Linear Regression with Hitters Data

İlkay Koyuncuoğlu

10/13/2020

Firstly, we load 'ISLR' package for Hitters data. And required packages must load.

```
library(DataExplorer) #for visulation NA
library(ISLR) # for hitters data
library("caret")
library("ISLR")
library("olsrr")
library("dplyr")
library("Hmisc")
```

Transfer to 'Hitters' data to R enviroment.

```
hitters <- Hitters
```

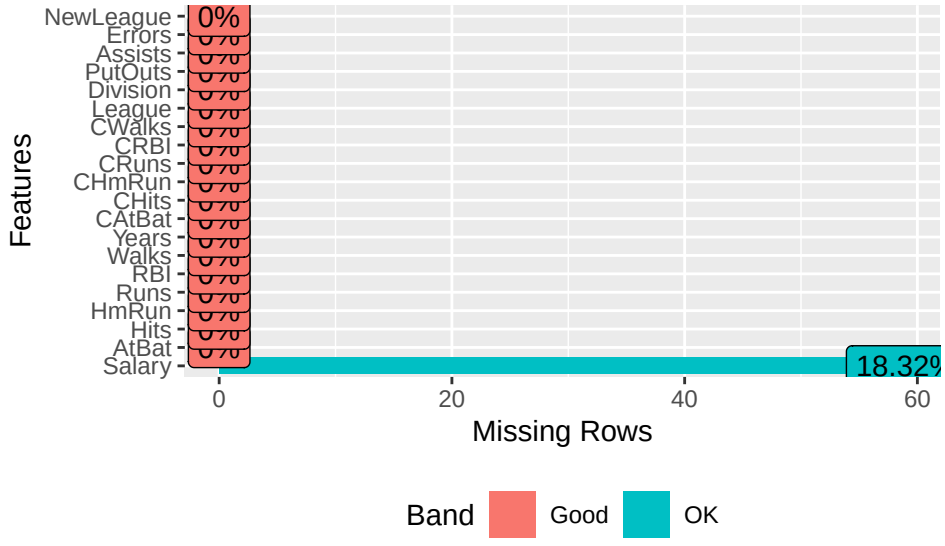
We will see if there is any 'NA'. Then, visulation to 'NA' data.

As can be seen in graphic, 59 observations are missing in the 'salary' variable.

```
sum(is.na(hitters))
```

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

```
plot_missing(hitters)
```



Missing data can be filled in with mean, median, or mode values. The only disadvantage is that the variable can be said to decrease the variance value. But it is the most preferred method.

```
hitters[is.na(hitters)]  <- mean(na.omit(hitters$Salary))
sum(is.na(hitters))
```

```
## [1] 0
```

We'r gonna build us first model for comparison to other models.

H_1 : β_{CATBAT}=β_{CRuns}=β_{Wlaks}=β_{Years}=$\beta_{devisionW}$=β_{CRBI}=β_{Walks}=$\beta_{NewLeaugeW}$=$\beta_{Putouts}$=0 H_0 : En az bir β_j!=0

H_0 is reject. The model is significant since f $<$ 0.05.

$H_0 : \beta_j = 0$ $H_1 : \beta_j != 0$ j= CRuns, Walks, Years, Division, CRBI, CWalks, NewLeague, CAtBat, PutOuts

With 95% confidence, the variables of CRUNS, Walks, Division, CRBI, CWALKS and PUTOUTS were significant. The explanatoryity of the model is 0.4096.

```
data=hitters[,-c(1,2,3,4,5,9,10,14,17,18)]
model <- lm(Salary~CRuns+Walks+Years+Division+CRBI+CWalks+NewLeague+CAtBat+PutOuts,data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Salary ~ CRuns + Walks + Years + Division + CRBI +
##     CWalks + NewLeague + CAtBat + PutOuts, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -787.07 -190.72  -29.27  175.47 1873.48
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  229.75372   61.42346   3.740 0.000219 ***
## CRuns          1.33872    0.36824   3.635 0.000324 ***
## Walks          3.88572    1.06464   3.650 0.000308 ***
## Years         -8.50461   10.51123  -0.809 0.419076
## DivisionW   -123.72407   35.80348  -3.456 0.000625 ***
## CRBI           0.41174    0.17594   2.340 0.019905 *
## CWalks        -0.62694    0.19796  -3.167 0.001693 **
## NewLeagueN    18.79198   36.35892   0.517 0.605629
## CAtBat        -0.09243    0.05843  -1.582 0.114674
## PutOuts        0.22687    0.06799   3.337 0.000950 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 317.6 on 312 degrees of freedom
## Multiple R-squared:  0.4096, Adjusted R-squared:  0.3926
## F-statistic: 24.05 on 9 and 312 DF,  p-value: < 2.2e-16
```

'set seed' function provides derivation of same random numbers.

smp_size variable is 70% of the data set. We took samples from the smp_size as much as sample size and transfered to 'train_ind'. ### We transfered as much as 'train_ind' to 'train' in the data. 'test' is what remains from the data set. ### Dimensions calculate with 'dim' function.

```
set.seed(123)
smp_size <- round(0.70 * nrow(data))
train_ind <- sample(nrow(data), size = smp_size, replace = FALSE)
train <- data[train_ind, ]
test <- data[-train_ind, ]

dim(train)
```

```
## [1] 225  10
```

```
dim(test)
```

```
## [1] 97 10
```

The Matrix Chart is used to evaluate the relationships between several pairs of variables at the same time. A matrix chart is a scatter chart series.

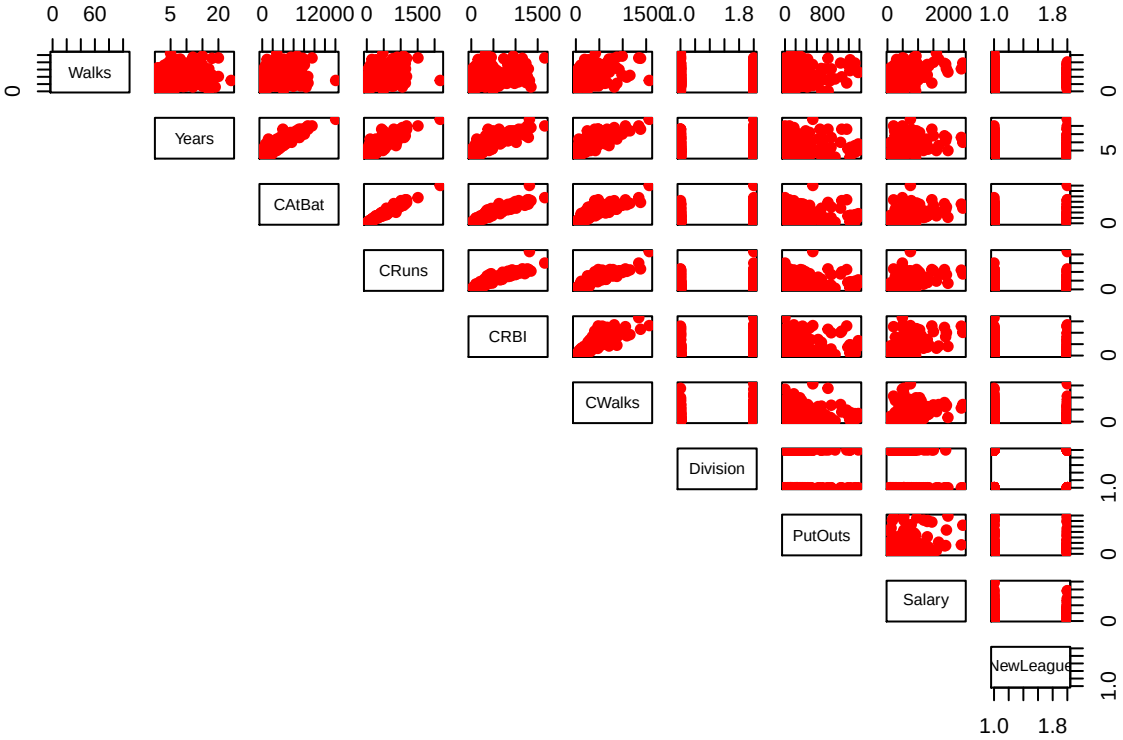
There is linear relationship between 'years, cArBat, CRuns, CRBI and CWalks' variables.

Desired situation, be relation between independent variables and dependent variable, but independent variables should not be related to each other.

The relationship between independent variables creates a multiple linear connection problem.

There is multiple linear connection problem. We'll solve.

```
pairs(train[,1:10], pch = 19, col='red', lower.panel = NULL)
```



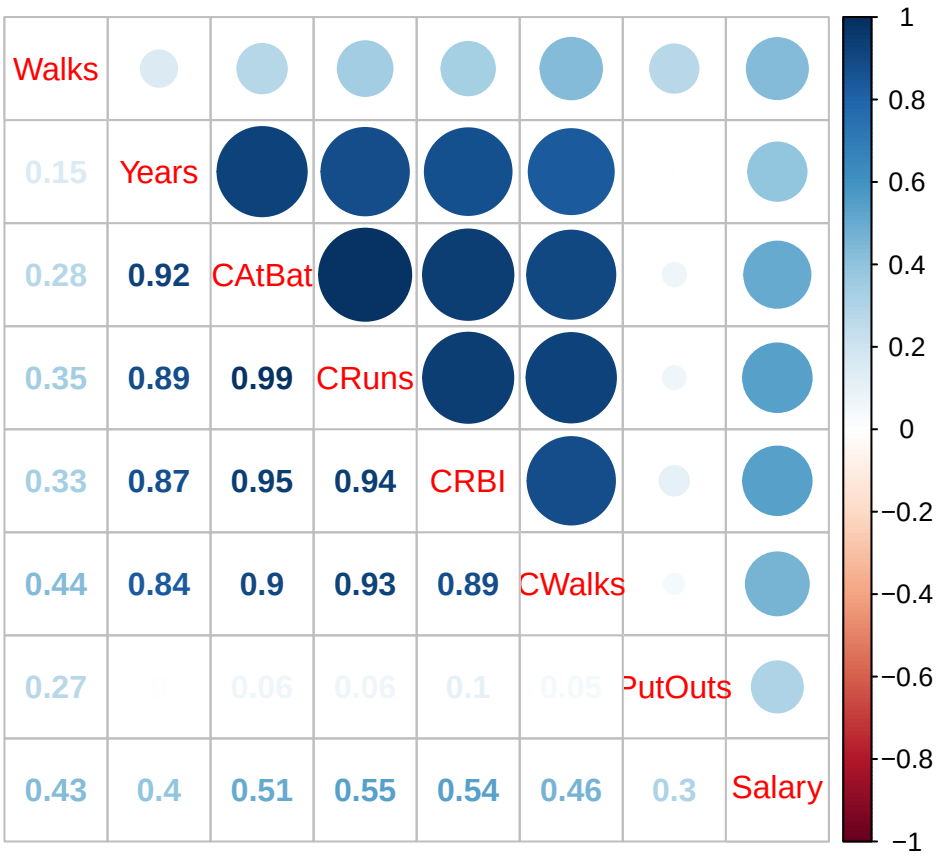
'corrplot' check relationship between variables as in matrix plot and as image, and 'rcorr' check relationship between variables as numbers.

As we said above, there is linear relationship between 'years, cArBat, CRuns, CRBI and CWalks' variables. We see in the image.

```
library("Hmisc")
data1 <- train[,-c(7,10)]
cor.data1=rcorr(as.matrix(data1))
cor.data1$r
```

```
##             Walks       Years     CAtBat     CRuns      CRBI     CWalks
## Walks   1.0000000 0.153154371 0.28342861 0.3472420 0.3310125 0.43908825
## Years   0.1531544 1.000000000 0.92438895 0.8867040 0.8722676 0.83907778
## CAtBat  0.2834286 0.924388948 1.00000000 0.9850518 0.9485960 0.90191679
## CRuns   0.3472420 0.886703975 0.98505179 1.0000000 0.9435783 0.92754482
## CRBI    0.3310125 0.872267639 0.94859597 0.9435783 1.0000000 0.88732515
## CWalks  0.4390882 0.839077783 0.90191679 0.9275448 0.8873251 1.00000000
## PutOuts 0.2712012 0.001815827 0.06442117 0.0616417 0.1010457 0.04823909
## Salary  0.4328496 0.395114813 0.50908221 0.5464626 0.5448355 0.46026376
##             PutOuts    Salary
## Walks   0.271201235 0.4328496
## Years   0.001815827 0.3951148
## CAtBat  0.064421168 0.5090822
## CRuns   0.061641702 0.5464626
## CRBI    0.101045668 0.5448355
## CWalks  0.048239087 0.4602638
## PutOuts 1.000000000 0.3032006
## Salary  0.303200560 1.0000000
```

```
library(corrplot)
corrplot.mixed(cor(data1))
```



The model is created with the data obtained from the 'training' data set.

H_1 : β_{CATBAT}=β_{CRuns}=β_{Wlaks}=β_{Years}=$\beta_{devisionW}$=β_{CRBI}=β_{Walks}=$\beta_{NewLeaugeW}$=$\beta_{Putouts}$=0 H_0 : En az bir β_j!=0

H_1 : β_j =0 H_1 : β_j !=0 j=CRuns,Walks,Division,CRBI,CWalks,NewLeague,CAtBat,PutOuts

years variable's significance value is 0.819 and NewLeagueN variable's significance value is 0.787.

Other varibales are significant with 95% confidence. Because, p-value's < 0.05.

The explanatoryity of the model is 0.4968.

```
mfulltrain=lm(Salary~CRuns+Years+Walks+Division+CRBI+CWalks+NewLeague+CAtBat+PutOuts, data=train)
summary(mfulltrain)
```

```
##
## Call:
## lm(formula = Salary ~ CRuns + Years + Walks + Division + CRBI +
##     CWalks + NewLeague + CAtBat + PutOuts, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -824.19 -192.56  -26.27  190.15 1077.21
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  176.12632   70.74954   2.489 0.013553 *
## CRuns          1.93624    0.45093   4.294 2.66e-05 ***
## Years          2.86540   12.56031   0.228 0.819761
## Walks          4.34076    1.19698   3.626 0.000359 ***
## DivisionW   -102.32986   41.59775  -2.460 0.014683 *
## CRBI           0.51087    0.19389   2.635 0.009032 **
## CWalks        -0.97618    0.22204  -4.397 1.73e-05 ***
## NewLeagueN   -11.36592   42.10091  -0.270 0.787444
## CAtBat        -0.17612    0.07298  -2.413 0.016650 *
## PutOuts        0.27929    0.07509   3.720 0.000255 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 304.8 on 215 degrees of freedom
## Multiple R-squared:  0.4968, Adjusted R-squared:  0.4757
## F-statistic: 23.59 on 9 and 215 DF,  p-value: < 2.2e-16
```

To determine the variables that create a multiple linear connection problem, vif values are checked. CRuns, Years, CRBI, CWalks and CAtBat is the variables that create a multiple linear connection problem.

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
## Loading required package: carData
## Warning: package 'carData' was built under R version 3.6.2
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
vif(mfulltrain)
```

```
##     CRuns     Years     Walks  Division      CRBI    CWalks NewLeague    CAtBat
## 61.720216  9.582501  1.708903  1.046844 10.937239  9.502808  1.044336 75.107524
##   PutOuts
##  1.139863
```

Not all of these variables are removed from the model. Variables which have multiple linear connection problem between them are combined to get the best model. For this, the 'ols_step_all_possible' function is used. '466., 502. and 503.' models are the best it has the value of R^2 and cp.

466. model inclued CRuns, Walks, Division, CRBI, CWalks, CAtBat and PutOuts variables. 'cp' value is 6.1399. R^2 value is 0.4964.
467. model inclued CRuns, Walks, Division, CRBI, CWalks,NewLeauge, CAtBat and PutOuts variables. 'cp' value is 8.05. R^2 value is 0.4966.

The purpose is to obtain the most explanatory information with less variables.

```
k=ols_step_all_possible(mfulltrain)

k$predictors[466]
```

```
## [1] "CRuns Walks Division CRBI CWalks CAtBat PutOuts"
```

```
k$predictors[502]
```

```
## [1] "CRuns Walks Division CRBI CWalks NewLeague CAtBat PutOuts"
```

These models are created with determinated variables.

Press values of the models are very close to each other. However, the model with the most explanations is the 466th model. The choose model is 466th model.

```
m_502_press <-lm(Salary~Walks+Division+CRBI+PutOuts+CRuns+CWalks+NewLeague+CAtBat, data=train)
m_466_press <-lm(Salary~Walks+Division+CRBI+CRuns+CWalks+PutOuts+CAtBat, data=train)
predictions1=predict(m_502_press,test)
predictions2=predict(m_466_press,test)
PRESS1 = RMSE(predictions1, test$Salary)
PRESS2 = RMSE(predictions2, test$Salary)
PRESS1 # 353.6381
```

```
## [1] 353.6381
```

```
PRESS2 # 352.6942
```

```
## [1] 352.6942
```

ASSUMPTION

Assumption 1: Linear relationship between Independent and dependent variables

H_0: There is Linear relationship between Independent and dependent variables. H_1: There is no Linear relationship between Independent and dependent variables.

p-value $= 0.01338 < 0.05$.

The fact that the residues are not suitable for normal distribution affects the analysis. However, it is difficult to obtain data that will fit the normal distribution. So when the p-value is very close to 0.05, it can be assumed to be normal.

```
shapiro.test(m_502_press$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m_502_press$residuals
## W = 0.98425, p-value = 0.01338
```

Assumption 2: Number of observations should be greater than number of independent variables. This is multiple linear connection problem.

```
vif(m_502_press)
```

```
##     Walks  Division      CRBI   PutOuts     CRuns    CWalks NewLeague    CAtBat
##  1.601517  1.039299 10.937206  1.131832 50.777397  8.394115  1.031975 47.419870
```