Proiect Probabilitate şi Statistică

Vîlculescu Mihai-Bogdan - Grupa 232
Martac Dana-Maria - Grupa 232
Drăguţescu Mihai-Valentin - Grupa 232

Set de date = randu

0. RANDU

Set-ul de date utilizat de echipa noastră a fost randu. Acest set de date este alcătuit din 400 de triplete de numere aleatoare succesive. Acestea sunt memorate ca un data frame cu 400 de observaţii pe 3 variabile: x, y şi z.

1. Operaţii de statistică descriptivă

Pentru primul exerciţiu, se cerea să efectuăm diverse operaţii de statistică descriptivă, precum: medie, varianţă, quantile, boxplot.

- **Media:** este valoarea medie a unui set de date. În R, am calculat media pentru fiecare din cele 3 variabile (x, y, z). Acest lucru se realizează utilizând funcţia **"mean".** Aceasta primeşte ca parametru un set de date, şi calculează media acelor date.

- **Mediana:** este valoarea din mijloc a unui set de date. Am procedat la fel ca la medie, şi anume am calculat mediana pentru x, y şi z. Funcţia utilizată se numeşte **"median"** şi se comportă la fel ca funcţia pentru calcul a mediei.

- **Deviaţia standard:** reprezintă deviaţia medie a valorilor dintr-un set de date faţă de valoarea medie. Aceasta se calculează în limbajul R, folosind o funcţie numită **"sd"**.

- **Varianţa / Dispersia:** reprezintă pătratul deviaţiei standard. Funcţia folosită se numeşte **"var"**.

- **Quartile:** reprezintă o modalitate de a împărţi un sample de date în mai multe subgrupuri de dimensiuni egale şi adiacente. Funcţia folosită se cheamă intuitiv **"quantile"**. Aceasta nu întoarce nimic, ci afişează în consolă distribuţia setului de date în aceste quartile.

- **Sapply:** o funcţie din R care permite afişarea în consolă a valorilor unui set de date în momentul aplicării unei funcţii asupra sa. Spre exemplu, se poate apela astfel: **"sapply(my_data, mean)"**, iar acest apel calculeaza media pentru toate variabilele set-ului my_data. Astfel, putem să vedem valorile mediei fără să apelăm manual funcţia pentru fiecare variabilă.

- **Summary:** o funcţie care afişează detaliile statistice cele mai importante ale unei anumite variabile dintr-un set de date furnizat ca parametru. (quartilele, media, mediana, min, max).

- **Boxplot:** este un tip de grafic folosit pentru a afişa modele de date cantitative. Ea oferă informaţii privind tendinţa centrală şi forma distribuţiei studiate. Pe această diagramă, sunt observabile următoarele elemente statistice importante: minimul, prima quartilă, mediana, a treia quartilă şi maximul.

- **Interpretare:** după apelul tuturor funcţiilor, putem încerca interpretarea valorilor obţinute.
    - **Mean** = 0.526 pentru x, 0.486 pentru y şi 0.481 pentru z. Deci putem afirma că mediile se apropie de 0.5 pentru toate cele 3 variabile.

    - **Median** = 0.540 (x), 0.483 (y), 0.463 (z). Astfel, medianele sunt foarte apropiate de valorile medii, la nu mai mult de 0.02 distanţă.

    - **Standard Deviation** = 0.285 (x), 0.293 (y), 0.279 (z). Deviaţia medie a valorilor faţă de medie este destul de mică, dar ţinând cont că valorile din setul de date randu sunt cuprinse între 0 şi 1, această deviaţie demonstrează că valorile sunt destul de răspândite.

    - **Variance** = 0.081 (x), 0.086 (y), 0.077 (z). Varianţa este mică dacă privim obiectiv, dar relativ la setul nostru de date, această varianţă confirmă ce am spus la deviaţia standard (cum era şi firesc), şi anume că valorile nu sunt tocmai apropiate una de cealaltă.

    - **Quantiles** = quartilele pentru cele 3 variabile, împart valorile în subgrupuri. Astfel, pentru x, 25% din valori sunt mai mici decât 0.3003. Pentru y, 75% din valori sunt mai mici decât 0.7399195. Quartilele ne ajută să observăm anumite distribuţii ale valorilor.

    - **Boxplot** = diagrama indică într-un mod grafic ceea ce putem citi din valorile regăsite în quartile. Mai exact, putem observa cum sunt distribuite valorile faţă de mediană şi faţă de minim şi maxim. Aici,

valorile lui x sunt mai mari decât y şi z, Inclusiv mediana lui x are o valoare mai mare. În schimb, y şi z sunt mai apropiate ca minim, maxim şi mediană, dar valorile din z sunt mai puţin răspândite, lucru observabil şi în rezultatul deviaţiei standard şi al varianţei.

2. Regresia liniară

Pentru al doilea exerciţiu, trebuie să creăm 2 modele de regresie liniară pentru setul nostru de date (unul de regresie liniară şi unul de regresie multiplă). Paşii efectuaţi pentru îndeplinirea cerinţei sunt:

A. Verificare

a. Pentru a putea crea modelele de regresie, trebuie, în primul rând, să ne asigurăm că setul nostru de date este unul care se pretează asupra regresiei liniare.

b. Apelăm funcţia **cor** având ca parametri variabilele x şi y, pe care intenţionăm să le folosim ca predicator şi răspuns. Această funcţie testează gradul de corelaţie dintre aceste 2 variabile. Mai exact, în funcţie de valoarea întoarsă de cor, ne putem da seama daca cât de dependente sunt cele două variabile una faţă de cealaltă.

c. Funcţia în cazul nostru întoarce **-0.048**. Această valoare este mai mică decât 0, fapt care îmi sugerează că variaţia variabilei răspuns nu poate fi susţinută de variabila predictor. Astfel, ar trebui să alegem variabile mai dependente una de cealaltă. Dar, dacă apelăm funcţia pentru toate combinaţiile celor 3 variabile, obţinem mereu o valoare scăzută şi apropiată de 0, ceea ce implică faptul că nu vom putea crea un model de regresie liniară care să funcţioneze corespunzător cu aceste variabile.

d. Am construit modelul liniar pentru întregul set de date apelând funcţia **lm**. Apoi, am aplicat **summary** pentru modelul creat. Apoi am extras din acel summary, coeficienţii pentru a putea calcula beta estimat şi eroarea standard. Cu aceste valori, am calculat valorile p şi t. Cunoaştem că, pentru o valoare mai mare a lui t, setul de date este mai semnificativ din punct de vedere statistic, dar noi am obţinut valoarea **-0.968**. În continuare, valoarea lui p este **0.3335**, iar noi ar fi trebuit să obţinem un p mai apropiat de 0.05. Astfel, rezultatele de la acest punct confirmă concluzia de la punctul c.

B. Construirea modelului

a. Pentru ambele tipuri de regresie efectuăm în mare parte aceleaşi operaţii. Prima dată, alegem seturile de date de training şi de test. Acestea vor fi alese random din datele iniţiale folosind funcţia **sample**. Am ales să împărţim datele iniţiale în raport 80:20 (training:test).

b. Creăm modelul liniar simplu, respectiv multiplu. Pentru cel simplu, specificăm întâi variabila răspuns apoi predicatorul (**response ~ predicator**). Pentru cel multiplu, adăugăm separaţi prin **+**, mai multe variabile de tip predicator (**response ~ predicator1 + predicator2**). Apoi, facem predicţiile apelând funcţia **predict**, care primeşte ca parametri modelul realizat în raport cu setul de date training şi al doilea parametru este setul de date test.

c. Următoarele instrucţiuni reprezintă verificări ale predicţiilor făcute, dar după cum am precizat anterior, acestea vor fi inexacte.

3. Repartiţia Beta

Pentru al treilea exerciţiu, am selectat repartiţia beta, pentru care trebuia să afişăm în 2 reprezentări alăturate funcţia de densitate şi cea de repartiţie. Astfel, am ales un şir de numere de la -5 la 5, cu pasul de 0.001 (-5, -4.999, …). Pentru acel şir am apelat funcţia **plot** cu şirul ca prim parametru şi funcţia corespunzătoare ca al doilea parametru. (**dbeta, pbeta)**. Acesta a fost procesul de construire a reprezentărilor.

Interpretări

- Un lucru care se observă imediat la cele 2 grafice este acela că funcţia de densitate nu este continuă, în timp ce cea de repartiţie este. Funcţia de densitate ia valori foartemici până într-un anumit punct, unde creşte brusc şi începe din nou să scadă.
- Funcţia de repartiţie creşte mai puţin brusc şi după nu mai scade, rămânând constantă.

Aplicaţii: Rule of Succession, Deducţia Bayesiană, Order Statistics, Logica subiectivă.