Probabilitati si Statistica

Grupa 231
Talmacel Sergiu
Iamandii Ana-Maria

1 Cerinta I

Interpretam urmatoarele variabile din setul de date quakes

1. **Magnitude**
   - Analizand boxplotul, putem observa ca majoritatea cutremurelor au o intensitate cuprinsa intre 4.30 si 4.90 de grade pe scara Richter,iar variatia este una destul de mica, ce ne indica faptul ca intensitatea cutremurelor nu difera foarte mult de media de 6.40. Numarul de outliers este destul de redus, sansa ca un cutremur cu intensitatea de peste 6.40 sa se produca fiind mica.
2. **Stations**
   - Numarul de outliers in acest caz este mai mare, posibil cauzat de cutremurele cu intensitatea ridicata care au fost resimtite de mai multe statii, Numarul mediu de statii este 27, iar prezenta unei variatii mai mari poate fi determinata de numarul de outliers corespunzator.
3. **Depth**
   - Din boxplot putem sesiza ca nu intalnim outliers, incat gradul de imprastiere este unul ridicat, fapt ce ne arata ca nu exista o adancime specifica la care se pot produce cutremure de intensitate mica sau mare. Media este una de 311 km si putem observa ca probabilitatea de a se produce un cutremur atat foarte aproape, cat si foarte departe de suprafata este redusa.
4. **Latitudinea** si **Longitudinea**, atat la nivel intuitiv, cat si numeric, au o variatie redusa si sunt concentrate in jurul mediei, indicand centrele seismice din zona analizata.

2 Cerinta II

2.1 Regresia liniara simpla

Pentru **regresia liniara** simpla am ales variabila aleatoare **mag** ca predictor, iar **stations** ca variabila raspuns. Pentru aceasta alegere am luat in considerare urmatorii factori:

- **Scatter plot-ul** si functia de **smoothness**, identificand o relatie liniara si pozitiva intre magnitudinea unui cutremur si numarul de statii care l-au inregistrat.
- Chiar daca o stransa legatura intre variabile nu implica cauzalitatea, avem un coeficient de corelatie de 0.8511, ceea ce indica ce tip de relatie poate exista intre cele doua variabile(intuitiv, aceasta relatie pare a fi existenta deoarece cu cat un cutremur este mai mare, distanta pe care este resimtit creste, deci mai multe statii l-ar inregistra).

Rezultatele acestui model sunt urmatoarele:

- **p-value <2.2e-16** <0.05(pragul stabilit) , **t-value** este destul de mare, rezulta ca null hypothesis este respinsa, deci fiecare coeficient este semnificativ pentru modelul nostru si nu se anuleaza.
- alt indicator semnificativ este **R-squared = 0.7167**, cee ce ne spune ca ecuatia regresiei poate explica 71% din varianta variabilei dependente **stations**.

2.2 Regresia liniara multipla

Pentru regresia liniara multipla,am decis sa adaugam variabila aleatoare predictor **depth** din urmatoarele motive:

> Analizand **scatter plot-ul** si functia de **smoothness**, observam ca seismele mai puternice s-ar produce la adancimi mai mici, deci, *intuitiv*, cu ajutorul acestei noi variabile aleatoare, am prezice "mai bine" cate statii vor inregistra cutremurul.Urmeaza sa decidem daca acest lucru va fi confirmat.

Rezultatele regresiei multiple sunt cele ce urmeaza:

- Analog modelului anterior, **p-value <2.2e-16** <0.05, **t-value** este destul de mare, rezulta ca null hypothesis este respinsa,iar fiecare coeficient e semnificativ in regresie.
- Evident ca valoarea lui **R-squared** creste, prin adaugarea unei noi variabile(**0.733**), asa ca este mai relevant sa luam in considerare **Adjusted R-squared = 0.7323** pentru a decide daca modelul nostru este semnificativ sau nu din punct de vedere statistic

2.3 Care dintre ele este "mai buna"?

Cum ambele modele au o semnificatie statistica, decidem sa alegem cel de-al doilea model din urmatoarele considerente:

- **RSquared1 <RSquared2**, deci al doilea model ar explica mai bine variabila raspuns.In plus, modelele fiind imbricate(regresia multipla contine aceleasi variabile aleatoare ca regresia simpla plus una noua), putem face testul *ANOVA* pentru a determina daca adaugarea unei noi variabile este semnificativa(lucru confirmat de valoarea lui P(>F), 6.36e-12 <0.05 ***)
- predictiile celui de-al doilea model sunt mai apropiate pentru acelasi test set,deci eroarea totala este mai mica; goodness of fit fiind specificat de indicatorul AIC care este mai mic in cazul regresiei multiple.

2.4 Crearea unei noi variabile aleatoare

La setul de date **quakes** am adaugat variabila aleatoare **distance** ce reprezinta raza pe care s-a resimtit cutremurul fiind intr-o relatie de dependenta liniara pozitiva cu numarul de statii ce au inregistrat un seism. Cum v.a.**magnitude** se afla intr-un tip de relatie similar cu **stations**,am construit **distance** folosind valorile din **mag** la care am inmultit cu o constanta pozitiva si am adunat un offset. Este evident ca, cu cat distanta de actiune creste, creste si numarul de statii, cu conditia ca seismul sa se produca intr-o zona in care exista statii.

3 Cerinta III

3.1 Repartitia Cauchy

Fie variabila aleatoare $X \sim \text{Cauchy}(a,b)$ .Functia densitate de probabilitate a v.a. X este

$$f(x) = \frac{1}{b\pi\left(1+\left(\frac{x-a}{b}\right)^2\right)}$$

Functia de repartitie a v.a. X este

$$F(x) = P(X \leq x) = 0.5 + \frac{1}{\pi}\arctan\frac{x-a}{b} \qquad -\infty < x < \infty \qquad -\infty < a < \infty \qquad b > 0.$$

Proprietati:

- “a” este parametrul locatie, “b” este parametrul scala
- Distributia este simetrica fata de parametrul a
- Parametrul b determina latimea distributiei
- Modul si mediana sunt egale cu parametrul a

– Distributia Cauchy standard este data de parametrii a=0 si b=1 si este un caz special pentru distributia t-Student cu un grad de libertate

– Media si varianta nu exista pentru repartitita Cauchy

Aplicatii:

– Poate aparea in biologie, legata de miscarea Browniana

– Populara pentru studiul robustetei

– Modeleaza punctele de impact a unei linii drepte, fixate de particule emise de la un punct sursa

– Caz special pentru distributia t-Student

– Modeleaza raportul a doua variabile aleatoare distribuite normal

– In mecanica cuantica, modeleaza distributia de energie a unei stari instabile

4 Problema bonus

4.1 Punctul a)

Ideea de generare a tabelului: Generam toate probabilitatile p_j, v.a. Y cu $i = j...m$. Pentru fiecare coloana din tabel vom genera $n-1$ valori, una dintre celule(aleasa random dupa indice) va fi lasata libera(egala cu 0).

Functii auxiliare:

- **create_line(limit,end)** - genereaza end probabilitati aleator astfel incat suma probabilitatilor sa fie limit
- **create_rc _names(char1,char2,n)** - creeaza etichete pentru linii si coloane pentru a reprezenta mai sugestiv tabelul(pentru lizibilitate nu se eticheteaza cu valorile v.a. , ci cu indicii acestora)
- **frepcomgen(n,m)** - genereaza valorile v.a. X si Y, eticheteaza matricea corespunzator, genereaza valorile si marcheaza cate un element de pe fiecare coloana ca fiind necompletat, **return value** = matricea si cele doua variabile aleatoare

4.2 Punctul b)

Cum s-a completat tabelul: Parcurgem tabelul, pentru fiecare celula necompletata(marcata cu 0) stim pentru coloana j p_j si calculam

$$mat[i,j] = p_j - \sum_{k=1k\neq i}^{n} mat[i,j]$$

La final, completam si probabilitatile pentru v.a X , calculand suma de pe fiecare linie. Functia **fcomplrepcom** returneaza matricea *mat* completata si v.a X si Y; in plus, se afiseaza matricea initial necompletata si matricea completata impreuna cu repartitiile marginale.

4.3 Punctul c)

Subpuncte:

- $Cov(5X, -3Y) = -15 * Cov(X, Y) = -15[\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)]$
- $\mathbb{P}(0 < X < 3 | Y > 2) = \frac{\mathbb{P}(0<X<3 \cap Y>2)}{1-\mathbb{P}(Y \leq 2)} = \frac{F_x(3)-F_x(0)-F_{xy}(3,2)}{1-F_y(2)}$
- $\mathbb{P}(X > 6, Y < 7) = \mathbb{P}(Y \leq 7) - \mathbb{P}(x \leq 6, Y \leq 7) = F_y(6) - F_{xy}(6,7)$

4.4 Punctul d)

- X, Y independente *ddaca* $\pi_{ij} = p_i * p_j \ , \forall i, j$ - **fverind** parcurge tabelul repartitiei comune si verifica aceasta proprietate; **return value**= TRUE, daca sunt independente, FALSE altfel
- $\rho(X, Y) = 0 \Rightarrow$ v.a X si Y sunt necorelate - **fvernecor** - conditia ce trebuie indeplinita este $Cov(X, Y) = 0$; functia calculeaza $Cov(X, Y)$ si returneaza raspunsul corespunzator