korelacije.