

Γραμμική Παλινδρόμηση με την R

Δεδομένα των Sams και Shadman (1986)

Για τη μελέτη απόδοσης σε φυσικό αέριο κοιτασμάτων άνθρακα έγινε ένα πείραμα στο οποίο μετρήθηκε η απόδοση (y) σε σχέση με την περιεκτικότητα σε άνθρακα (x) 22 δειγμάτων.

```
KperC <- c(0.05, 0.05, 0.25, 0.25, 0.50, 0.50, 0.50,  
          1.25, 1.25, 1.25, 1.25, 1.25, 2.10, 2.10,  
          2.10, 2.10, 2.10, 2.10, 2.50, 2.50, 2.50, 2.50)  
CO.des <- c(0.05, 0.10, 0.25, 0.35, 0.75, 0.85, 0.95,  
          1.42, 1.75, 1.82, 1.95, 2.45, 3.05, 3.19,  
          3.25, 3.43, 3.50, 3.93, 3.75, 3.93, 3.99, 4.07)  
samsad <- data.frame(CO.des, KperC)
```

Συσχέτιση δύο μεταβλητών

Στις διάφορες έρευνες, μας ενδιαφέρει συχνά η ένταση της σχέσης μεταξύ ποσοτικών μεταβλητών (π.χ. μας ενδιαφέρει όταν αυξάνεται η μία, αν αυξάνεται ή ελαττώνεται η άλλη και με ποιο τρόπο) .

Η πιο απλή μορφή σχέσης στα μαθηματικά είναι η γραμμική σχέση ($y=a+bx$). Ο συντελεστής συσχέτισης **r του Pearson** μετρά τη γραμμική συσχέτιση δύο ποσοτικών μεταβλητών. Παίρνει τιμές στο [-1,1].

- Τιμές κοντά στο 1 δηλώνουν ισχυρή θετική γραμμική συσχέτιση.
- Τιμές κοντά στο -1 δηλώνουν ισχυρή αρνητική γραμμική συσχέτιση.
- Τιμές κοντά στο 0 δηλώνουν ανυπαρξία γραμμικής συσχέτισης.

`cor (KperC , CO . des)`

Η συνάρτηση **cor.test** κάνει έλεγχο σημαντικότητας της συσχέτισης. Για να είναι αξιόπιστα τα αποτελέσματα του ελέγχου αυτού, θα πρέπει οι μεταβλητές να ακολουθούν κανονική κατανομή.

H_0 : Ο συντ. συσχέτισης $R=0$
 H_1 : Ο συντ. συσχέτισης $R\neq 0$

`cor . test (KperC , CO . des)`

Αν δεν ακολουθούν κανονική κατανομή, καταλληλότερος είναι ο συντελεστής συσχέτισης του **Spearman**.

`cor . test (KperC , CO . des , method="spearman")`

3

Απλή Γραμμική Παλινδρόμηση Αντικείμενα lm

```
lm (formula ,  
    data="πλαίσιο δεδομένων" ,  
    weights="διάνυσμα" ,  
    subset="έκφραση" ,  
    na.action=na.fail ,  
    method="qr" , model=T ,  
    x=F , y=F , qr=F ,  
    contrasts=NULL , ...)
```

$y \sim x_1 + x_2 + \dots$
(Πρώτα η Y!!!)

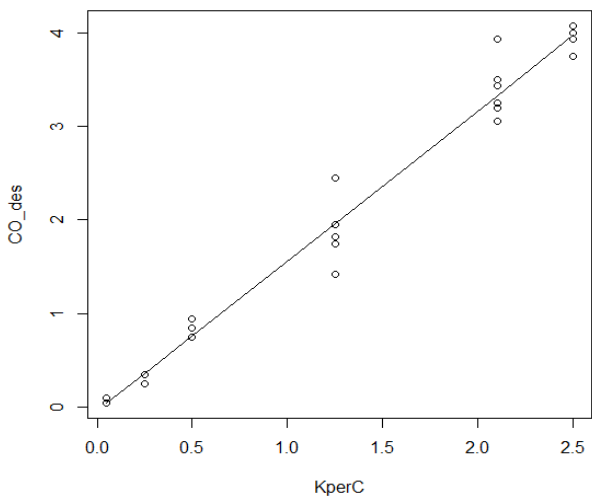
για σταθμισμένη
παλινδρόμηση

για περαιτέρω
μελέτη

Παλινδρόμηση με τα δεδομένα samsad

```
plot(KperC,CO.des, xlab="KperC",ylab="CO_des")
samsad.lm <- lm(CO.des ~ KperC)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(samsad.lm, las = 1)
par(opar)
anova(samsad.lm)
summary(samsad.lm)
```

Μουσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ



πίνακας ANOVA

συντελεστές
παλινδρόμησης,
προσδιορισμού,
τυπικά σφάλματα κλπ.

5

Τα βήματα της γραμμικής παλινδρόμησης αναλυτικά (1/2)

```
x <- KperC
y <- CO.des
X <- cbind(1,x)
XtX <- t(X) %*% X
XtXin <- solve(XtX)
XtXin
XtY <- t(X) %*% y
bhat <- XtXin %*% XtY
n <- length(x)
YtY <- sum(y^2)
sY1 <- (sum(y))^2/n
ssr <- t(bhat) %*% XtY - sY1
```

Σχηματισμός του πίνακα σχεδιασμού X
Υπολογισμός του πίνακα X'X

Υπολογισμός του πίνακα (X'X)⁻¹
Υπολογισμός του πίνακα X'Y
Εκτίμηση συντελεστών παλινδρόμησης
Το πλήθος παρατηρήσεων
Το άθροισμα τετραγώνων Y'Y
Το άθροισμα τετραγώνων (Y'1)²/n

Το άθροισμα τετραγώνων SSR

Μουσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

$$SSR = \hat{\beta}' X' Y - \frac{1}{n} (Y' \mathbf{1})^2$$

$$\hat{\beta} = (X' X)^{-1} X' Y$$

6

Τα βήματα της γραμμικής παλινδρόμησης αναλυτικά (2/2)

```
sst <- YtY - (sum(y))^2/n
c(ssr, sst, sst-ssr)
sse <- YtY - t(bhat) %*% XtY
```

Το άθροισμα τετραγώνων SST
Το άθροισμα τετραγώνων SSE

```
mse <- sse / (n-2)
ste <- sqrt(mse); c(mse, ste)
V.mat <- mse[1,1] * XtXin
```

Το άθροισμα τετραγώνων SSE απευθείας
Το άθροισμα τετραγώνων MSE (k=1=>n-k-1=n-2)
Τα MSE και s

```
x0 <- c(1, 2.05)
y0hat <- t(x0) %*% bhat
vary0 <- mse * t(x0) %*% XtXin %*% x0
vary0[1,1]
F.ratio <- ssr/mse; F.ratio[1,1]
qf(1-0.01, 1, 20)
```

Πίνακας διασπορών – συνδιασπορών
Διάνυσμα για πρόβλεψη
Πρόβλεψη
Var(Y0)
Παρατηρούμενο F
Κρίσιμο F για α=0.99

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

$$SST = \tilde{Y}'\tilde{Y} - \frac{1}{n}(\tilde{Y}'\tilde{1})^2$$

$$SSE = \tilde{Y}'\tilde{Y} - \hat{\beta}'X'\tilde{Y}$$

$$Var(\hat{\beta}) = \sigma^2 X'X^{-1}$$

$$Var(\hat{y}_0) = \sigma^2 \tilde{x}'_0 (X'X)^{-1} \tilde{x}_0$$

7

Επαναλαμβανόμενες μετρήσεις Καθαρά σφάλματα

Καθαρά σφάλματα

α/α	x	y	\bar{y}	SSE _i	β.ε
1	0.05	0.05	0.075	0.0012	1
2		0.10			
3	0.25	0.25	0.300	0.0050	1
4		0.35			
5		0.75			
6	0.50	0.85	0.850	0.0200	2
7		0.95			
8	1.25	1.42	1.878	0.5619	4
9		1.75			
10		1.82			
11		1.95			
12		2.45			
13	2.10	3.05	3.392	0.4805	5
14		3.19			
15		3.25			
16		3.43			
17		3.50			
18		3.93			
19	2.50	2.75	3.935	0.0555	3
20		3.93			
21		3.99			
22		4.07			
				1.1241	16

$$SSE_i = \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_i)^2 = (n_j - 1) \cdot s_i^2$$

$$\beta.ε. = n_j - 1$$

$$SS_e = \sum_{i=1}^r SSE_i = 1.1241$$

$$n_e = \sum_{i=1}^r (n_j - 1) = n - r = 22 - 6 = 16$$

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

8

Επαναλαμβανόμενες μετρήσεις Καθαρά σφάλματα

Μουσιτάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

```
ord <- order(KperC)
x <- KperC[ord]
y <- CO.des[ord]
lev <- levels(factor(x))
n <- length(lev)
lsd <- NULL
for(i in 1:n) lsd[i] <-
var(y[x==lev[i]]) * (length(x[x==lev[i]])-1) →  $SSE_i = (n_j - 1) s_i^2$ 
    Για κάθε διαφορετική τιμή της x το lsd περιέχει τα αθροίσματα
    τετραγώνων των αποκλίσεων από την αντίστοιχη μέση τιμή
ssp <- sum(lsd, na.rm=T)      Αθροίζουμε για να βρούμε τα καθαρά σφάλματα
ssf <- sse-ssp              Τα σφάλματα προσαρμογής
Msf <- ssf/(dff<-dfe-dfp) ; fr.ratio<-round(msf/msp,3)
ss<-round(c(ssr,sse,ssf,ssp,sst),5)
ms<-round(c(ssr/1,mse,msf,msp),5)
anov<-c(ss,1,dfe,dff,dfp,dfe+1,ms,"",F.ratio,"",Fr.ratio,"","")
dim(anov)<-c(5,4) ;
dimnames(anov)<-list(c("Παλινδρόμηση","Υπόλοιπα","Σφ. Προσαρμογής",
"Σφ. Καθαρά","Σύνολο"),c("Αθροίσματα","β.ε.",
"Μέσα Τετράγωνα","Λόγοι F"))
```

9

Ο πίνακας ANOVA για επαναλαμβανόμενες μετρήσεις

```
print.default(anov, quote=F)
```

Μουσιτάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

	Αθροίσματα	β.ε.	Μέσα Τετράγωνα	Λόγοι F
Παλινδρόμηση	42.69209	1	42.69209	697.614
Υπόλοιπα	1.22395	20	0.0612	
Σφ. Προσαρμογής	0.09983	4	0.02496	0.355
Σφ. Καθαρά	1.12411	16	0.07026	
Σύνολο	43.91604	21		

Συγκρίνετε με αυτό που βρήκαμε στο μάθημα

Πηγή	Αθρ. Τετραγώνων	β.ε.	Μέσα τετράγ.	F
Παλινδρόμηση	42.69209	1	42.69209	697.6
Υπόλοιπα	1.22395	20	0.06120	
Σφάλμ. Προσαρμ. Καθαρά Σφάλματα	0.099837	4	0.02496	0.355
Σύνολο	43.91604			

10

Συναρτήσεις που εφαρμόζονται σε ένα lm αντικείμενο

Οι συναρτήσεις που δέχονται ως όρισμα ένα lm αντικείμενο πχ
`dat.lm <- lm(y~x1+x2+..., ...)` είναι:

print(dat.lm) :

τυπώνεται μία απλή εκτύπωση

summary(dat.lm) :

τυπώνεται εκτενέστερη κατάσταση των παραμέτρων της παλινδρόμησης.

coef(dat.lm) :

τυπώνονται οι εκτιμώμενοι συντελεστές παλινδρόμησης.

resid(dat.lm) :

τυπώνονται τα υπόλοιπα.

fitted(dat.lm) :

υπολογίζονται οι εκτιμήσεις (προβλέψεις) των αρχικών δεδομένων.

deviance(dat.lm) :

υπολογίζεται το άθροισμα τετραγώνων των σφαλμάτων

anova(dat.lm) :

δίνεται ο πίνακας ανάλυσης της διασποράς.

predict(dat.lm, newdata) :

υπολογίζονται προβλέψεις για νέα δεδομένα

plot(dat.lm) :

τυπώνονται διαγνωστικά γραφήματα

11

Παράδειγμα με δεδομένα της βιβλιοθήκης MASS Σύγκριση δυο ευθειών παλινδρόμησης

```
library(mass)
attach(cats)
catsF <- lm(Hwt~Bwt,data=cats,
            subset=Sex=="F")
catsM <- update(catsF, subset = Sex == "M")

nF <- catsF$df.resid; nM <- catsM$df.resid
vF <- deviance(catsF)/nF
vM <- deviance(catsM)/nM
c(Male=vM,Female=vF, F=(f<-vM/vF), crit=qf(1-
  0.05,nM,nF), sig= 1-pf(f,nM,nF))
```

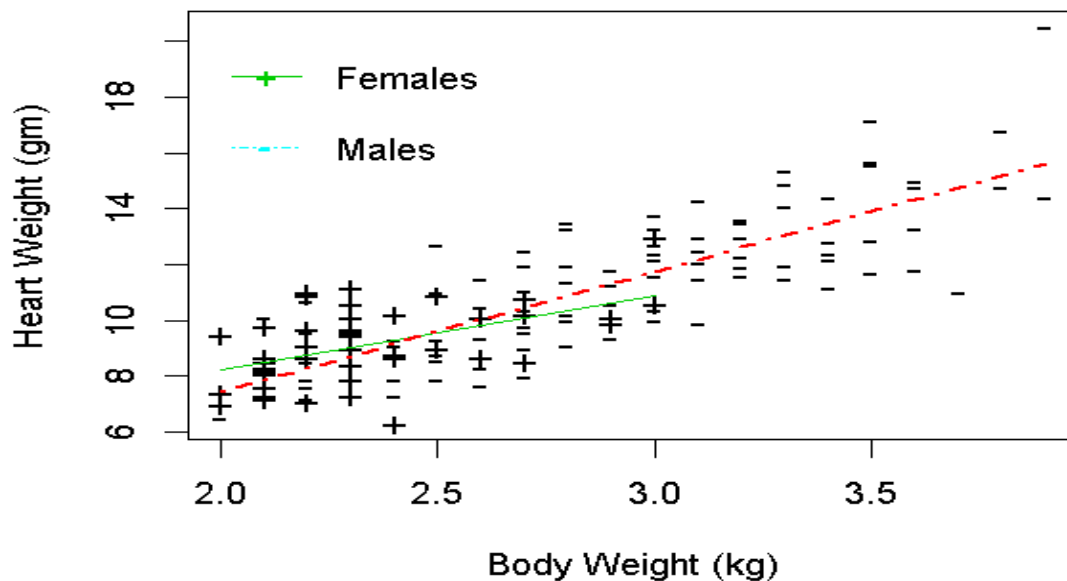
Male	Female	F	crit	sig
2.423778	1.350839	1.794276	1.55767	0.01552139

12

Πολλαπλές γραμμές παλινδρόμησης

```
plot(Bwt, Hwt, xlab="Body Weight (kg)",
      ylab="Heart Weight (gm)", type="n")
text(Bwt, Hwt, c("+", "-")[Sex])
legend(2.0, 20, c("Females", " ", "Males"),
      pch="+ -", lty = c(1, -1, 4),
      col=c(3, -1, 5), bty="n")
lines(Bwt[Sex == "F"], fitted(catsF),
      lty=1, col=3)
lines(Bwt[Sex == "M"], fitted(catsM),
      lty=4, col=5, lwd=2)
```

Συγκριτικό γράφημα των δύο μοντέλων



Πρόβλεψη με την predict

```
yhat <- fitted(samsad.lm)
```

1.223946

```
eres <- CO.des-yhat ; sum(eres^2)
```

```
samsad.newx <- data.frame(KperC=c(1,1.5,2,3))
```

```
samsad.newx
```

```
predict(samsad.lm, samsad.newx)
```

```
predict(samsad.lm, samsad.newx,se.fit=T)
```

δ.ε. για τη μέση τιμή της πρόβλεψης

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

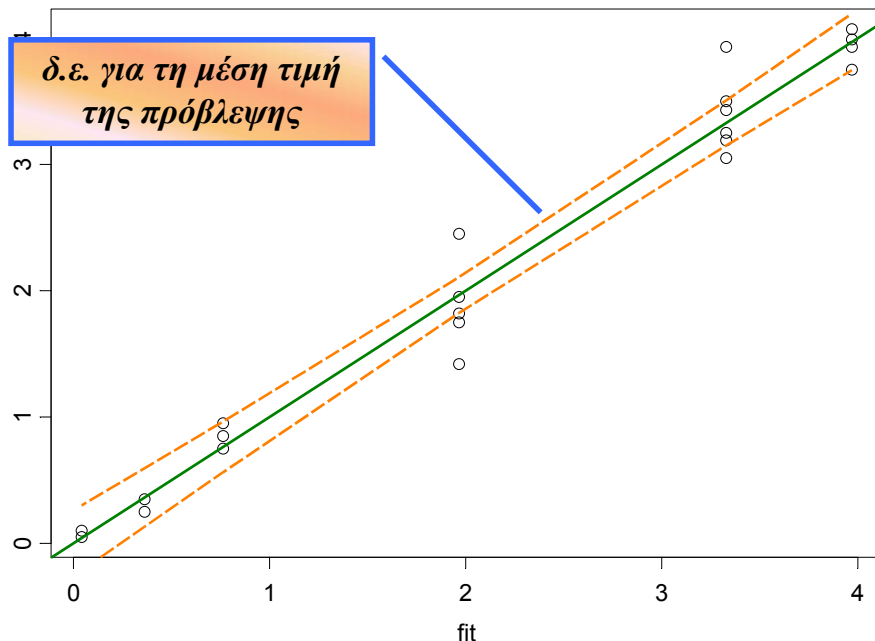
```
$fit:
      1      2      3      4
1.565089 2.366656 3.168222 4.771355
$se.fit:
      1      2      3      4
0.058236 0.053044 0.063860 0.110148
$residual.scale: [1] 0.2473809
$df: [1] 20
```

```
$upper:
      1      2      3      4
1.73079 2.517584 3.349923 5.084764
$fit:
      1      2      3      4
1.565089 2.366656 3.168222 4.771355
$lower:
      1      2      3      4
1.399388 2.215727 2.986521 4.457946
```

15

Η συνάρτηση confint.lm

Διάγραμμα διασποράς με ζώνη εμπιστοσύνης

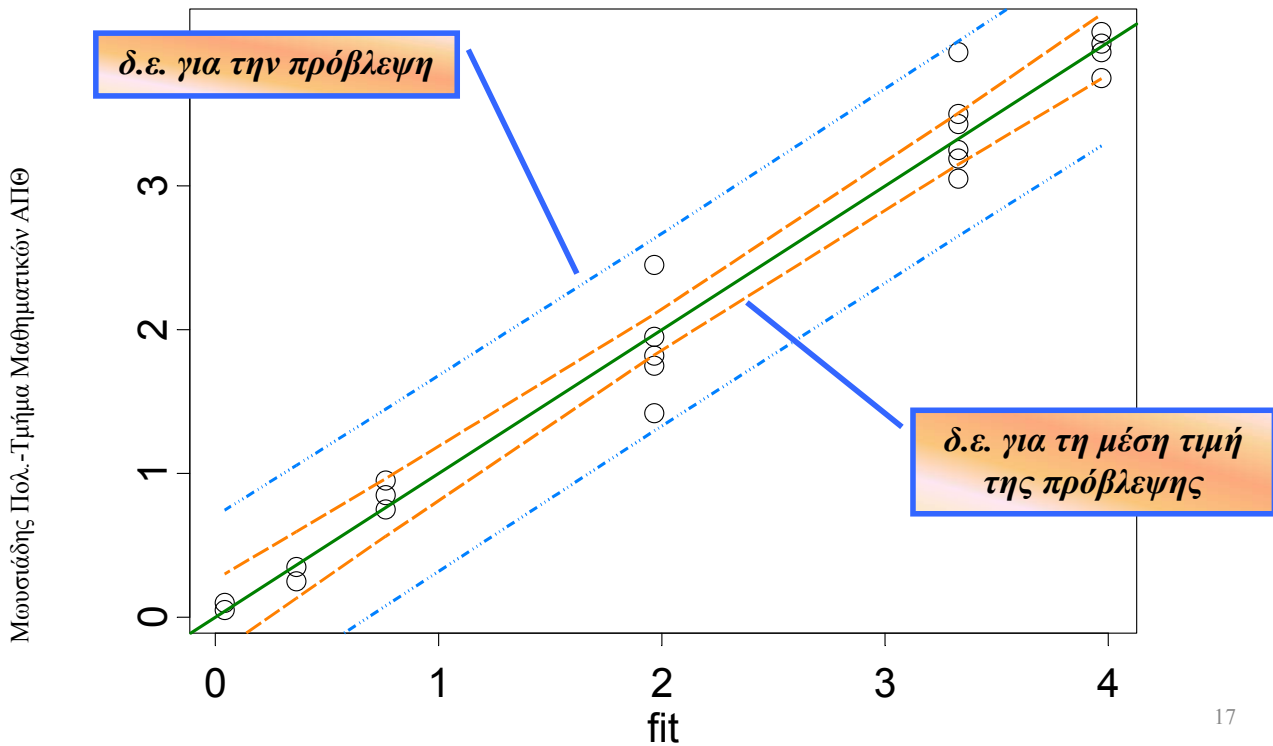


Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Από το σχήμα συμπεραίνουμε ότι το μοντέλο είναι αρκετά καλό, αφού έξω από τη ζώνη εμπιστοσύνης για τη μέση πρόβλεψη υπάρχουν λίγα σημεία.

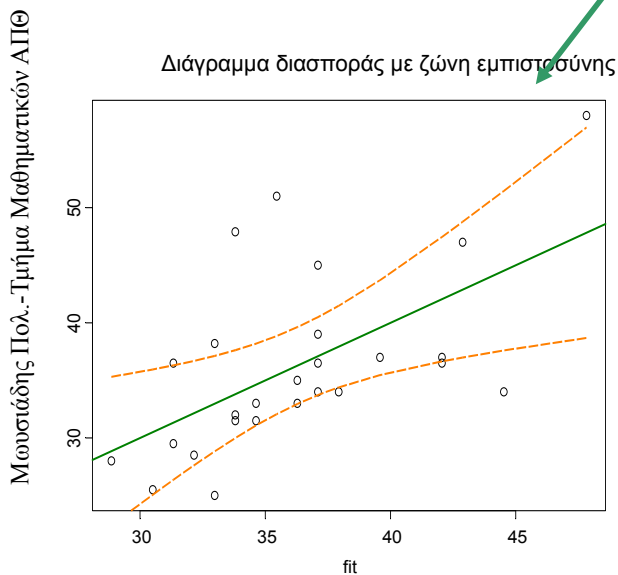
Η συνάρτηση confint.adj.lm

Διάγραμμα διασποράς με ζώνες εμπιστοσύνης - πρόβλεψης

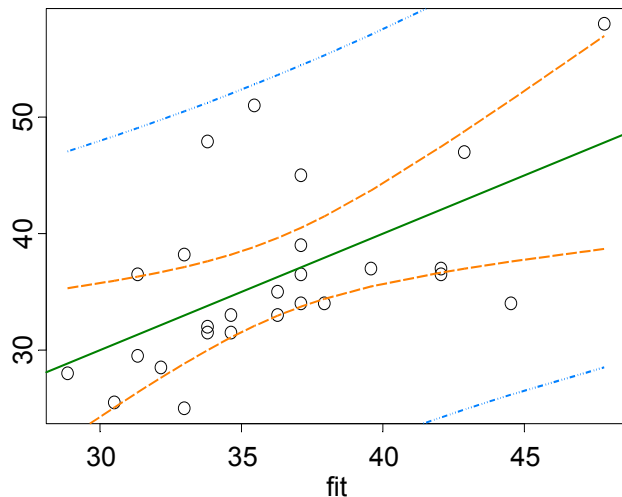


Άλλο παράδειγμα

```
#library(statlab)
#attach(talenta)
nm <- nomos==40
talenta.lm <- lm(
  baros[nm] ~ ypsos[nm])
confint.lm(talenta.lm)
confint.adj.lm(talenta.lm)
```



Διάγραμμα διασποράς με ζώνες εμπιστοσύνης - πρόβλεψης



Πολλαπλή Γραμμική Παλινδρόμηση

Δεδομένα του Hald

```
x1<-c(7,1,11,11,7,11,3,1,2,21,1,11,10)
x2<-c(26,29,56,31,52,55,71,31,54,47,40,66,68)
x3<-c(6,15,8,8,6,9,17,22,18,4,23,9,8)
x4<-c(60,52,20,47,33,22,6,44,22,26,34,12,12)
y<-c(78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5,
     93.1, 115.9,83.8,113.3,109.4)
hald.y<-y
hald.x<-cbind(x1,x2,x3,x4)
hald.df <- data.frame(hald.y,hald.x)
```

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

```
hald.lm <- lm(hald.df)   ή
hald.lm <- lm(hald.y~., data=hald.df)  ή
hald.lm <- lm(y ~ x1+x2+x3+x4)
```

19

Μοντέλα στο S-Plus

```
y ~ x1+x2+x3 ή y ~ 1+x1+x2+x3
```

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

```
y ~ -1+x1+x2+x3 χωρίς σταθερά
```

```
y ~ . Όλες οι μεταβλητές του data frame (προσοχή στην εξαρτημένη)
```

```
y ~ I(x1+x3)+x2  $y = \beta_0 + \beta_1(x_1 + x_3) + \beta_2 x_2 + \varepsilon$ 
```

```
y ~ I(x/z)+x+x*w  $y = \beta_0 + \beta_1(x/z) + \beta_2 x + \beta_3 xw + \varepsilon$ 
```

```
y ~ x+z+x*z
```

$z = 0$ ή 1
βωβή μεταβλητή

$$y = \begin{cases} \beta_{10} + \beta_{11}x + \varepsilon \\ \beta_{20} + \beta_{21}x + \varepsilon \end{cases}$$

```
y ~ x+x2+(1+x+x2)*z, x2=x^2
```

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

20

anova και summary

```
hald.lm <- lm(y ~ x1+x2+x3+x4)
anova(hald.lm)
Analysis of Variance Table
Response: hald.y      Terms added sequentially (first to last)

```

	Df	Sum of Sq	Mean Sq	F Value	Pr(>F)
x1	1	1450.076	1450.076	242.3679	0.0000003
x2	1	1207.782	1207.782	201.8705	0.0000006
x3	1	9.794	9.794	1.6370	0.2366003
x4	1	0.247	0.247	0.0413	0.8440715
Residuals	8	47.864	5.983		

```
summary(hald.lm)
.....
Coefficients:

```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	62.4054	70.0710	0.8906	0.3991
x1	1.5511	0.7448	2.0827	0.0708
x2	0.5102	0.7238	0.7049	0.5009
x3	0.1019	0.7547	0.1350	0.8959
x4	-0.1441	0.7091	-0.2032	0.8441

```
Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-Squared: 0.9824
F-statistic: 111.5 on 4 and 8 degrees of freedom, the p-value
is 4.756e-007
Correlation of Coefficients:
.....
```

21

Η συνάρτηση r.squared

Από τον πίνακα ANOVA που δίνει το S-Plus έχουμε το SSR κάθε μεταβλητής αλλά όχι το γνωστό συνολικό SSR.

```
r.squared <- function(x) {
  # x is lm type object (x<-lm(y~x1+...))
  y <- anova(x)[, 2]
  z<-anova(x)[, 1]
  sse <- y[length(y)]
  sst <- sum(y)
  rsq <- (sst - sse)/sst
  frat<-((sst-sse)/sum(z[-length(y)]))/
  (sse/z[length(z)])
  output <- c(sst - sse, sse, sst, rsq, frat)
  names(output) <- c("SSR ", " SSE ", " SST ",
    "Coefficient R^2", "F Ratio")
  output
}
```

```
r.squared(hald.lm)
```

SSR	SSE	SST	Coefficient R^2	F Ratio
2667.899	47.86364	2715.763	0.9823756	111.479

22

Σύγκριση πλήρους - περιορισμένου μοντέλου

$$F = \frac{(SSR - SSR_{\Pi\epsilon}) / k_1}{SSE / (n - k - 1)} = \frac{(SSE_{\Pi\epsilon} - SSE) / k_1}{SSE / (n - k - 1)}$$

$$F \sim F_{k_1, n-k-1}$$

Συγκρίνουμε το πλήρες μοντέλο με το περιορισμένο μοντέλο που περιέχει σαν ανεξάρτητες μεταβλητές μόνο την x3 και την x4.

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

```
hald.lm<-lm(hald.df)
hald.restricted<-lm(y~x3+x4)
n<-length(y)
k<-4
k1<-2 ##πλήθος μεταβλητών που φεύγουν
sse<-anova(hald.lm)[k+1,2]
sse.restricted<-anova(hald.restricted)[k-k1+1,2]
f.ratio<-(sse.restricted-sse)/k1/(sse/(n-k-1)); f.ratio
p.value<-1-pf(f.ratio,k1,n-k-1); p.value
#P(F>f.ratio)=1-P(F<=f.ratio)
```

k₁: πλήθος μεταβλητών που φεύγουν

μικρότερο του 0.05 άρα απορρίπτεται η υπόθεση της ισοδυναμίας των μοντέλων

23

Μέθοδοι forward και backward

`drop1(hald.lm)` ← Επιλογή μεταβλητής για διαγραφή

`hald0.lm <- lm(y~1)` ← Μοντέλο θέσης

`add1(hald.lm0, ~. +x1+x2+x3+x4)`

← Επιλογή μεταβλητής για προσθήκη

`hald.lm2 <- lm(y~x1+x3)`

προσθήκη ή

`update(hald.lm2, ~.+x2)` ←

διαγραφή

`update(hald.lm2, ~.-x3)`

μεταβλητής

Επιλογή του καλύτερου μοντέλου

τυπώνει τα

`hald.lm0 <- lm(y~1)`

ενδιάμεσα

`step(hald.lm0, ~x1+x2+x3+x4, trace=T)`

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

`direction="both"`
ή `"back"` ή `"for"`

η επιλογή στηρίζεται στο κριτήριο AIC (βασίζεται στο Cp), που όσο μικρότερο τόσο καλύτερη προσαρμογή

24

Επιλογή του καλύτερου μοντέλου

```
hald.lm0 <- lm(y~1)
step(hald.lm0, ~x1+x2+x3+x4,trace=T, direction = "forward")
```

```
step(hald.lm0, ~x1+x2+x3+x4,trace=T,direction = "both")
```

```
hald.lm <-lm(y~x1+x2+x3+x4)
step(hald.lm, ~x1+x2+x3+x4,trace=T,direction = "backward")
```

```
step(hald.lm, ~x1+x2+x3+x4,trace=T,direction = "both")
```

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

25

Επιλογή με το κριτήριο R^2

```
library(leaps)
lp<-leaps(hald.x,hald.y,nbest=2,method="r2")
#για κάθε κλάση βρίσκει τα καλύτερα nbest μοντέλα
# και για κάθε μοντέλο βρίσκει το  $R^2$  (ή το cp, κλπ)
k<-dim(hald.x)[2];a<-0.05
r0<-(1-(1-lp$r2[length(lp$r2)])*(1+k * qf(1-a,k,n-k-1)/(n-k-1)))*100

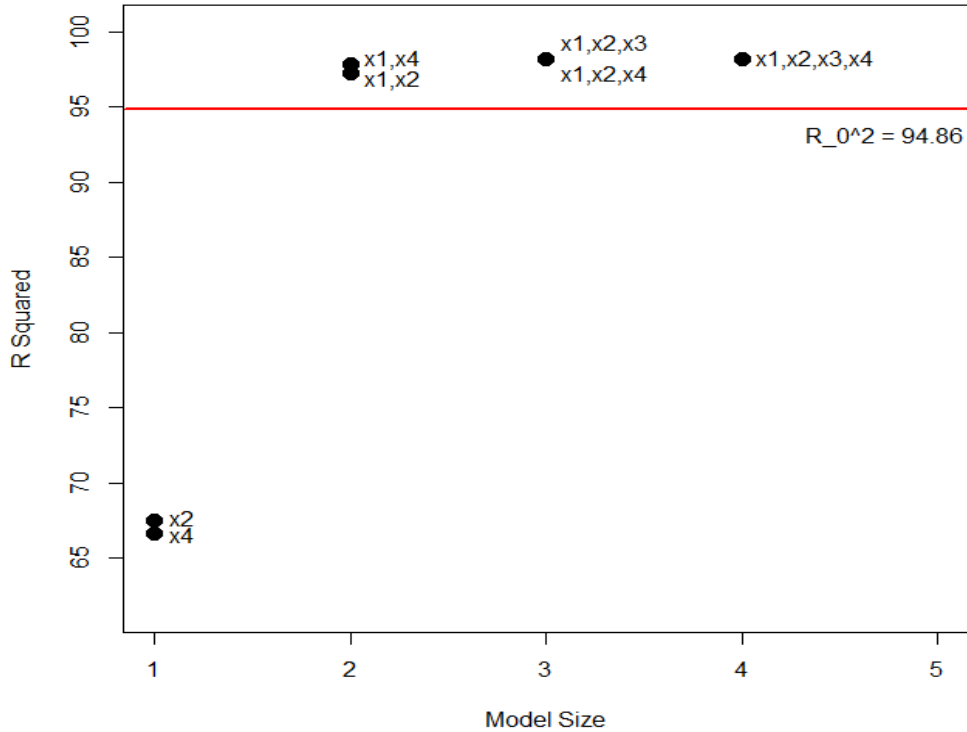
plot(lp$size-1,lp$r2*100,xlab="Model Size",ylab="R Squared",axes=F,
     xlim=c(1,k+1),ylim=c(min(lp$r2*100,r0)-5,max(lp$r2*100)+2),pch=16,cex=1.5)
m=lp$which
lab=NULL
for( i in 1:nrow(m) ){
  mod<-paste("x",colnames(m)[m[i,]],sep="")
  lab[i]<-paste(as.factor(mod),sep=" ",collapse="")
}
text(lp$size[c(1,3,5)]-1,100*lp$r2[c(1,3,5)]-1,lab[c(1,3,5)],pos=4)
text(lp$size[c(2,4,6)]-1,100*lp$r2[c(2,4,6)]+1,lab[c(2,4,6)],pos=4)
text(lp$size[7]-1,100*lp$r2[7],lab[7],pos=4)
axis(1,at=1:(k+1),labels=0:k+1)
axis(2,at=pretty(c(lp$r2*100,r0)))
title(main="Graph for the criterion R squared")
box()
lines(c(0.85,k+1.15),c(r0,r0),col=2,lwd=2)
legend(k+.1,r0,paste("R_0^2 =",as.character(round(r0,2))),bty="n")
```

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

26

Γράφημα με το κριτήριο R^2

Graph for the criterion R squared



Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Η συνάρτηση leaps και το C_p

```
> leaps(hald.x, hald.y, nbest=3)
```

```
$label
[1] "(Intercept)" "1" "2" "3" "4"

$size
[1] 2 2 2 3 3 3 4 4 4 5

$Cp
[1] 138.730833 142.486407 202.548769
[4] 2.678242 5.495851 22.373112
[7] 3.018233 3.041280 3.496824
[10] 5.000000

$which
      1  2  3  4
1 FALSE FALSE FALSE TRUE
1 FALSE TRUE FALSE FALSE
1 TRUE FALSE FALSE FALSE
2 TRUE TRUE FALSE FALSE
2 TRUE FALSE FALSE TRUE
2 FALSE FALSE TRUE TRUE
3 TRUE TRUE FALSE TRUE
3 TRUE TRUE TRUE FALSE
3 TRUE FALSE TRUE TRUE
4 TRUE TRUE TRUE TRUE
```

Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Μπορούμε να εργαστούμε με τη συνάρτηση αυτή για την επιλογή του καλύτερου μοντέλου με το κριτήριο C_p του Mallows

Επιλογή με το κριτήριο C_p του Malows

```

lp3=leaps(hald.x,hald.y,nbest=3)
# με το nbest=3 θα συμπεριλάβει τα 3 καλύτερα μοντέλα από κάθε
# περίπτωση, δηλ 3 καλύτερα με 1 μεταβλητή (με βάση το κριτήριο Cp),
# 3 καλύτερα με 2 μεταβλητές κ.τ.λ.
# στο size μετρά και τη σταθερά κ όχι μόνο το πλήθος των μεταβλητών
size <- lp3$size ; cp <- lp3$Cp
m=lp3$which
lab=NULL
for( i in 1:nrow(m) ){
  mod<-paste("x",colnames(m)[m[i,]],sep="")
  lab[i]<-paste(as.factor(mod),sep="," ,collapse =",")
}
lab
# [1] "x4"           "x2"           "x1"           "x1,x2"        "x1,x4"
# [6] "x3,x4"        "x1,x2,x4"    "x1,x2,x3"    "x1,x3,x4"    "x1,x2,x3,x4"

par(mai=c(1.2,1.2,.8,.5))#η mai=c(κάτω,αριστερά,πάνω,δεξιά) δίνει τα
περιθώρια στο γράφημα
plot(size, cp ,xlab="p", ylab="Cp of Malows",
xlim=c(0,6),ylim=c(0,6),pch=16,cex=1)
lines(c(0,5.5),c(0,5.5),lwd=2,type="l",col=2)
text(size[-c(7,8,9)],cp[-c(7,8,9)], lab[-c(7,8,9)],pos=1)
text(size[7],cp[7]+0.1, lab[7],pos=1)
text(size[8],cp[8]-0.1, lab[8],pos=3)
text(size[9],cp[9]-0.1, lab[9],pos=3)
title(main="Graph for the criterion Cp of Malows")

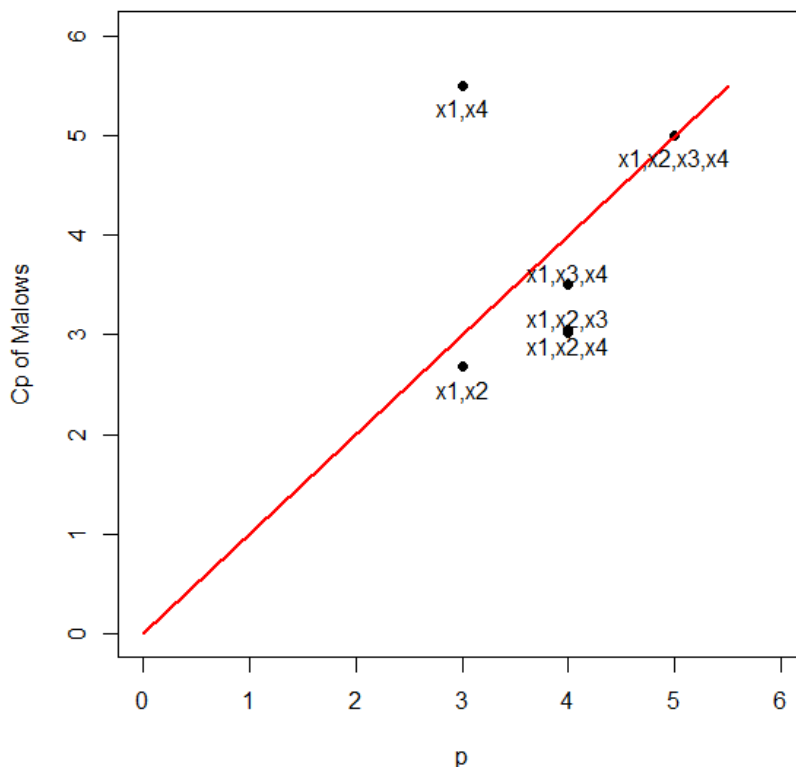
```

Μωυσσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

29

Γράφημα με το κριτήριο C_p του Malows

Graph for the criterion C_p of Malows



Μωυσσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

30

Παλινδρόμηση με βωβές μεταβλητές

Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. The object of the exercise was to assess the effect of the insulation on gas consumption.

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Insul

A factor, before or after insulation.

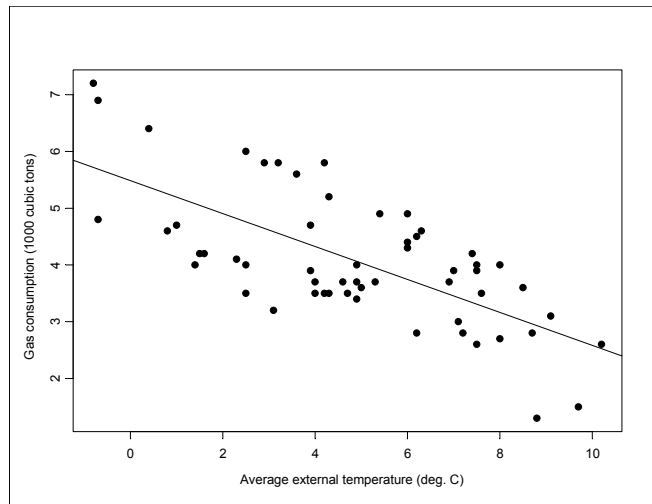
Temp

The weekly average outside temperature in degrees Celsius.

Gas

The weekly gas consumption in 1000s of cubic feet.

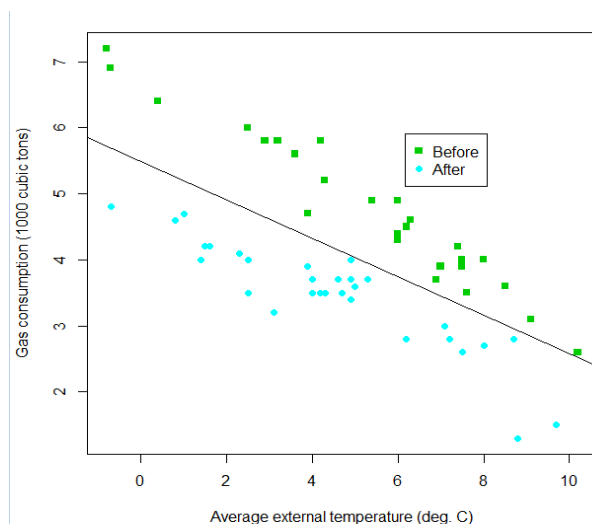
```
plot(whiteside$Temp,whiteside$Gas,xlab="Average external temperature (deg. C)", ylab="Gas consumption (1000 cubic tons)")  
abline(lm(Gas ~ Temp, data = whiteside))
```



Το γράφημα, λαμβάνοντας υπόψη τον παράγοντα

```
attach(whiteside)  
plot(Temp,Gas,xlab="Average external temperature (deg. C)", ylab="Gas consumption (1000 cubic tons)",type="n")  
points(Temp[Insul=="Before"],Gas[Insul=="Before"],pch=15,col=3)  
points(Temp[Insul=="After"],Gas[Insul=="After"],pch=16,col=5)  
abline(lm(Gas ~ Temp, data = whiteside))  
legend(locator(1),legend=c("Before", "After"),  
col=c(3,5),pch=15:16)
```

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

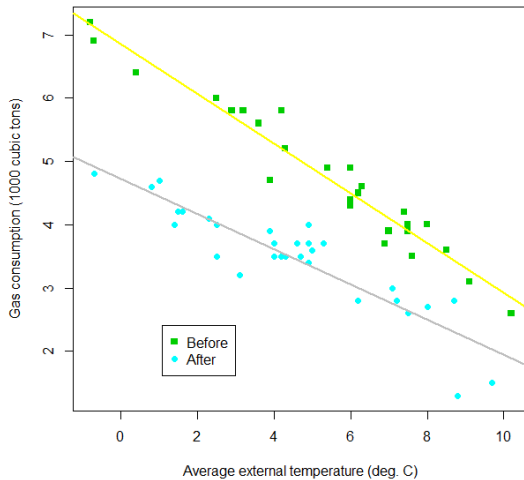


Υπάρχει εμφανής διαφορά μεταξύ **Before** και **After**.

Το γράφημα, με ξεχωριστές ευθείες παλινδρόμησης

```
plot(Temp, Gas, xlab="Average external temperature (deg. C)",
      ylab="Gas consumption (1000 cubic tons)", type="n")
points(Temp[Insul=="Before"], Gas[Insul=="Before"], pch=15, col=3)
abline(lm(Gas ~ Temp, data = whiteside,
          subset=Insul=="Before"), lwd=2, col=7)
points(Temp[Insul=="After"], Gas[Insul=="After"], pch=16, col=5)
abline(lm(Gas ~ Temp, data = whiteside,
          subset=Insul=="After"), lwd=2, col=8)
legend(locator(1), legend=c("Before", "After"),
      col=c(3, 5), pch=15:16)
```

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ



Θα εξετάσουμε αν οι δύο αυτές ευθείες ταυτίζονται, δηλαδή αν ο παράγοντας Insul επηρεάζει την σχέση των Temp και Gas.

33

Χρήση βωβών μεταβλητών

Μετασχηματίζουμε την Insul σε βωβή μεταβλητή (με τιμές 0 και 1).

```
Insul<-as.numeric(factor(Insul))-1
```

Θεωρούμε το πλήρες μοντέλο

$$\text{Gas} = b_0 + b_1 \text{Temp} + b_2 \text{Insul} + b_3 \text{Temp} * \text{Insul}$$

Για Insul = 0 (Before) παίρνουμε το μοντέλο

$$\text{Gas} = b_0 + b_1 \text{Temp}$$

Για Insul = 1 (After) παίρνουμε το μοντέλο

$$\text{Gas} = (b_0 + b_2) + (b_1 + b_3) \text{Temp}$$

Για να ελέγξουμε αν τα δύο αυτά μοντέλα είναι ισοδύναμα, θα ελέγξουμε τη μηδενική υπόθεση $H_0: b_0 = b_0 + b_2$ και $b_1 = b_1 + b_3$, δηλαδή την υπόθεση $H_0: b_2 = b_3 = 0$

Θα ελέγξουμε δηλαδή την ισοδυναμία του πλήρους μοντέλου

$$\text{Gas} = b_0 + b_1 \text{Temp} + b_2 \text{Insul} + b_3 \text{Temp} * \text{Insul}$$

και του περιορισμένου

$$\text{Gas} = b_0 + b_1 \text{Temp}$$

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

34

Οι εντολές για τη σύγκριση των μοντέλων στην R

```
> lm.ful<-lm(Gas ~ Temp+Insul+InsTemp) # Το πλήρες μοντέλο
> bt <- coef(lm.ful);bt
  (Intercept)      Temp      Insul  InsTemp
    6.853828 -0.3932388 -2.129978 0.1153039
> c(bt[1] + bt[3], bt[2] + bt[4])
  (Intercept)      Temp
    4.72385 -0.277935
> lm.restr<-lm(Gas ~ Temp) # Το περιορισμένο μοντέλο

d1_deviance(lm.restr) # το SSE του περιορισμένου
# anova(lm.restr)[2,2]
d2_deviance(lm.ful);d2 # το SSE του πλήρους

f.ratio_((d1-d2)/2)/(d2/lm.ful$df.resid);f.ratio
qf(1-0.05,2,52) # αν f.ratio μεγαλύτερο από αυτήν την τιμή
απορρίπτουμε
# την υπόθεση ισοδυναμίας πλήρους-περιορισμένου
1-pf(f.ratio,2,52) # το αντίστοιχο p-value
```

Το p-value είναι πολύ μικρό, επομένως απορρίπτεται η υπόθεση ισοδυναμίας των δύο μοντέλων, δηλαδή η επίδραση του παράγοντα Insul είναι σημαντική.

Έλεγχος παραλληλίας των δύο ευθειών παλινδρόμησης

Θεωρούμε το πλήρες μοντέλο

$$\text{Gas} = b_0 + b_1 \text{Temp} + b_2 \text{Insul} + b_3 \text{Temp} * \text{Insul}$$

$$\text{Για Insul} = 0 \quad \text{Gas} = b_0 + b_1 \text{Temp}$$

$$\text{Για Insul} = 1 \quad \text{Gas} = (b_0 + b_2) + (b_1 + b_3) \text{Temp}$$

Για να ελέγξουμε αν οι δύο ευθείες παλινδρόμησης είναι παράλληλες, ελέγχουμε τη μηδενική υπόθεση $H_0: b_3 = 0$

Αυτός ο έλεγχος μπορεί να γίνει όπως και προηγουμένως, συγκρίνοντας το πλήρες μοντέλο με το περιορισμένο

$$\text{Gas} = b_0 + b_1 \text{Temp} + b_2 \text{Insul}$$

Εναλλακτικά `summary(lm.ful)`

```
Coefficients:
      Value Std. Error t value Pr(>|t|)
(Intercept)  6.8538   0.1360   50.4091  0.0000
      Temp  -0.3932   0.0225  -17.4874  0.0000
      Insul  -2.1300   0.1801  -11.8272  0.0000
      InsTemp 0.1153   0.0321   3.5907  0.0007
```

Απορρίπτεται η υπόθεση $\beta_3 = 0$

ANOVA

με την R

37

Ανάλυση Διασποράς - Εισαγωγή

Η Ανάλυση Διασποράς (Analysis of Variance – ANOVA) εξετάζει αν η μέση τιμή μιας ποσοτικής μεταβλητής είναι ίδια για τις διάφορες ομάδες (περισσότερες από 2) στις οποίες χωρίζεται ο πληθυσμός με βάση μία ή περισσότερες ποιοτικές μεταβλητές (παράγοντες).

Η ANOVA αποτελεί γενίκευση του t-test. Χρησιμοποιείται γιατί η πολλαπλή χρήση του t-test οδηγεί σε αύξηση του σφάλματος τύπου II. Επίσης, τα ίδια αποτελέσματα που παίρνουμε με την ANOVA μπορούμε να τα πάρουμε και εκτελώντας παλινδρόμηση με ποιοτικές (βωβές) μεταβλητές.

Παράδειγμα: Έχουμε ένα δείγμα 50 ανθρώπων. Τους χωρίζουμε σε τρεις ομάδες (όχι κατ' ανάγκη ισοπληθείς) και σε κάθε ομάδα δίνουμε διαφορετική δίαιτα (**ποιοτική μεταβλητή**). Ενδιαφερόμαστε για το βάρος (**ποσοτική μεταβλητή**) που χάνουν σε 2 εβδομάδες ακολουθώντας τη δίαιτα που τους δόθηκε.

ANOVA στην R

Τρεις ομάδες των 8 ατόμων διδάχτηκαν την ίδια ύλη με 3 διαφορετικές μεθόδους. Στη συνέχεια υποβλήθηκαν σε κοινή εξέταση και μετρήσαμε τις επιδόσεις τους.

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

method1	method2	method3
3	4	6
5	4	7
2	3	8
4	8	6
8	7	7
4	4	9
3	2	10
9	5	9

```
m1<-c(3,5,2,4,8,4,3,9)
m2<-c(4,4,3,8,7,4,2,5)
m3<-c(6,7,8,6,7,9,10,9)
grades<-c(m1,m2,m3)
```

Ορίζουμε έναν παράγοντα (με στάθμες 1, 2, 3) για την μέθοδο που χρησιμοποιήθηκε

```
methodos<-rep(1:3,each=8)
```

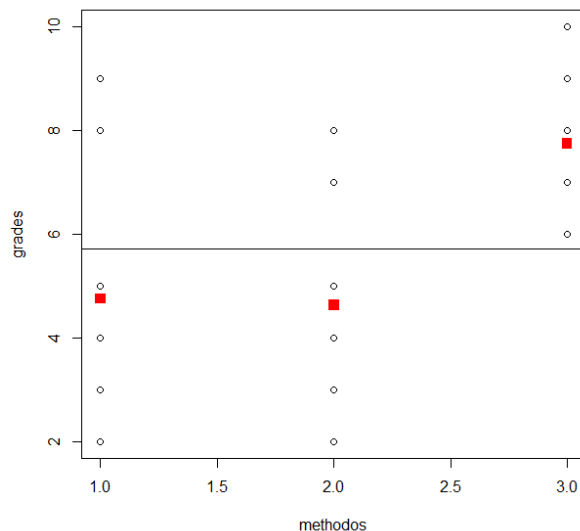
```
methodos<-factor(methodos,labels=paste("method",1:3),
levels=1:3)
```

39

Γράφημα

```
plot(methodos,grades,xlim=c(1,3),ylim=range(grades))
abline(mean(grades),0)
points(1:3,
c(mean(m1),mean(m2),mean(m3)),pch=15,col=2,cex=1.5)
```

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ



Συμβολισμοί

Κάθε παρατήρηση συμβολίζεται με Y_{ij} , όπου ο πρώτος δείκτης αναφέρεται στην ομάδα (στάθμη του παράγοντα) και ο δεύτερος στον αριθμό του ατόμου μέσα στην ομάδα.

πχ η Y_{23} παριστάνει την τιμή της (ποσοτικής) μεταβλητής για το τρίτο άτομο της δεύτερης ομάδας.

Με \bar{Y}_i συμβολίζουμε τη μέση τιμή της μεταβλητής

στην i ομάδα, ενώ με $\bar{Y}_{..}$ τη μέση τιμή για το σύνολο των παρατηρήσεων.

Υπολογίζουμε το γενικό μέσο όρο ($\bar{Y}_{..}$) της μεταβλητής grades και τους μέσους όρους για κάθε στάθμη του παράγοντα methodos ($\bar{Y}_{1..}$, $\bar{Y}_{2..}$, $\bar{Y}_{3..}$)

```
y..m <- mean(grades) ; y..m
```

```
means <- tapply(grades, methodos, mean) ; means
```

41

Μουσειάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

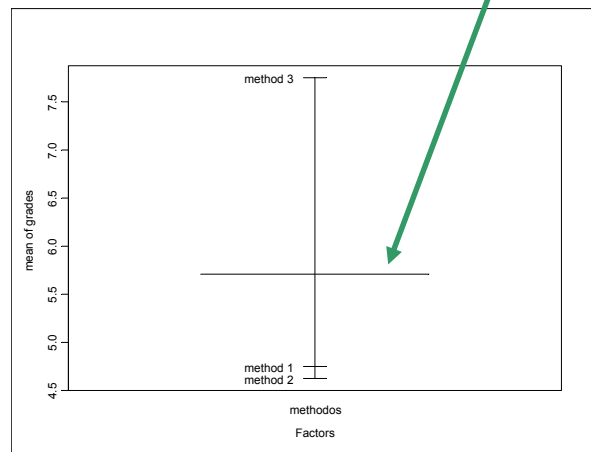
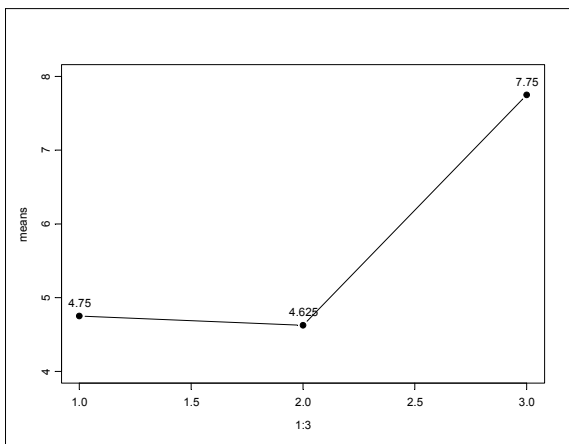
Διαγνωστικά γραφήματα

Κατασκευάζουμε ένα απλό γράφημα των επιμέρους μέσων όρων.

```
plot(1:3, means, type='b',  
     ylim=c(4, 8))  
text(1:3, means+.2, means)
```

Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε την εντολή `plot.design(grad.df)`.

Στο γράφημα τώρα φαίνεται και ο γενικός μέσος όρος.

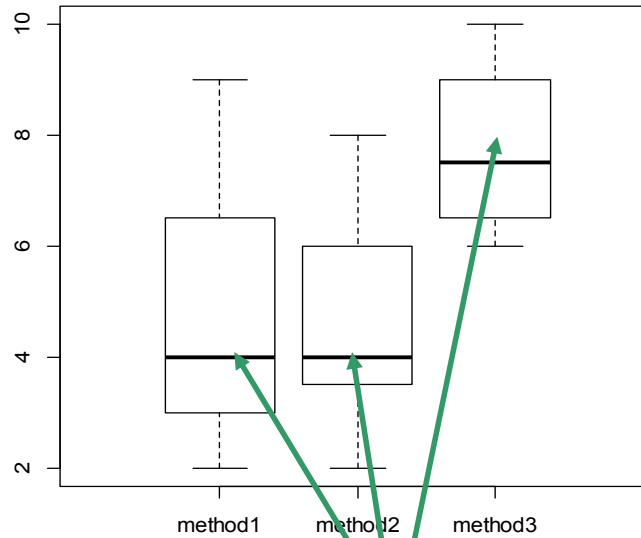


Μουσειάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

2

Διαγνωστικά γραφήματα (2)

```
boxplot(split(grades,methodos)) ;h
plot(methodos,grades)
```



διάμεσοι - όχι μέσες τιμές

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Ανάλυση διασποράς αναλυτικά

$$SSA = \sum_i r_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \quad \leftarrow \text{μεταξύ των ομάδων}$$

$$SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 \quad \leftarrow \text{ανάμεσα στις ομάδες}$$

$$SST = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$$

«ανάμεσα» = «μέσα»

```
means<-tapply(grades,methodos,mean)
y1.m<-means[1];y1.m
y2.m<-means[2];y2.m
y3.m<-means[3];y3.m
y..m<-mean(grades);y..m
```

```
Sizes<-table(methodos);sizes
```

```
sst<-sum((grades-y..m)^2);sst
ssa<-sum(sizes*(means-y..m)^2);ssa
vars<-tapply(grades,methodos,var);
```

```
sse<-sum(vars*(sizes-1));sse
```

```
# ή απλά
sse<-sst-ssa;sse
```

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Κατασκευή του πίνακα ANOVA

```
ss<-round(c(ssa,sse,sst),4)
```

```
df<-c(2,length(grades)-3,length(grades)-1)
```

```
ms<-round(c(ss[1]/df[1],ss[2]/df[2]),4)
```

```
f<-round(ms[1]/ms[2],4)
```

```
p.value<-round(1-pf(f,df[1],df[2]),4)
```

```
grades.anova<-c(ss,df,ms,"",f,"","",p.value,"","")
```

```
dim(grades.anova)<-c(3,5)
```

```
dimnames(grades.anova)<-
```

```
list(c("Παράγοντας","Σφάλματα","Σύνολο"),c("Άθροισμα
```

```
Τετραγώνων","Βαθμοί ελευθερίας","Μέσα τετράγωνα","F","p.value"))
```

```
print.default(grades.anova,quote=F)
```

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

Μουσιτάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

	Άθροισμα Τετραγώνων	Βαθμοί ελευθερίας	Μέσα τετράγωνα	F	p.value
Παράγοντας	50.0833	2	25.0416	6.0532	0.0084
Σφάλματα	86.875	21	4.1369		
Σύνολο	136.9583	23			

$H_0: \alpha_1 = \alpha_2 = \alpha_3$

$H_1: \alpha_1 \neq \alpha_2 \neq \alpha_3$

Απορρίπτεται η μηδενική υπόθεση ότι η μέση τιμή της grades για τις τρεις ομάδες είναι ίδια (δηλαδή οι κύριες επιδράσεις είναι διαφορετικές).

ANOVA με εντολές της R

Η εντολή της R για τη διεξαγωγή ανάλυσης διασποράς είναι η aov. Συντάσσεται με τρόπο εντελώς ανάλογο της lm.

```
grad.aov <- aov(grades ~ methodos); grad.aov
```

```
Call: aov(formula = grades ~ methodos)
```

```
Terms: methodos Residuals
```

```
Sum of Squares 50.08333 86.87500
```

```
Deg. of Freedom 2 21
```

```
Residual standard error: 2.033938
```

```
Estimated effects are balanced
```

```
summary(grad.aov)
```

```
      Df Sum of Sq Mean Sq F Value Pr(F)
methodos  2  50.08333 25.04167 6.053237 0.00839879
```

```
Residuals 21  86.87500  4.13690
```

```
model.tables(grad.aov) ##κύριες επιδράσεις
```

```
method 1 method 2 method 3
```

```
-0.9583 -1.0833  2.0417
```

```
model.tables(grad.aov, type="means") ##μέσοι όροι των ομάδων
```

```
method 1 method 2 method 3
```

```
4.750  4.625  7.750
```

Μουσιτάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

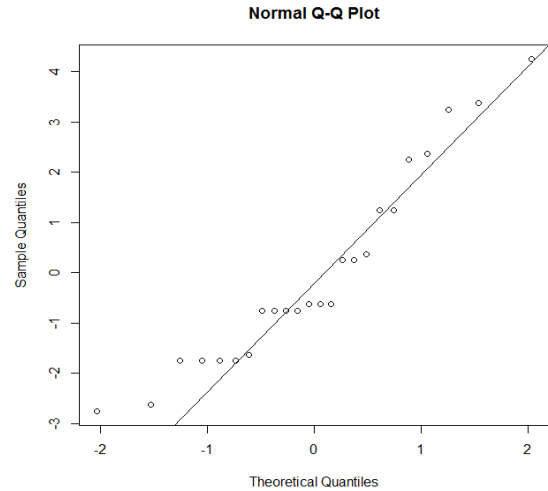
Έλεγχος του μοντέλου - Γραφήματα

```
res<-grad.aov$res  
qqnorm(res)  
qqline(res)
```

```
> shapiro.test(res)
```

Shapiro-Wilk Normality Test

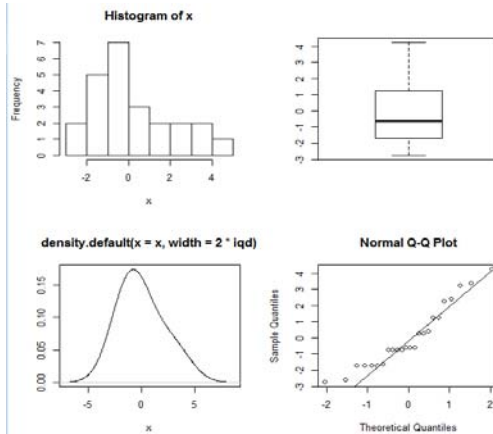
```
data: res  
W = 0.9221, p-value = 0.0648
```



```
eda.shape(res)
```

Επίσης, όπως και στην παλινδρόμηση μπορούμε να χρησιμοποιήσουμε την `plot(grad.aov)`

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

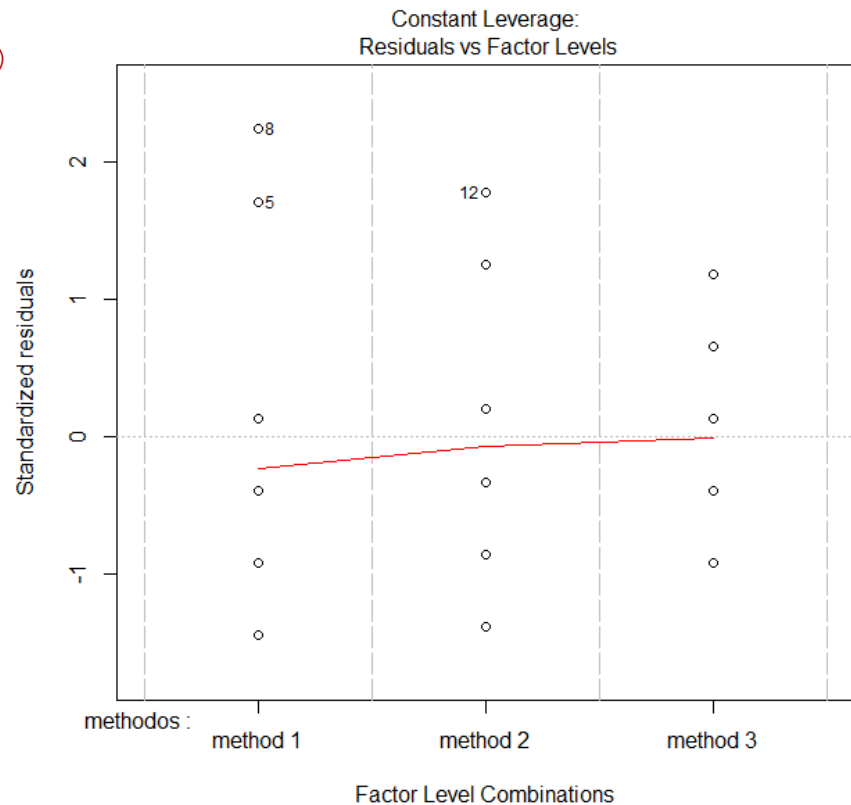


47

Οπτικός έλεγχος για ετεροσκεδαστικότητα

```
x1<-rep(1:3,each=8)  
res1<-res/stdev(res)  
plot(x1,res1)
```

Παρατηρώ ότι δεν έχω ετεροσκεδαστικότητα



Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Σχέση ανάλυσης διασποράς -παλινδρόμησης

Η ανάλυση διασποράς με έναν παράγοντα που έχει k στάθμες ισοδυναμεί με το μοντέλο παλινδρόμησης που έχει σαν προβλέπουσες k βωβές μεταβλητές X_i με $X_i = 1$ μόνο αν η παρατήρηση προέρχεται από την i στάθμη του παράγοντα

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Κατά την εφαρμογή του παραπάνω μοντέλου, ο πίνακας $X'X$ είναι μη αντιστρέψιμος και χρειάζεται η επιπλέον υπόθεση

$$r_1 \beta_1 + r_2 \beta_2 + \dots + r_k \beta_k = 0$$

Επομένως δεν μπορούμε να χρησιμοποιήσουμε στην R τη σύνταξη $\text{lm}(Y \sim X_1 + X_2 + \dots + X_k)$

Χρησιμοποιούμε εναλλακτικά τη σύνταξη $\text{lm}(Y \sim \text{factor})$ όπου factor ένας παράγοντας που δηλώνει από ποια στάθμη προέρχεται κάθε παρατήρηση

```
> anova(lm(grades ~ methodos))
Analysis of Variance Table

Response: grades

      Df Sum Sq Mean Sq F value    Pr(>F)
methodos  2 50.083 25.0417  6.0532 0.008399 **
Residuals 21 86.875  4.1369
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Μωρσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Παράδειγμα 4.1 – Αντοχή υλικών

```
antoxi_c(1.700,1.505,1.800,1.725,1.825,1.440,1.640,
1.735,1.935,1.815,1.890,2.025,2.075,2.110,1.970)
material_rep(paste("mat",LETTERS[1:4],sep=""),c(3,4,4,4))
antoxi.df_data.frame(material,antoxi);antoxi.df
```

```
> antoxi.df
  material antoxi
1    matA 1.700
2    matA 1.505
3    matA 1.800
4    matB 1.725
5    matB 1.825
6    matB 1.440
7    matB 1.640
8    matC 1.735
9    matC 1.935
10   matC 1.815
11   matC 1.890
12   matD 2.025
13   matD 2.075
14   matD 2.110
15   matD 1.970

> antox.aov <- aov(antoxi ~ material)
> summary(antox.aov)
              Df Sum of Sq  Mean Sq  F Value    Pr(F)
material     3  0.3785829  0.1261943  8.702487 0.00304325
Residuals   11  0.1595104  0.0145009

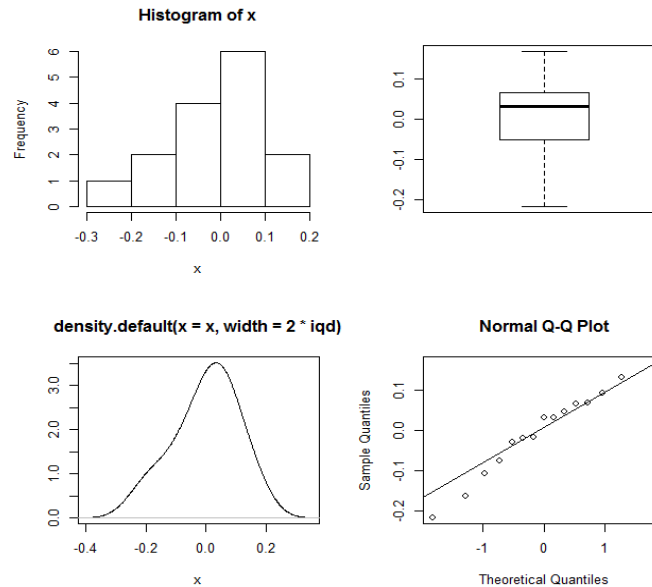
> antox.aov$coef
(Intercept) material1 material2 material3
 1.803646  -0.005416667  0.06027778  0.08045139

> antox.aov$fitted.values
      1      2      3      4      5      6
1.668333 1.668333 1.668333 1.6575 1.6575 1.6575
      7      8      9     10     11     12
1.6575 1.84375 1.84375 1.84375 1.84375 2.045
      13     14     15
2.045 2.045 2.045
```

Μωρσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Παράδειγμα 4.1 – Αντοχή υλικών (2)

```
res<-antox.aov$res
```



```
> shapiro.test(res)
```

Shapiro-Wilk Normality Test

data: res

W = 0.9689, p-value = 0.8415

Ισοδυναμία Ανονα με παλινδρόμηση

4 είδη υλικών => 3 βωβές μεταβλητές z_1, z_2, z_3 όπου:

$$z_1[i] = \begin{cases} 1 & \text{αν είναι το υλικό A} \\ 0 & \text{αλλού} \end{cases} \quad z_2[i] = \begin{cases} 1 & \text{αν είναι το υλικό B} \\ 0 & \text{αλλού} \end{cases}$$

$$z_3[i] = \begin{cases} 1 & \text{αν είναι το υλικό Γ} \\ 0 & \text{αλλού} \end{cases}$$

Το μοντέλο $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$

Στο υλικό A: $Y = \beta_0 + \beta_1 Z_1$ ($Y = \mu + \alpha_1$)

Στο υλικό B: $Y = \beta_0 + \beta_2 Z_2$ ($Y = \mu + \alpha_2$)

Στο υλικό Γ: $Y = \beta_0 + \beta_3 Z_3$ ($Y = \mu + \alpha_3$)

Στο υλικό Δ: $Y = \beta_0$ ($Y = \mu + \alpha_4$)

Το αντίστοιχο μη-παραμετρικό τεστ

Το αντίστοιχο μη-παραμετρικό τεστ της ανάλυσης διασποράς είναι το **Kruskal-Wallis test**. Το χρησιμοποιούμε όταν τα δεδομένα ανήκουν στην κλίμακα διάταξης ή αν τα σφάλματα δεν ακολουθούν κανονική κατανομή

```
>kruskal.test(grades, methodos)
```

```
Kruskal-Wallis rank sum test
```

```
data: grades and methodos
```

```
Kruskal-Wallis chi-square = 8.077, df = 2, p-value = 0.0176
```

```
alternative hypothesis: two.sided
```

Και πάλι απορρίπτεται η μηδενική υπόθεση της ισότητας των μέσων τιμών

53

Ανάλυση Διασποράς με 2 παράγοντες

ΠΑΡΑΔΕΙΓΜΑ

Δύο κατηγορίες ασθενών εξετάστηκαν ως προς την αντίδρασή τους σε τρία φάρμακα. Για κάθε φάρμακο και κάθε κατηγορία επιλέγησαν τυχαία τρεις ασθενείς. Πριν και μετά τη χορήγηση μετρήθηκε με κάποιο κριτήριο η αντίδραση των ασθενών και η διαφορά δίνεται παρακάτω.

Y_{ijk}	Φάρμακο 1 B_1	Φάρμακο 2 B_2	Φάρμακο 3 B_3
Κατηγ. 1 A_1	8, 4, 0	10, 8, 6	8, 6, 4
Κατηγ. 2 A_2	14, 10, 6	4, 2, 0	15, 12, 9

Τα δεδομένα κατά γραμμές

```
apotel<-c(8,4,0,10,8,6,8,6,4,14,  
10,6,4,2,0,15,12,9)
```

```
A<-rep(1:2,each=9)
```

```
A<-factor(A,labels=c("A1","A2"))
```

```
B<-rep(rep(1:3,each=3),2)
```

```
B<-factor(B,labels=c("B1","B2","B3"))
```

```
apoteld<-data.frame(A,B,apotel)
```

```
> apoteld
```

	A	B	apotel		A	B	apotel
1	A1	B1	8	10	A2	B1	14
2	A1	B1	4	11	A2	B1	10
3	A1	B1	0	12	A2	B1	6
4	A1	B2	10	13	A2	B2	4
5	A1	B2	8	14	A2	B2	2
6	A1	B2	6	15	A2	B2	0
7	A1	B3	8	16	A2	B3	15
8	A1	B3	6	17	A2	B3	12
9	A1	B3	4	18	A2	B3	9

Ανάλυση Διασποράς με 2 παράγοντες

Ανάλυση διασποράς χωρίς να λάβουμε υπόψη τις αλληλεπιδράσεις

```
> apot1 <- aov(apotel ~ A + B, data = apoteld)
```

```
> summary(apot1)
```

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
A	1	18	18.00000	1.008	0.3324207
B	2	48	24.00000	1.344	0.2924598
Residuals	14	250	17.85714		

Ανάλυση διασποράς λαμβάνοντας υπόψη τις αλληλεπιδράσεις

```
> apot2 <- aov(apotel ~ A * B, data = apoteld)
```

```
> summary(apot2)
```

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
A	1	18	18.00000	2.037736	0.1789399
B	2	48	24.00000	2.716981	0.1063435
A:B	2	144	72.00000	8.150943	0.0058103
Residuals	12	106	8.83333		

Οι κύριες επιδράσεις των δύο παραγόντων δεν είναι σημαντικές, αλλά η αλληλεπίδρασή τους είναι σημαντική

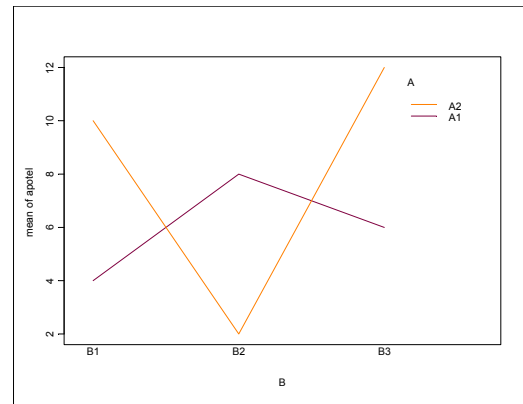
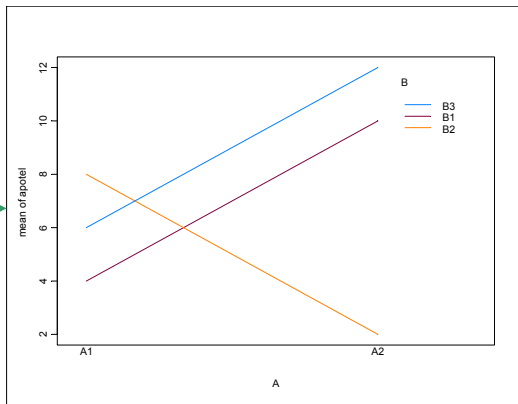
Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Γραφήματα αλληλεπιδράσεων

```
attach(apoteld)
```

```
interaction.plot(A,B,apotel,lwd=2,lty=1,col=c(3,5,6))
```

```
interaction.plot(B,A,apotel,lwd=2,lty=1,col=c(3,5))
```



Μουσιδάκης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Αν δεν υπάρχουν αλληλεπιδράσεις, περιμένουμε οι γραμμές να είναι περίπου παράλληλες.