Probabilitati si statistica

- documentatie proiect –

Setul de date “swiss” este alcatuit din 6 coloane si 47 de randuri, reprezentand informatii precum fertilitatea, agricultura, examinarea, educatia, catolicismul si mortalitatea infantila, corespunzatoare pentru 47 de orase elvetiene.

Cerinta 1:

Extragem vectorii pentru fiecare caracteristica in variabile noi, pentru a le putea modela mai usor ulterior.

Folosind functiile *mean* si *var* calculam media si dispersia pentru fiecare caracteristica.

Pentru a calcula quartilele avem nevoie sa sortam fiecare vector in parte. Utilizam functiile *sort* si *quantile*. Pentru a separa quartila in proportiile (25% - 75%, 50% - 50%, 75% - 25%) vom folosii valorile ¼, ½, si ¾.

Cu ajutorul functiei *boxplot* vom realiza diagrama boxplot pentru fiecare caracteristica a setului de date.

Diagrama Boxplot este un tip de grafic folosit pentru a afișa modele de date cantitative. Ea oferă informații privind tendința centrală și forma distribuției studiate.

Boxplot-urile celor 6 masuratori:

* + - Fertility - mediana = 70.4
    - Agriculture - mediana = 54.1
    - Examination - mediana = 16
    - Education - mediana = 8
    - Catholic - mediana = 15.1
    - Infant.Mortality - mediana = 20

Interpretare:

* + - Fertility,Infant Mortality - distributia este aproape simetrica, cutia aflandu-se la jumatatea distantei dintre cele 2 “mustati”, iar diagrama este de asemenea simetrica, mediana impartind cutia in 2 jumatati aproximativ egale.
    - Agriculture - distributia este asimetrica la stanga, cutia fiind apropiata de mustata superioara, iar mediana fiind apropiata de latura superioara a cutiei.
    - Examination, Education, Catholic - distributia este asimetrica la dreapta, cutia fiind apropiata de mustata inferioara, iar mediana fiind situata in partea inferioara a cutiei.

Cerinta 2:

Pentru regresia simpla am folosit criteriul de informare Akaike, calculand astfel pentru fiecare aranjament AIC- ul si BIC-ul, ajungand la concluzia ca cele mai bune caracteristici pentru realizarea regresiei sunt Infant.Mortality si Fertility (AIC ≈ 229 si BIC ≈ 235).

Pentru acest set de date nu s-a putut gasi o combinatie de date astfel incat testele AIC si BIC sa dea valori cat mai mici si corelatia sa fie cat mai mare.

Pentru regresia multipla am cautat trei variabile astfel incat una dintre ele sa fie dependenta de celelalte doua, celelalte doua fiind indepentente intre ele. Am ales astfel Fertilitatea in functie de Educatie si Agricultura.

Pentru generarea noii variabile am ales sa adaugam variabila Urbanizare (Urbanisation), folosind functia *runif* si am facut o regresie simpla cu aceasta (Educatia in functie de Urbanizare).

Cerinta 3

Am ales o repartitie Cauchy pentru a calcula si pentru a afisa graficele respective, folosind numarul de masuratori ale volumului si a unei probabilitati alese, 1/2.

Pentru functia de masa si functia de distributie folosim functiile *dcauchy* si *pcauchy*.