Εργασία 7

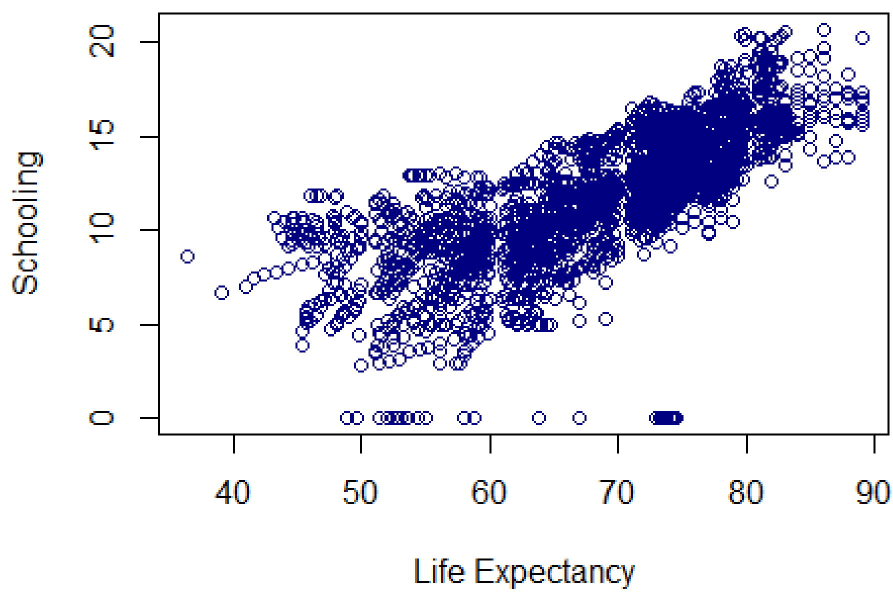
Marios

2024-04-30

Εργασία 5

Scatterplot - Educations impact on life expectancy

```
my_data <-  read.csv("C:/Users/mario/Downloads/Life Expectancy
Data.csv")
plot(x = my_data$Life.expectancy, y = my_data$Schooling, xlab = "Life
Expectancy", ylab = "Schooling",
col = "navy")
```

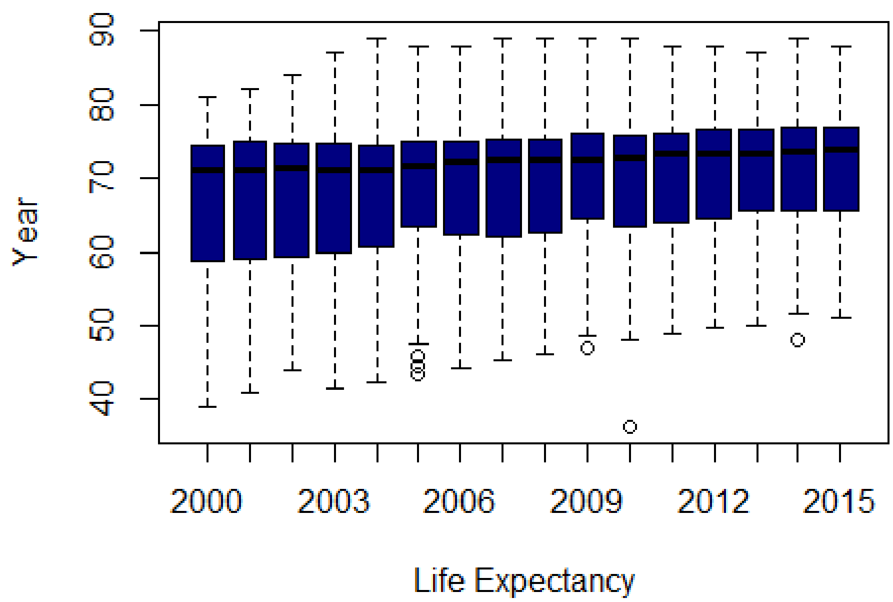


Παρατηρούμε ότι έχουμε μια ισχυρή συσχέτιση μεταξύ διάρκειας ζωής και επιπέδου εκπαίδευσης, αφού καθώς αυξάμεται το επίπεδο εκπαίδευσης αυξάνεται και η διάρκεια ζωής. Το επιχείρημα αυτό βλέπουμε και παρακάτω ότι όντως ισχύει, καθώς παίρνουμε μέσω της συνάρτησης cor τον αριθμό 0,75 ο οποίος είναι πολύ κοντά στο 1 και άρα έχουμε μια αρκετά ισχυρή θετική συσχέτιση. Ο λόγος είναι κατά πάσα πιθανότητα ότι οι ανρθωποί με χαμηλό επίπεδο μόρφωσης στρέφονται σε χειρονακτικές και ανθυγιεινες εργασίες με αποτέλεσμα να καταπονούν την υγείας τους και να μειώνουν το προσδόκιμο ζωής τους.

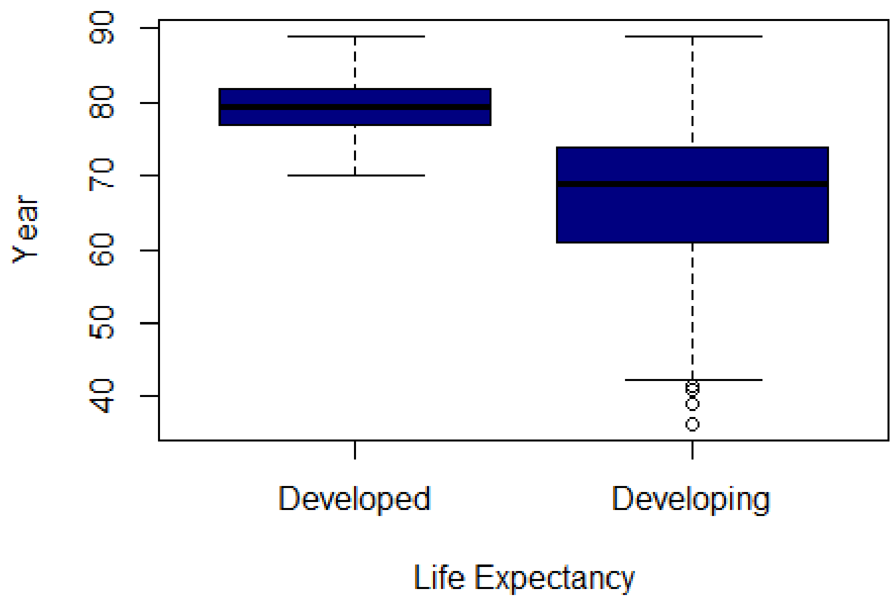
```r
complete_data <- na.omit(my_data[, c("Schooling", "Life.expectancy")])
correlation_result <- cor(complete_data$Schooling,
complete_data$Life.expectancy, method = "pearson")
print(correlation_result)

## [1] 0.7519755
```

Boxplot - Life expectancy through years

```r
boxplot(Life.expectancy ~ Year,  data = my_data, xlab = "Life
Expectancy", ylab ="Year",col="navy")
```



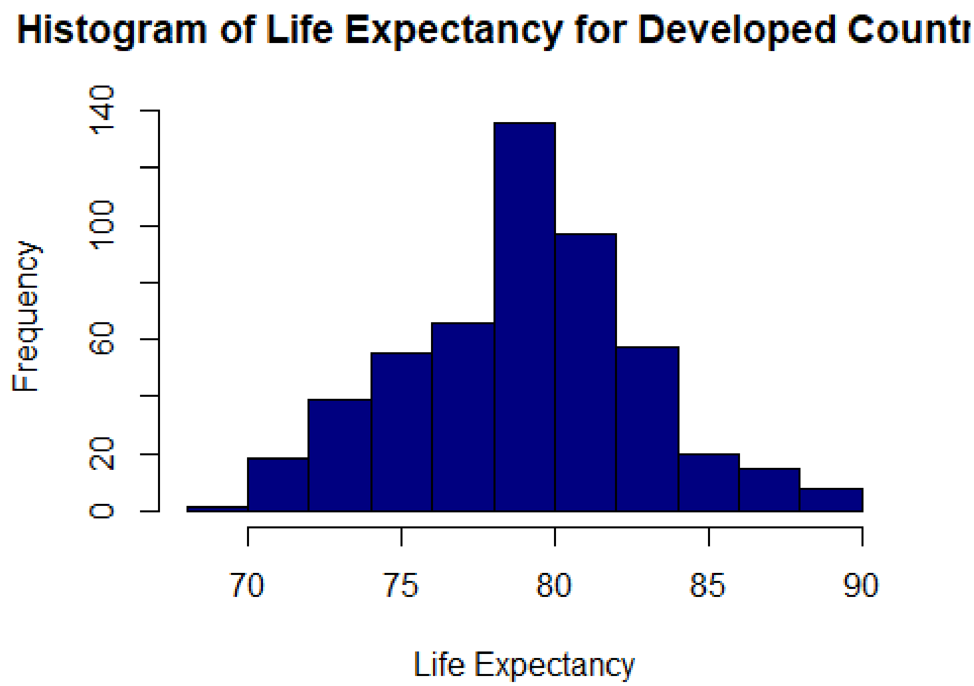

```r
boxplot(Life.expectancy ~ Status,  data = my_data, xlab = "Life
Expectancy", ylab ="Year",col="navy")
```



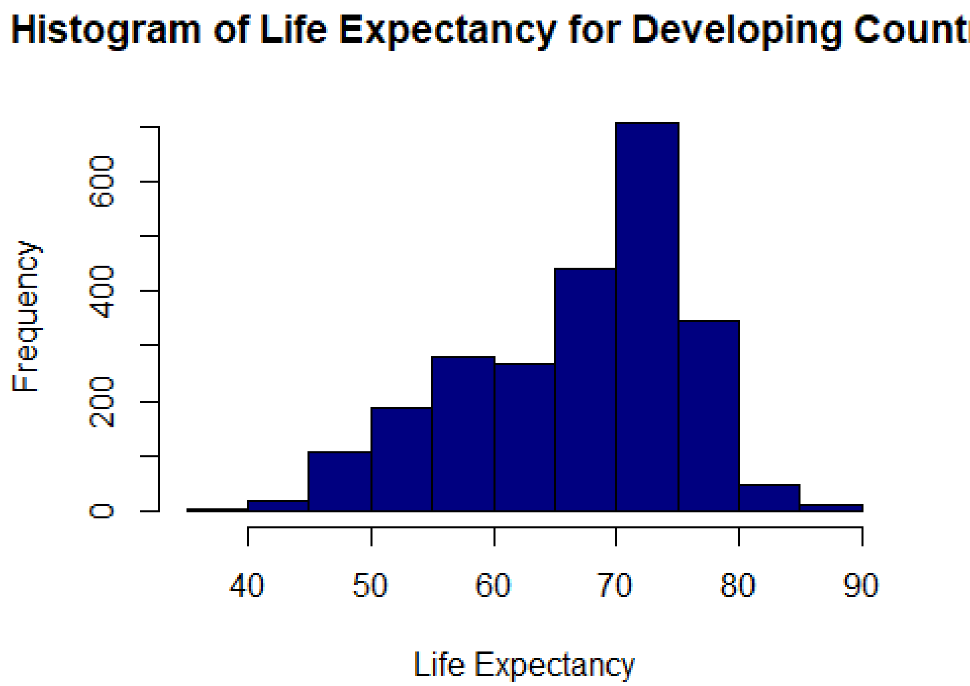
Παρατηρούμε το προσδόκιμο ζωής να αυξάνεται με την πάροδο του χρόνου αλλά να είναι και μεγαλυτερο για τις ανεπτυγμένες χώρες. Βλέπουμε επίσης τις τιμές να συσσωρεύονται σε τιμές υψηλότερες από αυτές που έχουμε στον πίνακα my_data και ο λόγος είναι κατά κύριο λόγο ότι οι τιμές που βρίσκονται εκτός από τα boxplots που έχουμε δημιουργήσει είναι outliers, ακραία σημεία που απέχουν από το κύριο σύνολο δεδομένων και γενικότερα παρατηρούμε ότι τα box plots δημιουργούνται εκεί που έχουμε το πλήθος των περισσότερων ηλικών.

Histogram - Life expectancy overview

```
developed_data <- subset(my_data, Status == "Developed")
hist(developed_data$Life.expectancy,
     xlab="Life Expectancy",
     main="Histogram of Life Expectancy for Developed Countries",
     col="navy")
```



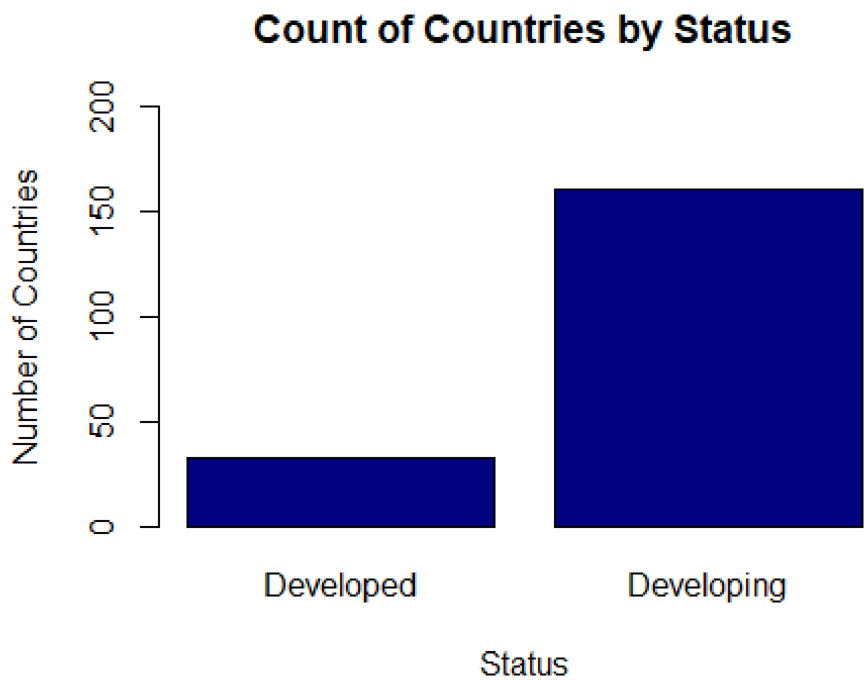

```
developing_data <- subset(my_data, Status == "Developing")
hist(developing_data$Life.expectancy,
     xlab="Life Expectancy",
     main="Histogram of Life Expectancy for Developing Countries",
     col="navy")
```



Στα παραπάνω διαγράμματα παρατηρούμε το προσδόκιμο ζωής για ανεπτυγμένες και αναπτυσσόμενες χώρες. Οι ανεπτυγμένες χώρες σε σχέση με τις αναπτυσσόμενες έχουν διαφορά κατά 30 περίπου χρόνια καθώς το προσδόκιμο ζωής ξεκινάει από το 70 ενώ οι για τις αναπτυσσόμενες από το 40.

Barchart - Status of countries

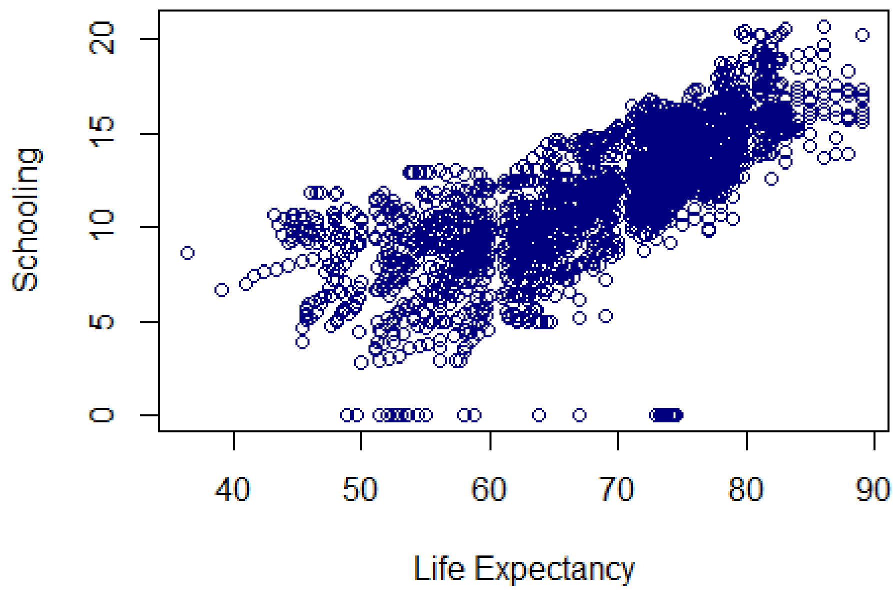
```
unique_data <- unique(my_data[c("Country", "Status")])
status_counts <- table(unique_data$Status)

barplot(status_counts, col="navy", ylim=c(0,200),
        main="Count of Countries by Status", xlab="Status",
ylab="Number of Countries")
```



1.Διερεύνηση του συνόλου δεδομένων(dataset)

```
my_data <-  read.csv("C:/Users/mario/Downloads/Life Expectancy
Data.csv")
library(ggplot2)
##A dataset with 2938 observations and 22 variables
str(my_data)

## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan"
"Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011
2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing"
"Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8
58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281
287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82
84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01
0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63
64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989
```

```
2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2
16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110
113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63
58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87
9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63
58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1
0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5
...
##  $ Population                     : num  33736494 327582 31731688
3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2
18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4
18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463
0.454 0.448 0.434 0.433 0.415 0.405 ...
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9
8.7 8.4 8.1 ...

plot(x = my_data$Life.expectancy, y = my_data$Schooling, xlab = "Life
Expectancy", ylab = "Schooling",
col = "navy")
```



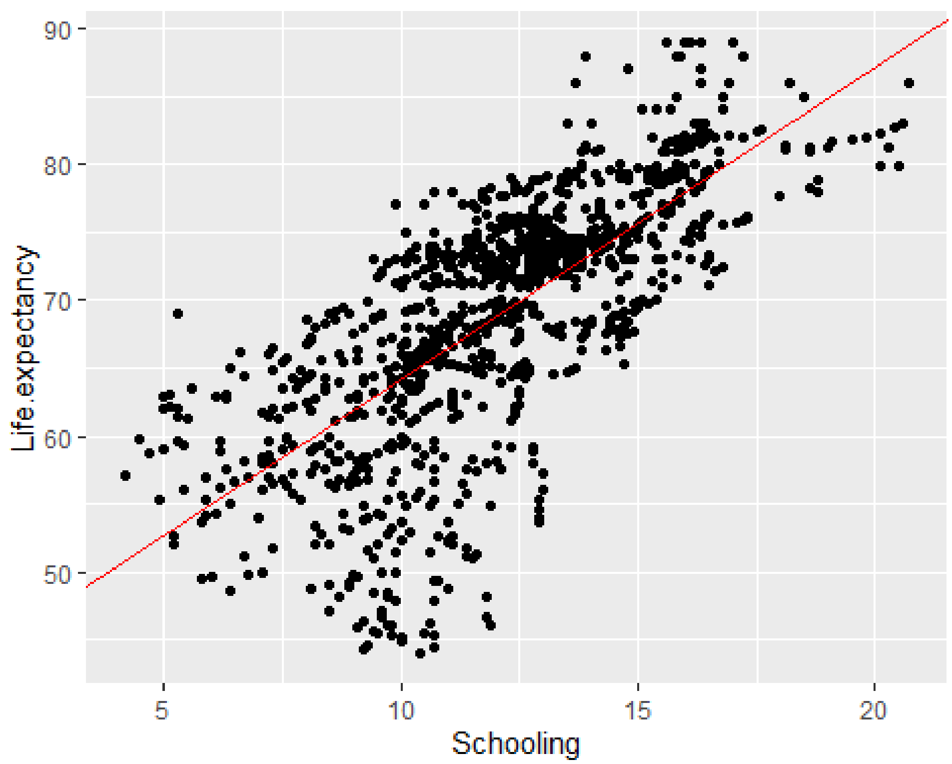
2α. Δημιουργία μοντέλου (γραμμικής) παλινδρόμησης

```
library(caTools)
library(ggplot2)
library(patchwork)
##Removing NA's
my_data_clean <- na.omit(my_data)
#Splitting Data into train and test
set.seed(64)
split <- sample.split(my_data_clean$Life.expectancy, SplitRatio=0.65)
train = subset(my_data_clean,split==TRUE)
test = subset(my_data_clean,split==FALSE)
##Creating 3 linear regression models
model1 <- lm(Life.expectancy ~ Schooling, data = train)
model2 <- lm(Life.expectancy ~  Income.composition.of.resources , data
= train)
model3 <- lm(Life.expectancy ~ Schooling + Alcohol + BMI + GDP +
HIV.AIDS + Income.composition.of.resources + Adult.Mortality +
Population, data = train)
##Projecting the data(summary) of each model and some diagrams
summary(model1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling, data = train)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -22.638   -3.184    0.894    4.000   15.531
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.28800    0.82312   50.16   <2e-16 ***
## Schooling    2.29831    0.06606   34.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.143 on 1085 degrees of freedom
## Multiple R-squared:  0.5273, Adjusted R-squared:  0.5269
## F-statistic:  1211 on 1 and 1085 DF,  p-value: < 2.2e-16
```

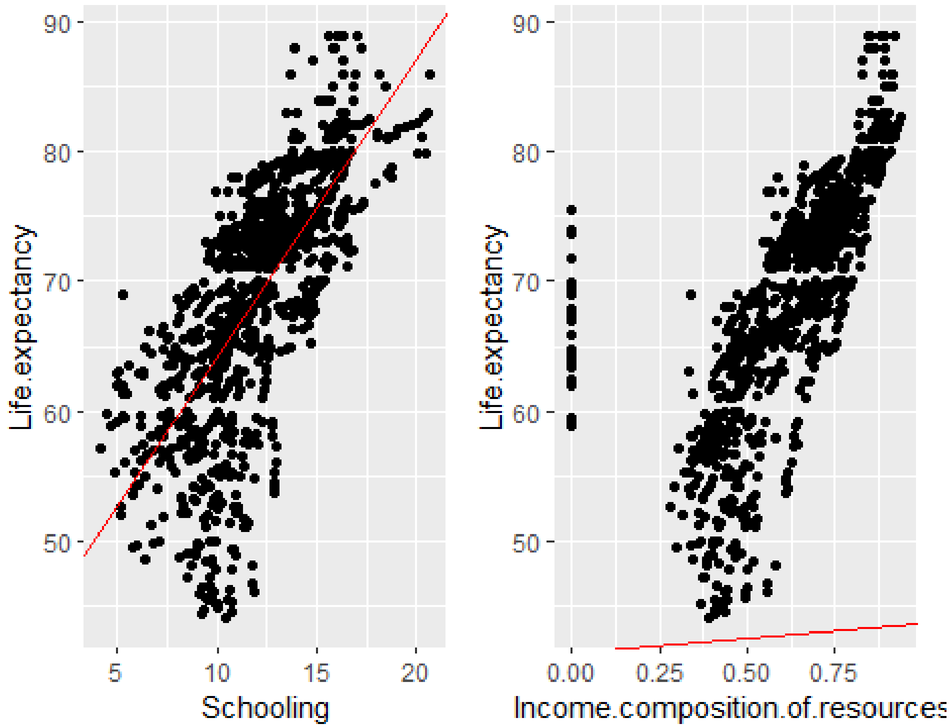
```
ggplot(train, aes(Schooling, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")
```




```
summary(model2)

##
## Call:
## lm(formula = Life.expectancy ~ Income.composition.of.resources,
##     data = train)
##
## Residuals:
```

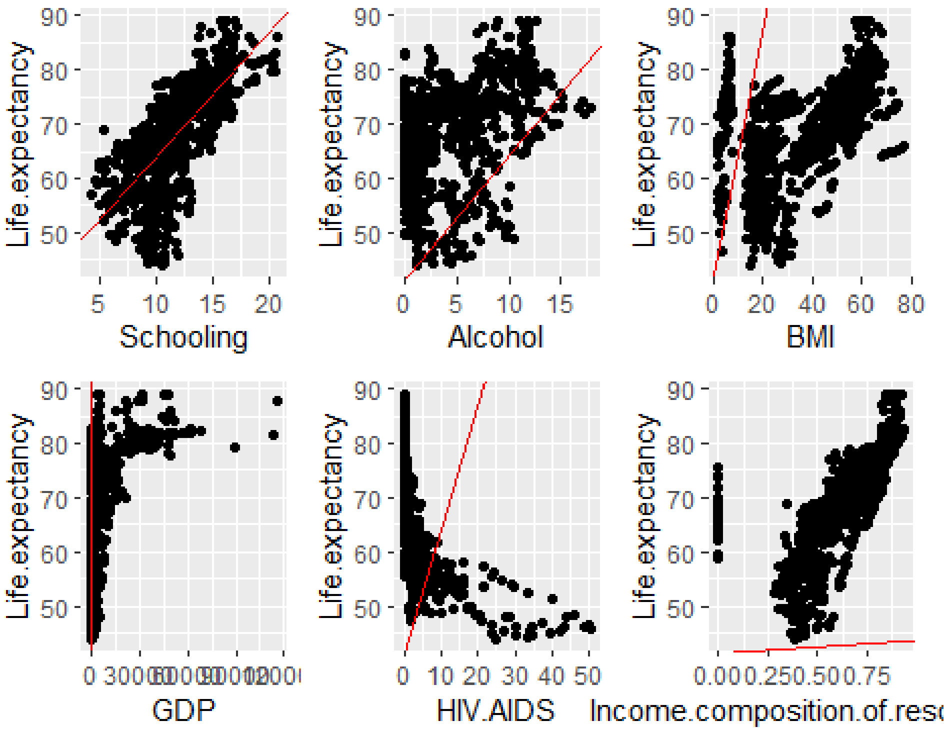
```
##      Min       1Q   Median       3Q      Max
## -20.4923  -2.7404   0.6802   3.0399  29.0803
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      46.420      0.671   69.18   <2e-16
***
## Income.composition.of.resources   35.972      1.019   35.29   <2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.097 on 1085 degrees of freedom
## Multiple R-squared:  0.5344, Adjusted R-squared:  0.534
## F-statistic:  1245 on 1 and 1085 DF,  p-value: < 2.2e-16
```

```
m1 <- ggplot(train, aes(Schooling, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

m2 <- ggplot(train, aes(Income.composition.of.resources,
Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

m1 + m2
```




```
summary(model3)

##
## Call:
## lm(formula = Life.expectancy ~ Schooling + Alcohol + BMI + GDP +
##     HIV.AIDS + Income.composition.of.resources + Adult.Mortality +
##     Population, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7784  -2.0982   0.0648   2.2902  11.0299
##
## Coefficients:
##                                    Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                       5.324e+01  7.203e-01  73.917  < 2e-
16 ***
## Schooling                         9.711e-01  7.530e-02  12.895  < 2e-
16 ***
## Alcohol                          -1.016e-01  3.740e-02  -2.718
0.00668 **
## BMI                               3.342e-02  6.960e-03   4.801 1.80e-
06 ***
## GDP                               8.040e-05  1.206e-05   6.668 4.13e-
11 ***
## HIV.AIDS                         -4.385e-01  2.111e-02 -20.775  < 2e-
16 ***
```

```
## Income.composition.of.resources  1.102e+01  1.108e+00   9.941  < 2e-
16 ***
## Adult.Mortality                 -1.823e-02  1.163e-03 -15.679  < 2e-
16 ***
## Population                      -1.661e-09  1.354e-09  -1.226
0.22036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.73 on 1078 degrees of freedom
## Multiple R-squared:  0.8269, Adjusted R-squared:  0.8256
## F-statistic: 643.6 on 8 and 1078 DF,  p-value: < 2.2e-16

p1 <- ggplot(train, aes(Schooling, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

p2 <- ggplot(train, aes(Alcohol, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

p3 <- ggplot(train, aes(BMI, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

p4 <- ggplot(train, aes(GDP, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

p5 <- ggplot(train, aes(HIV.AIDS, Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

p6 <- ggplot(train, aes(Income.composition.of.resources,
Life.expectancy)) +
      geom_point() +
      geom_abline(aes(intercept = coef(model1)[1],
slope = coef(model1)[2]), colour = "red")

p1 + p2 + p3 + p4 + p5 + p6
```



2β. Αξιολόγηση μοντέλου παλινδρόμησης

```
##A decrease of errors as we move to models with more variables
SSE1 = sum(model1$residuals^2)
SSE2 = sum(model2$residuals^2)
SSE3 = sum(model3$residuals^2)
SSE1

## [1] 40944.99

SSE2

## [1] 40332.82

SSE3

## [1] 14997.1

##A look at the more "realistic" projection of errors (Decreasing as we
add variables)
RMSE1 <- sqrt(SSE1/nrow(train))
RMSE1

## [1] 6.137417

RMSE2 <- sqrt(SSE2/nrow(train))
RMSE2

## [1] 6.091363
```

```
RMSE3 <- sqrt(SSE3/nrow(train))
RMSE3

## [1] 3.714402
```

2γ. Τροποποίηση μοντέλου παλινδρόμησης

```
#Removal of Population variable
model4 <- lm(Life.expectancy ~ Schooling + Alcohol + BMI + GDP +
HIV.AIDS + Income.composition.of.resources + Adult.Mortality, data =
train)

##R-squared increase due to Population's variable removal(Low
significance).
summary(model4)

##
## Call:
## lm(formula = Life.expectancy ~ Schooling + Alcohol + BMI + GDP +
##     HIV.AIDS + Income.composition.of.resources + Adult.Mortality,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9588  -2.1011   0.0444   2.3185  11.0389
##
## Coefficients:
##                                  Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                     5.317e+01  7.182e-01  74.037  < 2e-
16 ***
## Schooling                       9.742e-01  7.528e-02  12.942  < 2e-
16 ***
## Alcohol                        -1.019e-01  3.741e-02  -2.725
0.00654 **
## BMI                             3.417e-02  6.934e-03   4.928 9.59e-
07 ***
## GDP                             8.062e-05  1.206e-05   6.685 3.69e-
11 ***
## HIV.AIDS                       -4.378e-01  2.111e-02 -20.745  < 2e-
16 ***
## Income.composition.of.resources  1.096e+01  1.108e+00   9.899  < 2e-
16 ***
## Adult.Mortality                -1.819e-02  1.163e-03 -15.647  < 2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.731 on 1079 degrees of freedom
## Multiple R-squared:  0.8266, Adjusted R-squared:  0.8255
## F-statistic:   735 on 7 and 1079 DF,  p-value: < 2.2e-16
```

```
##Creating the new SSE and we can see a slightly expected increase of SSE4 BUT a new RMSE that remained the same.
SSE4 = sum(model4$residuals^2)
SSE4

## [1] 15018.02

RMSE4 <- sqrt(SSE4/nrow(train))
RMSE4

## [1] 3.716991
```

3. Εφαρμογή πρόβλεψης

```
LifePrediction <- predict(model4,newdata=test)
LifePrediction

##        3        5        6       11       16       17       18
24
## 63.71923 62.95021 62.54179 60.35783 56.77791 75.80267 77.02429
73.87912
##       28       30       38       39       40       41       43
52
## 72.30774 71.96525 74.19718 73.44801 72.51654 70.80037 71.42310
61.39088
##       54       57       87       93       94       98       99
104
## 59.12755 57.91836 77.44825 76.20923 76.30060 75.15737 73.07794
72.53564
##      106      107      109      110      111      112      115
120
## 70.50833 70.06667 69.75803 69.68916 69.82229 69.82658 88.72263
85.69000
##      123      131      132      133      137      139      140
141
## 82.99071 77.68587 82.55272 81.54217 81.53695 79.01567 78.61449
75.80773
##      144      146      147      149      151      153      156
157
## 76.32348 73.48917 71.29687 72.42266 71.45216 73.71227 69.82623
69.57887
##      198      199      203      205      226      227      228
236
## 65.66092 64.97918 64.23343 63.63265 74.88910 77.96951 72.10307
70.96325
##      238      245      246      255      258      259      261
262
## 69.90626 79.84330 78.20284 83.15356 69.95913 74.45923 70.60915
70.72343
##      263      267      275      281      284      286      291
297
```

```
## 70.96831 70.61415 64.13769 61.24771 59.59857 59.24307 68.45215
59.28948
##      299      300      303      331      343      345      347
349
## 57.97932 57.29730 60.48927 65.04967 62.30846 59.04865 52.85119
45.16538
##      352      354      355      363      367      368      388
390
## 42.68667 74.74958 74.46332 72.17788 72.36139 72.01322 74.71941
74.03171
##      392      395      398      400      410      418      425
427
## 73.66723 75.44328 71.73568 71.47788 59.90721 62.86495 55.67207
53.08641
##      455      458      460      470      474      482      483
486
## 70.37015 69.12705 70.21647 65.87696 64.22159 61.67037 61.48301
65.16059
##      489      498      499      500      504      506      516
518
## 56.00337 78.92266 83.00447 82.94375 82.07324 78.55886 54.64552
52.39914
##      530      531      533      535      546      547      548
550
## 57.35297 57.14778 56.02928 61.00032 79.24272 78.59640 78.42401
74.77413
##      551      554      564      567      581      583      585
587
## 76.69059 76.71918 71.32578 71.54446 72.86525 74.31290 71.18156
68.55146
##      590      591      595      600      602      603      604
605
## 69.94254 69.75719 65.43357 63.62164 63.64818 63.17939 58.15403
57.80695
##      627      629      638      641      646      650      682
686
## 75.15850 75.35348 74.31229 71.48370 77.88328 75.49825 77.57634
74.48634
##      687      757      758      760      772      773      774
775
## 74.86313 59.94228 58.89663 57.77372 72.78598 72.20510 72.11820
74.67837
##      776      780      783      784      785      790      794
797
## 71.62525 70.20585 72.25878 68.79122 68.94583 73.06335 72.56361
71.58523
##      798      799      801      826      830      833      857
858
## 71.43796 71.22233 70.70126 71.30358 69.76050 68.53065 57.01915
56.84255
```

```
##      864      868      869      877      878      887      889 
901 
## 50.75156 80.04934 78.44872 74.50562 74.12399 60.96502 59.94613 
74.60302 
##      902      904      906      907      914      932      934 
939 
## 72.50263 77.12719 76.35767 75.84735 69.32783 81.64131 78.31249 
77.55850 
##      940      942      946      950      951      955      956 
957 
## 80.34283 79.25046 79.72856 66.64304 65.14224 63.49648 62.72667 
66.71057 
##      983      991      993      994      997     1002     1003 
1005 
## 73.18548 71.07784 71.07124 63.18059 79.73543 83.59810 81.85259 
80.82770 
##     1013     1015     1016     1019     1022     1023     1024 
1029 
## 66.68862 65.36334 64.51470 66.76076 64.66764 64.64855 63.71879 
81.10856 
##     1031     1036     1041     1042     1061     1062     1065 
1066 
## 80.91365 78.25811 76.34887 75.04479 68.44834 68.36030 67.64492 
70.80957 
##     1068     1069     1078     1079     1080     1082     1094 
1113 
## 69.81724 69.33635 60.46836 64.91243 59.27808 64.01913 59.92827 
66.95724 
##     1114     1115     1144     1147     1148     1150     1154 
1189 
## 65.68505 65.42950 69.44596 70.98247 68.07979 67.26306 65.92610 
67.85180 
##     1190     1193     1205     1206     1208     1209     1210 
1213 
## 70.59795 69.72317 70.63311 70.45510 69.64486 69.00159 68.81089 
71.14631 
##     1214     1216     1238     1239     1240     1243     1244 
1255 
## 67.51996 67.32112 70.36583 70.14178 69.62923 67.98558 67.45858 
86.07402 
##     1269     1274     1275     1278     1281     1287     1291 
1292 
## 82.13166 81.21554 78.63573 78.40986 77.57953 83.03788 81.33435 
78.33720 
##     1296     1297     1298     1303     1304     1307     1308 
1311 
## 78.56673 76.27588 76.07241 72.19305 70.92412 71.40184 73.78449 
69.79300 
##     1342     1344     1346     1356     1357     1359     1368 
1371 
```

```
## 73.84729 73.24058 72.58795 71.18761 70.58109 69.48792 63.03668
58.42564
##     1375     1381     1383     1384     1387     1389     1446
1447
## 52.20946 73.53816 73.33180 73.27457 69.62746 63.63909 76.28364
75.43454
##     1448     1450     1452     1453     1457     1460     1465
1466
## 79.25115 76.46425 75.27377 74.90777 72.30703 75.40848 74.43363
74.24434
##     1467     1471     1472     1473     1474     1476     1480
1486
## 73.91727 69.52761 69.88269 69.74371 67.78450 56.65276 54.25797
41.94420
##     1492     1495     1496     1498     1525     1527     1531
1534
## 62.05676 61.88854 61.69943 61.29947 76.84350 78.21593 79.21358
78.06172
##     1537     1538     1540     1546     1547     1550     1551
1552
## 76.31061 76.10348 83.97783 85.89500 75.31562 80.87838 80.08598
74.97852
##     1558     1559     1562     1575     1581     1583     1590
1591
## 64.80315 64.64568 63.30144 55.97180 56.12143 45.84032 73.22477
75.24314
##     1592     1610     1617     1621     1626     1630     1636
1640
## 72.82655 70.54768 69.06862 60.70903 57.67907 56.50440 77.05631
76.88390
##     1641     1645     1653     1655     1661     1669     1673
1676
## 76.70327 77.99861 66.37848 66.19453 61.62879 74.40132 71.69827
72.43985
##     1685     1686     1687     1688     1696     1699     1723
1725
## 73.66519 75.42721 73.10617 71.80105 72.09684 70.83310 71.05604
73.81310
##     1727     1728     1729     1736     1737     1750     1752
1753
## 67.86265 66.97917 65.17723 77.11924 76.42850 72.36242 71.32320
72.33308
##     1756     1759     1767     1774     1777     1778     1783
1784
## 69.33038 68.06918 64.04017 50.26175 49.23052 49.09816 68.16081
67.97388
##     1787     1791     1793     1799     1801     1815     1816
1819
## 63.56168 62.22468 61.62775 67.61736 64.46864 69.01283 68.78264
66.89041
```

```
##     1820     1823     1824     1825     1831     1834     1865
1868
## 66.01353 67.77363 67.35513 63.71498 86.13558 80.29256 69.86954
69.31268
##     1873     1874     1875     1876     1882     1883     1898
1901
## 67.85268 67.70667 67.31821 67.00771 57.31885 56.60548 59.21808
58.07881
##     1902     1944     1945     1948     1950     1952     1954
1963
## 57.13257 64.67195 64.47850 63.76374 62.74180 62.43121 61.13200
75.08738
##     1965     1968     1969     1970     1973     1977     1979
1989
## 73.45255 72.74219 73.17353 71.60368 72.43823 64.96500 64.31208
59.33360
##     1990     1993     1994     1995     1996     2000     2003
2009
## 58.91774 72.00367 71.32899 71.26487 73.57999 70.16721 69.69229
74.36599
##     2010     2012     2013     2017     2019     2025     2026
2028
## 73.75822 73.40547 74.89439 72.33938 71.63631 68.66141 68.28244
68.17457
##     2032     2039     2041     2042     2043     2044     2045
2048
## 67.83860 67.16361 80.35928 76.24536 76.90209 76.91768 78.54712
74.89502
##     2051     2057     2061     2066     2124     2131     2132
2133
## 75.88006 76.66559 78.57174 78.59733 75.69597 71.84118 71.19336
70.39065
##     2140     2149     2162     2164     2205     2209     2215
2216
## 72.89033 68.73372 62.05248 59.92639 73.35519 70.36626 70.75554
73.55433
##     2219     2220     2226     2230     2251     2257     2258
2259
## 66.83736 67.56787 65.00585 63.49861 64.50806 60.57063 63.78522
60.28167
##     2270     2283     2286     2287     2299     2300     2305
2306
## 73.78499 75.43894 72.18358 74.27013 59.23104 66.35610 56.12495
63.00354
##     2364     2367     2368     2371     2372     2377     2395
2400
## 66.33843 65.73889 65.63688 67.87340 67.23005 62.09152 65.09214
57.26097
##     2403     2404     2407     2431     2438     2439     2441
2446
```

```
## 52.19251 51.48240 53.44848 78.84476 78.44363 76.89031 77.86784
71.79458
##     2447     2449     2450     2451     2454     2475     2476
2478
## 72.54083 71.40995 71.79194 71.62986 70.92754 72.14625 72.09538
71.61539
##     2479     2484     2491     2498     2504     2507     2551
2553
## 73.66903 72.41169 61.42948 50.66042 36.41163 79.22147 68.66653
68.29474
##     2556     2559     2561     2564     2566     2572     2575
2576
## 69.12828 68.67658 67.99484 67.24402 66.72559 72.53571 71.16136
73.81130
##     2577     2580     2581     2585     2603     2605     2608
2619
## 70.96387 69.26897 68.82510 66.97061 69.86928 69.83208 69.05011
65.26203
##     2621     2623     2624     2625     2635     2636     2643
2644
## 64.67407 61.91900 60.36286 59.47520 75.34720 75.32164 72.57298
77.10886
##     2645     2646     2653     2655     2663     2667     2668
2669
## 74.54805 73.33327 72.93595 72.50059 69.04351 75.56169 77.34052
77.17002
##     2672     2674     2677     2678     2680     2683     2688
2689
## 76.76968 76.48150 75.35245 73.56150 74.59252 78.32842 72.95379
70.91976
##     2700     2701     2704     2708     2709     2711     2717
2719
## 68.19493 72.34654 60.72469 58.53497 58.79509 60.06203 61.14537
59.11247
##     2721     2722     2725     2732     2733     2740     2741
2813
## 57.97212 63.70091 56.06461 75.08624 71.83085 71.28142 70.92829
77.65728
##     2819     2821     2823     2825     2831     2832     2834
2835
## 75.47068 75.63431 75.44966 74.60938 70.02253 68.54554 69.46473
69.21770
##     2837     2838     2846     2848     2851     2854     2908
2916
## 68.88820 68.87344 68.08201 69.30514 68.72703 62.04212 64.92289
52.92697
##     2925     2935
## 57.90630 38.47044
```

Παρατηρούμε ότι οι τιμές που έχουμε είναι κοντά στο διάστημα 60-75 κατά μέσω όρο που είναι και η πρόβλεψη ουσιαστικά ότι βρισκόμαστε κοντά σε αυτό τον αριθμό όσον αφόρα το προσδόκιμο ζωής.

Εργασία 6

1.Διερεύνηση του συνόλου δεδομένων(dataset)

```
##Loading-Cleaning-Splitting | Data-Libraries
Data <-  read.csv("C:/Users/mario/Downloads/framingham.csv")
library(ggplot2)
library(caTools)
CleanData <- na.omit(Data)
set.seed(964)
split <- sample.split(CleanData$TenYearCHD,SplitRatio=0.65)
train = subset(CleanData,split==TRUE)
test = subset(CleanData,split==FALSE)
nrow(train)

## [1] 2378

nrow(test)

## [1] 1280
```

2α.Δημιουργία μοντέλου (λογιστικής) παλινδρόμησης

```
framinghamLog <- glm(TenYearCHD ~ ., data = train, family=binomial)
```

2β. Αξιολόγηση μοντέλου παλινδρόμησης

```
summary(framinghamLog)

##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = train)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.043e+00  8.866e-01  -9.072  < 2e-16 ***
## male             4.594e-01  1.345e-01   3.416 0.000635 ***
## age              6.213e-02  8.248e-03   7.533 4.96e-14 ***
## education       -2.795e-02  6.074e-02  -0.460 0.645421
## currentSmoker    8.256e-02  1.921e-01   0.430 0.667417
## cigsPerDay       1.528e-02  7.560e-03   2.021 0.043290 *
## BPMeds           8.769e-02  2.936e-01   0.299 0.765190
## prevalentStroke  8.524e-01  5.952e-01   1.432 0.152130
## prevalentHyp     1.684e-01  1.747e-01   0.964 0.335130
## diabetes         2.809e-01  4.319e-01   0.650 0.515461
## totChol          2.284e-05  1.494e-03   0.015 0.987805
## sysBP            1.900e-02  4.902e-03   3.876 0.000106 ***
## diaBP           -7.285e-03  8.044e-03  -0.906 0.365095
```

```
## BMI               2.045e-02  1.579e-02   1.295 0.195406    
## heartRate        -5.887e-03  5.294e-03  -1.112 0.266155    
## glucose           7.192e-03  2.806e-03   2.563 0.010375 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2028.7  on 2377  degrees of freedom
## Residual deviance: 1787.2  on 2362  degrees of freedom
## AIC: 1819.2
## 
## Number of Fisher Scoring iterations: 5
```

Οι μεταβλητές που έχουν ισχυρή συσχέτιση με την εξαρτημένη μεταβλητή είναι οι παρακάτω:

- age, με significance *** που είναι ιδιαιτέρα σημαντικό.
- male με significance *** που είναι ιδιαιτέρα σημαντικό.
- cigsPerDay με significance * που δεν έχει τόσο μεγάλη σημαντικότητα για το μοντέλο μας.
- sysBP με significance *** που είναι ιδιαιτέρα σημαντικό.
- glucose με significance *** που είναι ιδιαιτέρα σημαντικό.

Επίσης το Residual deviance είναι 1787.2 σε σύγκριση με το Null deviance των 2028.7, υποδεικνύοντας ότι το μοντέλο έχει μειωμένη απόκλιση (σφάλμα) και έτσι βελτιώνει την πρόβλεψη σε σύγκριση με ένα μοντέλο χωρίς προβλεπτικούς παράγοντες. Το AIC του μοντέλου είναι 1819.2, το οποίο βοηθά στη σύγκριση μοντέλων. Όσο χαμηλότερο είναι το AIC, τόσο καλύτερα ταιριάζει το μοντέλο στα δεδομένα, λαμβάνοντας υπόψη τον αριθμό των προβλεπτικών παραγόντων.

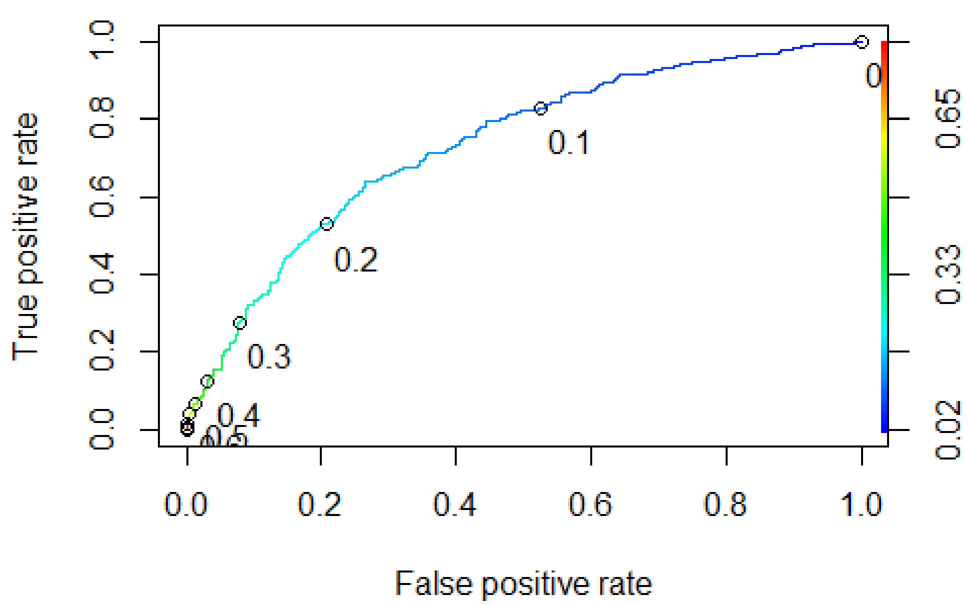
3. Εφαρμογή πρόβλεψης

```r
predictTest <- predict(framinghamLog, type="response", newdata=test)
summary(predictTest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.01700 0.06394 0.11459 0.15213 0.20357 0.81105
```

- (Min.): Η ελάχιστη προβλεπόμενη πιθανότητα (0.01700).
- (1st Qu.): Το 25% των προβλεπόμενων δεδομένων είναι μικρότερο του (0.06394) ενώ το 75% είναι μεγαλύτερο από αυτόν τον αριθμό
- (Median): Το 50% είναι μικρότερο του (0.11459) και το άλλο μισό μεγαλύτερο.
- (Mean): Η μέση προβλεπόμενη πιθανότητα (0.15213).
- (3rd Qu.): Το 75% είναι μικρότερο του (0.20357) και το υπόλοιπο μεγαλύτερο.
- (Max.): Η μέγιστη προβλεπόμενη πιθανότητα (0.81105).

```r
##Confusion Matrix showing us, 1070 false negatives and 15 true negatives and 182 false positives and 13 true positives.
table(test$TenYearCHD, predictTest > 0.5)

##
##     FALSE TRUE
##   0  1070   15
##   1   182   13

##Data already cleaned in the beginning no need for cleaning. No missing values.
library(ROCR)
ROCRpred <- prediction(predictTest, test$TenYearCHD)
ROCRperf <- performance(ROCRpred, "tpr", "fpr")
plot(ROCRperf,colorize = TRUE)
plot(ROCRperf,colorize = TRUE,
print.cutoffs.at=seq(0,1,0.1),text.adj=c(-0.2,1.7))
```




```r
as.numeric(performance(ROCRpred, "auc")@y.values)

## [1] 0.7346426

##Number of registrations in the "new" set are answered above and they are the following.
nrow(train)

## [1] 2378
```

```r
nrow(test)

## [1] 1280
```