doğrusal regresyon giriş

ilke

2022-08-04

```
library(pls)
library(tidyverse)
library(elasticnet)
library(broom)
library(glmnet)
library(MASS)
library(ISLR)
library(PerformanceAnalytics)
library(funModeling)
library(Matrix)
library(readxl)
```

makina öğrenmesi nedir? Bilgisayarları insan gibi düşünebilen ve analiz edebilen bir yapı haline getirmekiçin, istatistiksel yöntemlere dayanan öğretme tekniklerinin tamamıdır. Bu sekilde, insanların analiz edemeyeceği miktardaki büyük ve karmaşık haldeki verileri analiz edebilir hale getimektir.

doğrusal regresyon: Bir bağımlı değişken ile bir veya birden fazla değişken arasındaki doğrusal ilişkiyi modellemeye yarayan istatistiksel bir yöntemdir.

```
getwd()
```

```
## [1] "/home/ilke/Documents/github/r"
```

```
setwd("/home/ilke/Downloads")

veri_<- read.csv("Hitters.csv",sep=",",header=TRUE,stringsAsFactors = FALSE)

head(veri_)
```

```
##   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks
## 1   293   66     1   30  29    14     1    293    66      1    30   29     14
## 2   315   81     7   24  38    39    14   3449   835     69   321  414    375
## 3   479  130    18   66  72    76     3   1624   457     63   224  266    263
## 4   496  141    20   65  78    37    11   5628  1575    225   828  838    354
## 5   321   87    10   39  42    30     2    396   101     12    48   46     33
## 6   594  169     4   74  51    35    11   4408  1133     19   501  336    194
##   League Division PutOuts Assists Errors Salary NewLeague
## 1      A        E     446      33     20     NA         A
## 2      N        W     632      43     10  475.0         N
## 3      A        W     880      82     14  480.0         A
## 4      N        E     200      11      3  500.0         N
## 5      N        E     805      40      4   91.5         N
## 6      A        W     282     421     25  750.0         A
```

```
son_maaş<- as.numeric(veri_$Salary)
vuruş_sayısı<- as.numeric(veri_$CAtBat)
isabet_sayısı<- as.numeric(veri_$CHits)
değerli_vuruş<- as.numeric(veri_$CHmRun)
kazanılan_sayı<- as.numeric(veri_$CRuns)
yaptırılan_hata<- as.numeric(veri_$CWalks)
major_lig<- as.numeric(veri_$Years)
yardımlaşma<- as.numeric(veri_$PutOuts)

veri<- data.frame(son_maaş, vuruş_sayısı, isabet_sayısı, değerli_vuruş, kazanılan_sayı, yaptırılan_hata, major_lig, yardımlaşma)
```

eksik verilerin tespit edilmesi

```
sum(is.na(veri))
```

```
## [1] 59
```

```
veri <- na.omit(veri)
```

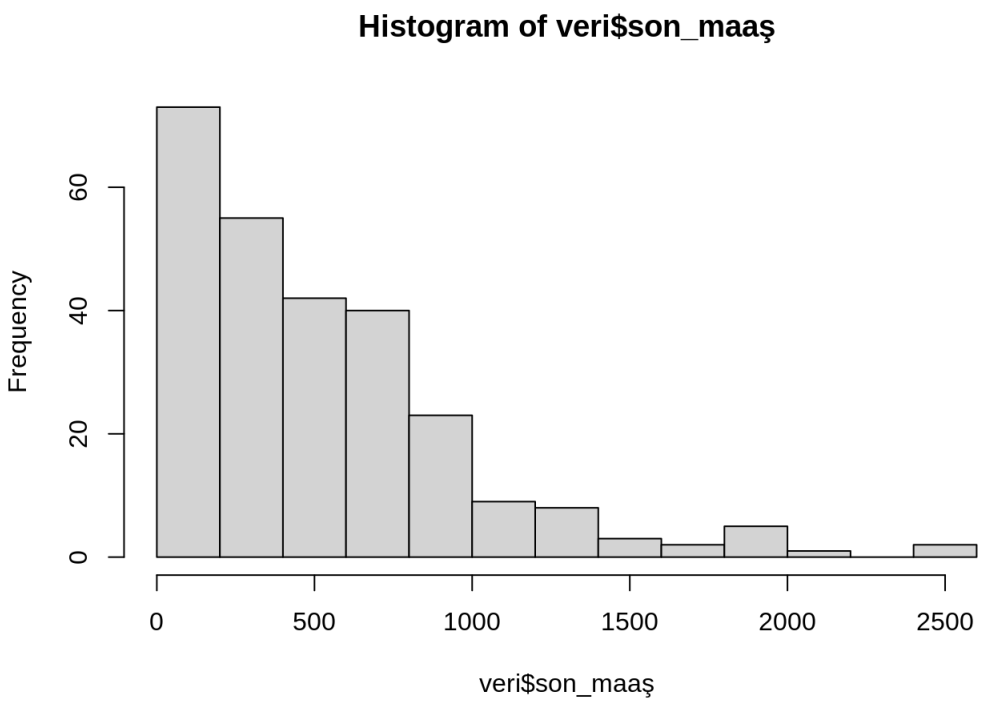
bağımlı değişken için aykırı gözlem tespiti

```
summary(son_maaş)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    67.5   190.0   425.0   535.9   750.0  2460.0      59
```

```
#normal veya yakın davranması: min mod median değerleri doğru üzerinde yakın Normal dağılımın en bilindik özelliklerinden bir tanesi, bu dağılımın simetrik olmasıdır. Dolayısıyla Mean, Mode, Median istatistikleri birbirine yakın değerler almaktadır.
```

```
hist(x = veri$son_maaş,freq = T)
```




```
which(son_maaş>800.000)
```

```
##  [1]  10  30  36  50  51  54  60  66  73  75  76  83  85  87  92  97 101 109 111
## [20] 113 116 130 137 143 146 149 164 171 178 179 180 181 185 190 218 219 230 235
## [39] 244 249 261 272 279 287 294 296 301 305 311 314 319 321 322
```

```
veri<- veri[-which(son_maaş>700.000),]
```

Keşifci veri analizi

```
glimpse(veri)  #veri seti genel yapı hakkında bilgi, dbl sürekli
```

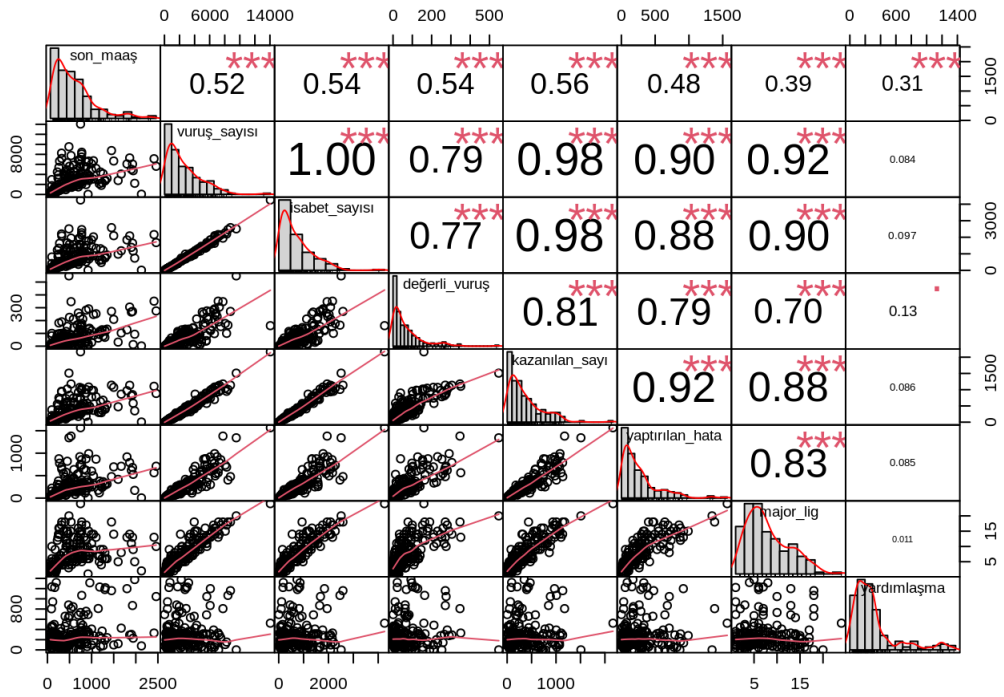
```
## Rows: 201
## Columns: 8
## $ son_maaş        <dbl> 475.000, 480.000, 500.000, 91.500, 750.000, 100.000, 7…
## $ vuruş_sayısı    <dbl> 3449, 1624, 5628, 396, 4408, 509, 341, 5206, 1876, 151…
## $ isabet_sayısı   <dbl> 835, 457, 1575, 101, 1133, 108, 86, 1332, 467, 392, 51…
## $ değerli_vuruş   <dbl> 69, 63, 225, 12, 19, 0, 6, 253, 15, 41, 4, 36, 177, 5,…
## $ kazanılan_sayı  <dbl> 321, 224, 828, 48, 501, 41, 32, 784, 192, 205, 309, 37…
## $ yaptırılan_hata <dbl> 375, 263, 354, 33, 194, 12, 8, 866, 161, 203, 207, 238…
## $ major_lig       <dbl> 14, 3, 11, 2, 11, 3, 2, 13, 9, 4, 6, 13, 15, 5, 1, 1, …
## $ yardımlaşma     <dbl> 632, 880, 200, 805, 282, 121, 143, 0, 304, 211, 121, 8…
```

```
profiling_num(veri) #sürekli değişkenlerin özet istatistiklerini verir
```

```
##          variable        mean     std_dev variation_coef p_01  p_05 p_25 p_50
## 1        son_maaş  567.709104  483.859272      0.8523014   70  87.5  191  450
## 2    vuruş_sayısı 2688.507463 2330.654258      0.8668952  196 278.0  822 1941
## 3   isabet_sayısı  732.278607  665.857389      0.9092952   44  68.0  210  516
## 4   değerli_vuruş   70.213930   82.733844      1.1783110    1   2.0   13   41
## 5  kazanılan_sayı  366.447761  341.444430      0.9317684   16  27.0   99  250
## 6 yaptırılan_hata  266.134328  270.109953      1.0149384    8  18.0   76  178
## 7       major_lig    7.378109    4.768262      0.6462715    1   1.0    4    6
## 8     yardımlaşma  308.124378  297.352568      0.9650407    0  28.0  118  227
##       p_75 p_95     p_99  skewness kurtosis      iqr       range_98    range_80
## 1  773.333 1600 2127.333 1.5020690 5.393123  582.333 [70, 2127.333] [100, 1200]
## 2 3949.000 7127 8759.000 1.3194890 5.189294 3127.000    [196, 8759] [396, 6100]
## 3 1077.000 2081 2510.000 1.4911439 6.258190  867.000     [44, 2510] [101, 1661]
## 4   97.000  259  347.000 2.1686828 9.222213   84.000       [1, 347]    [4, 177]
## 5  518.000 1019 1175.000 1.5394653 6.331454  419.000     [16, 1175]   [45, 897]
## 6  340.000  820 1342.000 1.8774305 7.289176  264.000      [8, 1342]   [33, 644]
## 7   10.000   16   18.000 0.7826254 2.965017    6.000        [1, 18]     [2, 14]
## 8  331.000 1067 1314.000 1.8956193 6.200813  213.000      [0, 1314]   [65, 732]
```

```
chart.Correlation(veri)
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```




```
model <- lm(veri$son_maaş~
            veri$vuruş_sayısı+
            veri$isabet_sayısı+
            veri$değerli_vuruş+
            veri$kazanılan_sayı+
            veri$yaptırılan_hata+
            veri$major_lig+
            veri$yardımlaşma)

summary(model)
```

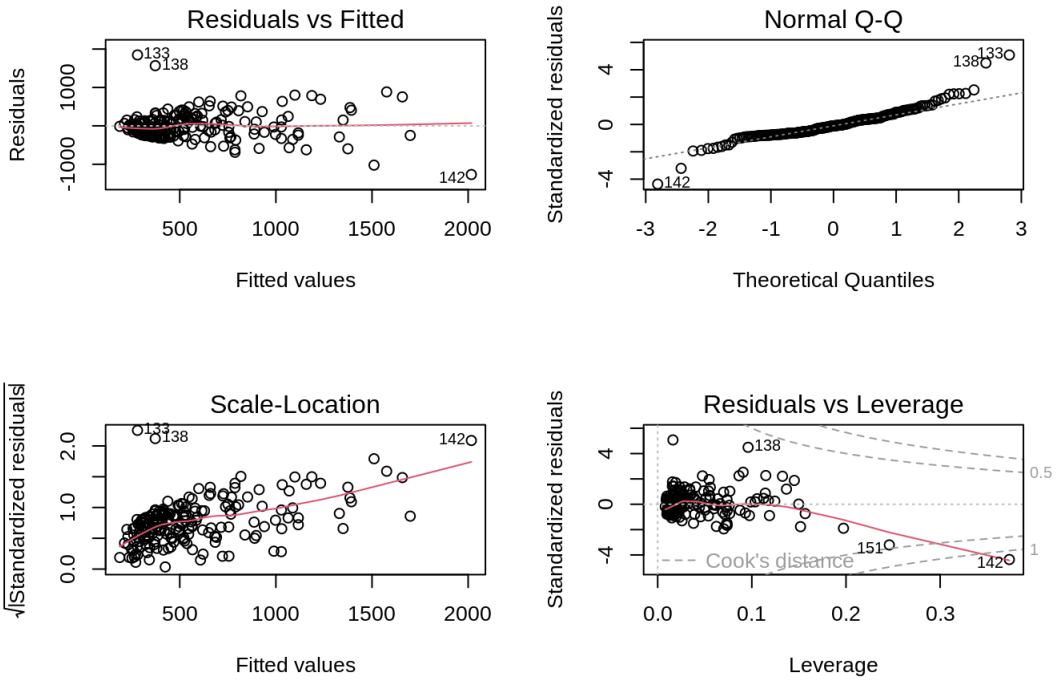
```
##
## Call:
## lm(formula = veri$son_maaş ~ veri$vuruş_sayısı + veri$isabet_sayısı +
##     veri$değerli_vuruş + veri$kazanılan_sayı + veri$yaptırılan_hata +
##     veri$major_lig + veri$yardımlaşma)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1266.67  -234.84   -40.79   159.50  1847.09
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          293.15421   61.02906   4.804 3.12e-06 ***
## veri$vuruş_sayısı     -0.33638    0.15043  -2.236  0.02648 *
## veri$isabet_sayısı     1.20693    0.56885   2.122  0.03514 *
## veri$değerli_vuruş     1.63321    0.57732   2.829  0.00516 **
## veri$kazanılan_sayı    0.75943    0.65193   1.165  0.24549
## veri$yaptırılan_hata  -0.15573    0.31403  -0.496  0.62051
## veri$major_lig       -21.94486   14.92506  -1.470  0.14310
## veri$yardımlaşma       0.34241    0.09119   3.755  0.00023 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 366.9 on 193 degrees of freedom
## Multiple R-squared:  0.445,  Adjusted R-squared:  0.4249
## F-statistic: 22.11 on 7 and 193 DF,  p-value: < 2.2e-16
```

```
confint(model) #modele eklenen değişkenlerin anlamlılıklarını incelemek mümkün. Bu fonksiyon bir sayı aralığı döndürür ve bu aralığın 0 değerini içermem
esi beklenir. Eğer 0 değerini içeriyorsa değişken o güven düzeyinde anlamsızdır yorumu yapılabilir.
```

```
##                          2.5 %       97.5 %
## (Intercept)         172.78466402 413.52376311
## veri$vuruş_sayısı    -0.63307408  -0.03969166
## veri$isabet_sayısı    0.08496593   2.32889274
## veri$değerli_vuruş    0.49455067   2.77186716
## veri$kazanılan_sayı  -0.52638101   2.04524878
## veri$yaptırılan_hata -0.77510002   0.46363439
## veri$major_lig      -51.38203497   7.49230581
## veri$yardımlaşma      0.16254579   0.52227384
```

#vuruş sayısı ve isabet sayısı, yardımlaşma anlamlıdır diyebiliriz.

```
par(mfrow=c(2,2))
plot(model)
```



#1. Grafik için; Varyans homojenliği var mı yok mu diye bakıyoruz. Noktaların 0 etrafında rasgele dağılması istenir – gözlem sayısının az olması sebebi ile grafikler subjektiftir, yapılan yorumlar yanlıdır.
#2. Grafik için; Artıkların normal dağılıp dağılmadığını belirtir. Bir doğru üzerinde olması istenir. Görsel olarak normal dağılıyor denilebilir ama test yapılmas ı gerekmektedir.
#3. Grafik için; Standartlaştırılmış artık değerler için ve fitted valuelar için inceliyoruz. 1. grafikle benzer yapıdalar.
*#4. Grafik için; Standardizied residuals kısmı uç değerlerin veya aykırı gözlem etrafında olup olmadığını gösterir. Standardized gözlemlerin +3 ve -3 değerleri arasında olup olmadığı incelenir. 4. Grafik için; Leverage ile etkin gözlem olup olmadığına bakıyoruz.- Leverege için kriter 2*p/n. Hesaplanan değ er bu kriteri geçiyor ise etkin gözlemdir. Yani modelin başarısını doğrudan etkiler yorumu yapılabilir.*