Proiect Probabilitati si Statistica

Echipa: Condurachi Corina
Racovita Andra Georgiana
Grupa: 232
dataset: Seatbelts

1.

Am calculate pentru fiecare coloana din tabel media(Mean), variatia(Var), Quantile(quantile), boxplot.

Boxplot

Functia boxplot prezinta intr-o maniera compacta modul in care este reprezentata o variabila.

Boxplot-ul pentru DiversKilled:

- **Range-ul** este 138, datasetul avand valoarea minima egala cu 60, iar maxima cu 198.
- **Interquartile range** este 23.3 (Q3-Q1=138-104.8=23.2), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.= 138.0+104.8 = 242.8, 242.8/2=121.4!= Mediana, 121.4>118.5), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed left.

Boxplot-ul pentru drivers:

- **Range-ul** este 1597, datasetul avand valoarea minima egala cu 1057, iar maxima cu 2654.
- **Interquartile range** este 389 (Q3-Q1=1851-1462=389), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.= 1462+1851 = 3131, 3131/2=1656.5!= Mediana, 1656.5>1631), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed left.

Boxplot-ul pentru rear:

- **Range-ul** este 422, datasetul avand valoarea minima egala cu 224, iar maxima cu 646.
- **Interquartile range** este 120.4(Q3-Q1=465.2-344.8=120.4), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.= 465.2+344.8 = 810, 810/2=405!= Mediana, 405>401.5), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed left.

Boxplot-ul pentru front:

- **Range-ul** este 873, datasetul avand valoarea minima egala cu 426, iar maxima cu 1299.
- **Interquartile range** este 235.3(Q3-Q1=950.8-715.5=235.3), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.= 950.8+715.5 = 1666.3, 1666.3/2=833.15!= Mediana, 833.15>828.5), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed left.

Boxplot-ul pentru kms:

- **Range-ul** este 13941, datasetul avand valoarea minima egala cu 7685, iar maxima cu 21626.
- **Interquartile range** este 235.3(Q3-Q1=17203-12685=4518), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.= 12685+17203 = 29888, 29888/2=14944!= Mediana, 14944<14987), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed right.

Boxplot-ul pentru PetrolPrice:

- **Range-ul** este 0.08118, datasetul avand valoarea minima egala cu 0.08118, iar maxima cu 0.13303.
- **Interquartile range** este 0.02148(Q3-Q1=0.11406-0.09258=4518), dimensiunea box-ului.

- **Mediana** nu este la mijlocul box-ului (Q1+Q3.= 0.09258+0.11406 = 0.20664, 0.20664/2=0.10332!= Mediana, 0.10332<0.10448), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed right.

Boxplot-ul pentru VanKilled:

- **Range-ul** este 15000, datasetul avand valoarea minima egala cu 2000, iar maxima cu 17000.
- **Interquartile range** este 6000(Q3-Q1=12000-6000=6000), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.=6000+12000 = 18000, 18000/2=9000!= Mediana, 9000>8000), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed left.

Boxplot-ul pentru law in functie de DriversKilled:

I. pentru law=1:

- **Range-ul** este 94, datasetul avand valoarea minima egala cu 60, iar maxima cu 154.
- **Interquartile range** este 24(Q3-Q1=119-85=24), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.=119+85 = 204, 204/2=102!= Mediana, 102>92), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed left.

II. pentru law=0:

- **Range-ul** este 119, datasetul avand valoarea minima egala cu 79, iar maxima cu 198.
- **Interquartile range** este 32(Q3-Q1=140-108=32), dimensiunea box-ului.
- **Mediana** nu este la mijlocul box-ului (Q1+Q3.=140+108 = 248, 248/2=124!= Mediana, 124>121), fiind pozitionata mai aproape de marginea de sus a box-ului, deci dataset-ul este skewed left.

In graficul, **Drivers killed vs law**, facut cu boxplot se poate observa ca dupa introducerea legii privind obligativitatea soferilor de a purta centura de siguranta numarul deceselor a scazut seminficativ. Acest lucru este sugerat de faptul ca al 2-lea boxplot (cand legea este in vigoare) este pozitionat mai jos fata de primul boxplot. Medianele celor 2 reprezentatii difera, insa ambele sunt pozitionate mai aproape de magina de jos a box-ului, deci dataset-urile nu sunt simetrice, ambele fiind skewed right(coada distributiei fiind mai lunga in partea dreapta).

2.

Regresia Liniara Simpla

Am facut mai intai regresia liniara simpla, impartind mai intai setul de date in training set(2/3) si test set(1/3). Am construit un regresor folosind datele din training_set. Am folosit acest regresor pentru a prezice rezultatele din test_set, acestea fiind retinute in y_pred. Am ales sa afisam si un grafic cu rezultatele obstinate. Din figura se vede clar ca modelul de regresie ales este unul potrivit.

Regresia Liniara Multipla

La regresia multipla am inclus initial toate variabilele (dupa ce am construit regresorul). Am afisat la fiecare pas valorile obtinute, eliminand de fiecare data valoarea cu p-value cea mai mare, intrucat aceea nu avea un impact semnificativ asupra regresiei. Acesta metoda pe care am folosit-o se numeste ‘backward elimination’.

Observatii - Initial doar campul drivers era marcat cu ***, iar kms cu ‘.’. Dupa cea de-a treia eliminare, Intercept era marcat cu **, drivers cu ***, iar kms cu *. Metoda se executa atata timp cat p-value>0.05 (alegem aceasta ca fiind valoarea prag- SL). De aceea, nu am mai eliminat alte date intrucat celelalte au p-value 0.003, <2e-16, respectiv 0.01, deci sunt relevante pentru modelul construit.

Testul -R squared

Am calculat pentru ambele testul R patrat. Pentru regresia liniara simpla am obtinut valoarea 0.8000215, iar pentru cea multipla - 0.7968785. Deci, regresia care este mai potrivita pentru setul nostru de date este cea simpla, intrucat are valoarea R^2 mai mare.

Testul - Adjusted R squared

Am calculat pentru ambele testul Adjusted R squared. Pentru regresia liniara simpla am obtinut valoarea 0.7984713, iar pentru cea multipla - 0.7947291. Deci, regresia care este mai potrivita pentru setul nostru de date este cea simpla, intrucat are valoarea adjusted R^2 mai mare.

Testul AIC

Am calculat pentru ambele testul AIC. Pentru regresia liniara simpla am obtinut valoarea 1025.379, iar pentru cea multipla -1487.671. Deci, regresia care este mai potrivita pentru setul nostru de date este cea simpla, intrucat are valoarile AIC sunt mai mici.

Testul BIC

Am calculat pentru ambele testul BIC. Pentru regresia liniara simpla am obtinut valoarea 1034.005, iar pentru cea multipla -1500.701. Deci, regresia care este mai potrivita pentru setul nostru de date este cea simpla, intrucat are valoarile BIC sunt mai mici.

Testul -F Statistic

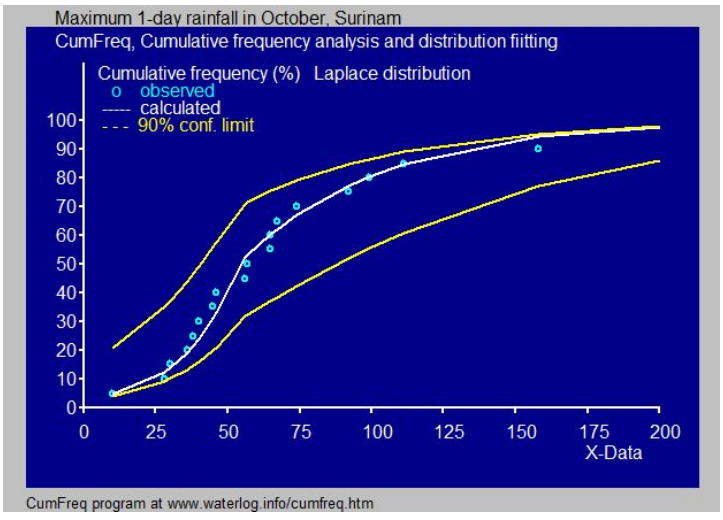
Am calculat pentru ambele testul F statistic. Pentru regresia liniara simpla am obtinut valoarea 516.0694, iar pentru cea multipla - 370.7388. Deci, regresia care este mai potrivita pentru setul nostru de date este cea simpla, intrucat are valoarea F statistic mai mare.

Standard Error

Am calculat pentru ambele Standard Error. Pentru regresia liniara simpla am obtinut valoarea 0.003540667, iar pentru cea multipla - 0.00320763. Deci, regresia care este mai potrivita pentru setul nostru de date este, in acest caz, cea multipla, intrucat are valoarea standard error mai mica.

3. Aplicand boxplot observam ca datele nu sunt distribuite simetric.

Distributia Laplace este aplicata in situatiile in care valorile inferioare provin din conditii externe diferite fata de cele superioae, prin urmare acestea urmeaza un model diferit. Studiind diverse fenomene, se constata ca desi acestea apartin unor stiinte diferite repartitia in frecventa a acestora este asemanatoare respectiv ca histogramele au aceeasi forma. Spre exemplu 90% din fenomenele fizice se supun legii normale de repartitie (legea Gauss-Laplace).



Distributia Laplace este folosita in hidrologie pentru a reprezenta evenimente extrem (inundatii, ploi abudente). Imaginea albastra ilustreaza un exemplu de incadrare a distributiei Laplace pentru a clasa cantitatea maxima de precipitatii din fiecare zi, aratand ca 90% din confidence belt este bazat pe distributia binomiala.

Distributia Laplace a fost folosita in recunoasterea vocala pentru a modela coeficientii DFT si in compresia imaginii JPEG pentru a modela coeficientii AC generati de DCT.

Bonus

1. Functia frepcomgen(m,n)

Construieste initial doua matrici x (m linii, 2 coloane) si y (n linii, 2 coloane) care contin doar 0. Se construieste mai intai matricea x. Se genereaza cate un numar aleatoriu distinct pentru primul rand, iar pentru cel de-al doilea - un numar random intre 0 si val(initializata cu 1) avand 2 zecimale. Intrucat suma probabilitatilor trebuie sa fie 1, luam o variabila val si o scadem de fiecare data cand adaugam o probabilitate noua, astfel incat sa nu generam un numar mai mare si astfel suma probabilitatilor sa depaseasca 1. Pe ultima pozitie adaugam diferenta ramasa. Astfel am constuit vsriabila x. Facem acelasi lucru si pentru y.

Construim tabelul cu repartitia comuna a celor doua variabile aleatoare x si y. Mai intai construim o matrice cu n linii si m coloane care contine doar 0. Pe fieace linie generam un nr random intre 1 si m astfel incat spatiul acela va ramane necompletat. Pe fiecare pozitie generam un nr random intre 0 si minimul dintre cele doua probabilitati (x[2,i]; y[2,j]) pentru a fi siguri ca nu depasim valoarea minima. Scad mereu valoarea adaougata pentru a ma asigura ca nu voi depasi pe viitor valorile lui pi, respectiv qi. Afisam in final matricea rezultata.

2. Functia fcomplrepcom(matrice)

Are ca parametru matricea returnata de functia precedenta. Cauta simultan pe linii si coloane locurile unde valoarea este 0 si le numara. Daca pe o linie, respectiv pe o coloana, se afla doar o singura valoare de 0, putem calcula direct acea valoare ca fiind pi (qj pentru coloana)- suma celorlalte elemente (intrucat suma de pe o linia i=pi; suma de pe colona j=qj). Astfel parcurgem matricea pana cand nu vom mai avea nicio valoare egala cu 0. Daca avem mai multe linii decat coloane, avem un while separat in care verificam doar liniile. Analaog pentru cazul cand avem mai multe coloane decat linii.

3. Punctul c

1) **Cov(5X,-3Y):** In variabila auxx am calculat 5*x, iar in variabila auxy am calculat -3*y. Am calculat mediile pentru cele doua variabile aleatore ex, respectiv ey. Am calculat media lui xy in variabila exy, adunanad la fiecare pas xiyi*piqi. Am calculat covarianta dupa formula covxy=exy-ex*ey.
2) **P(0<X<3/Y>2):** P(0<X<3/Y>2)=P(0<X<3 ∩ Y>2)/P(Y>2)=P(0<X<3)*P(Y>2)/P(Y>2)=P(0<X<3). In sumx am adunat elementele care aveau probabilitatea intre (0,2). Suma aceasta este chiar raspunsul
3) **P[X>6,Y<7]:** P[X>6,Y<7]=P[X>6 ∩ Y<7]=P[X>6]*P[Y>7]. Am calculat separat P[X>6] si P[Y<7], iar raspunsul este produsul lor.

4. Punctul d

1) Am facut functia **fverind** care verifica daca cele 2 variabile aleatoare sunt independente. Pentru ca X si Y sa fie independente trebuie ca elementele aflate pe pozitia ij in matricea comuna (matri[i,j]) sa fie egale cu x[2,i]*y[2,j].
2) Am facut functia **fvernecor** care verifica daca cele 2 variabile aleatoare sunt necorelate. Pentru ca X si Y sa fie necorelate trebuie ca covalenta lui X si Y sa fie egala cu 0, daca e diferita de 0 sunt corelate. Am calculat mediile pentru cele doua variabile aleatore ex, respectiv ey. Am calculat media lui xy in variabila exy, adunanad la fiecare pas xiyi*piqi. Am calculat covarianta dupa formula covxy=exy-ex*ey.