Εργασία 7

2024-04-22

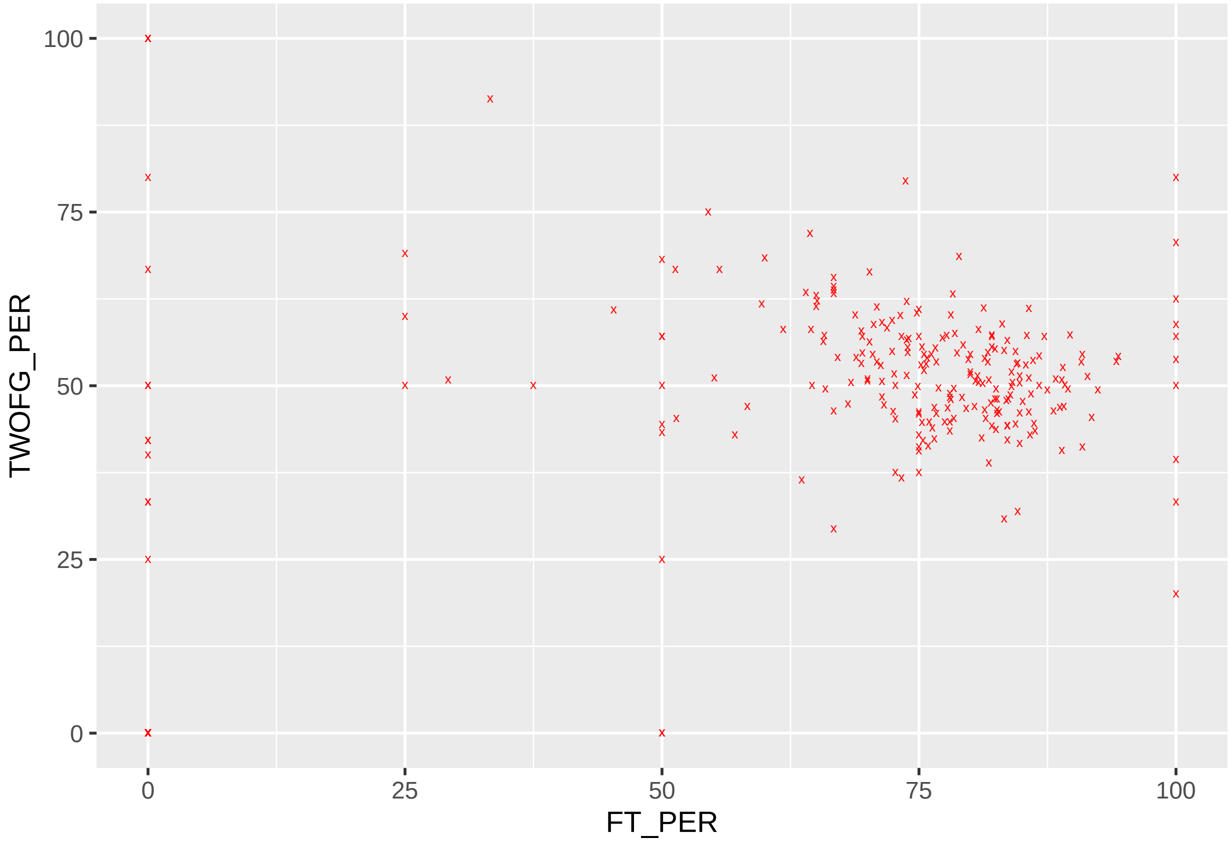
Άσκηση 005

1. Διερεύνηση του συνόλου δεδομένων

```
library(ggplot2)
library(readr)
euroleaguePlayers_average <- read_csv("F:/M    _Σ    /E       _A    /Exercises/Exercise_005/euroleaguePl
```

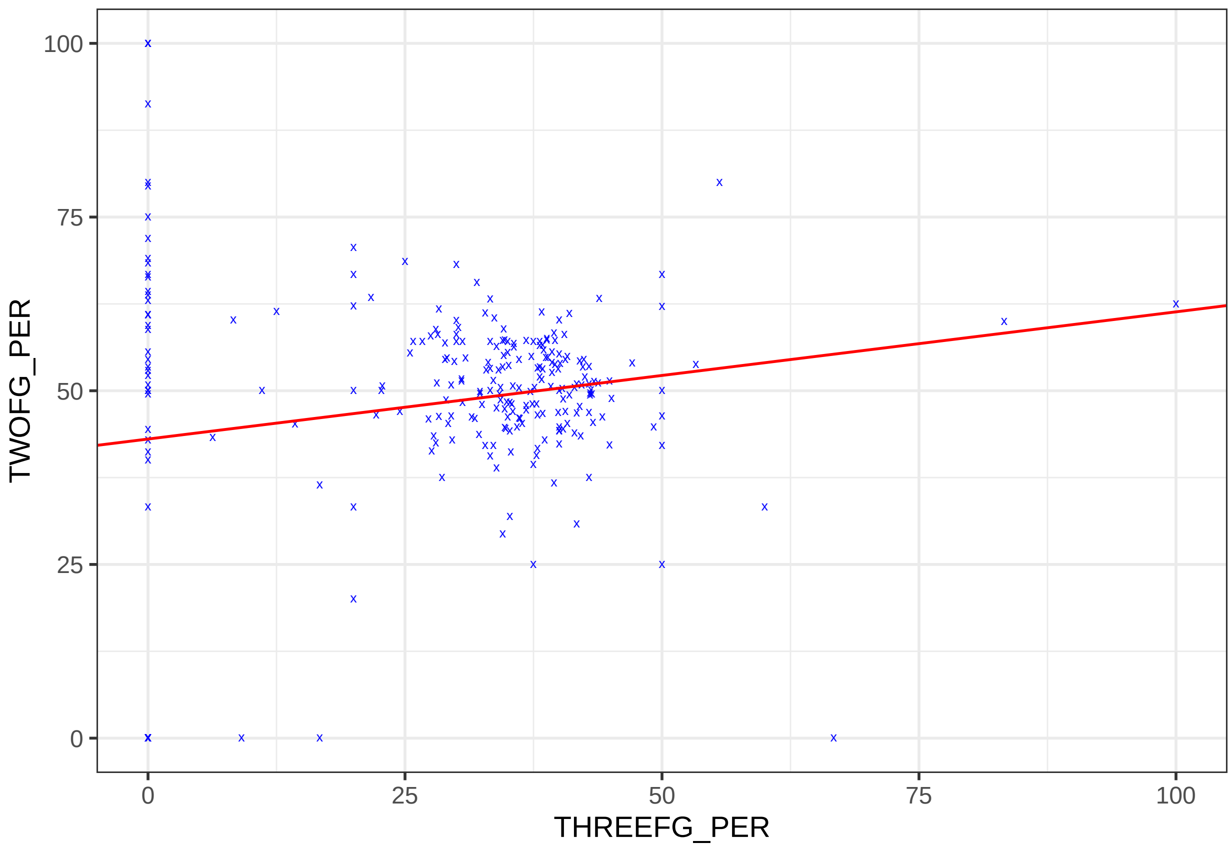
```
## Rows: 281 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): full_name, surname, name, Club, Position, Born, Nationality, 2FG, ...
## dbl (12): Dorsal, Height, G, Pts, Avg, 2FG_%, 3FG_%, FT_%, Reb, St, As, Bl
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
colnames(euroleaguePlayers_average) <- c('full_name', 'surname', 'name', 'Club', 'Dorsal', 'Position',
ggplot(euroleaguePlayers_average, aes(FT_PER, TWOFG_PER))+geom_point(shape="x",color="red", na.rm = TRU
```



2α. Δημιουργία μοντέλου (γραμμικής) παλινδρόμησης

```
model1 <- lm(TWOFG_PER ~ THREEFG_PER, euroleaguePlayers_average)
ggplot(euroleaguePlayers_average, aes(THREEFG_PER, TWOFG_PER))+ geom_point(shape="x",color="blue", na.r
```



2β. Αξιολόγηση μοντέλου παλινδρόμησης

```
summary(model1)$r.squared
```

```
## [1] 0.03185629
```

```
SSE1 <- sum(model1$residuals^2)
SSE1
```

```
## [1] 75376.06
```

Η τιμή του R-squared μας δείχνει ότι η συσχέτιση μεταξύ της μεταβλητής TWOFG_PER και της ανεξάρτητης μεταβλητής THREEFG_PER είναι πολύ χαμηλή.

```
RMSE <- sqrt(SSE1/nrow(euroleaguePlayers_average))
RMSE
```

```
## [1] 16.3781
```

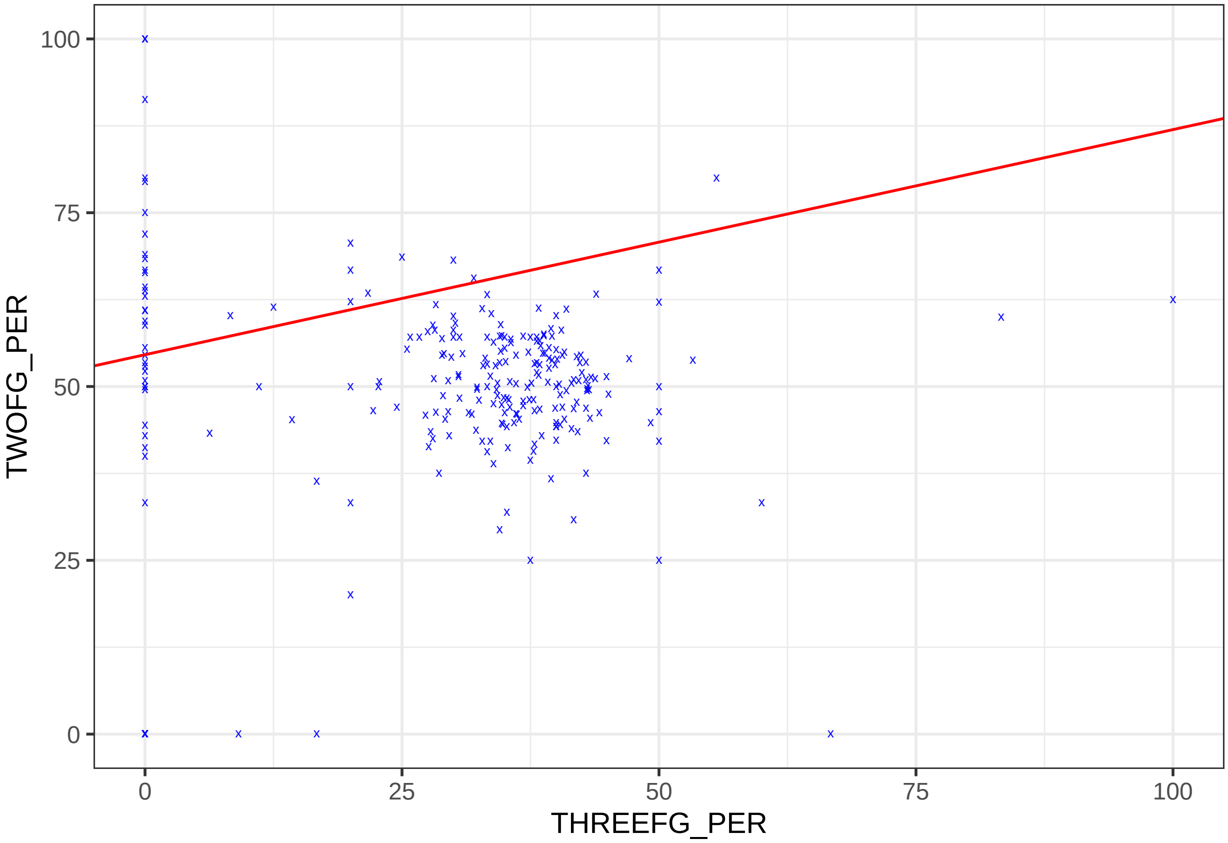
```
mean(euroleaguePlayers_average$TWOFG_PER, na.rm = TRUE)
```

```
## [1] 48.32703
```

Κατα μέσο όρο έχουμε σφάλμα 16.38 % για τις προσπάθειες δύο πόντων και ο μέσος όρος του ποσοστού των προσπαθειών δύο πόντων είναι 48.33% το οποίο φαίνεται πολύ αλλά δεν είναι.

2γ. Τροποποίηση μοντέλου παλινδρόμησης

```
model2 <- lm(TWOFG_PER ~ THREEFG_PER + Position, euroleaguePlayers_average)
ggplot(euroleaguePlayers_average, aes(THREEFG_PER, TWOFG_PER))+ geom_point(shape="x",color="blue", na.
```



```
summary(model2)
```

```
##
## Call:
## lm(formula = TWOFG_PER ~ THREEFG_PER + Position, data = euroleaguePlayers_average)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.124  -4.607   1.857   7.588  66.012
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        54.57510    2.42845  22.473  < 2e-16 ***
## THREEFG_PER         0.32381    0.05979   5.416 1.41e-07 ***
## PositionForward   -18.04947    2.78754  -6.475 4.84e-10 ***
## PositionGuard     -20.58733    2.73404  -7.530 8.78e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.48 on 255 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.2154, Adjusted R-squared:  0.2061
## F-statistic: 23.33 on 3 and 255 DF,  p-value: 2.234e-13
```

Αξιολόγηση μοντέλου

```
summary(model2)$r.squared
```

```
## [1] 0.2153671
```

```
SSE2 <- sum(model2$residuals^2)
SSE2
```

```
## [1] 61088.59
```

```
RMSE2 <- sqrt(SSE2/nrow(euroleaguePlayers_average))
RMSE2
```

```
## [1] 14.74439
```

```
mean(euroleaguePlayers_average$TWOFG_PER, na.rm = TRUE)
```

```
## [1] 48.32703
```

Παρατηρούμε ότι με την προσθήκη τη μεταβλητής Position βελτιώθηκε το μοντέλο καθώς αυξήθηκε το R-squared και μειώθηκαν το SSE και το RMSE.

Άσκηση 006

1. Διερεύνηση του συνόλου δεδομένων

```
library(caTools)
library(readr)
framingham <- read_csv("F:/M   _Σ    /E       _A    /Exercises/Exercise_006/framingham.csv",show_col_ty
set.seed(25)
split <- sample.split(framingham$TenYearCHD,SplitRatio=0.65)
train <- subset(framingham,split==TRUE)
test <- subset(framingham,split==FALSE)
```

Καταχωρίσεις training set:

```
nrow(train)
```

```
## [1] 2756
```

Καταχωρίσεις test set:

```r
nrow(test)
```

```
## [1] 1484
```

2α. Δημιουργία μοντέλου (λογιστικής) παλινδρόμησης

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

2β. Αξιολόγηση μοντέλου

```r
summary(framinghamLog)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = framingham)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.328186   0.715449 -11.641  < 2e-16 ***
## male             0.555279   0.109033   5.093 3.53e-07 ***
## age              0.063515   0.006679   9.509  < 2e-16 ***
## education       -0.047767   0.049395  -0.967  0.33353
## currentSmoker    0.071601   0.156752   0.457  0.64783
## cigsPerDay       0.017914   0.006238   2.872  0.00408 **
## BPMeds           0.162496   0.234326   0.693  0.48802
## prevalentStroke  0.693660   0.489569   1.417  0.15652
## prevalentHyp     0.234208   0.138026   1.697  0.08973 .
## diabetes         0.039167   0.315506   0.124  0.90120
## totChol          0.002332   0.001127   2.070  0.03850 *
## sysBP            0.015403   0.003808   4.044 5.24e-05 ***
## diaBP           -0.004159   0.006438  -0.646  0.51831
## BMI              0.006672   0.012758   0.523  0.60097
## heartRate       -0.003246   0.004211  -0.771  0.44082
## glucose          0.007127   0.002234   3.190  0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3121.2  on 3657  degrees of freedom
## Residual deviance: 2754.5  on 3642  degrees of freedom
##   (582 observations deleted due to missingness)
## AIC: 2786.5
##
## Number of Fisher Scoring iterations: 5
```

3. Εφαρμογή πρόβλεψης

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

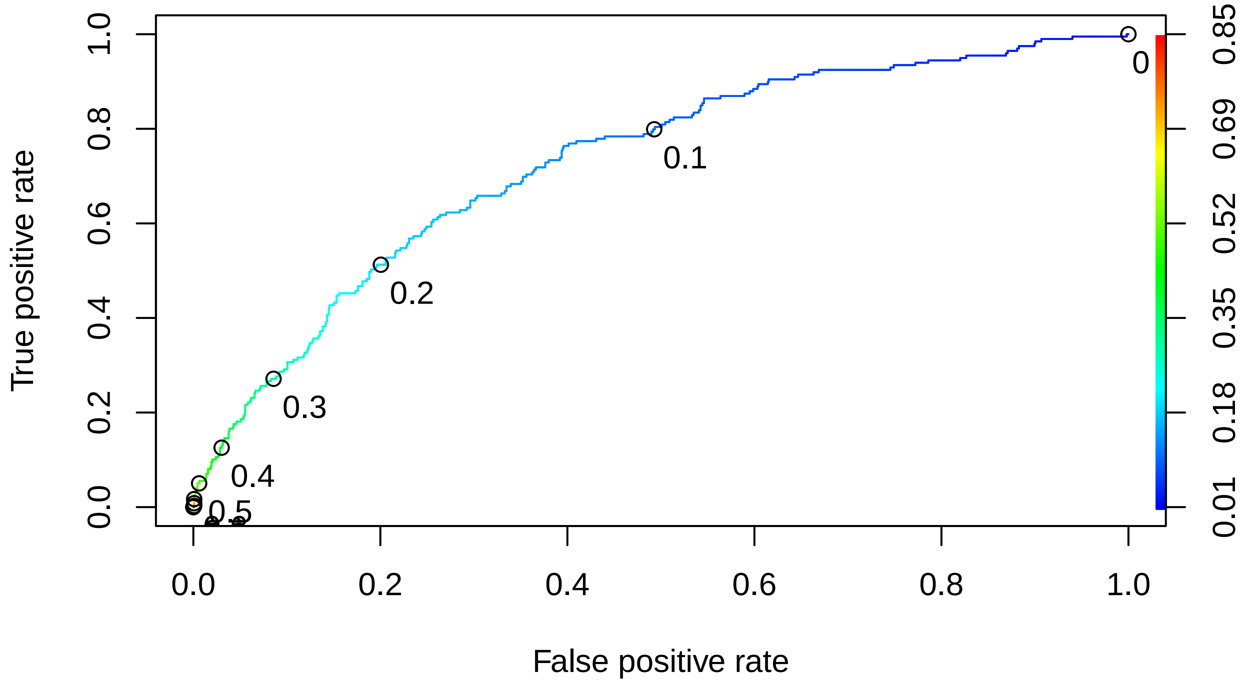
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01504 0.06503 0.11601 0.14915 0.20097 0.85298     201
```

```
table(test$TenYearCHD, predictTest > 0.5)
```

```
##
##     FALSE TRUE
##   0  1078    6
##   1   183   16
```

Area Under the Curve (AUC)

```
library(ROCR)
train2 = na.omit(train)
test2 = na.omit(test)
framinghamLog2 = glm(TenYearCHD ~ ., data = train2, family = binomial)
predictTest2 = predict(framinghamLog2, type = "response", newdata = test2)
ROCRpred2 <- prediction(predictTest2, test2$TenYearCHD)
ROCRperf2 <- performance(ROCRpred2, 'tpr', 'fpr')
plot(ROCRperf2,colorize = TRUE, print.cutoffs.at=seq(0,1,0.1),text.adj=c(-0.2,1.7))
```



```
as.numeric(performance(ROCRpred2, "auc")@y.values)
```

```
## [1] 0.7292412
```