

2. MODELE REGRESIVE

Modelele regresive sunt modele statistice care stabilesc relația între valorile a două sau mai multe variabile aleatoare, în vederea prognozării valorilor uneia în raport cu valorile celeilalte sau celorlalte. Această problemă se pune atunci când între variabilele aleatoare considerate există o legătură consistentă, bazată pe natura intimă a fenomenelor care stau la baza lor.

Intuitiv, având o mulțime de valori experimentale din $\mathbf{R}^2$, considerate a fi valorile a două variabile aleatoare, încercăm să determinăm curba care este „cea mai apropiată” de punctele respective. Metoda cea mai des utilizată în acest scop este *metoda celor mai mici pătrate*. Dacă relația care stabilește legătura între variabila dependentă și variabilele independente este una liniară, se vorbește despre *regresia liniară*, altfel fiind vorba de *regresia neliniară* (polinomială, exponențială, logaritmică etc.).

Regresia liniară

Regresia liniară este tehnica statistică ce permite modelarea relației liniare între o variabilă explicativă (notată cu X) și variabila ce urmează a fi explicată (notată cu Y). Vom începe prin a pune în evidență mecanismele cu care se pot evidenția legăturile între secvențe de date, provenind de la două sau mai multe variabile aleatoare.

În cazul a două serii statistice $\{x_i\}_{1 \leq i \leq n}$ și $\{y_i\}_{1 \leq i \leq n}$ definite pe același lot de obiecte putem considera seria cuplurilor de observații $\{(x_i, y_i)\}_{1 \leq i \leq n}$ definite de cele două variabile statistice pe același obiect i . Acest cuplu de observații este reprezentat grafic folosind *norul* de puncte definit de reprezentarea bidimensională a punctelor (x_i, y_i) – așa numita *diagramă de împrăștiere*.

Exemplu:

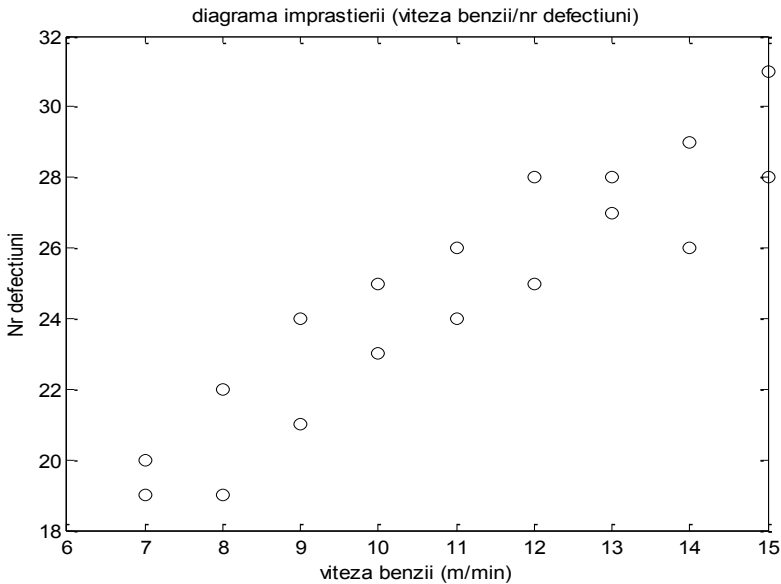
2.1 Într-o fabrică s-a constatat că viteza (m/min) benzii de lucru afectează numărul de defecțiuni descoperite în timpul verificării. Următorul tabel prezintă diferite viteze ale benzii și numărul defectelor găsite.

viteza (m/min)	nr de defecțiuni
6	21
6	18
7	19
7	20
8	22
8	19
9	21
9	24
10	23

10	25
11	24
11	26
12	25
12	28
13	27
13	28
14	26
14	29
15	28
15	31

Prezentăm diagrama „norului” de împrăștiere a cuplurilor (viteza, număr defectiuni) pentru lotul de 20 de observații:

```
>> X=[6 6 7 7 8 8 9 9 10 10 11 11 12 12 13 13 14 14 15 15];
>> Y=[21 18 19 20 22 19 21 24 23 25 24 26 25 28 27 28 26 29 28 31];
>> plot(X,Y,'ok')
```



Pentru o analiză statistică diagrama împrăștierii este un instrument util, dând informații importante privind legătura între cele două serii statistice. Dacă norul are forma unei figuri geometrice alungite, simetrică față de o axă, există o legătură liniară între variabile. Forma de cerc sau pătrat a norului implică independența variabilelor. Este importantă alegerea unităților de măsură a celor două variabile pentru a nu fi modificată forma norului.

Prezentarea numerică a datelor multiple se face prin tabele ținând seama că este indicat ca datele de aceeași natură să fie puse pe coloane. Tabelele pot conține fie date neprelucrate, adică datele reale, neprocesate, fie date prelucrate, rezultate ale procesării statistice. În articole, cărți sau cursuri, este preferabilă prezentarea rezultatelor procesării, datele neprelucrate nefiind totdeauna interesante.

O mare parte a studiilor statistice uzuale se ocupă cu analiza relației între două variabile statistice ce corespund aceluiași grup de subiecți/obiecte. Cel mai cunoscut exemplu se referă la relația ce există între înălțimea și greutatea unui individ ce corespunde unor anumite standarde geografice, rasiale etc. Pentru a o identifica, se studiază relația dintre cele două caracteristici măsurate pe indivizii dintr-un anumit lot. Suntem interesați în a descrie relația care ar putea exista între cele două variabile, analizând legătura între cele două serii de observații. Concret, se analizează dacă tendința ascendentă a uneia implică o tendință ascendentă, descendentă sau nici o tendință a celeilalte. Scopul final, în ipoteza existenței unei legături reale între cele două variabile, este *prognoza* valorilor uneia în raport cu valorile celeilalte pe baza ecuației de regresie.

Posibilele asociații între valorile a două variabile statistice continue, prelevate de la același grup de subiecți, sunt date de *coeficientul de corelație*. Acest coeficient poate fi calculat pentru orice set de date. Pentru ca el să aibă relevanță statistică, trebuie să fie îndeplinite următoarele două condiții, cerință satisfăcută în cadrul acestei cărți:

1. cele două variabile sunt definite de același lot de obiecte, cuplurile de date corespunzând aceluiași obiect;
2. cel puțin una din variabile să aibă o repartiție aproximativ normală.

Prezentăm formulele matematice ce stau la baza calculării coeficientului de corelație și a intervalului de încredere corespunzător.

Consideră, două serii statistice $\{x_i\}_{1 \leq i \leq n}$ și $\{y_i\}_{1 \leq i \leq n}$ corespunzătoare variabilelor statistice X și Y , generate de un grup de obiecte. Prin *coeficientul de corelație* r al celor două variabile, numit și *Pearson's r* , vom înțelege numărul real r , cuprins între -1 și 1, definit de formula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}},$$

$$\text{unde } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Formula poate fi scrisă sub forma:

Data Mining: Modele predictive și de clasificare Implementare în MATLAB și Java

Smaranda Belciug, Marina Gorunescu

- 1. NOȚIUNI INTRODUCTIVE**
- 2. METODE REGRESIVE**
- 3. SERII TEMPORALE**
- 4. CLASIFICARE BAYESIANĂ**
- 5. K-NEAREST NEIGHBOR**
- 6. ARBORI DE CLASIFICARE ȘI DECIZIE**
- 7. REȚELE NEURONALE**