regresyon\_analizi

## Soru1 Regression

library(caret)

library(Matrix)  
library(olsrr)  
library("car")  
library("ISLR")  
library(broom) #tidy model icin  
library(glmnet)  
library(MASS)  
library(ISLR)  
library("knitr")  
library("rmarkdown")  
library("corrplot")  
library("Hmisc")  
library("corrplot")  
library("car")  
library("olsrr")  
library("GGally")  
library(nortest)

## veri işleme \_veriye genel bakış

# 

df <- ISLR::Hitters  
rownames(df) <- c()#row namesleri sildim veriye daha güzel bakabiliriz.  
glimpse(df)

## Observations: 322  
## Variables: 20  
## $ AtBat <int> 293, 315, 479, 496, 321, 594, 185, 298, 323, 401, 574,…  
## $ Hits <int> 66, 81, 130, 141, 87, 169, 37, 73, 81, 92, 159, 53, 11…  
## $ HmRun <int> 1, 7, 18, 20, 10, 4, 1, 0, 6, 17, 21, 4, 13, 0, 7, 3, …  
## $ Runs <int> 30, 24, 66, 65, 39, 74, 23, 24, 26, 49, 107, 31, 48, 3…  
## $ RBI <int> 29, 38, 72, 78, 42, 51, 8, 24, 32, 66, 75, 26, 61, 11,…  
## $ Walks <int> 14, 39, 76, 37, 30, 35, 21, 7, 8, 65, 59, 27, 47, 22, …  
## $ Years <int> 1, 14, 3, 11, 2, 11, 2, 3, 2, 13, 10, 9, 4, 6, 13, 3, …  
## $ CAtBat <int> 293, 3449, 1624, 5628, 396, 4408, 214, 509, 341, 5206,…  
## $ CHits <int> 66, 835, 457, 1575, 101, 1133, 42, 108, 86, 1332, 1300…  
## $ CHmRun <int> 1, 69, 63, 225, 12, 19, 1, 0, 6, 253, 90, 15, 41, 4, 3…  
## $ CRuns <int> 30, 321, 224, 828, 48, 501, 30, 41, 32, 784, 702, 192,…  
## $ CRBI <int> 29, 414, 266, 838, 46, 336, 9, 37, 34, 890, 504, 186, …  
## $ CWalks <int> 14, 375, 263, 354, 33, 194, 24, 12, 8, 866, 488, 161, …  
## $ League <fct> A, N, A, N, N, A, N, A, N, A, A, N, N, A, N, A, N, A, …  
## $ Division <fct> E, W, W, E, E, W, E, W, W, E, E, W, E, E, E, W, W, W, …  
## $ PutOuts <int> 446, 632, 880, 200, 805, 282, 76, 121, 143, 0, 238, 30…  
## $ Assists <int> 33, 43, 82, 11, 40, 421, 127, 283, 290, 0, 445, 45, 11…  
## $ Errors <int> 20, 10, 14, 3, 4, 25, 7, 9, 19, 0, 22, 11, 7, 6, 8, 0,…  
## $ Salary <dbl> NA, 475.000, 480.000, 500.000, 91.500, 750.000, 70.000…  
## $ NewLeague <fct> A, N, A, N, N, A, A, A, N, A, A, N, N, A, N, A, N, A, …

# NA's olanları silelim  
df <- na.omit(df)  
sum(ifelse(is.na(df), 1, 0))

## [1] 0

#Columnları filtreyelim  
# CRuns-11,Walks-6,Years-7,Division,CRBI-12,CWalks-13,NewLeague-20,CAtBat,PutOuts en sona Salarayi attım  
df <- df[,c(6, 7, 8, 11, 12, 13, 15, 16, 20, 19)]  
head(df)

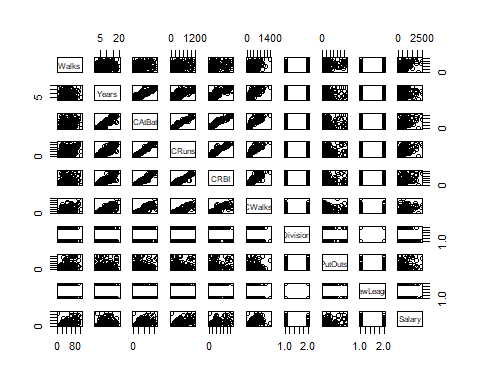
## Walks Years CAtBat CRuns CRBI CWalks Division PutOuts NewLeague Salary  
## 2 39 14 3449 321 414 375 W 632 N 475.0  
## 3 76 3 1624 224 266 263 W 880 A 480.0  
## 4 37 11 5628 828 838 354 E 200 N 500.0  
## 5 30 2 396 48 46 33 E 805 N 91.5  
## 6 35 11 4408 501 336 194 W 282 A 750.0  
## 7 21 2 214 30 9 24 E 76 A 70.0

Train-TEST

set.seed(123)  
smp\_size <- round(0.7 \* nrow(df))  
train\_ind <- sample(x = nrow(df),size = smp\_size,replace = F)  
train <- df[train\_ind,]  
test <- df[-train\_ind,]

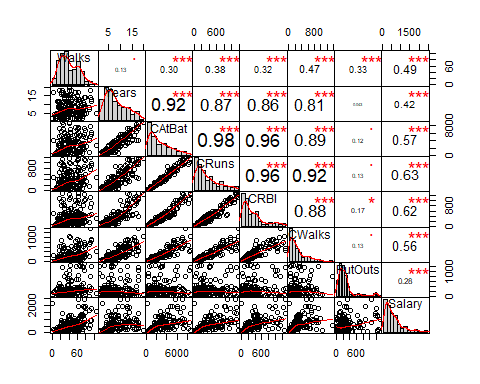
### 1) Matris Plot oluşturarak yorumlayınız

plot(train)



Gelişmiş Scatter plot:

chart.Correlation(train %>% dplyr::select(-c("NewLeague","Division")), histogram=TRUE, pch=19)



Yorum: Salary bağımlı değişkenim ile Cruns bağımsız değişknenim arasında pozitif yönlü güçlü ilşkivar.Scatter plottanda ikilinin yönünü ve şeklini görebiliyoruz doğrusalilerlemiş noktalar.

Bağımsız değişkenlerim arasında güçlü bir korelasyon ilşkisi var.Bu benim istemediğim bir durum.Bağımsız değişkenler arsaında güçlü ilşki olmamalı.İleride çoklu doğrusal bağlantı sorununa neden olabilir.

###2) Korelasyon matrisi ve korelasyon katsayılarına ilişkin p değerlerini matrisini elde ederek, %5 önem düzeyinde anlamlı korelasyon katsayılarını belirleyiniz.

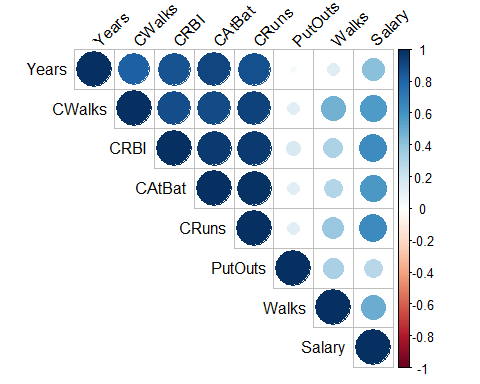
#ilk başta korelasyon .  
cormatris <- as.matrix(train[,-c(7,9)])  
matris1 <-Hmisc::rcorr(cormatris)  
matris1$r

## Walks Years CAtBat CRuns CRBI CWalks  
## Walks 1.0000000 0.13403070 0.2951302 0.3776355 0.3194478 0.4726044  
## Years 0.1340307 1.00000000 0.9164487 0.8711206 0.8607106 0.8105461  
## CAtBat 0.2951302 0.91644869 1.0000000 0.9806450 0.9572556 0.8903359  
## CRuns 0.3776355 0.87112059 0.9806450 1.0000000 0.9580767 0.9208436  
## CRBI 0.3194478 0.86071058 0.9572556 0.9580767 1.0000000 0.8847016  
## CWalks 0.4726044 0.81054610 0.8903359 0.9208436 0.8847016 1.0000000  
## PutOuts 0.3299292 0.04331099 0.1218142 0.1263307 0.1681641 0.1295375  
## Salary 0.4946395 0.41654500 0.5729843 0.6327602 0.6204514 0.5608001  
## PutOuts Salary  
## Walks 0.32992923 0.4946395  
## Years 0.04331099 0.4165450  
## CAtBat 0.12181418 0.5729843  
## CRuns 0.12633068 0.6327602  
## CRBI 0.16816410 0.6204514  
## CWalks 0.12953747 0.5608001  
## PutOuts 1.00000000 0.2797524  
## Salary 0.27975237 1.0000000

matris1 <-Hmisc::rcorr(cormatris)  
matris1$r

## Walks Years CAtBat CRuns CRBI CWalks  
## Walks 1.0000000 0.13403070 0.2951302 0.3776355 0.3194478 0.4726044  
## Years 0.1340307 1.00000000 0.9164487 0.8711206 0.8607106 0.8105461  
## CAtBat 0.2951302 0.91644869 1.0000000 0.9806450 0.9572556 0.8903359  
## CRuns 0.3776355 0.87112059 0.9806450 1.0000000 0.9580767 0.9208436  
## CRBI 0.3194478 0.86071058 0.9572556 0.9580767 1.0000000 0.8847016  
## CWalks 0.4726044 0.81054610 0.8903359 0.9208436 0.8847016 1.0000000  
## PutOuts 0.3299292 0.04331099 0.1218142 0.1263307 0.1681641 0.1295375  
## Salary 0.4946395 0.41654500 0.5729843 0.6327602 0.6204514 0.5608001  
## PutOuts Salary  
## Walks 0.32992923 0.4946395  
## Years 0.04331099 0.4165450  
## CAtBat 0.12181418 0.5729843  
## CRuns 0.12633068 0.6327602  
## CRBI 0.16816410 0.6204514  
## CWalks 0.12953747 0.5608001  
## PutOuts 1.00000000 0.2797524  
## Salary 0.27975237 1.0000000

corrplot::corrplot(matris1$r, type = "upper", order = "hclust",   
 tl.col = "black", tl.srt = 45)



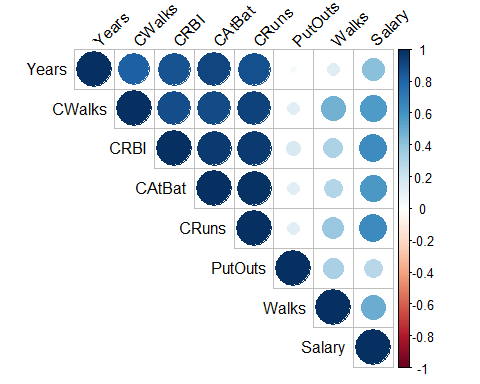
matris1$P

## Walks Years CAtBat CRuns CRBI  
## Walks NA 6.969558e-02 4.761757e-05 1.259916e-07 9.868250e-06  
## Years 6.969558e-02 NA 0.000000e+00 0.000000e+00 0.000000e+00  
## CAtBat 4.761757e-05 0.000000e+00 NA 0.000000e+00 0.000000e+00  
## CRuns 1.259916e-07 0.000000e+00 0.000000e+00 NA 0.000000e+00  
## CRBI 9.868250e-06 0.000000e+00 0.000000e+00 0.000000e+00 NA  
## CWalks 1.256639e-11 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00  
## PutOuts 4.794136e-06 5.593756e-01 9.950718e-02 8.748658e-02 2.250008e-02  
## Salary 9.561241e-13 4.070794e-09 0.000000e+00 0.000000e+00 0.000000e+00  
## CWalks PutOuts Salary  
## Walks 1.256639e-11 4.794136e-06 9.561241e-13  
## Years 0.000000e+00 5.593756e-01 4.070794e-09  
## CAtBat 0.000000e+00 9.950718e-02 0.000000e+00  
## CRuns 0.000000e+00 8.748658e-02 0.000000e+00  
## CRBI 0.000000e+00 2.250008e-02 0.000000e+00  
## CWalks NA 7.967949e-02 0.000000e+00  
## PutOuts 7.967949e-02 NA 1.200806e-04  
## Salary 0.000000e+00 1.200806e-04 NA

1987 yılında oyuncuların açılış maaşları(salary) bağımlı olduğu için onunla ilgili değerlerin yüksek çıkması normal olarak görülmektedir. p degerleri oldukça iyidir. Hepsi sıfıra yakınsamaktadır yani yaklaşık sıfır değeri almıştır önem düzeyinde anlamlı korelasyon katsayılarına sahiptir.Bağımlı “Salary” değişkenimiz ile diğer bağımsız değişkenlerin korelasyonu anlamlıdır diyebirim.

aşağıda görüğümüz Corrplotta yine maviler yüksek korelesyonlu pozitif yönlü ilişikiyi temsil etmektedir. Salary bağımsız değişknemile diğer bağımlı değişkenlerim arasında örneğin CWalksarasında pozitif yönlü güçlü bir ilşki vardır. Maaşın ile 1986’daki çıkış sayısı(Putouts)arasında pozitif yönlü güçlü olmayan bir ilişki var.

corrplot::corrplot(matris1$r, type = "upper", order = "hclust",   
 tl.col = "black", tl.srt = 45)



3)Çoklu doğrusal regresyon modelini elde ediniz ve model geçerliliğini sıfır ve alternatif hipotezleri belirterek %5 önem düzeyinde test ediniz.

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

##   
## Call:  
## lm(formula = Salary ~ Walks + Years + CAtBat + CRuns + CRBI +   
## CWalks + Division + PutOuts + NewLeague, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -803.91 -176.81 -2.54 155.23 1925.88   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 116.51658 86.03901 1.354 0.177419   
## Walks 5.16882 1.47090 3.514 0.000563 \*\*\*  
## Years -16.50774 14.59295 -1.131 0.259522   
## CAtBat -0.18520 0.08067 -2.296 0.022878 \*   
## CRuns 2.17272 0.54671 3.974 0.000103 \*\*\*  
## CRBI 0.79763 0.27953 2.853 0.004850 \*\*   
## CWalks -0.86064 0.26864 -3.204 0.001614 \*\*   
## DivisionW -138.64609 49.00632 -2.829 0.005216 \*\*   
## PutOuts 0.15030 0.09087 1.654 0.099928 .   
## NewLeagueN 98.63966 49.18452 2.006 0.046459 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 319.8 on 174 degrees of freedom  
## Multiple R-squared: 0.5892, Adjusted R-squared: 0.568   
## F-statistic: 27.73 on 9 and 174 DF, p-value: < 2.2e-16

1987 yılı lig açılış gününde yıllık maaş(Salary) tahmin etmek için; yürüyüş sayısı(walks),,years,Catbat (Kariyeri boyunca catbat sayısı),Crun(Kariyeri sırasındaki koşu sayısı),CRBI,CWalks ,Putouts, DivisionW,NewLeagueN değişkenleri ile full bir model oluşturuldu.

Değişkenlerden “Years”, “Putouts” ,b0 önem düzeyinde anlamlı değildir. Diğer değişkenlerin t istatistiklerinin olasılık değeri 0.05 aşağısında olduğundan anlamlı gözükmektedir.Modelde bulunan diğer değişkenler anlamlıdır. F istatistiği <2e-16<0.05 olduğu için H0 hipotezi red edilir ve en az bir tane ß 0 dan farklıdır yorumu yapılır.

Yıllık maaştaki(salary) değişimin %58’i açıklayıcı değişkenler tarafından açıklanmaktadır.

###4)Hipotezleri yazarak, elde ettiğiniz modeldeki anlamlı katsayıları belirleyiniz. Katsayıların Yorumları:

H0 : β0 (Sabit terimin modele katkısı anlamlı değildir.) H1 : β0 (Sabit terimin modele katkısı anlamlıdır.) p- value=0.177419 değeri 0.05’den büyük olduğundan dolayı H0 hipotezimiz reddedilemez yani sabit terimin *modele katkısı anlamlı değildir*.

Walks H0: ß1 = 0 H1: ß1 != 0 p = 0.000563 p-value değeri=0.000563 , 0.05’den küçük olduğu için H0 hipotezimiz reddedilir.yani Walksın *modele katkısı anlamlıdır.*

Years H0: ß2 = 0 H1: ß2 != 0  
p = 0.259522

p- value=0.259522 değeri 0.05’den büyük olduğundan dolayı H0 hipotezimiz reddedilemez yani Years modele*katkısı anlamlı değildir.*

CAtBat H0: ß3 = 0 H1: ß3 != 0 p = 0.022878 H0 hipotezimiz red. Değişken anlamlıdır.

CRuns H0: ß4 = 0 H1: ß4 != 0 p = 0.000103 \*\*\* p-value değeri=0.000103 , 0.05’den küçük olduğu için H0 hipotezimiz reddedilir.yani Crun *modele katkısı anlamlıdır.*

CRBI H0: ß5 = 0 H1: ß5 != 0 p = 0.004850 p-value değeri=0.0048 , 0.05’den küçük olduğu için H0 hipotezimiz reddedilir. *Değişken anlamlıdır.*

CWalks H0: ß6 = 0 H1: ß6 != 0 p = 0.001614 p-value değeri 0.001614 0.05’den küçük olduğu için H0 hipotezimiz reddedilir. *Değişkenimiz anlamlıdır.*

DivisionW H0: ß7 = 0 H1: ß7 != 0 p = 0.005216 p-value değeri=0.005216 \*\* , 0.05’den küçük olduğu için H0 hipotezimiz reddedilir.yani Divisionın *modele katkısı anlamlıdır.*

PutOuts H0: ß8 = 0 H1: ß8 != 0 p = 0.099928 p-value değeri 0.099928 0.05’den büyük olduğu için H0 hipotezimiz reddedilemez. Değişken anlamlı değildir *modele katkısı yoktur.* .

NewLeagueN H0: ß9 = 0 H1: ß9 != 0 p = 0.046459 p-value değeri 0.046459 0.05’den küçük olduğu için H0 hipotezimiz reddedilir. *Değişkenimiz anlamlıdır.*

##5. VIF değerlerini hesaplayınız ve yorumlayınız.

vif(model)

## Walks Years CAtBat CRuns CRBI CWalks Division   
## 1.940138 8.416346 55.930008 51.977872 14.395539 8.388450 1.074037   
## PutOuts NewLeague   
## 1.223284 1.059235

Walks,Divison,Putouts,Newleauqe açıklayıcı değişkenlerimiz hariç diğer bütn değişkenlerin Vif değeri>5tir.Yani diğer bağımsız değişkenlerimde*çoklu doğrusal bağlantı*sorunu vardır.

##6En iyi olası alt küme değişken seçim yöntemini uygulayarak alternatif iki model belirleyiniz. Gerekçelerini belirtiniz.

olsrr::ols\_step\_best\_subset(model)

## Best Subsets Regression   
## ------------------------------------------------------------------------------  
## Model Index Predictors  
## ------------------------------------------------------------------------------  
## 1 CRuns   
## 2 Walks CRBI   
## 3 Walks CRBI Division   
## 4 Walks Years CRBI Division   
## 5 Walks CAtBat CRuns CRBI CWalks   
## 6 Walks CAtBat CRuns CRBI CWalks Division   
## 7 Walks CAtBat CRuns CRBI CWalks Division NewLeague   
## 8 Walks CAtBat CRuns CRBI CWalks Division PutOuts NewLeague   
## 9 Walks Years CAtBat CRuns CRBI CWalks Division PutOuts NewLeague   
## ------------------------------------------------------------------------------  
##   
## Subsets Regression Summary   
## -----------------------------------------------------------------------------------------------------------------------------------------------  
## Adj. Pred   
## Model R-Square R-Square R-Square C(p) AIC SBIC SBC MSEP FPE HSP APC   
## -----------------------------------------------------------------------------------------------------------------------------------------------  
## 1 0.4004 0.3971 0.3794 73.9868 2709.9781 2186.5546 2719.6229 144280.5295 144263.3889 788.4650 0.6128   
## 2 0.4828 0.4771 0.4526 41.0682 2684.7646 2161.5835 2697.6244 125827.2660 125789.8953 687.6215 0.5343   
## 3 0.5093 0.5011 0.4749 31.8563 2677.0980 2153.9841 2693.1727 120721.3821 120656.8443 659.7189 0.5125   
## 4 0.5324 0.5220 0.4946 24.0544 2670.2099 2147.3208 2689.4995 116320.9852 116224.2526 635.6716 0.4937   
## 5 0.5482 0.5355 0.4961 19.3953 2665.9169 2143.2809 2688.4215 113680.1498 113545.0976 621.2399 0.4823   
## 6 0.5669 0.5522 0.5114 13.4571 2660.1226 2137.9947 2685.8421 110203.4450 110026.7005 602.2404 0.4674   
## 7 0.5791 0.5624 0.5202 10.2869 2656.8627 2135.2040 2685.7971 108321.6712 108096.4702 591.9569 0.4592   
## 8 0.5862 0.5673 0.5192 9.2796 2655.7326 2134.4537 2687.8819 107718.6075 107437.0738 588.6612 0.4564   
## 9 0.5892 0.5680 0.518 10.0000 2656.3843 2135.3578 2691.7486 108168.4083 107821.4470 591.1193 0.4580   
## -----------------------------------------------------------------------------------------------------------------------------------------------  
## AIC: Akaike Information Criteria   
## SBIC: Sawa's Bayesian Information Criteria   
## SBC: Schwarz Bayesian Criteria   
## MSEP: Estimated error of prediction, assuming multivariate normality   
## FPE: Final Prediction Error   
## HSP: Hocking's Sp   
## APC: Amemiya Prediction Criteria

Çoklu doğrusal bağlantı sorunu olmayan ve cp değeri en az yanlı olan, adjusted r-squared ve r-squared birbirine yakın, AIC ve SBIC mümkün olan en düşük, parametrelerin p-değerlerinin %5 önem düzeyinde anlamlı olduğu mümkün olan en iyi model aşağıdaki çıkmıştır.

model1<-lm(Salary~Walks+CAtBat+CRuns+CRBI+CWalks+Division+PutOuts+NewLeague,data=train)  
summary(model1)

##   
## Call:  
## lm(formula = Salary ~ Walks + CAtBat + CRuns + CRBI + CWalks +   
## Division + PutOuts + NewLeague, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -787.64 -187.77 -8.15 147.84 1947.81   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 58.95047 69.43150 0.849 0.397016   
## Walks 5.63071 1.41422 3.981 0.000100 \*\*\*  
## CAtBat -0.23943 0.06492 -3.688 0.000302 \*\*\*  
## CRuns 2.37143 0.51813 4.577 8.92e-06 \*\*\*  
## CRBI 0.81163 0.27948 2.904 0.004158 \*\*   
## CWalks -0.94995 0.25698 -3.697 0.000292 \*\*\*  
## DivisionW -133.88839 48.86451 -2.740 0.006781 \*\*   
## PutOuts 0.15723 0.09074 1.733 0.084898 .   
## NewLeagueN 102.01806 49.13298 2.076 0.039322 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 320 on 175 degrees of freedom  
## Multiple R-squared: 0.5862, Adjusted R-squared: 0.5673   
## F-statistic: 30.99 on 8 and 175 DF, p-value: < 2.2e-16

vif(model1)

## Walks CAtBat CRuns CRBI CWalks Division PutOuts   
## 1.790636 36.172020 46.611669 14.367327 7.663762 1.066126 1.217733   
## NewLeague   
## 1.055330

best\_yeni kurmuş olduğum en iyi modelde çoklu doğrusal bağlantı sorunnu var.

model2=lm(Salary~Years+CRuns+Division+PutOuts,data=train)  
 #formul <- train$Salary~train$Years + train$CRuns + train$Division + train$PutOuts  
#model2 <- lm(formul)  
summary(model2)

##   
## Call:  
## lm(formula = Salary ~ Years + CRuns + Division + PutOuts, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1000.81 -191.16 -36.51 169.13 1845.91   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 351.62227 62.47806 5.628 6.93e-08 \*\*\*  
## Years -50.87520 10.97883 -4.634 6.89e-06 \*\*\*  
## CRuns 1.59331 0.16746 9.515 < 2e-16 \*\*\*  
## DivisionW -148.31945 50.60430 -2.931 0.00382 \*\*   
## PutOuts 0.28193 0.08859 3.183 0.00172 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 338.8 on 179 degrees of freedom  
## Multiple R-squared: 0.5257, Adjusted R-squared: 0.5151   
## F-statistic: 49.6 on 4 and 179 DF, p-value: < 2.2e-16

#bütün katsayılar anlamlı vif()bakalım  
vif(model2)

## Years CRuns Division PutOuts   
## 4.244458 4.345032 1.020381 1.035772

çoklu doğrusal bağlantı sorunum bu değişkenleri model2 ye alarak çözülmüş oldu. Son bir model daha deneyelim.

model3 <- lm(Salary~Walks+CRuns+CWalks+Division+PutOuts, data = train)  
summary(model3)

##   
## Call:  
## lm(formula = Salary ~ Walks + CRuns + CWalks + Division + PutOuts,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -770.68 -196.48 -2.57 129.16 2091.58   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.27299 65.93379 0.232 0.81708   
## Walks 6.67128 1.34661 4.954 1.68e-06 \*\*\*  
## CRuns 1.37188 0.20743 6.614 4.23e-10 \*\*\*  
## CWalks -0.82119 0.26684 -3.077 0.00242 \*\*   
## DivisionW -149.79249 50.21020 -2.983 0.00325 \*\*   
## PutOuts 0.19914 0.09157 2.175 0.03097 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 335.3 on 178 degrees of freedom  
## Multiple R-squared: 0.5379, Adjusted R-squared: 0.5249   
## F-statistic: 41.44 on 5 and 178 DF, p-value: < 2.2e-16

car::vif(model3)

## Walks CRuns CWalks Division PutOuts   
## 1.478751 6.804554 7.526579 1.025287 1.129568

CRuns,CWalks vif değerlerim>5 olduğundan model3 te çoklu doğrusal bağlantı sorunu var. En iyi elde ettiğim model2 olmuş oldu bağımsız değişkenlerin vif değerleri <5 oldu.

Sonuç olarak olabilecek en iyi modeli bulabilme için ols\_step\_best\_subset(model) yöntemini uygulayarak bana alternatif modeller sundu.ols\_step\_best\_subset(model) verdiği AIC, R^,adj\_r2, Cp değerlerine bakarak olabilecek alternatif 3 model oluşturdum.R^2 değeri yüksek ve R^2 adjusted R^2 değerleri birbirine yakın,CP değeri düşük,Aıc değeri düşükolan mode seçimi en doğrusu olacaktır. Model1 ,model3 te çoklu doğrusal bağlantı sorunu vardı.model2 ise summary(model2) de tüm değişkenler anlamlı,model anlamlı ve vif değerleri bütün bağımsız değişkenlerin 5 altındaydı.

### 7) Alternatif modellerin tahmin performansını test seti üzerinde PRESS değerini dikkate alarak inceleyiniz ve en uygun modele karar veriniz.

pred2 <- predict(model2, test)  
press2 <- Metrics::rmse(pred2, test$Salary)  
print(press2)

## [1] 384.9864

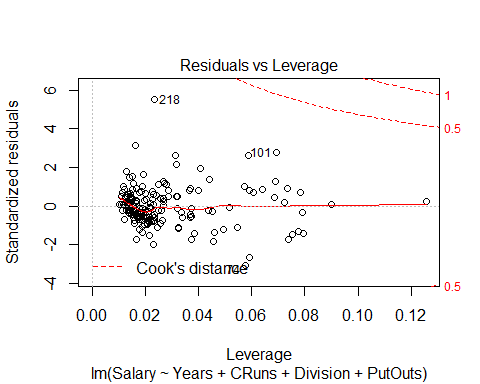
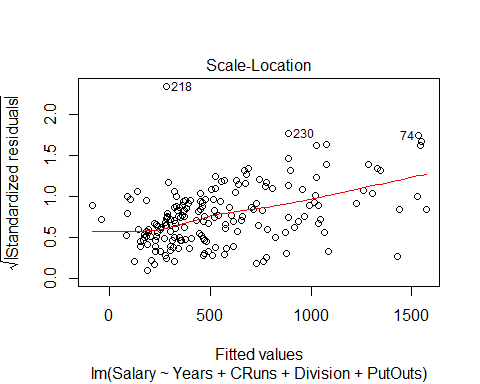
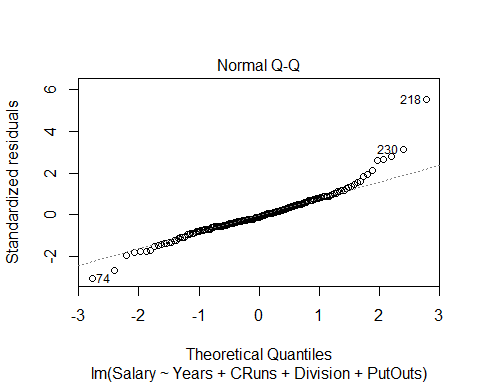
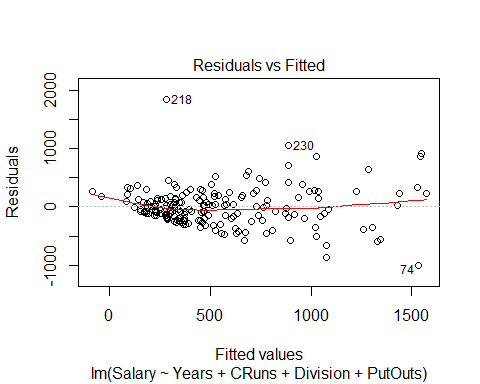
pred3 <- predict(model3, test)  
press3 <- Metrics::rmse(pred3, test$Salary)  
print(press3)

## [1] 331.3998

Sonuç:Pres değeri MOdel3=331,3998dir. Press değerimodel2’nin 384.9864tür. Fakat Model3 te çoklu doğrusal bağlantı sorunu olduğundan model2 ile devam ederim.

### 8) Hipotezleri yazarak, hataların normal dağıldığı varsayımını grafikle ve uygun istatistiksel test ile kontrol ediniz. (α=0.05) (4 p.)

plot(model2)

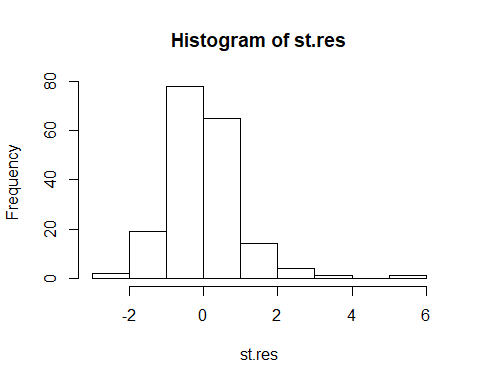


#####Normallik varsayımı Normallik varsayımı: Gözlem sayımız 50den az olmadığı için shapiro-whilk testi kullanılamamaktadır. Kolmogorov-Smirnov testi kullanılacaktır. H0:Hatalar normal dağılım varsayımına uyar. H1:Hatalar normal dağılım varsayımına uymaz.

st.res <- model2$residuals / sd(model2$residuals)   
ks.test(st.res, y = "pnorm")

##   
## One-sample Kolmogorov-Smirnov test  
##   
## data: st.res  
## D = 0.07124, p-value = 0.3078  
## alternative hypothesis: two-sided

hist(st.res)

 One-sample Kolmogorov-Smirnov testine göre hatalar normal dağılır.p-value>0.05 Normal dağılır Varsayımı sağlanır.

#### 9) Hipotezleri yazarak, hataların sabit varyanslı olup olmadığını grafikle ve uygun istatistiksel test ile kontrol ediniz. (α=0.05)

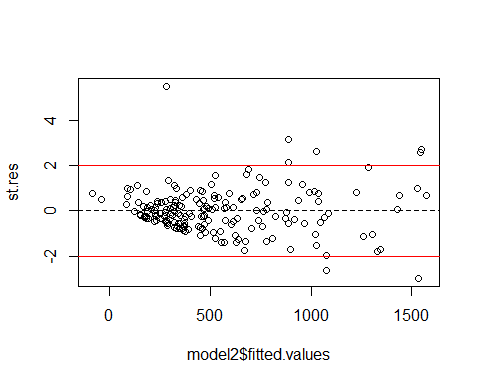
H0:Hatalar sabit varyanslıdır. H1:Hatalar sabit varyanslı değildir.

car::ncvTest(model2)

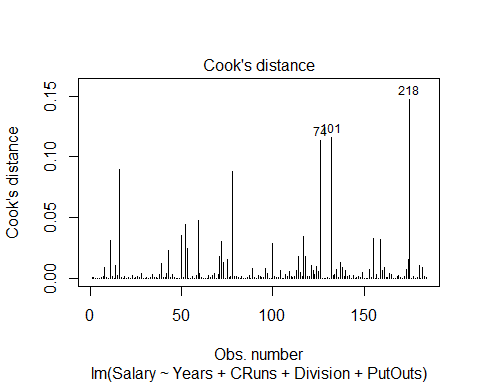
## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 40.37289, Df = 1, p = 2.0983e-10

Modelde ki Hatalar sabit varyanslı değildir. HO red. #### 10) Uç değer ve etkin gözlem olup olmadığını grafiklerle ve ilgili değerlerle belirleyiniz.

plot(model2$fitted.values, st.res)  
abline(h = 0, lty = 2)  
abline(h = c(-2,2), col = "red")

 uç değerli gösteren grafik çizdik.uç değerlerimiz mevcut.

cutoff <- 4/(length(train$Salary)-5)  
plot(model2, which=4, cook.levels=cutoff)

 74,101,218. gözzler Cooks distansela ortaya çıkmıştır. Bunlar benim etkin gözlemlerim.

|  |
| --- |
| *Hat-Value* ve *Cook’s Distance* kriterlerine bakılarak sorun yaratacak gözlemler modelden çıkartılmalıdır. |
| r bad\_lev <- which(abs(st.res)>2 & hatvalues(model2)> 2\*mean(hatvalues(model2))) bad\_lev |
| ## 122 164 74 101 ## 16 78 126 132 Bu gözlemleri çıkartarak bütün aşamaları Tekrarlıyacağız. |
| r require(dplyr) df2 <- df[-bad\_lev,] df2 <- df2[-c(128),] set.seed(123) smp\_size2 <- round(0.7 \* nrow(df2)) train\_ind2 <- sample(x = nrow(df2),size = smp\_size2,replace = F) train2 <- df2[train\_ind2,] test2 <- df2[-train\_ind2,] Belirlediğimiz gözlemleri çıkardıktan sonra modeli tekrar kuracak ve varsayımları kontrol edeceğiz. |
| r model4 <- lm(Salary~Years + CRuns + Division + PutOuts, data=train2) summary(model4) |
| ## ## Call: ## lm(formula = Salary ~ Years + CRuns + Division + PutOuts, data = train2) ## ## Residuals: ## Min 1Q Median 3Q Max ## -758.8 -196.1 -45.0 176.6 1851.2 ## ## Coefficients: ## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 301.28455 57.57026 5.233 4.70e-07 \*\*\* ## Years -28.77881 10.53425 -2.732 0.00694 \*\* ## CRuns 1.05580 0.16510 6.395 1.40e-09 \*\*\* ## DivisionW -98.63848 48.47173 -2.035 0.04335 \* ## PutOuts 0.33410 0.07941 4.207 4.11e-05 \*\*\* ## --- ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Residual standard error: 322.8 on 176 degrees of freedom ## Multiple R-squared: 0.3936, Adjusted R-squared: 0.3798 ## F-statistic: 28.56 on 4 and 176 DF, p-value: < 2.2e-16 |
| r plot(model4) |
|  |
| r st.res4 <- model4$residuals / sd(model4$residuals) stats::ks.test(st.res4, y = "pnorm") |
| ## ## One-sample Kolmogorov-Smirnov test ## ## data: st.res4 ## D = 0.069413, p-value = 0.3477 ## alternative hypothesis: two-sided |
| r car::ncvTest(model4) |
| ## Non-constant Variance Score Test ## Variance formula: ~ fitted.values ## Chisquare = 20.12997, Df = 1, p = 7.2354e-06 Normallik varsayımı sağlanmaktadır fakat değişen varyans sorunu devam etmektedir. Ağırlıklandırılmış regresyon deneyeceğiz. |
| ```r # artıkların mutlak değeri abs\_res <- abs(model4$residuals) # fitted valuelar fitted\_val <- model4$fitted.values # ağırlıklandırılmak için böyle bir method izleyeceğiz. wts değişkenini modelimizi ağırlıklandırmak için kullanacağız. wts <- 1/fitted(lm(abs\_res ~ fitted\_val))^2 # ağırlıklandırılmış model: model5 <- lm(Salary~Years + CRuns + Division + PutOuts, data = train2, weights = wts) |
| summary(model5) ``` |
| ## ## Call: ## lm(formula = Salary ~ Years + CRuns + Division + PutOuts, data = train2, ## weights = wts) ## ## Weighted Residuals: ## Min 1Q Median 3Q Max ## -2.4406 -0.8952 -0.2156 0.7076 11.1677 ## ## Coefficients: ## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 242.25505 52.69257 4.598 8.14e-06 \*\*\* ## Years -15.43905 9.14003 -1.689 0.0930 . ## CRuns 0.96909 0.16542 5.858 2.25e-08 \*\*\* ## DivisionW -77.41752 45.06075 -1.718 0.0875 . ## PutOuts 0.26978 0.09301 2.900 0.0042 \*\* ## --- ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Residual standard error: 1.398 on 176 degrees of freedom ## Multiple R-squared: 0.3221, Adjusted R-squared: 0.3067 ## F-statistic: 20.9 on 4 and 176 DF, p-value: 4.083e-14 |
| r plot(model5) |
|  |
| r st.res5 <- model5$residuals / sd(model5$residuals) |
| r stats::ks.test(st.res5, y = "pnorm") |
| ## ## One-sample Kolmogorov-Smirnov test ## ## data: st.res5 ## D = 0.11227, p-value = 0.02087 ## alternative hypothesis: two-sided |
| r car::ncvTest(model5) |
| ## Non-constant Variance Score Test ## Variance formula: ~ fitted.values ## Chisquare = 4.553621, Df = 1, p = 0.032849 ###11. Modelde yer alan iki değişkene ait katsayıyı yorumlayınız. |
| r summary(model5) |
| ## ## Call: ## lm(formula = Salary ~ Years + CRuns + Division + PutOuts, data = train2, ## weights = wts) ## ## Weighted Residuals: ## Min 1Q Median 3Q Max ## -2.4406 -0.8952 -0.2156 0.7076 11.1677 ## ## Coefficients: ## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 242.25505 52.69257 4.598 8.14e-06 \*\*\* ## Years -15.43905 9.14003 -1.689 0.0930 . ## CRuns 0.96909 0.16542 5.858 2.25e-08 \*\*\* ## DivisionW -77.41752 45.06075 -1.718 0.0875 . ## PutOuts 0.26978 0.09301 2.900 0.0042 \*\* ## --- ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Residual standard error: 1.398 on 176 degrees of freedom ## Multiple R-squared: 0.3221, Adjusted R-squared: 0.3067 ## F-statistic: 20.9 on 4 and 176 DF, p-value: 4.083e-14 |
| yorum: Modelde bulunan Intecrpt,Cruns,PutOutputs değişkenler yüksek oranda anlamlı çıkmıştır. F istatistiği 4.083e-14<0.05 olduğu için hipotezi red edilir ve en az bir tane ß 0 dan farklıdır yorumu yapılır. SAlary bağımlı değişkenindeki değişimin %32.93’ü açıklayıcı değişkenler tarafından açıklanmaktadır. |
| r model5$coefficients |
| ## (Intercept) Years CRuns DivisionW PutOuts ## 242.2550484 -15.4390538 0.9690865 -77.4175208 0.2697802 Yorum: Katsayı anamlılığı: |
| Byears- değişkenim bir birim arttığında, diğer değişkenler modelde ve sabitken Maaş ortalama 15.4390 azalır. |
| BCRuns değişkenim bir birim arttığında ,diğer değişkenler modelde ve sabitken Salary ortalama 0.9690 artar. |
| PutOutputs değikenim bir birim arttığında diğer değişkenler modelde ve sabitken ortalama SAlary bağımlı değişkenim 0.2697802 artar. |
| ### 12.Yeni bir gözlem değeri için %95’lik güven aralığını ve kestirim aralığını bularak yorumlayınız. (3 p.) |
| r library(stats) confint(model5, level = 0.95) |
| ## 2.5 % 97.5 % ## (Intercept) 138.26443983 346.2456569 ## Years -33.47720920 2.5991015 ## CRuns 0.64262225 1.2955507 ## DivisionW -166.34646271 11.5114210 ## PutOuts 0.08621687 0.4533435 *Byears* için yorum; Years bir birim arttığında, diğer değişkenler modelde ve sabitken Oyuncuların maaşları(salary) %95 güvenle ortalama -33.477 ile 0.259 arasındadır. |
| BCruns için yorum: CRuns bir birim arttığında , diğer değişkenler modelde ve sabitken Oyuncuların Maaşları %95 güvenle 0.64262225- 1.2955507 arasındadır. |
| Bputoutputs: PutOuts bir birim arttığında diğer değişkenler model de ve sabitken Oyuncuların maaşları %95 güvenle ortalama 0.08621687 - 0.4533435 değerleri arasındadır. |
| Yeni gözlem için test seti kullanılacaktır. |
| r #x0 Y0 için Güven aralığı tahmini. pred\_confint <- predict(model5, test2, interval = "confidence") test\_confint <- cbind(test2, pred\_confint) head(test\_confint) |
| ## Walks Years CAtBat CRuns CRBI CWalks Division PutOuts NewLeague Salary ## 3 76 3 1624 224 266 263 W 880 A 480.000 ## 4 37 11 5628 828 838 354 E 200 N 500.000 ## 11 59 10 4631 702 504 488 E 238 A 517.143 ## 13 47 4 1512 205 204 203 E 211 N 550.000 ## 17 73 15 8068 1045 993 732 W 105 N 775.000 ## 22 62 1 426 55 43 62 W 361 N 115.000 ## fit lwr upr ## 3 573.0023 438.8562 707.1484 ## 4 928.7851 787.8269 1069.7433 ## 11 832.3709 715.6715 949.0703 ## 13 436.0852 362.1255 510.0449 ## 17 974.2740 803.3287 1145.2193 ## 22 300.0889 218.8385 381.3393 Tahmin değeri %95 güvenle lwr ve upr sütunlarındaki değerler arasında olacaktır Walks 76,yearsı 3 , CatBAt 1621 Cruns 266 iken %95 güvenle Salary 483.8562 ile 707.1484 arasında değişir. |

Tahmin değeri %95 güvenle lwr ve upr sütunlarındaki değerler arasında olacaktır.

##X0 icin Y0'a ait tahmin araligi##  
pred\_confint1 <- predict(model5, test2, interval = "predict")

## Warning in predict.lm(model5, test2, interval = "predict"): Assuming constant prediction variance even though model fit is weighted

test\_confint1 <- cbind(test2, pred\_confint1)  
head(test\_confint)

## Walks Years CAtBat CRuns CRBI CWalks Division PutOuts NewLeague Salary  
## 3 76 3 1624 224 266 263 W 880 A 480.000  
## 4 37 11 5628 828 838 354 E 200 N 500.000  
## 11 59 10 4631 702 504 488 E 238 A 517.143  
## 13 47 4 1512 205 204 203 E 211 N 550.000  
## 17 73 15 8068 1045 993 732 W 105 N 775.000  
## 22 62 1 426 55 43 62 W 361 N 115.000  
## fit lwr upr  
## 3 573.0023 438.8562 707.1484  
## 4 928.7851 787.8269 1069.7433  
## 11 832.3709 715.6715 949.0703  
## 13 436.0852 362.1255 510.0449  
## 17 974.2740 803.3287 1145.2193  
## 22 300.0889 218.8385 381.3393

Tahmin değeri lwr ve upr sütunlarındaki değerler arasında olacaktır Walks 76,yearsı 3 , CatBAt 1621 Cruns 266 iken %95 güvenle Salary 438.8562 ile 707.1484 arasında değişir.

```