CURS 6 – REGRESIA LINIARA SIMPLA

Rezumat

1. **Formularea problemei**
2. **Definitia regresiei liniara simple**
3. **Estimarea dreptei de regresie**
4. **Teste asupra coeficientilor si semnificatiei modelului**
5. **Test asupra dreptei de regresie in R**
6. **Analiza calitatii modelului**
7. **Predictia**
8. **Codul in R**

1. Formularea problemei

- Fisier de lucru: data1.csv
- Care este relatia intre inaltime (height) si greutate (weight)
- Cand inaltimea creste, creste si greutatea?
- Cunoscand greutatea (weight), puteti prezice inaltimea?

Exemplu:

Studiati relatia dintre greutate si inaltime pentru persoanele din esantionul dat in data1.csv

- Incarcati fisierul
- Calculati media greutatilor so inaltimilor: m=?
- Calculati dispersia greutatilor so inaltimilor: s^2=?
- Construiti histogramele si boxplot-urile pentru greutate si inaltime
- Faceti graficul greutatii fata de inaltime
- Cod R:

```
data1<-read.csv("E:\\ data1.csv", sep=",", header=TRUE)
data1
weight<-data1[,1]
height<-data1[,2]
par(mfrow=c(2,2))
hist(weight, col="blue")
boxplot(weight, col="blue")
hist(height, col="red" )
boxplot(height, col="red" )
mean(weight)
var(weight)
mean(height)
var(height)
plot(weight, height)
```

2. Definitia regresiei liniara simple

- Regresia lui Y in functie de X
  - Y= height(cm)
  - X= weight (Kg)
- Se determina relatia lui Y fata de X ⟹ **height= *f(weight)***
- Functie liniara **E(height|weight) = α+β*weight**
- Pentru fiecare individ

height = α+β*weight +ε → Eroare individuala

3. Estimarea dreptei de regresie

- Dreapta de regresie se determina astfel ca suma patratelor errorilor (SCE) sa fie minima

$$y_i = \alpha + \beta \times x_i + \varepsilon_i$$
$$E(Y/X) = \alpha + \beta \times X \Rightarrow \boxed{\varepsilon_i = y_i - E(Y/X)}$$
$$SCE = \sum_{i=1}^{n} (\varepsilon_i)^2$$

- Estimator al pantei (β): $b = \frac{cov(X,Y)}{var(X)}$
- Estimator al intercept (α): $a = E(Y) - b \times E(Y|X)$
- Covarianta dintre inaltime si greutate: in R: cov(height, weight)
- Estimator al lui β: in R: b<-cov(height, weight)/var(weight); b

4. Teste asupra coeficientilor si semnificatiei modelului

- *Testul t asupra pantei:* $H_0: \beta = 0$, $H_0: \beta \neq 0$.
- *Testul t asupra intercept:* $H_0: \alpha = 0$, $H_0: \alpha \neq 0$.
- *Testul F asupra modelului in ansamblu:* $H_0: \alpha = \beta = 0$, $H_1: \alpha, \beta$ *nu sunt nuli simultan*

5. Test asupra dreptei de regresie in R: *functia lm*

- Se construieste modelul liniar: mod1<-lm(height~1 + weight)
- Se vizualizeaza: mod1

  Output-ul este dat in chemarul albeastru de mai jos.

6. Analiza calitatii modelului

- Intervalul de incredere pentru parametri: *confint(mod1)*

	2.5 %	97.5 %
(Intercept)	95.5045320	117.459904
weight	0.6789554	1.010671

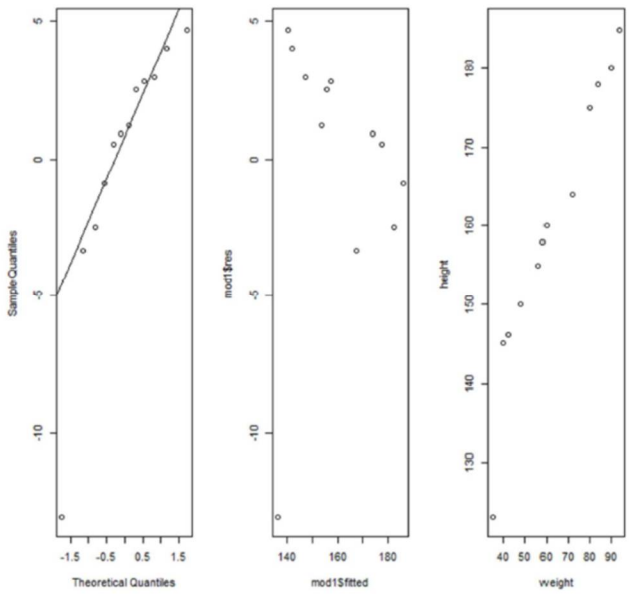
- Coeficientul de corelatie: *r = cor(weight, height)*; *r*
- Coeficientul de determinatie:R^2: *R^2=var(mod1$fitted.value)/var(height)*

Obs: Coeficientul de determinatie se determina din: summary(mod1)

```
summary(mod1)

Call:
lm(formula = height ~ 1 + weight)

Residuals:
Min     1Q  Median     3Q    Max
-13.051  -1.300   1.071   2.863   4.725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 106.48222    4.92684   21.61 1.00e-09 ***
weight        0.84481    0.07444   11.35 4.93e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.028 on 10 degrees of freedom
Multiple R-squared:  0.928,    Adjusted R-squared:  0.9208
F-statistic: 128.8 on 1 and 10 DF,  p-value: 4.926e-07
```

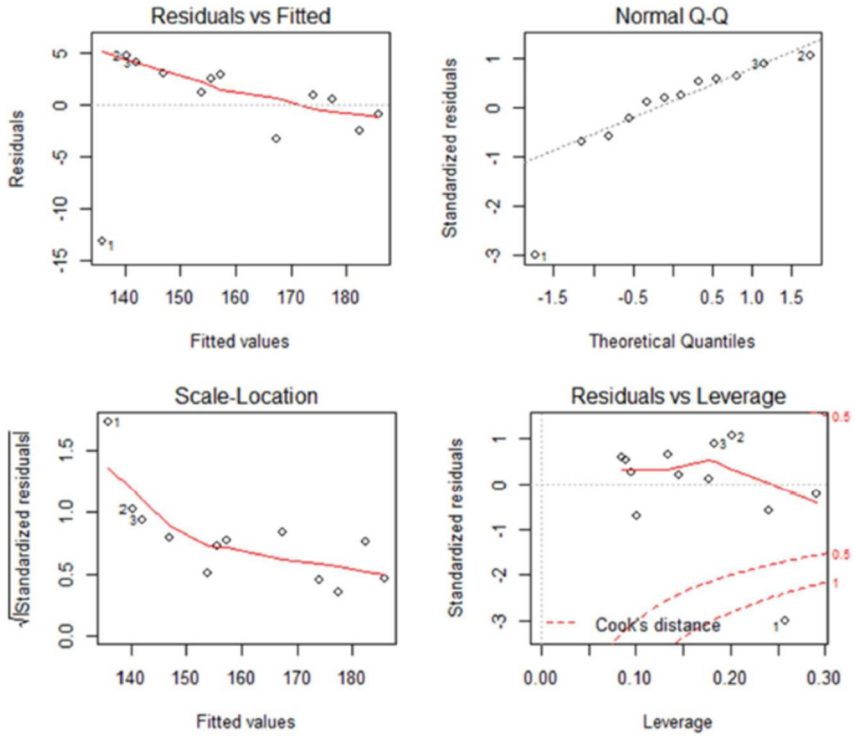
- Reziduurile sunt obtinute cu instructiunea: *mod1$res*
- *Teste asupra seriei reziduurilor*:
  - Normalitate: H_0: seria este normal distribuita; H_1: seria nueste normal distribuita
  - Homoscedasticitate: H_0: seria este homoscedastica (are varianta constanta);
    H_1: seria este heteroscedastica
  - Necorelare
  - Reziduurile nu sunt corelate cu variabilele explicative

```
par(mfrow=c(1,3))
qqnorm(mod1$res)
qqline(mod1$res)
plot(mod1$fitted,mod1$res)
plot(weight,height)
```



par(mfrow=c(2,2))

plot(mod1)



7. Predictia

- Greutatea (weight)

  new.x=data.frame(weight=60.2)
- Predictia/estimarea unei anumite inaltimi (height)

  Tx<-predict(mod1,newdata=new.x,se.fit=TRUE)

  Tx
- Intervalul de incredere al greutatii estimate (weight)

  Confint<-predict(mod1,newdata=new.x, interval="confidence")

  Confint$fit
- Intervalul de incredere al inaltimii estimate (height)

  Predint<-predict(mod1,newdata=new.x, interval="prediction")

  Predint$fit

8. Codul in R

```
data<-read.csv("data1.csv", sep=",", header=TRUE)
data1
```

	weight	height
1	35	123
2	40	145
3	42	146

```
4     48   150
5     56   155
6     58   158
7     60   160
8     72   164
9     80   175
10    84   178
11    90   180
12    94   185
```

```
weight<-data1[,1]
height<-data1[,2]
hist(height, col="red")
hist(weight, col="blue")
boxplot(height, col="red")
boxplot(weight, col="blue")
mean(weight)
var(weight)
mean(height)
var(height)
plot(weight, height)
cov(height, weight)
b<- cov(height, weight)/var(weight)
b
mod1<-lm(height~1+weight)
mod1
summary(mod1)
par(mfrow=c(1,3))
qqnorm(mod1$res)
qqline(mod1$res)
plot(mod1$fitted,mod1$res)
plot(weight,height)
par(mfrow=c(2,2))
X<-mod1$res
library(fBasics)
ksnormTest(X)
shapiroTest(X)
library(nortest)
```

```
ad.test(X)
library(dplyr)
CO<-c(1,1,1,1,1,1,2,2,2,2,2,2)
B<-bartlett.test(X~CO)
```