# ΕΦΑΡΜΟΣΜΕΝΗ ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ
# Εργασία $1^{\eta}$

Παντελεήμων Μαθιουδάκης
ΜΕΣ20022

April 20, 2021

## Contents

# 1 Άσκηση 1

## 1.1 a

Για κάθε μεταβλητή, ενας συνοπτικός τρόπος παρουσίασης των περιγραφικών μέτρων είναι η δημιουργία διαγραμμάτων Box-Plots .

```
library(formatR)
```

```
### 1-a
library(foreign)
library(ggplot2)
df <- read.spss("/home/user/POLYMETABLHTH/ergasies/labEx1Dat.sav",
    header = T)
df <- as.data.frame(df)
head(df, 1)
```

```
##   Year Temperature  sun heat rain quality
## 1 1913        3308 1376   27  319    good
```

```
fqual <- factor(df$quality, levels = c("bad", "medium", "good"), labels = c(1,
    2, 3))
attach(df)
df <- cbind(df, fqual)

library(reshape2)
dfn <- df
dfn <- melt(dfn)
```

```
## Using quality, fqual as id variables
```

```
ggplot(dfn, aes(x = fqual, y = value, color = quality)) + geom_boxplot() +
    theme(aspect.ratio = 1) + facet_wrap(~variable, scales = "free_y")
```

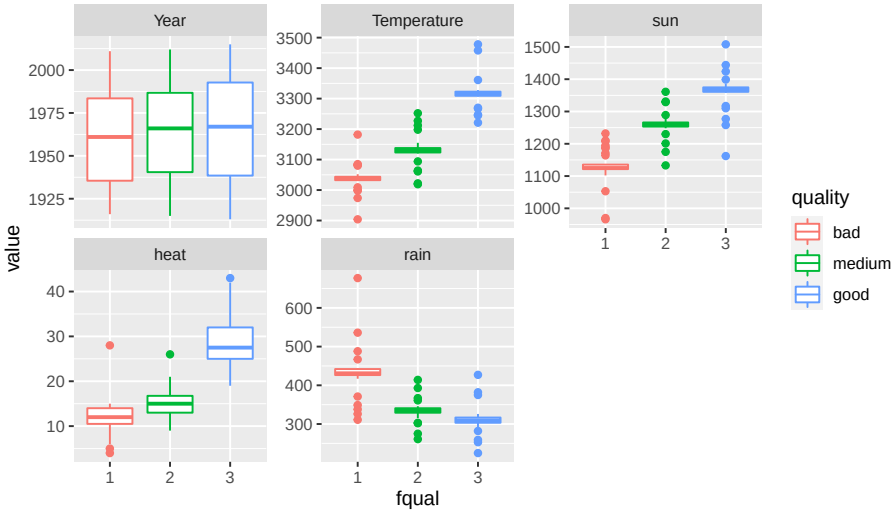


Figure 1: Διάγραμμα Θηκογραμμάτων ανα μεταβλητή και κατηγορία ||χαμηλό - μέτριο - καλό ||

Πέραν της μεταβλητής **χρόνος** οι οποία διαθέτει κατανομή χωρίς μεγάλες ουρές ή ακραίες τιμές αλλα επίσης φαίνεται να μην επηρεάζονται οι τιμές της ανα ποιότητα, οι υπόλοιπες μεταβλητές χαρακτηρίζονται απο αρκετές ακραίες τιμές (πιο μεγάλες ουρές των κατανομών τους) καθώς και απο μια εμφανή αλλαγή των τιμών ανα ποιότητα.

```
library(heplots)
```

```
## Loading required package:  car
## Loading required package:  carData
```

```
pdf("ellipses.pdf")
```

```r
ellipses <- heplots::covEllipses(df[, -c(6, 7)], df$quality, fill = TRUE,
    pooled = FALSE, col = c("blue", "red", "purple"), variables = c(1:5),
    fill.alpha = 0.05)
print(ellipses)
```

```
## NULL
```

```r
dev.off()
```

```
## pdf
##   2
```

```r
library(psych)
```

```
##
## Attaching package: 'psych'
## The following object is masked from 'package:car':
##
##     logit
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
pdf("pairs.pdf")
pairs <- psych::pairs.panels(df[, -c(6, 7)], gap = 0, bg = c("blue",
    "red", "green")[quality], pch = 21)  # Πακέτο psych για διάγραμμα
# συσχετίσεων ανα δύο
print(pairs)
```

```
## NULL
```

```r
dev.off()
```

```
## pdf
##   2
```

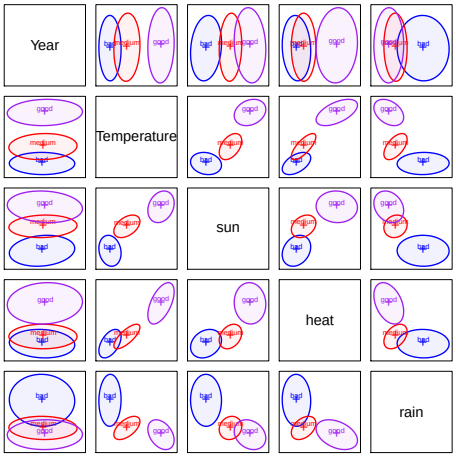


Figure 2: Διάγραμμα κατανομής των επιπέδων ανα δύο μεταβλητές

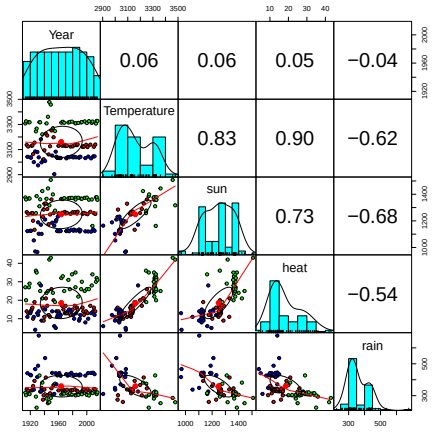


Figure 3: Διάγραμμα συσχετίσεων ανα δύο

Απο το Σχήμα 2 διακρίνονται οι μορφές των κατανομών των επιπέδων (κακό-μέτριο-καλό) ανα 2 μεταβλητές.Γενικώς όταν τα ελλειπτικά των επιπέδων εχουν το ίδιο σχήμα, δεν φαίνεται τότε να παρουσιάζεται διαφορά σοτυς πίνακες διακύμανσης των επιπέδων.
Στο παρόν σχήμα, φαίνονται διαφορές στις ελλείψεις των Year - Rain, Heat - Rain, Temperature -Rain, Sun - Heat. Συνεπώς, ίσως υπάρχει διαφορά στους πίνακες διακυμάνσεων
Στο Σχήμα 3, παρουσιάζονται οι συσχετίσεις των μεταβλητών. Πέραν του Χρόνου με όλες τις άλλες, διαφαίνονται αρκετά ισχυρές συσχετίσεις

## 1.2 β

Εκτελώντας τον παρακάτω κώδικα στην R λαμβάνονται τα εξής :

```
dfn1 <- df[, -c(7)]
dfn1 <- melt(dfn1)

## Using quality as id variables

# αλγόριθμος που υπολογίζει τους μέσους όρους κάθε μίας απο τις
# κατηγορίες :
aggregate(value ~ quality + variable, data = dfn1, mean)

##    quality    variable      value
## 1      bad        Year 1962.17143
## 2   medium        Year 1963.38235
## 3     good        Year 1965.91176
```

```
## 4      bad Temperature 3037.57143
## 5   medium Temperature 3130.29412
## 6     good Temperature 3317.26471
## 7      bad         sun 1129.11429
## 8   medium         sun 1258.82353
## 9     good         sun 1362.47059
## 10     bad        heat   12.02857
## 11  medium        heat   15.00000
## 12    good        heat   28.79412
## 13     bad        rain  432.08571
## 14  medium        rain  334.41176
## 15    good        rain  310.61765
```

Πέραν της μεταβλητής Χρόνος με περιθώριους μέσους όρους(ανα ποιότητα) 1962-1963-1965, οι μέσοι όροι των άλλων μεταβλητών ανα επίπεδο διαφέρουν αισθητά. Αυτό επιβεβαιώθηκε και απο το Σχήμα 1 των Box-Plots .

## 1.3 γ

Συνεχίζοντας στο SPSS, εκτελείται ο έλεγχος Box M :

**Box's Test of Equality of Covariance Matrices**

**Log Determinants**

| quality | Rank | Log Determinant |
|---|---|---|
| bad | 5 | 32.356 |
| medium | 5 | 29.252 |
| good | 5 | 32.400 |
| Pooled within-groups | 5 | 32.202 |

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

**Test Results**

| | | |
|---|---|---|
| Box's M | | 85.583 |
| F | Approx. | 2.647 |
| | df1 | 30 |
| | df2 | 31640.742 |
| | Sig. | .000 |

Tests null hypothesis of equal population covariance matrices.

Figure 4: Έλεγχος Ισότητας Πινάκων Διακύμανσης

$P\_value \simeq 0.000 \Rightarrow$ απορρίπτεται ισότητα πινάκων διακ.-συνδιακύμανσης.

# 2 Άσκηση 2

**Classification Function Coefficients**

| | quality | | |
|---|---|---|---|
| | bad | medium | good |
| Year | 2.055 | 2.050 | 2.046 |
| Temperature | 2.129 | 2.185 | 2.263 |
| sun | .250 | .295 | .324 |
| heat | -12.324 | -12.607 | -12.389 |
| rain | .098 | .044 | .032 |
| (Constant) | -5338.366 | -5532.234 | -5812.033 |

Fisher's linear discriminant functions

Figure 5: Γραμμική Διαχωριστική Ανάλυση

Η 3 συναρτήσεις του διαχωρισμού είναι

$$f_1(x) = -5.338 + 0.098 \cdot rain - 12.324 \cdot heat + 0.25 \cdot sun + 2.129 \cdot Temp + 2.055 \cdot Year$$

$$f_2(x) = -5.332 + 0.044 \cdot rain - 12.607 \cdot heat + 0.295 \cdot sun + 2.185 \cdot Temp + 2.050 \cdot Year$$

$$f_3(x) = -5.812 + 0.032 \cdot rain - 12.389 \cdot heat + 0.324 \cdot sun + 2.263 \cdot Temp + 2.046 \cdot Year$$

Η κατάταξη δοθέντος παρατήρησης $\underset{\sim}{x} = (rain, heat, sun, Temp, Year)$ γίνεται στο επίπεδο με το μέγιστο $f_i(\underset{\sim}{x}) : max(f_1(\underset{\sim}{x})), f_2(\underset{\sim}{x})), f_3(\underset{\sim}{x}))$

# 3 Άσκηση 3

## 3.1 α

**Casewise Statistics**

| | Case Number | Actual Group | Highest Group: Predicted Group | Highest Group: P(D>d \| G=g) p | Highest Group: P(D>d \| G=g) df | Highest Group: P(G=g \| D=d) | Squared Mahalanobis Distance to Centroid |
|---|---|---|---|---|---|---|---|
| Original | 1 | 3 | 3 | .952 | 2 | 1.000 | .098 |
| | 2 | 3 | 3 | .317 | 2 | 1.000 | 2.297 |
| | 3 | 2 | 2 | .978 | 2 | 1.000 | .045 |
| | 4 | 1 | 1 | 1.000 | 2 | 1.000 | .000 |
| | 5 | 2 | 2 | .936 | 2 | 1.000 | .133 |
| | 6 | 3 | 3 | .932 | 2 | 1.000 | .141 |
| | 7 | 3 | 3 | .993 | 2 | 1.000 | .013 |
| | 8 | 1 | 1 | .990 | 2 | 1.000 | .020 |
| Cross-validated[a] | 1 | 3 | 3 | .585 | 5 | 1.000 | 3.759 |
| | 2 | 3 | 3 | .033 | 5 | 1.000 | 12.108 |
| | 3 | 2 | 2 | .686 | 5 | 1.000 | 3.093 |
| | 4 | 1 | 1 | .732 | 5 | 1.000 | 2.789 |
| | 5 | 2 | 2 | .717 | 5 | 1.000 | 2.887 |
| | 6 | 3 | 3 | .676 | 5 | 1.000 | 3.155 |
| | 7 | 3 | 3 | .749 | 5 | 1.000 | 2.679 |
| | 8 | 1 | 1 | .821 | 5 | 1.000 | 2.197 |

Figure 6: Leave One Out CV

Οπως παρουσιάζεται, η μέθοδος Leave One Out CV, και η κανονική με όλες τις παρατηρήσεις κατέταξαν επιτυχώς και τα 8 πρώτα δεδομένα

## 3.2 β

**Classification Results[a,b]**

| | | | | Predicted Group Membership: bad | medium | good | Total |
|---|---|---|---|---|---|---|---|
| Cases Selected | Original | Count | bad | 22 | 2 | 0 | 24 |
| | | | medium | 1 | 19 | 1 | 21 |
| | | | good | 0 | 0 | 24 | 24 |
| | | % | bad | 91.7 | 8.3 | .0 | 100.0 |
| | | | medium | 4.8 | 90.5 | 4.8 | 100.0 |
| | | | good | .0 | .0 | 100.0 | 100.0 |
| Cases Not Selected | Original | Count | bad | 11 | 0 | 0 | 11 |
| | | | medium | 0 | 13 | 0 | 13 |
| | | | good | 0 | 1 | 9 | 10 |
| | | % | bad | 100.0 | .0 | .0 | 100.0 |
| | | | medium | .0 | 100.0 | .0 | 100.0 |
| | | | good | .0 | 10.0 | 90.0 | 100.0 |

a. 94.2% of selected original grouped cases correctly classified.

b. 97.1% of unselected original grouped cases correctly classified.

Figure 7: Cross Validation train, test = 70%, 30%

```
***ask-3-ii

USE ALL.
COMPUTE filter_$=(
      uniform(1)<=.70).
VARIABLE LABELS
      filter_$ '
      Approximately 70% of
       the cases (SAMPLE)
      '.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

FILTER OFF.
USE ALL.
EXECUTE.

DISCRIMINANT
  /GROUPS=quality(1 3)
  /VARIABLES=Year
       Temperature sun
       heat rain
  /SELECT=filter_$(1)
  /ANALYSIS ALL
  /SAVE=CLASS
  /PRIORS EQUAL
  /STATISTICS=COEFF
       TABLE
  /CLASSIFY=
       NONMISSING
       POOLED.
```

Αρχικά, μέσω του Select Cases επιλέχθηκε τυχαία 70% των παρατηρήσεων στις οποίες βασίστηκε ο υπολογισμός του μοντέλου. Στο Σχήμα 7, Cases Not Selected αφορά την προσαρμογή του μοντέλου στο υπόλοιπο σετ δεδομένων (δηλαδή αυτά που δεν επιλέχθηκαν, 30%)
Τέλεια είναι η προσαρμογη, βάσει πίνακα Σχήματος 7, στις κατηγορίες κακό και μέτριο, ενω 90% η επιτυχία κατάταξης στην κατηγορία καλό,

## 3.3 γ

**Classification Results[a,c]**

| | | quality | Predicted Group Membership bad | medium | good | Total |
|---|---|---|---|---|---|---|
| Original | Count | bad | 34 | 1 | 0 | 35 |
| | | medium | 1 | 32 | 1 | 34 |
| | | good | 0 | 1 | 33 | 34 |
| | % | bad | 97.1 | 2.9 | .0 | 100.0 |
| | | medium | 2.9 | 94.1 | 2.9 | 100.0 |
| | | good | .0 | 2.9 | 97.1 | 100.0 |
| Cross-validated[b] | Count | bad | 33 | 2 | 0 | 35 |
| | | medium | 1 | 31 | 2 | 34 |
| | | good | 0 | 1 | 33 | 34 |
| | % | bad | 94.3 | 5.7 | .0 | 100.0 |
| | | medium | 2.9 | 91.2 | 5.9 | 100.0 |
| | | good | .0 | 2.9 | 97.1 | 100.0 |

a. 96.1% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 94.2% of cross-validated grouped cases correctly classified.

Figure 8: Μοντέλο όλων των δεδομένων και Leave One Out CV

**Classification Results[a,b]**

| | | | quality | Predicted Group Membership bad | medium | good | Total |
|---|---|---|---|---|---|---|---|
| Cases Selected | Original | Count | bad | 22 | 2 | 0 | 24 |
| | | | medium | 1 | 19 | 1 | 21 |
| | | | good | 0 | 0 | 24 | 24 |
| | | % | bad | 91.7 | 8.3 | .0 | 100.0 |
| | | | medium | 4.8 | 90.5 | 4.8 | 100.0 |
| | | | good | .0 | .0 | 100.0 | 100.0 |
| Cases Not Selected | Original | Count | bad | 11 | 0 | 0 | 11 |
| | | | medium | 0 | 13 | 0 | 13 |
| | | | good | 0 | 1 | 9 | 10 |
| | | % | bad | 100.0 | .0 | .0 | 100.0 |
| | | | medium | .0 | 100.0 | .0 | 100.0 |
| | | | good | .0 | 10.0 | 90.0 | 100.0 |

a. 94.2% of selected original grouped cases correctly classified.

b. 97.1% of unselected original grouped cases correctly classified.

Figure 9: Cross Validation μοντέλο

| % κατατάξεων | Πλήρες Μοντέλο | LOOCV | CV |
|---|---|---|---|
| bad | 97.1 | 94.3 | 100 |
| medium | 94.1 | 91.2 | 100 |
| good | 97.1 | 97.1 | 90 |

## 3.4 δ

Για την κατάταξη θα χρησιμοποιηθούν οι εξισώσεις της γραμμικής διαχωριστικής ανάλυσης :

$$Year = 1940, temperature = 3300, sun = 1100, heat = 12, rain = 300$$

$$f_1(x) = -5.338 + 0.098 \cdot rain - 12.324 \cdot heat + 0.25 \cdot sun + 2.129 \cdot Temp + 2.055 \cdot Year = 5830.546$$

$$f_2(x) = -5.332 + 0.044 \cdot rain - 12.607 \cdot heat + 0.295 \cdot sun + 2.185 \cdot Temp + 2.050 \cdot Year = 5841.682$$

$$f_3(x) = -5.812 + 0.032 \cdot rain - 12.389 \cdot heat + 0.324 \cdot sun + 2.263 \cdot Temp + 2.046 \cdot Year = 5842.439$$

Συνεπώς η παρατήρηση κατατάσσεται στην κατηγορία ΚΑΛΟ.