ΔΙΑΛΕΞΗ 29/2/24

σχόλιο

```
> # Input data
> data <- data.frame(c(152,     183,171,165,158,161,149,158,170,153,164,190,185),
+                    c( 120,    141,124,126,117,129,123,125,132,123,132,155,147),
+                    c(50,  20,  20,  30,  30,  50,  60,  50,  40,  55,  40,  40,  20)
+ )
> colnames(data) <- c("Y", paste0("X",1:2))
> data
    Y  X1 X2
1  152 120 50
2  183 141 20
3  171 124 20
4  165 126 30
5  158 117 30
6  161 129 50
7  149 123 60
8  158 125 50
9  170 132 40
10 153 123 55
11 164 132 40
12 190 155 40
13 185 147 20
```

print - εκτύπωση του αρχείου data

```
> #*******************************************************

>  # INPUT DATA FROM EXCEL FILE
> install.packages("readxl")
> library(readxl)
> mydata<-read_xlsx("Proodos.xlsx")
> mydata
    Y    X1    X2
  <dbl> <dbl> <dbl>
 1  120  152   50
 2  141  183   20
 3  124  171   20
 4  126  165   30
 5  117  158   30
 6  129  161   50
 7  123  149   60
 8  125  158   50
 9  132  170   40
10  123  153   55
11  132  164   40
12  155  190   40
13  147  185   20
```

πακέτο για διάβασμα αρχείων excel

εντολή για διάβασμα

```
> Y = mydata$Y
> A = mydata$X1
> B = mydata$X2
> Y
 [1] 120 141 124 126 117 129 123 125 132 123 132 155 147
> A
 [1] 152 183 171 165 158 161 149 158 170 153 164 190 185
> B
 [1] 50 20 20 30 30 50 60 50 40 55 40 40 20
```

```
> cor.matrix<-cor(cbind(Y,A,B))
> cor.matrix
          Y          A          B
Y  1.0000000  0.9025094 -0.3619206
A  0.9025094  1.0000000 -0.7002831
B -0.3619206 -0.7002831  1.0000000
> cov.matrix<-cov(mydata)
> cov.matrix
           Y          X1          X2
Y  123.89744   132.2244   -55.86538
X1 132.22436   173.2436  -127.82051
X2 -55.86538  -127.8205   192.30769
```

$$-1 \le \text{Correlation} = \text{συσχέτιση} = \frac{Cov(W,Z)}{\sqrt{VarW \cdot VarZ}} \le +1$$

Cov: Συνδιασπορά (Covariance)

$$Cor(Y,Y) = \frac{Cov(Y,Y)}{\sqrt{VarY \, VarY}} = \frac{VarY}{VarY} = 1$$

Συμμετρικός πίνακας

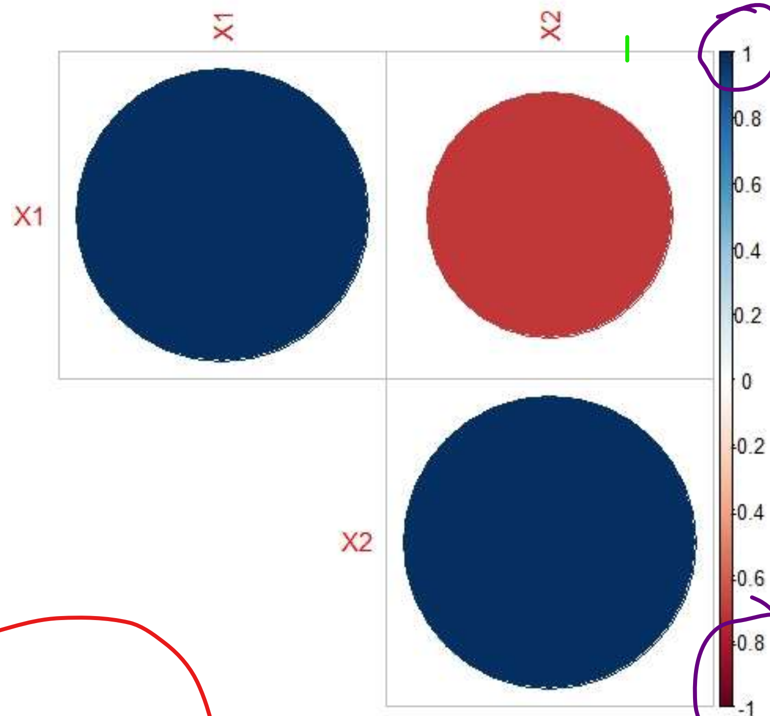
$$Cor(A,B) = -0.7002831$$

VarY = 123.89   VarX1 = VarA = 173.2

VarB = VarX2 = 192.3

$$Cov(W,Z) = E(WZ) - E(W)E(Z)$$

```
> library(corrplot)
> # visualize pairwise correlation
> corrplot(cor(mydata[ , -1]), type = "upper")
```



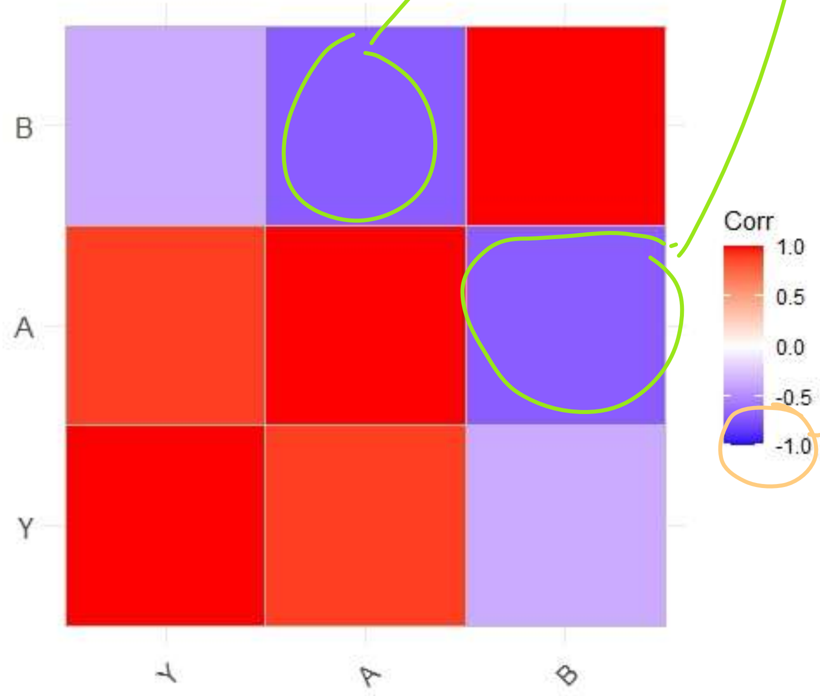
όσο πιο μπλε τόσο πιο θετική συσχέτιση

όσο πιο κόκκινο τόσο πιο αρνητική συσχέτιση

Corr = 0 ⇒ ανεξ/ασυσχέτιστα

Corr ≈ 1 ή -1 συσχετισμένες

```
> install.packages("ggcorrplot")
> library(ggcorrplot)
> ggcorrplot(cor.matrix)
```

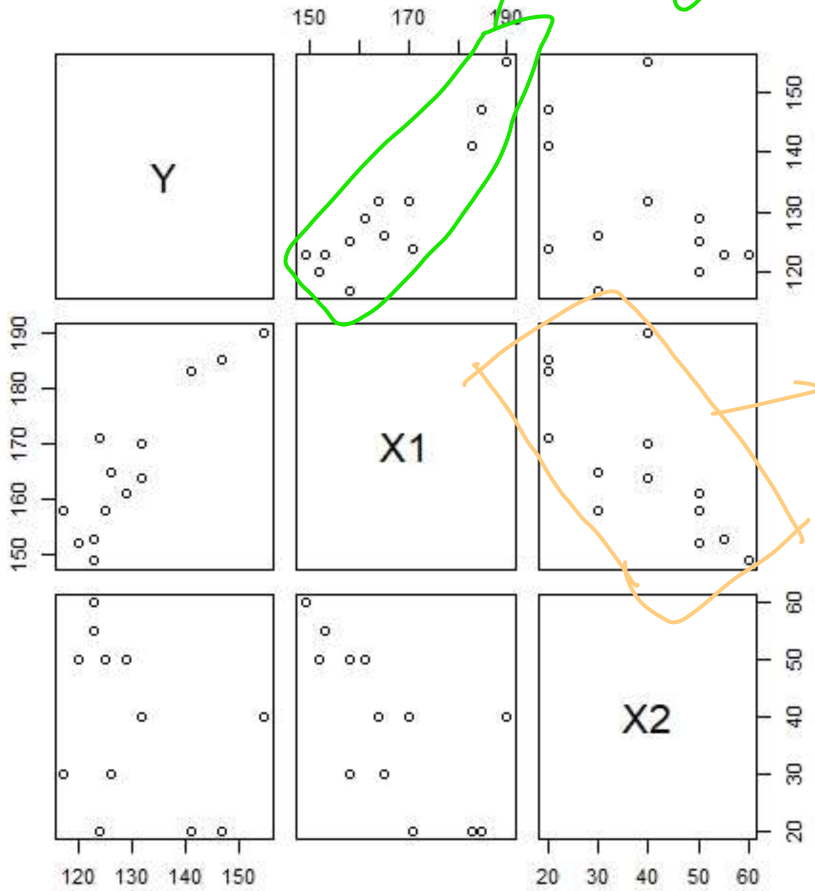


> pairs(mydata)

Pearson correlation

Spearman correlation

Kendall τ correlation



εκτιμώ με παλινδρόμηση: $Y_i = a_0 + a_1 \cdot A_i + a_2 \cdot B_i + \varepsilon_i$

$Y \sim A + B$

```
> regr<-lm(Y~A+B)
> summary(regr)
```

a_0, a_1, a_2 } συντελεστές παλινδρόμησης

θέλουμε να τα εκτιμήσουμε

$\varepsilon \sim N(0, \sigma^2)$

σφάλματα ΑΓΝΩΣΤΑ!

Call:

lm(formula = Y ~ A + B)

Residuals:

Min	1Q	Median	3Q	Max
-3.5929	-1.3530	-0.0946	0.6155	4.3276

1st quantile

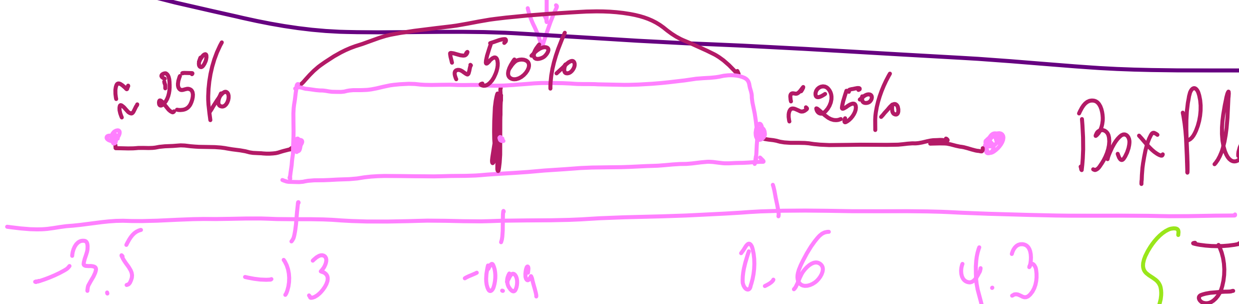
50% below
50% above

3rd quantile

Μετά από την παλινδρόμηση βρίσκω:

$\hat{Y}_i = \hat{a}_0 + \hat{a}_1 \cdot A_i + \hat{a}_2 \cdot B_i$

$Y_i - \hat{Y}_i = \hat{\varepsilon}_i$ = Εκτίμηση σφάλματος = Residual



Box Plot - 5 point summary

$IQR = 3Q - 1Q$ = Interquantile range

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-65.09968	14.94458	-4.356	0.001430	**
A	1.07710	0.07707	13.975	6.89e-08	***
B	0.42541	0.07315	5.815	0.000169	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.509 on 10 degrees of freedom

Multiple R-squared: 0.9577, Adjusted R-squared: 0.9492

F-statistic: 113.1 on 2 and 10 DF, p-value: 1.359e-07

```
> anova(regr)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
A          1 1211.01 1211.01 192.433 7.393e-08 ***
B          1  212.83  212.83  33.819 0.0001694 ***
Residuals 10   62.93    6.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR = 1211.01 + 212.83 = 1423.84

SSE = Residuals

$SSTO = \sum (Y_i - \bar{Y})^2 \approx 1486.77 = 1211 + 212.8 + 62.93$

$\frac{SSR}{SSTO} = \frac{1423.84}{1486.77} = 0.957673 \approx 0.9577 = R^2$

> library(car)

> #create vector of VIF values

> vif_values <- vif(regr)

> vif_values

A	B
1.96231	1.96231

Είναι οι A, B ανεξάρτητες; Έχουμε βρει $Corr(A,B) \approx -0.7$ πάνω από 1 άρα δεν είναι ανεξάρτητες

VIF: VARIANCE INFLATION FACTOR

Προϋποθέσεις Παλινδρόμησης

(α) ανεξαρτησία → θέλουμε
(β) γραμμικότητα (Y, X)
(γ) ομοσκεδαστικότητα $= Var(\varepsilon) = \sigma^2$
(δ) κανονικότητα → θέλουμε
(ε) $E(\varepsilon) = 0$

$y = a_0 + a_1 A + a_2 B + \varepsilon$

$\varepsilon \sim iid\ N(0, \sigma^2)$

ΑΝΑΛΥΣΗ ΚΑΤΑΛΟΙΠΩΝ
Residual Analysis

με βάση τα ε

NO correlation $vif \equiv 1$

μικρή/μέτρια συσχέτιση $1 < vif < 2$

μεγάλη συσχέτιση > 2 (5 ή 10 practitioners)

$A = \gamma + \delta B + \varepsilon$

"X" παλινδρόμηση της μιας ως προς την άλλη

$$VIF = \frac{1}{1-R^2}$$

- ∄ συσχέτιση $R^2=0 \Rightarrow VIF = \frac{1}{1-0} = 1$
- ∃ [illegible] $R^2=0.9 \rightarrow VIF = \frac{1}{1-0.9} = \frac{1}{0.1} = 10$
- ∃ μέτρια $R^2 \approx 0.5 \rightarrow VIF = \frac{1}{1-0.5} = 2$

```
> #new data set
> mydata2<-read_xlsx("Proodos 2.xlsx")
> mydata2
      Y    X1    X2
  <dbl> <dbl> <dbl>
 1  120   152    50
 2  141   183   120
 3  124   171   115
 4  126   165   100
 5  117   158    80
 6  129   161    90
 7  123   149    50
 8  125   158    85
 9  132   170   110
10   123   153    55
11   132   164   100
12   155   190   120
13   147   185   120
```

```
> Y1 = mydata2$Y
> A1 = mydata2$X1
> B1 = mydata2$X2
> Y1
 [1] 120 141 124 126 117 129 123 125 132 123 132 155 147
> A1
 [1] 152 183 171 165 158 161 149 158 170 153 164 190 185
> B1
 [1]  50 120 115 100  80  90  50  85 110  55 100 120 120
>
```

```
> cor1.matrix<-cor(cbind(Y1,A1,B1))
> cor1.matrix
          Y1        A1        B1
Y1 1.0000000 0.9025094 0.7154185
A1 0.9025094 1.0000000 0.9050344
B1 0.7154185 0.9050344 1.0000000
> cov1.matrix<-cov(mydata2)
> cov1.matrix
          Y       X1       X2
Y  123.8974 132.2244 211.0256
X1 132.2244 173.2436 315.6731
X2 211.0256 315.6731 702.2436
```

```
> regr1<-lm(Y1~A1+B1)
> summary(regr1)

Call:      lm(formula = Y1 ~ A1 + B1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7446 -2.8733  0.7424  2.7817  6.0695
```

```
Coefficients:
            Estimate Std.   Error    t value   Pr(>|t|)
(Intercept) -46.0412   28.5000   -1.615    0.137280
A1            1.1921    0.2255    5.286    0.000355 ***
B1           -0.2354    0.1120   -2.101    0.061932 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.374 on 10 degrees of freedom
Multiple R-squared:  0.8713,    Adjusted R-squared:  0.8456
F-statistic: 33.86 on 2 and 10 DF,  p-value: 3.526e-05
```

> anova(regr1)

Analysis of Variance Table

Response: Y1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
A1	1	1211.01	1211.01	63.3074	1.235e-05	***
B1	1	84.47	84.47	4.4159	0.06193	.
Residuals	10	191.29	19.13			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
> library(car)
> #create vector of VIF values
> vif_values1 <- vif(regr1)
> vif_values1
      A1       B1
5.527524 5.527524
```

```
> Rm<-cov2cor(cov(mydata2))
> phiindex=sqrt((sum(Rm^2)-2)/(2*(2-1)))
> Rm
          Y        X1        X2
Y  1.0000000 0.9025094 0.7154185
X1 0.9025094 1.0000000 0.9050344
X2 0.7154185 0.9050344 1.0000000
> Rm<-cov2cor(cov(cbind(A,B)))
> phiindex=sqrt((sum(Rm^2)-2)/(2*(2-1)))
```

```
> Rm
          X1        X2
X1 1.0000000 0.9050344
X2 0.9050344 1.0000000


> phiindex
[1] 0.9050344
> phiindex
[1] 0.7002831
```

```
# Input data
data <- data.frame(c(152,     183,    171,    165,    158,    161,    149,    158,    170,    153,    164,    190,    185),
              c( 120,      141,    124,    126,    117,    129,    123,    125,    132,    123,    132,    155,    147),
              c(50,        20,     20,     30,     30,     50,     60,     50,     40,     55,     40,     40,     20)
)
colnames(data) <- c("Y", paste0("X",1:2))
data

#**************************************************
 # INPUT DATA FROM EXCEL FILE
install.packages("readxl")
library(readxl)
mydata<-read_xlsx("Proodos.xlsx")
mydata
Y = mydata$Y
A = mydata$X1
B = mydata$X2
Y
A
B

cor.matrix<-cor(cbind(Y,A,B))
cor.matrix
cov.matrix<-cov(mydata)
cov.matrix

library(corrplot)
# visualize pairwise correlation
```

```
corrplot(cor(mydata[ , -1]), type = "upper")

install.packages("ggcorrplot")
library(ggcorrplot)
ggcorrplot(cor.matrix)
pairs(mydata)

regr<-lm(Y~A+B)
summary(regr)
anova(regr)

library(car)
#create vector of VIF values
vif_values <- vif(regr)
vif_values

#new data set

mydata2<-read_xlsx("Proodos 2.xlsx")
mydata2
Y1 = mydata2$Y
A1 = mydata2$X1
B1 = mydata2$X2
Y1
A1
B1

cor1.matrix<-cor(cbind(Y1,A1,B1))
```

```
cor1.matrix
cov1.matrix<-cov(mydata2)
cov1.matrix

library(corrplot)
# visualize pairwise correlation
corrplot(cor(mydata[ , -1]), type = "upper")


install.packages("ggcorrplot")
library(ggcorrplot)
ggcorrplot(cor.matrix)
pairs(mydata)

regr1<-lm(Y1~A1+B1)
summary(regr1)
anova(regr1)

library(car)
#create vector of VIF values
vif_values1 <- vif(regr1)
vif_values1

Rm<-cov2cor(cov(cbind(A,B)))
phiindex=sqrt((sum(Rm^2)-2)/(2*(2-1)))
phiindex
```