3.3. Utilizarea regresiei si corelatiei liniare pentru analiza si luarea deciziilor

3.3.1. Mod de tratare si rezolvare specific

Despre metoda de obtinerea unei drepte de regresie - metoda celor mai mici patrate

Metoda celor mai mici patrate este cea mai utilizata metoda de obtinere a dreptei de regresie.

In cazul oricarui set de date bivariabile se pot obtine doua tipuri de drepte de regresie.

a) dreapta de regresie de forma "y in raport cu x" reprezinta acea dreapta de regresie care se foloseste pentru estimarea lui y in functie de valoarea lui x(care se da).

b) dreapta de regresie de forma "x in raport cu y" reprezinta acea dreapta de regresie care se foloseste pentru estimarea lui x in functie de valoarea lui y(care se da).

Aceste doua drepte de regresie sunt complet diferite.

In cazul in care se utilizeaza o diagrama de dispersie pentru a reprezenta grafic datele unui set bivariabila, pot fi trasate mai multe drepte de regresie. Dintre acestea, dreapta de regresie de forma "y in raport cu x" determinata prin metoda celor mai mici patrate este cea pentru care suma patratelor abaterilor verticale (ale tuturor punctelor fata de dreapta) este minima. Orice alta dreapta care trece prin punctul mediu al datelor de mai sus are suma patratelor abaterilor mai mare decat cea inregistrata in cazul dreptei de regresie.

In cazul metodei celor mai mici patrate (folosita pentru obtinerea dreptei de regresie de forma "y in raport cu x"), ecuatia dreptei este de forma $y = a + b \cdot x$, putem determina valorile celor doi parametrii necunoscuti a si b cu ajutorul urmatoarelor formule:

$$b = \frac{n \cdot \sum x \cdot y - \sum x \cdot \sum y}{n \sum x^2 - \left(\sum x\right)^2} \quad \text{si} \quad a = \frac{\sum y}{n} - b \cdot \frac{\sum x}{n}$$

Una din propr-le dreptei de regresie de tip "y in raport cu x" determ. pe baza acestei metode este urmat.: in cazul oricarui set de date bivariabile, dreapta de regresie a celor mai mici patrate trece obligatoriu prin punctul mediu (x,y) al datelor.

Despre corelatia liniara

Scopul analizei regresiei este de a identifica relatia care exista intre elementele unui set de date bivariabile, si totusi ea nu furnizeaza nici o informatie referitoare la cat de buna este aceasta relatie.

Corelatia arata intensitatea rationalizarii dintre doua variabile prin masurarea gradului de "imprastiere" a datelor inregistrate, in jurul dreptei de regresie determinate prin metoda celor mai mici patrate.

Este un numar cuprins in intervalul [-1,+1] si, de obicei, se noteaza cu r. Se poate scrie $(-1 \le r \le 1)$ Daca $r = 0$ inseamna ca intre cele doua variabile nu exista nici o corelatie. Cu cat r este mai indepartat de 0 (catre -1 sau +1), cu atat corelatia este mai puternica.

In cazul unui set de date bivariabile de forma (x,y), coeficientul de corelatie se calculeaza folosind formula:

$$r = \frac{n \cdot \sum x \cdot y - \sum x \cdot \sum y}{\sqrt{n \sum x^2 - \left(\sum x\right)^2} \cdot \sqrt{n \sum y^2 - \left(\sum y\right)^2}}$$

In cazul in care cresterea valorii unei variabile determina scaderea valorii celeilalte variabile (si viceversa), corelatia es negativa. In aceasta situatie, coeficientul de corelatie r va lua o valoare in intervalul [-1,0], iar in cazul in care va avea chiar valoarea -1($r = -1$) se va putea spune ca exista o corelatie negativa "perfecta".

Cateva exemple de date bivariabile intre care ar putea exista o corelatie negativa:

a) numarul de saptamani de experienta si numarul de rebuturi (cu cat o persoana are mai multa experienta in ceea c priveste realizarea unei anumite munci, cu atat va da mai putine rebuturi);

b) varsta la care a survenit decesul si varsta de pensionare (documentele statisticienilor din societatile de asigurari arata ca, in general, cu cat o persoana se pensioneaza la o varsta mai inaintata, cu atat si decesul poate surveni mai curand)

c) cantitatea de produse si costul minim unitar al produselor (daca cantitatea creste, costul mediu unitar scade).

In cazul in care cresterea valorii unei variabile determina cresterea valorii celeilalte variabile (si viceversa), corelatia es pozitiva. In aceasta situatie, coeficientul de corelatie r va lua o valoare in intervalul [0,1], iar in cazul in care va avea chiar valoarea 1($r = 1$) se va putea spune ca exista o corelatie pozitiva "perfecta".

Cateva exemple de date bivariabile intre care ar putea exista o corelatie pozitiva:

a) varsta angajatilor si salariul (inaintarea in varsta inseamna atat sporirea calificarii, cat si a experientei, ambele fiind reflectate printr-un salariu mare);

b) numarul de contractari telefonice ralizate de un comis-voiajor si numarul de vanzari realizate (numarul de vanzari care pot fi efectuate creste direct proportional cu numarul de contractari telefonice realizate de agentul respectiv);

c) varsta unei persoane asigurate si valoarea ratei de asigurare (cu cat o persoana este mai in varsta, cu atat este m mare si sansa de a se imbolnavii sau de a deceda)

d) cheltuielile cu intretinerea utilajelor si vechimea acestora e) numarul de vehicule inmatriculate si accidente mortale.

3.3.2. Aplicatii

Pe baza metodei de determinare a regresiei liniare si a coeficientului de corelatie, de mai sus, care se gaseste in lucrarea [12] si [13], mai jos se incearca scrierea unor secvente de program care sa duca la apelarea concomitenta a datelor pentru rezolvarea celor 3 aplicatii similare. Secventele de program sunt incadrate si dublate de cateva explicatii.

Aplicatie 1

Datele alaturate se refera la transportul rutier de marfuri din Marea Britanie. Considerand ca "numar de vehicule comerciale" reprezinta variabila dependenta:

- Sa se determine ecuatia dreptei de regresie;
- Sa se determine coeficientul de corelatie;
- Sa se estimeze: numarul de vehicule comerciale care ar fi necesare in cazul in care s-ar transporta 80mii de milioane tone-kilometri; Sa se estimeze cantitatea de marfuri transportate atunci cand numarul de vehicule comerciale este 1800 de mii.

Obs. nu exista relatie cauzala

col 0 - marfuri transportate [mii de milioane tone-kilometrii]
col 1 - numar de vehicule comerciale inregistrate [mii]

$date_1 :=$

	0	1
0	79	1640
1	83	1640
2	85	1630
3	86	1632
4	88	1660
5	90	1736
6	90	1778
7	95	1791
8	96	1773
9	98	1712

Aplicatie 2

Datele din tabelul alaturat se refera la dependenta existenta intre costurile de intretinere saptamanale (in unitati monetare) si vechimea (in luni) a zece utilaje de acelasi tip ale unei intreprinderi productive.

- Sa se determine ecuatia dreptei de regresie a costurilor de intretinere in functie de vechimea utilajelor prin metoda celor mai mici patrate;
- Sa se determine coeficientul de corelatie;
- Sa se estimeze cu ajutorul dreptei de regresie, costurile de intretinere ale unui utilaj care are vechimea de 40 de luni; Sa se estimeze cand se va ajunge ca costurile cu intretinerea saptamanala a unui utilaj, sa fie 420 u.m.

col 0 - vechime utilaj [luni]
col 1 - costul saptamanal de intretinere [u.m./sapt]

$date_2 :=$

	0	1
0	5	190
1	10	240
2	15	250
3	20	300
4	30	310
5	30	335
6	30	300
7	50	300
8	50	350
9	60	395

Aplicatie 3

Directorii unei intreprinderi constituita din zece fabrici aflate in functiune, de aceleasi dimensiuni, care produc componente mici, au observat urmatorul model legat de cheltuielile pentru controlul de calitate si numarul de rebuturi care au fost livrate clientilor.

- Ei isi pun problema cat de puternica este relatia dintre cheltuielile pentru controlul de calitate si numarul de rebuturi livrate clientilor si in ce masura se poate anticipa numarul de rebuturi livrate, cunoscandu-se cheltuielile pentru controlul de calitate. Pentru aceasta sa se determine ecuatia dreptei de regresie si coeficientul de corelatie;
- Sa se estimeze cate defecte la 1000 de unitati livrate pentru o anumita suma cheltuita si sa se estimeze ce suma de bani trebuie cheltuita pentru un anumit numar de defecte la 1000 de unitati livrate.

col 0 - cheltuieli de control de calitate la 1000 de unitati, pence in um
col 1 - unitati defecte la 1000 de unitati livrate

$date_3 :=$

	0	1
0	25	50
1	30	35
2	15	60
3	75	15
4	40	46
5	65	20
6	45	28
7	24	45
8	35	42
9	70	22

- calculul parametrilor regresiei liniare

$$\text{dreapta}(\text{date}) :=$$

$$x \leftarrow \text{date}^{\langle 0 \rangle}$$

$$y \leftarrow \text{date}^{\langle 1 \rangle}$$

$$n \leftarrow \text{rows}(\text{date}) - 1$$

$$b \leftarrow \frac{(n+1)\cdot\sum_{i=0}^{n} x_i \cdot y_i - \sum_{i=0}^{n} x_i \cdot \sum_{i=0}^{n} y_i}{(n+1)\sum_{i=0}^{n} (x_i)^2 - \left(\sum_{i=0}^{n} x_i\right)^2}$$

$$a \leftarrow \frac{\sum_{i=0}^{n} y_i}{n+1} - b \cdot \frac{\sum_{i=0}^{n} x_i}{n+1}$$

$$X_0 \leftarrow \min(x) - \frac{\max(x) - \min(x)}{5}$$

$$X_1 \leftarrow \max(x) + \frac{\max(x) - \min(x)}{5}$$

$$\text{for } pct \in 0..1$$

$$Y_{pct} \leftarrow a + b \cdot X_{pct}$$

$$Xmed \leftarrow \frac{\sum_{i=0}^{n} x_i}{(n+1)}$$

$$Ymed \leftarrow \frac{\sum_{i=0}^{n} y_i}{(n+1)}$$

$$\begin{pmatrix} X \\ Y \\ a \\ b \end{pmatrix}$$

< retinerea datelor de corelatie in variabile locale locale

< indicele ultimului termen

< calculul parametrilor dreptei

< valorile lui x_minim si x_maxim alese pentru afisarea dreptei

< valorile lui y calculate pentru afisarea dreptei

< determinarea punctului mediu al datelor care se afla totdeauna pe dreapta de regresie

- calculul coeficientului de corelatie

$$r(\text{date}) :=$$

$$x \leftarrow \text{date}^{\langle 0 \rangle}$$

$$y \leftarrow \text{date}^{\langle 1 \rangle}$$

$$n \leftarrow \text{rows}(\text{date}) - 1$$

$$r \leftarrow \frac{(n+1)\cdot\sum_{i=0}^{n} x_i \cdot y_i - \sum_{i=0}^{n} x_i \cdot \sum_{i=0}^{n} y_i}{\sqrt{(n+1)\sum_{i=0}^{n} (x_i)^2 - \left(\sum_{i=0}^{n} x_i\right)^2} \cdot \sqrt{(n+1)\sum_{i=0}^{n} (y_i)^2 - \left(\sum_{i=0}^{n} y_i\right)^2}}$$

$$r$$

< retinerea datelor de corelatie in var. locale

< indicele ultimului termen

< formula de determinare a coeficientului de corelatie -1=<r<=1

Aplicatie 1

- indicele aplicatiei $i := 1$
- dreapta de regresie

$a := \text{dreapta}(date_i)_2$

$a = 954.027$

$b := \text{dreapta}(date_i)_3$

$b = 8.373$

$y(x) := b \cdot x + a$

pentru $x := 80$ mii mil tone-km

$y(x) = 1.624 \times 10^3$ vehicule

pentru $y := 1800$ vehicule $x := \frac{y - a}{b}$

$x = 101.039$ mii mil tone-km

- coeficientul de corelatie $r(date_i) = 0.766$

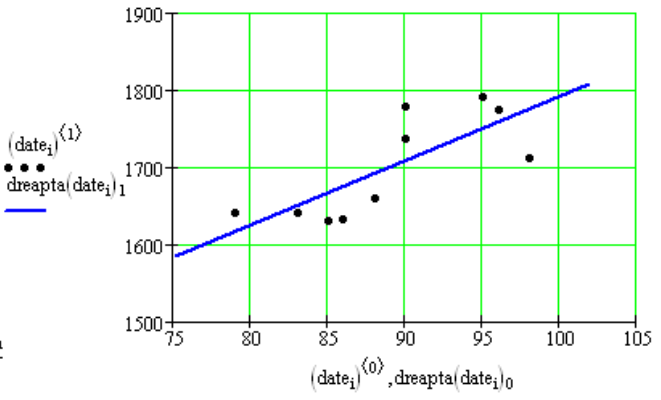

Figura 3.4

x - marfuri transportate [mii de milioane tone-kilometrii]
y - numar de vehicule comerciale inregistrate [vehicule]

Aplicatie 2

- indicele aplicatiei $i := 2$
- dreapta de regresie

$a := \text{dreapta}(date_i)_2$

$a = 212.902$

$b := \text{dreapta}(date_i)_3$

$b = 2.803$

$y(x) := b \cdot x + a$

pentru $x := 40 \cdot \text{luni}$

$\text{round}(y(x)) = 325\,\text{um}$

pentru $y := 420 \cdot \text{um}$ $x := \frac{y - a}{b}$

$x = 73.877\,\text{luni}$ $\text{round}(x) = 74\,\text{luni}$

- coeficientul de corelatie $r(date_i) = 0.88$

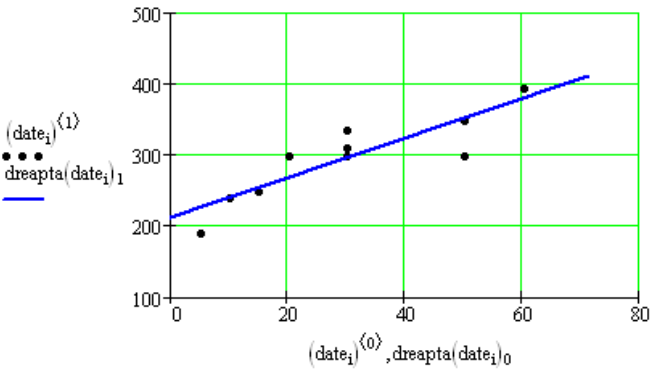

Figura 3.5

x - vechime utilaj [luni]
y - costul saptamanal de intretinere [u.m./sapt.]

Aplicatie 3

- indicele aplicatiei $i := 3$
- dreapta de regresie

$a := \text{dreapta}(date_i)_2$

$a = 63.965$

$b := \text{dreapta}(date_i)_3$

$b = -0.652$

$y(x) := b \cdot x + a$

pentru $x := 50 \cdot \text{um}$

$\text{round}(y(x)) = 31\,\text{defecte}$

pentru $y := 25 \cdot \text{defecte}$ $x := \frac{y - a}{b}$

$x = 59.719\,\text{um}$ $\text{round}(x) = 60\,\text{um}$

- coeficientul de corelatie $r(date_i) = 0.88$

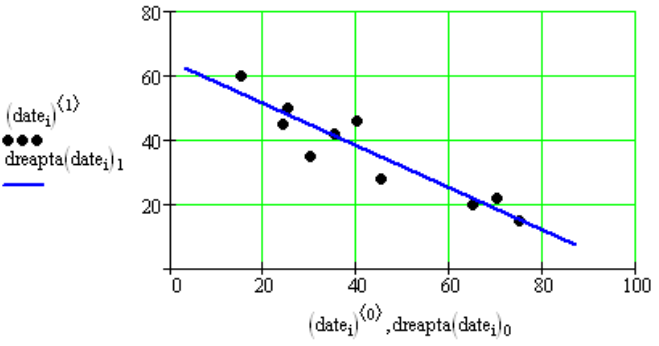

Figura 3.6

x - cheltuieli de control de calitate la 1000 de unitati, pence in um
y - unitati defecte la 1000 de unitati livrate