Proiect Biostatistică și programare în R

Rebeca Oriana Bâscă

Setul de date ales pentru a fi folosit în proiect se numește **Seatbelts** și reprezintă o serie cronologică care prezintă totalul lunar al șoferilor de mașini din Marea Britanie uciși sau răniți grav din ianuarie 1969 până în decembrie 1984, având în vedere faptul că purtarea obligatorie a centurilor de siguranță a fost introdusă la 31 ianuarie 1983. Datele sunt organizate în 9 coloane, și anume DriversKilled (numărul șoferilor decedați), drivers (numărul total al șoferilor), front (pasagerii din locul din față răniți sau uciși), rear (pasagerii din spate răniți sau uciși), kms (distanța totală parcursă de șofer), PetrolPrice (prețul combustibilului la momentul respectiv), VanKilled (numarul șoferilor de furgonete uciși) și law (o variabilă binară ce reprezintă dacă legea purtării centurii era sau nu pusă în aplicare la momentul respectiv).

Deoarece setul de date este original o serie cronologică, pentru a putea realiza acțiunile dorite și a prelucra informațiile furnizate, a fost necesară convertirea acestuia într-un data frame, structură asemănătoare unei matrici cu coloane bine definite, unde am adăugat și coloana responsabilă cu momentul cronologic respective fiecărei înregistrări:

```
Seatbelts <- data.frame(as.matrix(Seatbelts), date=time(Seatbelts))
```

După această transformare au urmat să fie aplicate tehnici de statistică descriptivă pentru anumite coloane, s-a împărțit setul de date în două cadre pe baza unui criteriu definit, au fost create anumite ipoteze statistice cu privire la mai multe variabile și de asemenea a fost realizat un model predictiv pentru obținerea unor variabile pe baza altora deja existente.

Pentru început, am atașat datasetul ales, Seatbelts, la sesiunea de lucru curentă, pentru a ajunge mai ușor la coloanele aferente și am afișat informații despre conținut, precum și primele câteva rânduri pentru a fi mai ușor de vizualizat structura datelor.

```
dataset = Seatbelts
View(dataset)
attach(dataset)
str(dataset)
head(dataset)
```

```
> str(dataset)
'data.frame':   192 obs. of  9 variables:
 $ DriversKilled: num  107 97 102 87 119 106 110 106 107 134 ...
 $ drivers      : num  1687 1508 1507 1385 1632 ...
 $ front        : num  867 825 806 814 991 ...
 $ rear         : num  269 265 319 407 454 427 522 536 405 437 ...
 $ kms          : num  9059 7685 9963 10955 11823 ...
 $ PetrolPrice  : num  0.103 0.102 0.102 0.101 0.101 ...
 $ VanKilled    : num  12 6 12 8 10 13 11 6 10 16 ...
 $ law          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ date         : Time-Series  from 1969 to 1985: 1969 1969 1969 1969 1969 ...
> head(dataset)
  DriversKilled drivers front rear   kms PetrolPrice VanKilled law     date
1           107    1687   867  269  9059   0.1029718        12   0 1969.000
2            97    1508   825  265  7685   0.1023630         6   0 1969.083
3           102    1507   806  319  9963   0.1020625        12   0 1969.167
4            87    1385   814  407 10955   0.1008733         8   0 1969.250
5           119    1632   991  454 11823   0.1010197        10   0 1969.333
6           106    1511   945  427 12391   0.1005812        13   0 1969.417
```

Realizarea unor statistici descriptive la nivel de coloane a fost realizată atât cu funcții precum plot sau hist, cât și mean sau max. Am ales să calculez media distanței parcurse de șoferi, valoarea maximă la care a ajuns prețul combustibilului:

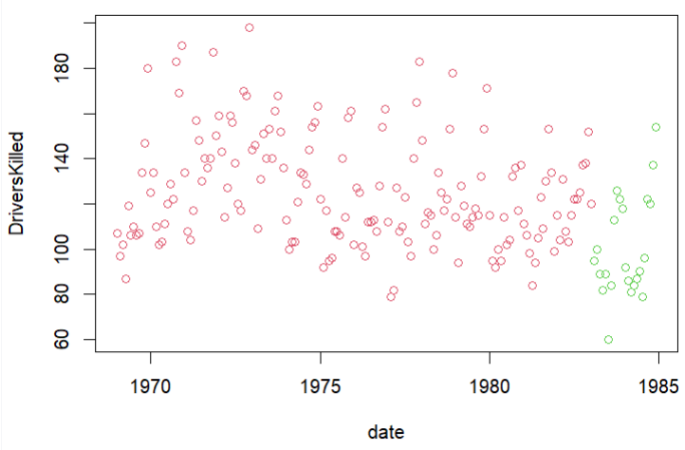
```
mean(kms)
max(PetrolPrice)
```

```
> mean(kms)
[1] 14993.6
> max(PetrolPrice)
[1] 0.1330274
```

,

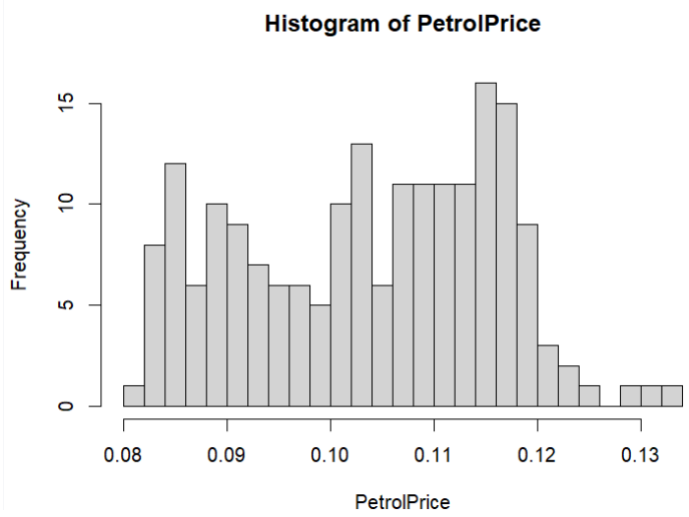
iar ca și reprezentări grafice, am ales să proiectez efectul legii centurilor de siguranță printr-o diagramă de dispersie a șoferilor uciși de-a lungul timpului, fiind scos în evidență momentul în care s-a introdus acea lege:

```
plot(date, DriversKilled, col=(law+2))
```

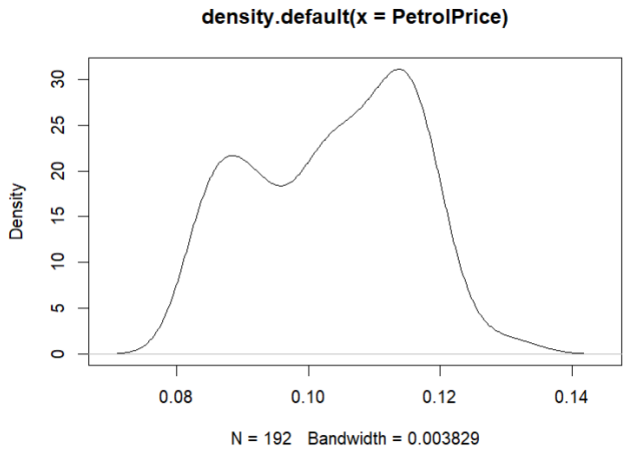


De aemenea, am ales să proiectez o histogramă ce prezintă prețul combustibilului și o diagramă a estimării densității acestuia:

```
hist(PetrolPrice,breaks = 20)
```



```
plot(density(PetrolPrice))
```



Pentru a vedea câteva detaii despre coloana drivers, am aplicat asupra acesteia funcțiile skewness() și kurtosis() și am obținut valorile de asimetrie a distribuției și gradul de concentrație pe care îl prezintă valorile acestei variabile în jurul zonei centrale a distribuției de frecvență:

```
library(moments)
skewness(drivers)
kurtosis(drivers)
```

```
> library(moments)
> skewness(drivers)
[1] 0.5344923
> kurtosis(drivers)
[1] 3.103583
```

Separarea setului de date în două seturi diferite a fost relizată pe baza condiției de diferențiere a datelor înainte și după ce legea purtării centurii de siguranță a fost implementată. Adică, separarea a fost bazată pe valoarea din coloana law, respective 0 sau 1.

```
Law0 = subset(dataset,law == 0)
Law1 = subset(dataset,law != 0)
```

Fiecare subset a fost mai apoi analizat cu ajutorul funcției summary(), iar astfel s-au putut observa ușor pentru fiecare coloană în parte indicatorii static numerici, cum ar fi quartilele, mediana sau valoarea medie.

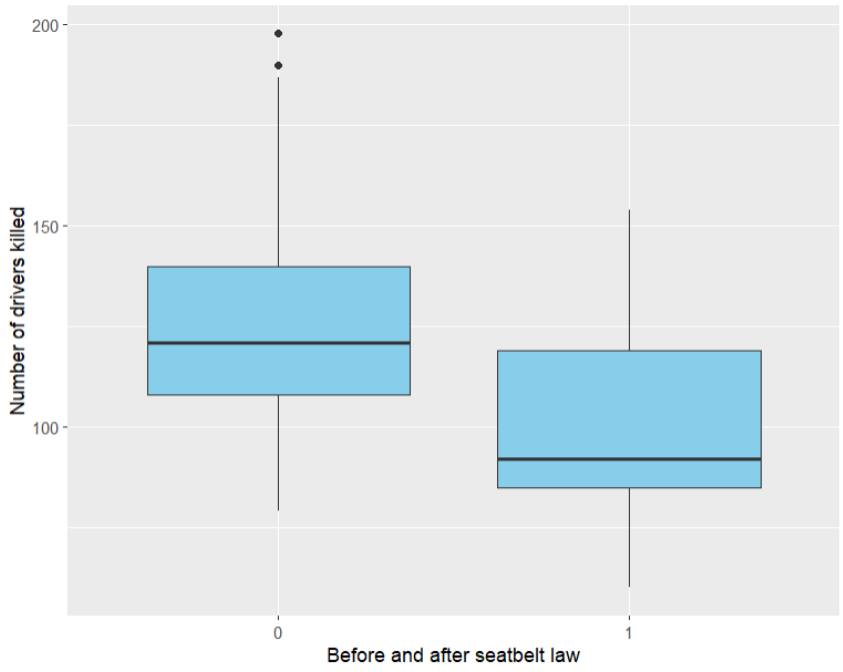
```
summary(Law0Sub)
summary(Law1Sub)
```

```
> summary(Law0Sub)
      Year          Month     DriversKilled      drivers         front            rear
 Min.   :1969   Jan    :15   Min.   : 79.0   Min.   :1309   Min.   : 567.0   Min.   :224.0
 1st Qu.:1972   Feb    :14   1st Qu.:108.0   1st Qu.:1511   1st Qu.: 767.0   1st Qu.:344.0
 Median :1976   Mar    :14   Median :121.0   Median :1653   Median : 860.0   Median :401.0
 Mean   :1976   Apr    :14   Mean   :125.9   Mean   :1718   Mean   : 873.5   Mean   :400.3
 3rd Qu.:1979   May    :14   3rd Qu.:140.0   3rd Qu.:1926   3rd Qu.: 986.0   3rd Qu.:454.0
 Max.   :1983   Jun    :14   Max.   :198.0   Max.   :2654   Max.   :1299.0   Max.   :646.0
                (Other):84
      kms          PetrolPrice        VanKilled          law
 Min.   : 7685   Min.   :0.08118   Min.   : 2.000   Min.   :0
 1st Qu.:12387   1st Qu.:0.09078   1st Qu.: 7.000   1st Qu.:0
 Median :14455   Median :0.10273   Median :10.000   Median :0
 Mean   :14463   Mean   :0.10187   Mean   : 9.586   Mean   :0
 3rd Qu.:16585   3rd Qu.:0.11132   3rd Qu.:13.000   3rd Qu.:0
 Max.   :21040   Max.   :0.13303   Max.   :17.000   Max.   :0
```

```
> summary(Law1Sub)
      Year          Month    DriversKilled      drivers        front           rear
 Min.   :1983   Feb    : 2   Min.   : 60.0   Min.   :1057   Min.   :426.0   Min.   :296.0
 1st Qu.:1983   Mar    : 2   1st Qu.: 85.0   1st Qu.:1171   1st Qu.:516.0   1st Qu.:347.0
 Median :1984   Apr    : 2   Median : 92.0   Median :1282   Median :585.0   Median :408.0
 Mean   :1984   May    : 2   Mean   :100.3   Mean   :1322   Mean   :571.0   Mean   :407.7
 3rd Qu.:1984   Jun    : 2   3rd Qu.:119.0   3rd Qu.:1464   3rd Qu.:629.5   3rd Qu.:471.5
 Max.   :1984   Jul    : 2   Max.   :154.0   Max.   :1763   Max.   :721.0   Max.   :521.0
                (Other):11
      kms          PetrolPrice       VanKilled          law
 Min.   :15511   Min.   :0.1131   Min.   :2.000   Min.   :1
 1st Qu.:17971   1st Qu.:0.1148   1st Qu.:3.500   1st Qu.:1
 Median :19162   Median :0.1161   Median :5.000   Median :1
 Mean   :18890   Mean   :0.1165   Mean   :5.174   Mean   :1
 3rd Qu.:19952   3rd Qu.:0.1180   3rd Qu.:7.000   3rd Qu.:1
 Max.   :21626   Max.   :0.1201   Max.   :8.000   Max.   :1
```

Din aceste atatistici obținute am observat că numărul șoferilor uciși este considerabil mai mic în subsetul Law1, adică după ce legea a fost aplicată. Astfel, am realizat un grafic de tip boxplot, pentru a observa această diferență:

```
law_comparison <-ggplot(Seatbelts, aes(x=factor(law), y =DriversKilled)) +geom_boxplot(fill = "skyblue")
  +theme_grey()+ylab ("Number of drivers killed")+xlab("Before and after seatbelt law")
law_comparison
```



De asemenea, am aplicat și un test de statistică *t test* pentru a verifica ipoteza anterioară ce susține că numărul deceselor în rândul șoferilor a scăzut în urma reglementării acelei legi.

```
t.test(Law0$DriversKilled, Law1$DriversKilled, mu = 0, alternative = "greater")
```

```
        Welch Two Sample t-test

data:  Law0$DriversKilled and Law1$DriversKilled
t = 5.1253, df = 29.609, p-value = 8.467e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 17.12488      Inf
sample estimates:
mean of x mean of y
 125.8698  100.2609
```

Se poate observa că valoarea lui t este pozitivă, iar cele două grupuri au valori destul de similare. Totuși, valoarea probabilității p-value este diferită de 0 și putem concluziona, cu ajutorul ipotezei alternative, că o valoare negativă înseamnă că grupul experimental este asociat cu o scădere a valorii rezultatului. Deci, ipoteza realizată inițial este corectă, iar numărul șoferilor decedați a scăzut în urma inițierii legii purtatului de centură.

Aceste teste se pot realiza folosind orice coloană din cele două seturi de date, cum ar fi numărul pasagerilor din față sau din spatele mașinii ce au fost uciși, sau distanța parcursă de mașini.

Testul Shapiro este un test statistic folosit pentru a verifica dacă datele luate în considerare sunt date distribuite în mod normal sau nu. Ipoteza nulă afirmă că populația este distribuită normal, adică dacă valoarea p este mai mare de 0,05, atunci ipoteza nulă este acceptată. Ipoteza alternativă afirmă că populația nu este distribuită în mod normal, adică dacă valoarea p este mai mică sau egală cu 0,05, iar atunci ipoteza nulă este respinsă.

Am aplicat acest test Shapiro pe datele privind distanța parcursă de șoferi, separat pentru subsetul Law0 și Law1 :

```
> shapiro.test(Law0$kms)

        Shapiro-Wilk normality test

data:  Law0$kms
W = 0.99089, p-value = 0.3567

> shapiro.test(Law1$kms)

        Shapiro-Wilk normality test

data:  Law1$kms
W = 0.9766, p-value = 0.8409
```

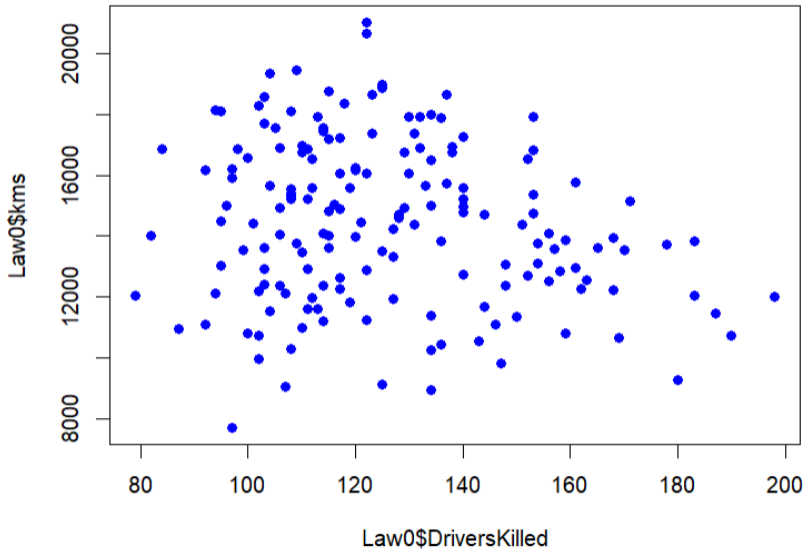
,

Și se poate observa că în ambele cazuri valoare p este mai mare de 0.05, deci distribuția valorilor este normală.

Modelele predictive din R utilizează operațiuni statice pentru a analiza fapte deja existente pentru a putea prezice evenimente viitoare. Am ales să fac un model pentru a analiza relația dintre numărul kilometrilor parcurși de un șofer și rata mortalității șoferilor, în contextul în care legea nu a fost încă implementată, deci purtatul centurii de siguranță nu era obligatoriu.

Am început cu o reprezentare grafică a acestor date, precum și calcularea valorii de corelație dintre cei doi vectori:

```
plot(Law0$DriversKilled, Law0$kms, pch=16, col= "Blue")
cor(Law0$DriversKilled, Law0$kms)
```




```
> cor(Law0$DriversKilled, Law0$kms)
[1] -0.1914247
```

Modelul liniar a fost realizat cu ajutorul funcției lm(), iar astfel am putut să găsim formula corespunzătoare cu care putem calcula numărul morților șoferilor folosindu-ne de distanța parcursă în kilometri:

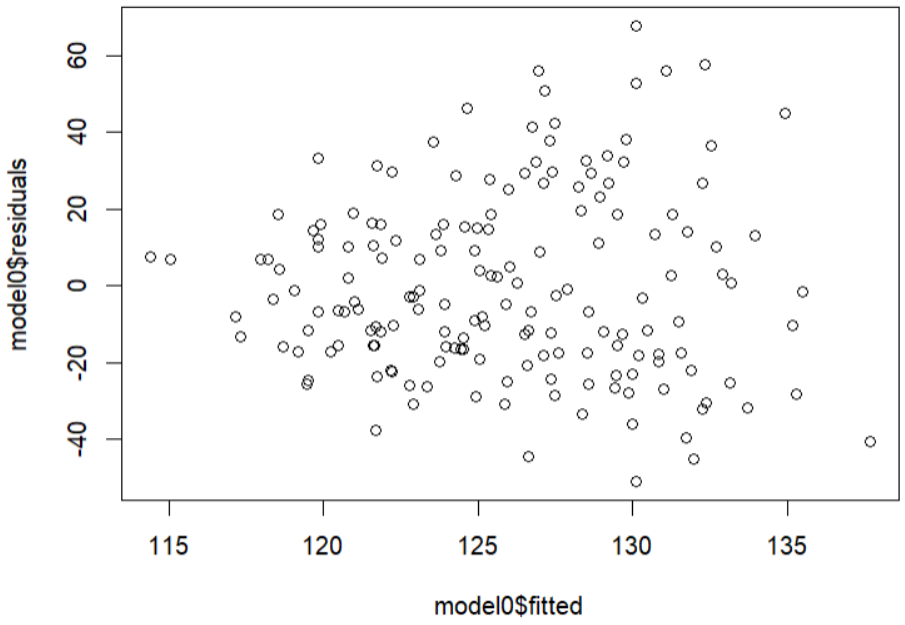
```
model0 = lm(Law0$DriversKilled ~ Law0$kms, data= Law0)
model0
```

```
> model0

Call:
lm(formula = Law0$DriversKilled ~ Law0$kms, data = Law0)

Coefficients:
(Intercept)     Law0$kms
 151.091397    -0.001744
```

Astfel, putem spune că DriversKilled ~~ (-0.001744)*kms + 151.091397.

Datele reziduale ale modelului de regresie liniară simplă sunt diferența dintre datele observate ale variabilei dependente y și valorile ajustate ŷ, iar pentru modelul de mai sus am realizat un grafic ce ne ajută la observarea mai ușoară a acestor date reziduale:

```
rez=model0$residuals
plot(model0$fitted,model0$residuals)
```


model0$residuals
60
40
20
0
-20
-40
115
120
125
130
135
model0$fitted