Grupa 233

Set de date: “quakes”

Năstase Alexandru

Avădănei Antonia

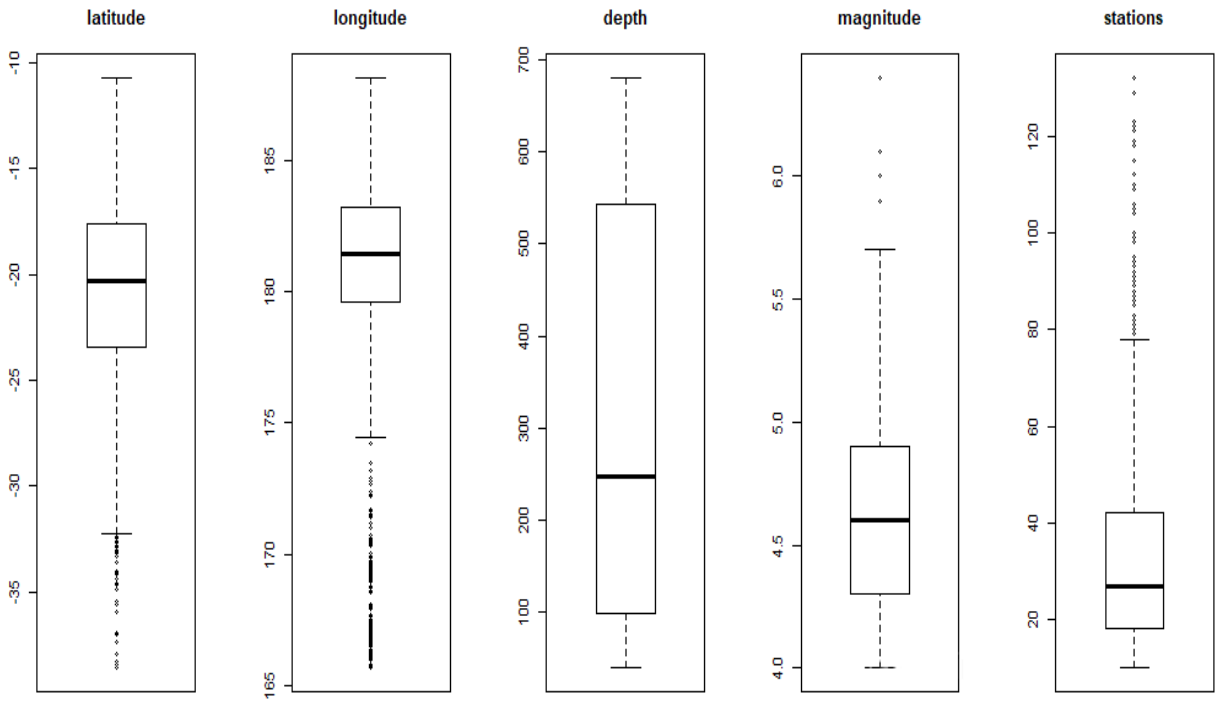
Cincă Adrian

Problema 1:

Ce observam in urma statisticilor dataset-ului "quakes" (interpretari) :

lat	long	depth	mag	stations
Min. :-38.59	Min. :165.7	Min. : 40.0	Min. :4.00	Min. : 10.00
1st Qu.:-23.47	1st Qu.:179.6	1st Qu.: 99.0	1st Qu.:4.30	1st Qu.: 18.00
Median :-20.30	Median :181.4	Median :247.0	Median :4.60	Median : 27.00
Mean :-20.64	Mean :179.5	Mean :311.4	Mean :4.62	Mean : 33.42
3rd Qu.:-17.64	3rd Qu.:183.2	3rd Qu.:543.0	3rd Qu.:4.90	3rd Qu.: 42.00
Max. :-10.72	Max. :188.1	Max. :680.0	Max. :6.40	Max. :132.00

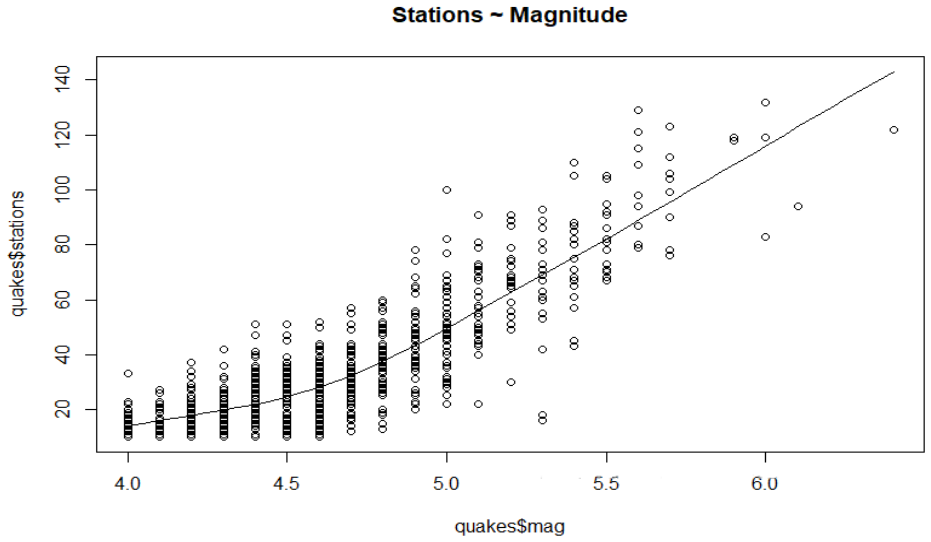
1. Latitudinea maxima - 10.72
2. Au fost foarte putine cutremure care s-au intamplat la o adancime intre 250-500 km comparat cu restul adancimilor de pana in 250 si respectiv peste 500.
3. Aproximativ 1/3 din toate cutremurele au avut o longitudine de 180-182
4. Cea mai mare magnitudine a fost de 6.4 iar majoritatea au avut peste 5



Problema 2:

REGRESIA SIMPLA

- Pentru regresia simpla am ales drept variabila predictor magnitudea, iar drept variabila raspuns numarul de statii care au raportat cutremurul. Astfel, vom examina dacă există o relație liniară între magnitudinea unui cutremur și numărul de stații care au raportat activitatea. Alegerea noastra a plecat de la presupunerea ca pe măsură ce magnitudinea unui cutremur se schimbă, la fel se intampla și cu numărul de stații care îl raportează.

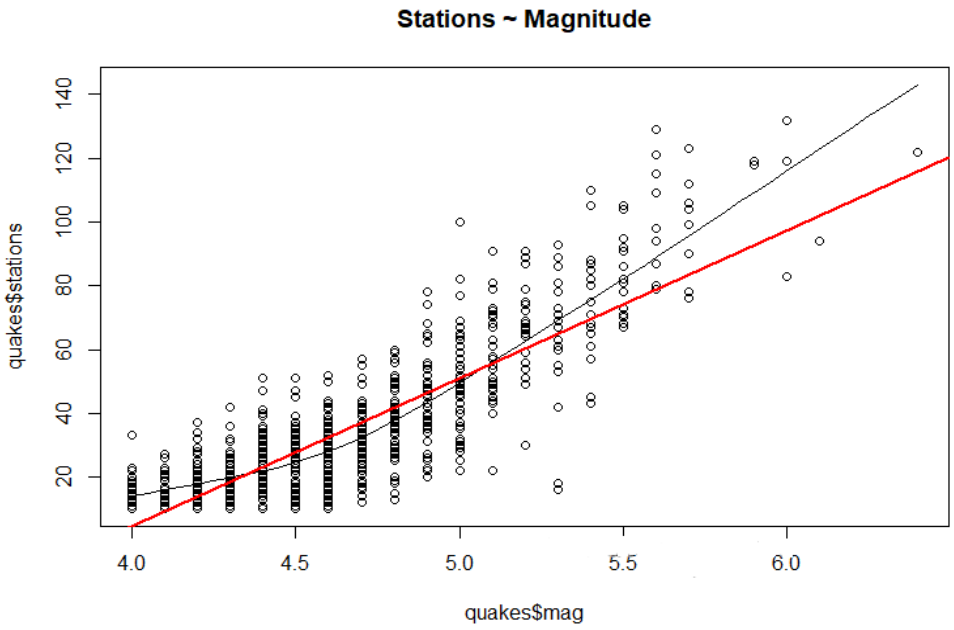


- Scatter.plot-ul de mai sus sugerează o relație în creștere liniară între variabilele „mag” și „stations”. Acesta este un lucru bun, deoarece, una dintre ipotezele care stau la baza regresiei liniare este că relația dintre variabilele răspuns și predictor este liniară și aditivă.
- Corelatia dintre magnitudine si numarul de statii este egala cu 0.8511824, deci cele doua variabile au o *corelație pozitivă.*
- Modelul nostru liniar ne oferă următoarele:
  - **Numarul de stații = -180.42 + 46,28* (Magnitudinea).**

```
Call:
lm(formula = stations ~ mag, data = quakes)

Coefficients:
(Intercept)          mag
    -180.42        46.28
```

- Din coeficientul de pantă (mag), aflăm că o modificare de 1 pe scara Richter va modifica, în medie, numărul de stații de raportare cu 46,28. Deoarece panta este pozitivă, modelul prezice că există o asociere pozitivă între magnitudine și numărul de stații care raportează un cutremur. Coeficientul de interceptare (Intercept) ne spune că dacă magnitudinea cutremurului ar fi zero, -180,42 stații l-ar raporta.

- Adaugam un abline cu parametri Intercept si panta pentru a vizualiza linia regresiei.

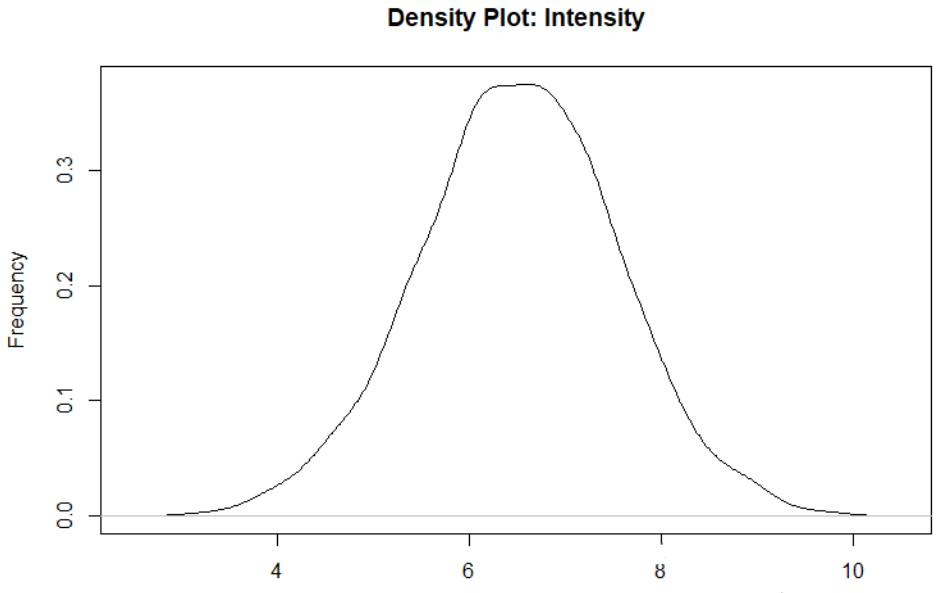


- Dupa ce afisam summary-ul modelului avem urmatoarele:

```
Call:
lm(formula = stations ~ mag, data = quakes)

Residuals:
    Min      1Q  Median      3Q     Max
-48.871  -7.102  -0.474   6.783  50.244

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -180.4243     4.1899  -43.06   <2e-16 ***
mag           46.2822     0.9034   51.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.5 on 998 degrees of freedom
Multiple R-squared:  0.7245,    Adjusted R-squared:  0.7242
F-statistic:  2625 on 1 and 998 DF,  p-value: < 2.2e-16
```

REGRESIA MULTIPLA

- Generam folosind rnorm datele unei variabile noi (“intensity” - ia valori intre 1 si 12) calculata pe scara Mercalli
- Folosim rnorm deoarece variabila, în mod ideal, trebuie să aibă o distribuție aproape normală (o curbă în formă de clopot), fără a fi înclinată spre stânga sau spre dreapta.



- Vom examina dacă există o relație liniară multiplă între intensitatea si adancimea unui cutremur și numărul de stații care au raportat activitatea.
- Modelul nostru liniar ne oferă următoarele:
    - **Numarul de stații = 41.548575 -0.883683 * (Intensity) -0.007617*(Depth).**

```
Call:
lm(formula = stations ~ intensity + depth, data = quakes)

Coefficients:
(Intercept)    intensity        depth
  41.548575    -0.883683    -0.007617
```

- Dupa ce afisam summary-ul modelului avem urmatoarele:

```
Call:
lm(formula = stations ~ intensity + depth, data = quakes)

Residuals:
    Min      1Q  Median      3Q     Max
-26.397 -15.583  -6.649   7.926  99.076

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.548575   4.571853   9.088  <2e-16 ***
intensity   -0.883683   0.671011  -1.317   0.1882
depth       -0.007617   0.003208  -2.374   0.0178 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.84 on 997 degrees of freedom
Multiple R-squared:  0.007132,  Adjusted R-squared:  0.00514
F-statistic: 3.581 on 2 and 997 DF,  p-value: 0.02822
```

Dintre cele două modele construite consideram că primul este mai potrivit pentru setul nostru de date deoarece:

STATISTICA	CRITERIU	**PRIMUL MODEL**	AL DOILEA MODEL
p-value model	<0.05	<2.2e-16	0.02822
R-squared	Higher the better *(> 0.70)*	0.7245	0.007132
AIC	Lower the better	7726.676	9010.728
BIC	Lower the better	7741.399	9030.359

- Preziceri pentru primul model:

```
Call:
lm(formula = stations ~ mag, data = trainingData)

Residuals:
    Min      1Q  Median      3Q     Max
-48.704  -7.084  -0.480   6.916  50.486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -179.277      4.740  -37.82   <2e-16 ***
mag           46.034      1.025   44.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.57 on 798 degrees of freedom
Multiple R-squared:  0.7167,    Adjusted R-squared:  0.7163
F-statistic:  2019 on 1 and 798 DF,  p-value: < 2.2e-16
```

- Din rezumatul modelului, p-value a modelului și p-value a predictorului sunt mai mici decât 0.05, deci știm că avem un model semnificativ statistic. De asemenea, R-squares este mai mare decat 0.7
- Corelația mare (0.8666185) implică faptul că valorile actuale și cele prezise au o mișcare direcțională similară, adică atunci când valorile reale cresc, valorile prezise cresc și vice-versa.

```
> head(actuals_preds)
   actuals predicteds
3       43  69.307604
5       11   4.859640
8       15  23.273344
22      12  14.066492
25      57  69.307604
39      17   9.463066
```

Problema 3:

Repartitia f este in stransa legatura cu repartitia chi. Daca repartitia chi are nevoie de un singur parametru pentru gradul de libertate(df), pentru f vom ultiliza mai multi. F test, care foloseste repartia, este o metoda de a compara multiple variabile independente din mai multe grupuri distribuite normal.Un exemplu concret: Sa presupunem ca vrem sa testam un nou medicament. In acest caz, dorim sa determinam efectele adverse in functie de dozajul utilizat pe un grup de x pacienti. Pentru acest exemplu se va utiliza repartitia f(cei 2 parametrii fiind dozajul si numarul de pacienti).