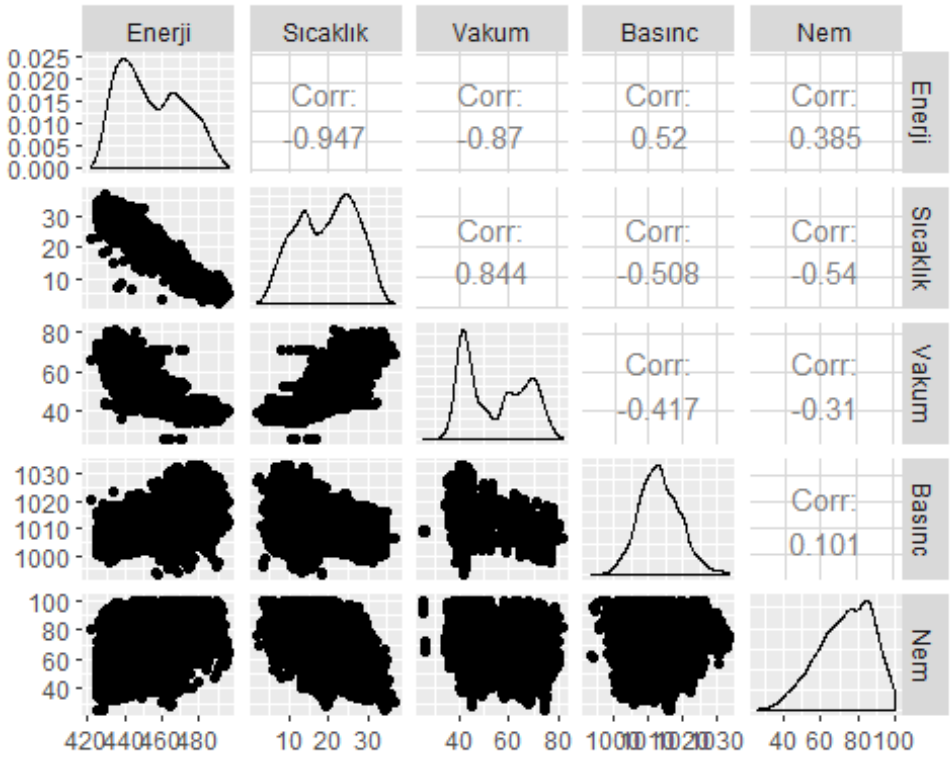
Arasınav_doğrusal

Buse Baltacıoğlu

03 01 2021

1-Tanımlayıcı istatistikleri grafiklerle destekleyerek elde ediniz ve yorumlayınız.



Enerji: Net saatlik elektrik enerjisi çıkışı 421.6 MW ile 495.8 MW arasında 454.4 MW ortalama ve 17 standart sapmayla değişmektedir.Grafikten iki tepeli ve sağa çarpık bir dağılıma sahip olduğunu söyleyebiliriz.

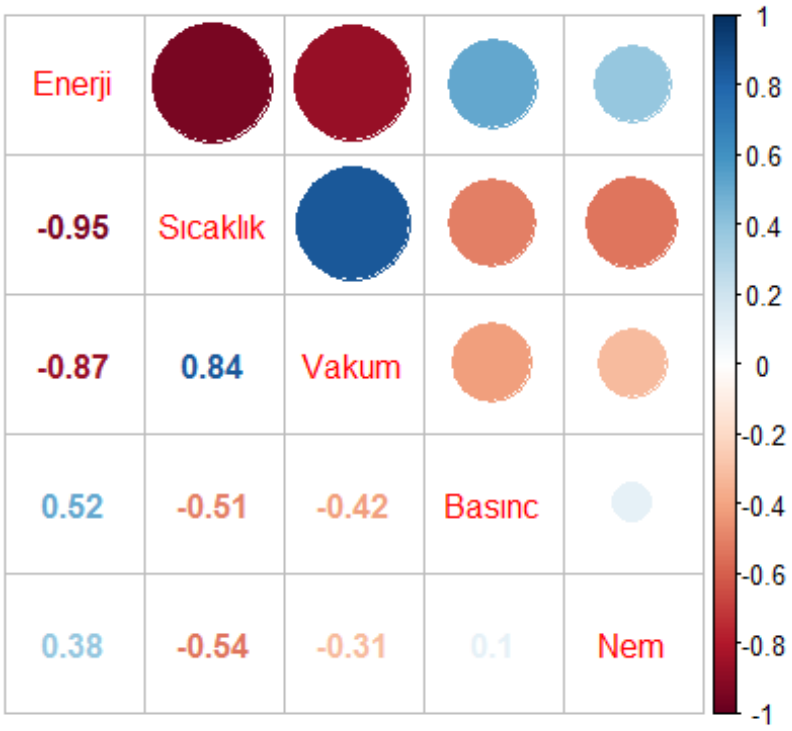
Sıcaklık: Sıcaklık 1.81 C ile 37.11 C arasında 19.63 C ortalama ve 7.42 standart sapmayla değişmektedir. İki tepeli bir dağılım olduğunu söyleyebiliriz.

Vakum: Egzoz vakum 25.36 cm Hg ile 81.56 cm Hg arasında 54.33 cm Hg ortalama ve 12.70 standart sapmayla değişmektedir. Bu değişkenimiz üç tepeli ve sağa çarpık bir dağılıma sahip olduğunu söyleyebiliriz.

Basınç: Ortam basıncı 993.1 milibar ile 1033.3 milibar arasında 1013.3 milibar ortalama ve 5.94 standart sapmayla değişmektedir. Ortam basıncı değişkenimizin normal dağıldığını söyleyebiliriz yinede test edilmelidir.

Nem: Bağıl nem %25.56 ile %100.15 arasında 73.45 ortalama ve 14.53 standart sapmayla değişmektedir.Sola çarpık bir dağılıma sahiptir.

2-Matris Plot oluşturarak yorumlayınız.



Korelasyon matrisine baktığımızda;

-Net saatlik elektrik enerjisi çıkışı ile sıcaklık arasında doğrusal negatif yönlü güçlü bir ilişki bulunmaktadır.

-Net saatlik elektrik enerjisi çıkışı ile egzoz vakum arasında doğrusal negatif yönlü güçlü bir ilişki bulunmaktadır.

-Net saatlik elektrik enerjisi çıkışı ile ortam basıncı arasında doğrusal pozitif yönlü bir ilişki bulunmaktadır.

-Net saatlik elektrik enerjisi çıkışı ile bağıl nem arasında doğrusal pozitif yönlü zayıf bir ilişki bulunmaktadır.

-Sıcaklık ile egzoz vakum arasında doğrusal pozitif yönlü güçlü bir ilişki bulunmaktadır.

-Sıcaklık ile basınç arasında doğrusal negatif yönlü bir ilişki bulunmaktadır.

-Sıcaklık ile nem arasında doğrusal negatif yönlü bir ilişki bulunmaktadır.

-Vakum ile basınç arasında doğrusal negatif yönlü bir ilişki bulunmaktadır.

-Vakum ile nem arasında doğrusal negatif yönlü zayıf bir ilişki bulunmaktadır.

-Basınç ile nem arasında doğrusal bir ilişki bulunmamaktadır.

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

```
model<-lm(train$Enerji~train$Sıcaklık+train$Vakum+train$Basınc+train$Nem)
```

H0: Bj=0

H1: En az bir Bj farklıdır, j=1,2,3

```
summary(model)
```

```
##
## Call:
## lm(formula = train$Enerji ~ train$Sıcaklık + train$Vakum + train$Basınc +
##     train$Nem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.428  -3.118  -0.103   3.171  16.775
##
## Coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    451.756462  11.666751   38.722  < 2e-16 ***
## train$Sıcaklık  -1.970609   0.018317 -107.581  < 2e-16 ***
## train$Vakum     -0.237202   0.008718  -27.208  < 2e-16 ***
## train$Basınc     0.065034   0.011319    5.746 9.56e-09 ***
## train$Nem       -0.159527   0.004999  -31.911  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.572 on 6692 degrees of freedom
## Multiple R-squared:  0.9278, Adjusted R-squared:  0.9277
## F-statistic: 2.149e+04 on 4 and 6692 DF,  p-value: < 2.2e-16
```

-F istatistiğine karşılık gelen 2.2e-16<0.05 olduğu için %95 güven düzeyinde H0 red edilir; en az bir katsayı model için anlamlıdır.Model geçerlidir.

#4-Hipotezleri yazarak elde ettiğiniz modeldeki anlamlı katsayıları belirleyiniz.

H0: B0=0 2e-16<0.05 olduğu için B0 parametresi %95 güvenle anlamlı çıkmıştır. -b0:Bağımsız değişkenlerin değerleri 0 olduğunda net saatlik elektrik enerjisi çıkışı ortalama 451.76 MW'dur.

H0: B1=0 2e-16<0.05 olduğu için B1 parametresi %95 güvenle anlamlı çıkmıştır. -b1:Diğer değişkenler modelde ve sabitken sıcaklık, 1 C arttığında net saatlik elektrik enerjisi çıkışını ortalama 1.97 MW azaltır.

H0: B2=0 2e-16<0.05 olduğu için B2 parametresi %95 güvenle anlamlı çıkmıştır. -b2:Diğer değişkenler modelde ve sabitken egzoz vakumu, 1 cm Hg arttığında net saatlik elektrik enerjisi çıkışını ortalama 0.24 MW azaltır.

H0: B3=0 9.56e-09<0.05 olduğu için B3 parametresi %95 güvenle anlamlı çıkmıştır. -b3:Diğer değişkenler modelde ve ortam basıncı, 1 milibar arttığında net saatlik elektrik enerjisi çıkışını ortalama 0.07 MW arttıracaktır.

H0: B4=0 2e-16<0.05 olduğu için B4 parametresi %95 güvenle anlamlı çıkmıştır. -b4:Diğer değişkenler modelde ve sabitken bağıl nem, 1 birim arttığında net saatlik elektrik enerjisi çıkışını ortalama 0.16 MW azaltır.

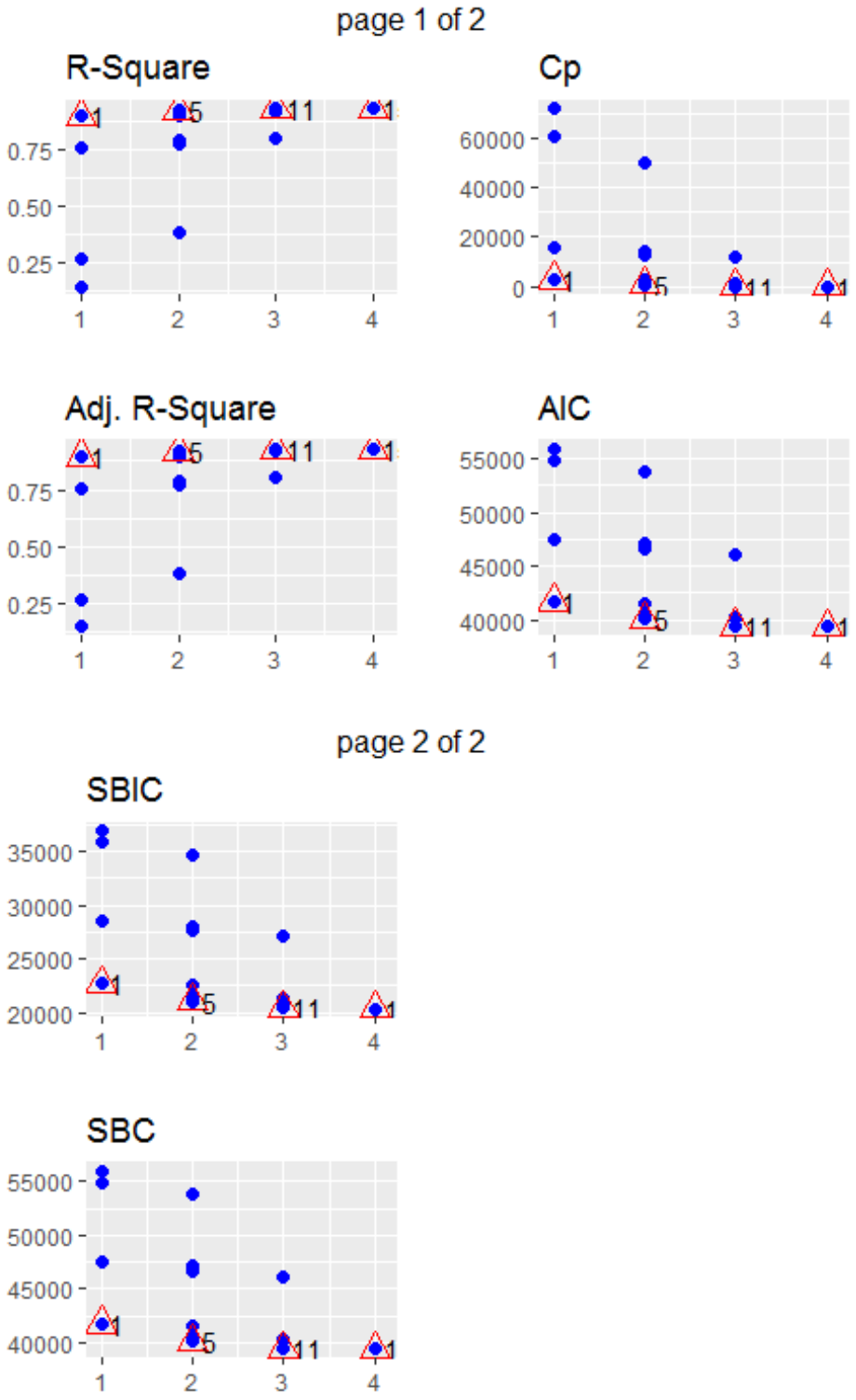
#5-VIF değerlerini hesaplayınız ve yorumlayınız.

```
vif(model)
```

```
## train$Sıcaklık    train$Vakum   train$Basınc      train$Nem
##       5.916272       3.929245       1.449351       1.689888
```

Bu modelde vif değerleri sıcaklık dışında 5'ten küçük çıkmıştır.

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



-ilk elemeyi Cp üzerinden yapılırsa modeldeki yanlılığı ortadan kaldırabiliriz. Modeldeki parametre sayısına eşit çıkması istenir. Bu kritere baktığımızda full (15.) model yansız çıkmıştır. Bununla birlikte (sıcaklık+vakum+nem)) 11. modelde göze alınabilecek bir yanlılık bulunmaktadır. 5. model (sıcaklık+nem) de incelenmelidir.

-Düzeltilmiş R^2 üzerinden karşılaştırma yapmak daha uygun ve ikisininde büyük ve yakın çıkması(modele alınan değişkenlerin anlamlı olduğu anlamına gelir) bu sebepten 11. ve 15. model en yüksek açıklamaya sahiptir.

-Akaike Bilgi Kriteri (AIC) ve Bayesian Bilgi Kriteri (BIC,SBIC) kriterleri için de değerlerin küçük çıkması istenir. Bu kriterler 11. ve 15. modelleri önerir. 5. model SBIC açısından değerlendirilebilir.

-Bunlarla birlikte ne kadar az değişken o kadar iyi olduğu için R^2'ler ve Cp açısındanda 11. model alternatif bir modeldir.

#7-Alternatif modellerin tahmin performansını test seti üzerinde PRESS, RMSE ve MAE değerlerini dikkate alarak inceleyiniz ve en uygun modele karar veriniz.

```
mdl_11<-lm(train$Enerji~train$Sıcaklık+train$Vakum+train$Nem)
summary(mdl_11)
```

```
##
## Call:
## lm(formula = train$Enerji ~ train$Sıcaklık + train$Vakum + train$Nem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.753  -3.107  -0.081   3.147  16.972
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    518.736267   0.466195 1112.70   <2e-16 ***
## train$Sıcaklık  -2.013238   0.016788 -119.92   <2e-16 ***
## train$Vakum     -0.231543   0.008683  -26.67   <2e-16 ***
## train$Nem       -0.167065   0.004835  -34.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.583 on 6693 degrees of freedom
## Multiple R-squared:  0.9274, Adjusted R-squared:  0.9274
## F-statistic: 2.851e+04 on 3 and 6693 DF,  p-value: < 2.2e-16
```

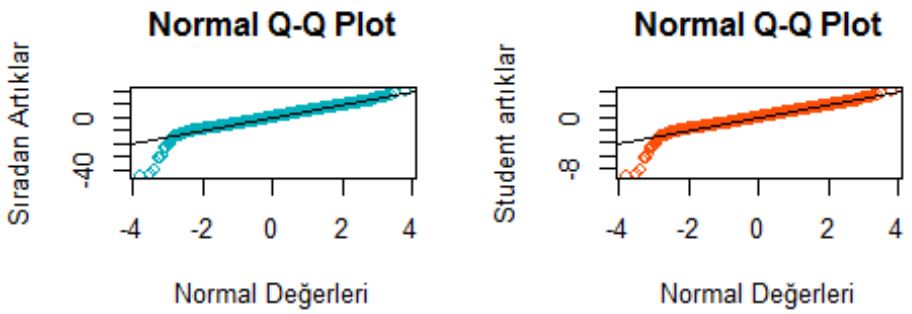
```
mdl_5<-lm(train$Enerji~train$Sıcaklık+train$Nem)
summary(mdl_5)
```

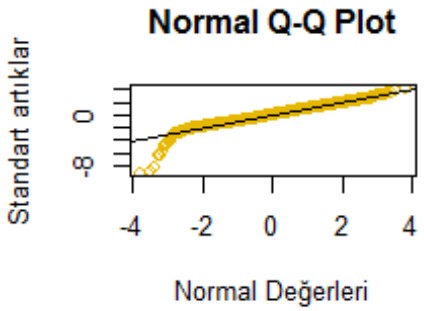
```
##
## Call:
## lm(formula = train$Enerji ~ train$Sıcaklık + train$Nem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.174  -3.278   0.022   3.239  20.601
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

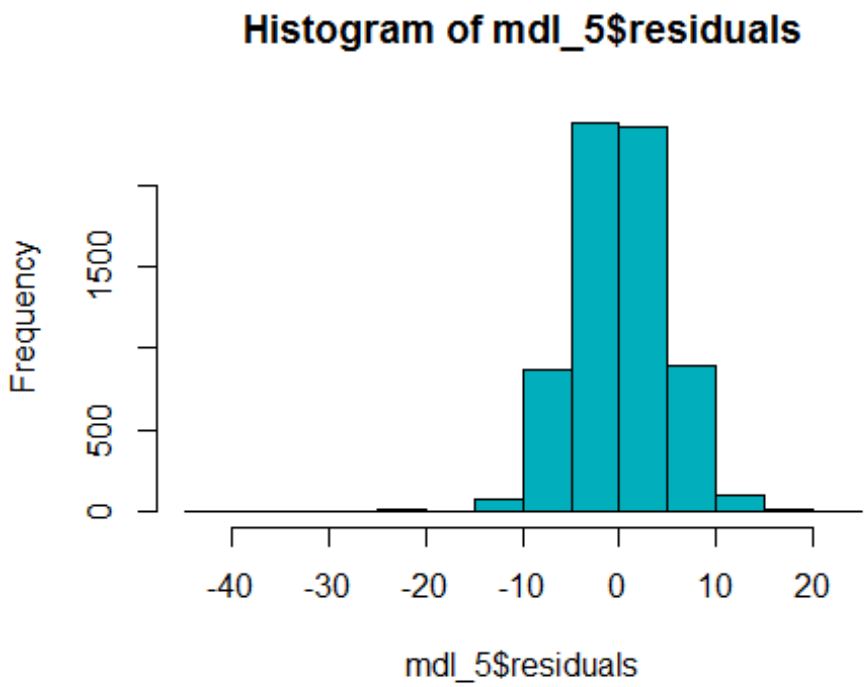
```
## (Intercept)    516.629630   0.483208  1069.2   <2e-16 ***
## train$Sıcaklık  -2.391706   0.009430  -253.6   <2e-16 ***
## train$Nem       -0.208525   0.004815   -43.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.819 on 6694 degrees of freedom
## Multiple R-squared:  0.9197, Adjusted R-squared:  0.9197
## F-statistic: 3.834e+04 on 2 and 6694 DF,  p-value: < 2.2e-16


cbind(RMSE_5, RMSE_11, RMSE_full, MAE_5, MAE_11, MAE_full)

##        RMSE_5  RMSE_11 RMSE_full   MAE_5   MAE_11 MAE_full
## [1,] 23.86116 23.93444  23.93683 19.3547 19.43967  19.4433
```

Model 5 (sıcaklık+nem), Model 11 (sıcaklık+vakum+nem), model 15(full) modellerinin tahmin doğruluğu üzerinde RMSE ve MAE değerlerini karşılaştırdığımızda 11 ve full modelin neredeyse aynı değerleri verdiğini bununla birlikte model 5'in en küçük değerleri vermesi üzerine biz sıcaklık ve nemin bulunduğu model 5'e karar verdik.

#8-Hipotezleri yazarak, hataların normal dağıldığı varsayımını grafikle ve uygun istatistiksel test ile kontrol ediniz.







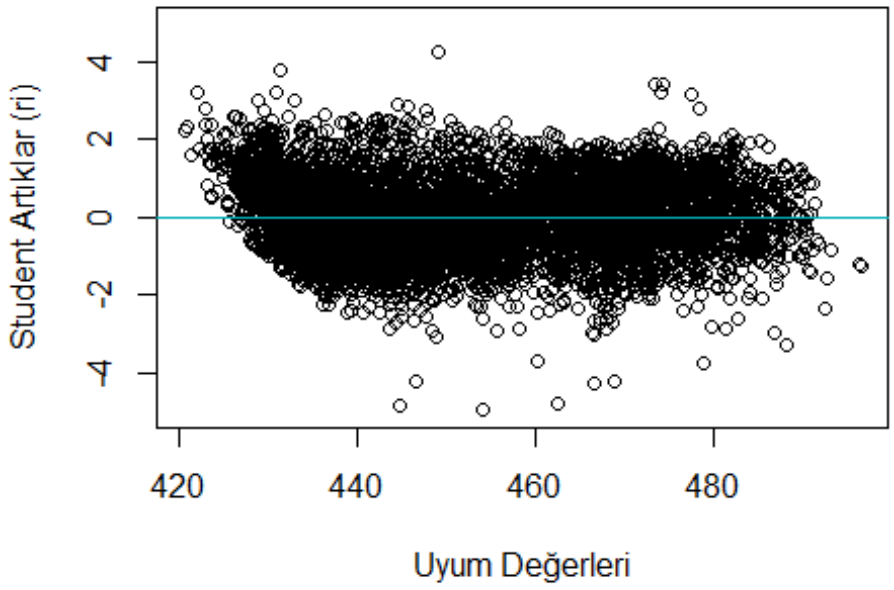
H_0:Artıklar normal dağılır

H_1:Artıklar normal dağılmaz

```
ad.test(mdl_5$residuals)

##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: uniform distribution
##  Parameters assumed to be fixed
##
## data:  mdl_5$residuals
## An = Inf, p-value = 8.959e-08

ks.test(model$residuals, alternative="two.sided",
        "pnorm",mean=0,sd=1)

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  model$residuals
## D = 0.3279, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

8.959e-08<0.05 → H_0 red artıklar normal dağılmaz.

9-Hipotezleri yazarak, hataların sabit varyanslı olup olmadığını grafikle ve uygun istatistiksel test ile kontrol ediniz.



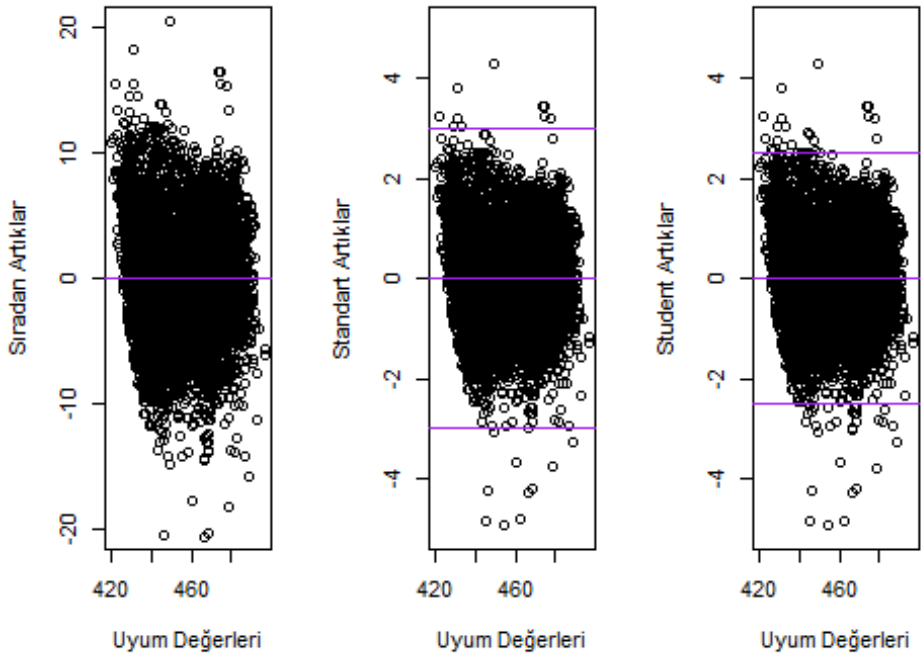
H_0:Artıkların varyansı homojendir

H_1:Artıkların varyansı heterojendir

```
bptest(mdl_5)

##
##  studentized Breusch-Pagan test
##
## data:  mdl_5
## BP = 7.7895, df = 2, p-value = 0.02035
```

0.02035<0.05 → H_0 red artıkların varyansı sabit değildir.

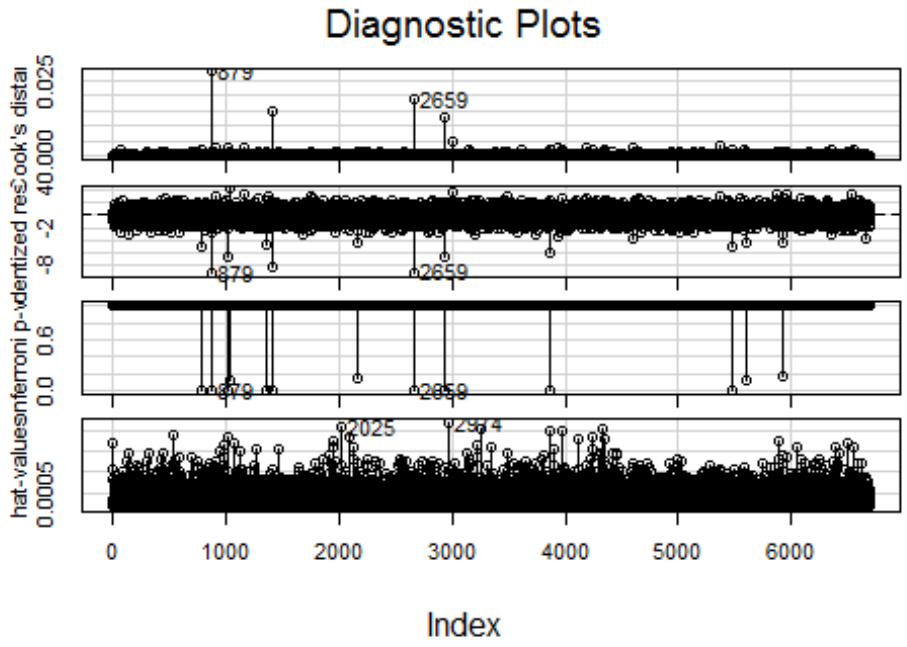
10-Uç değer ve etkin gözlem olup olmadığını grafiklerle ve ilgili değerlerle belirleyiniz.

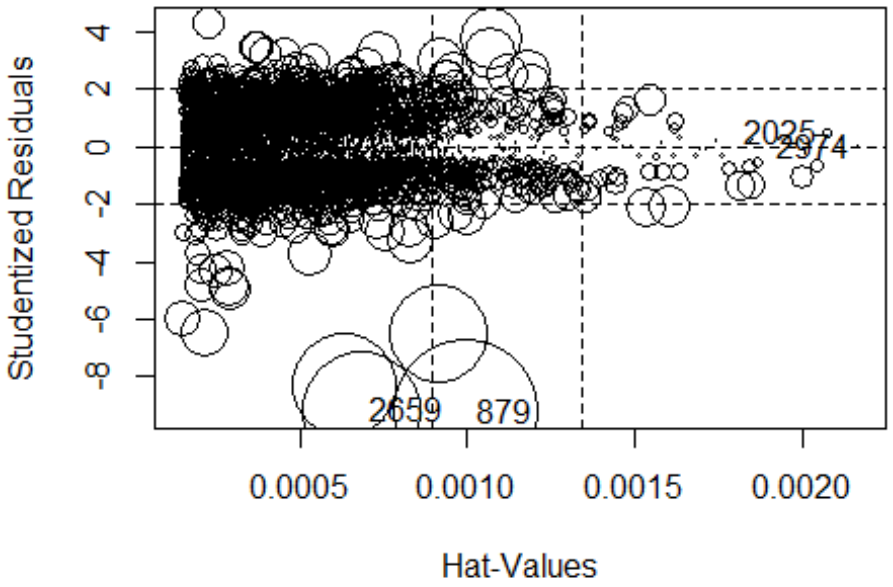


Student artık (|ri|>2.5) kriterimize baktığımızda 459 918 1035 1170 1293 1384 1757 1778 2827 3004 3487 3964 4181 4346 4370 4655 5375 5443 5550 5878 5936 5951 6087 6537 6609 74 137 332 522 584 792 875 879 1027 1361 1409 1679 1970 2159 2242 2264 2462 2551 2659 2791 2873 2936 3204 3207 3585 3735 3872 3931 3962 4607 4663 5015 5128 5485 5594 5918 5924 6346 6555 6665 gözlemler uç değer çıkmıştır.

```
outlierTest(mdl_5)
```

```
##       rstudent unadjusted p-value Bonferroni p
## 879  -9.227710         3.6394e-20   2.4373e-16
## 2659 -9.131764         8.7732e-20   5.8754e-16
## 1409 -8.325045         1.0126e-16   6.7811e-13
## 2936 -6.505089         8.3254e-11   5.5755e-07
## 1027 -6.468131         1.0622e-10   7.1139e-07
## 3872 -5.971246         2.4751e-09   1.6576e-05
## 792  -4.935805         8.1752e-07   5.4749e-03
## 5485 -4.844992         1.2947e-06   8.6704e-03
## 1361 -4.821580         1.4557e-06   9.7489e-03
```





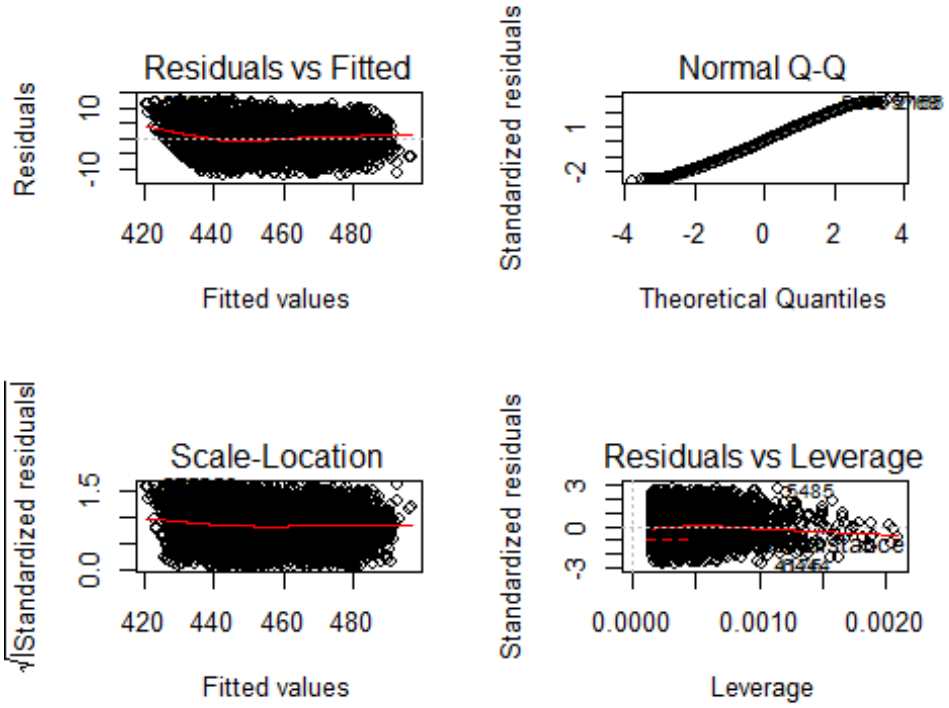

```
##           StudRes          Hat        CookD
## 879  -9.227710003 0.0009984864 2.801671e-02
## 2025  0.478777736 0.0020700508 1.585176e-04
```

```
## 2659 -9.131764113 0.0006812002 1.871743e-02
## 2974 -0.007116848 0.0021683320 3.669336e-08
```

Bu veri seti için bağımlı değişkenin normal dağılım varsayımı sağlamaması üzerine karekök, küpkök, logaritmik, ters, kare, küp, vs. dönüşümleri denenmiş uç değer ve kaldıraç değerleri çıkartılarak yeni modeller oluşturulmuş F testine göre model anlamlı olmasına rağmen bağımlı değişkenin normal dağılması yani artıkların normal dağılması, artıkların varyansının sabit olması varsayımlarını sağlayamamıştır. Bu sebepten yeterli iyileştirme yapılamamıştır. Yinede model seçmemiz gerektiğinden biz model 5'e yani sıcaklık ve nem değişkenlerinin oluşturduğu ve açıklama oranının %92.99 ve düzeltilmiş R^2'sininde %92.99 olan modele karar verilmiştir.

```
summary(final_model)
```

```
##
## Call:
## lm(formula = data$Enerji ~ data$Sıcaklık + data$Nem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8788  -3.2683  -0.0344   3.1305  12.1037
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   516.886029   0.455074 1135.83   <2e-16 ***
## data$Sıcaklık  -2.405735   0.008884 -270.80   <2e-16 ***
## data$Nem       -0.207487   0.004533  -45.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.495 on 6584 degrees of freedom
## Multiple R-squared:  0.9299, Adjusted R-squared:  0.9299
## F-statistic: 4.37e+04 on 2 and 6584 DF,  p-value: < 2.2e-16
```



11-Yeni bir gözlem değeri için %95'lik güven aralığını ve/veya kestirim aralığını bularak yorumlayınız.

```
new <- data.frame(Sıcaklık = test$Sıcaklık, Nem = test$Nem)

head(predict(final_model, new))

##        1        2        3        4        5        6
## 448.6097 443.9494 462.1044 437.4699 427.9936 470.8745
```

Sıcaklık ve nem için enerji çıkısının dağılımının ortalamasının güven aralığı

```
predictnewconf <- predict(final_model,newdata = new,interval="confidence")

lower<-predictnewconf[2,]
upper<-predictnewconf[3,]
fit<-predictnewconf[1,]
cbind(new,lower,upper,fit,test$Enerji)

##      Sıcaklık    Nem    lower    upper      fit test$Enerji
## 1       23.64  74.20 443.9494 462.1044 448.6097      445.75
## 2       13.97  84.60 443.8024 461.9730 448.4772      470.96
## 3       22.10  75.38 444.0965 462.2359 448.7422      442.35
## 4       31.25  36.83 443.9494 462.1044 448.6097      428.77
## 5       22.99  49.42 443.8024 461.9730 448.4772      451.41
## 6       29.30  61.23 444.0965 462.2359 448.7422      426.25
## 7       22.72  60.34 443.9494 462.1044 448.6097      453.13
```

-Sıcaklık 23.64 C ve nem %74.20 iken %95 güvenle net saatlik elektrik enerjisi çıkışı ortalama 7.73 dolar443.95 MW ile 462.10 MW arasında değişir.

x1 ve x2 için y'a ait kestirim aralığı

```
predictnewpred<-predict(final_model,new,interval="prediction")

## Warning: 'newdata' had 2871 rows but variables found have 6587 rows

fit_k<-predictnewpred[1,]
lower_k<-predictnewpred[2,]
upper_k<-predictnewpred[3,]
cbind(new, lower_k, upper_k, fit_k, test$Enerji)

##      Sıcaklık    Nem  lower_k  upper_k    fit_k test$Enerji
## 1       23.64  74.20 443.9494 462.1044 448.6097      445.75
## 2       13.97  84.60 435.1368 453.2921 439.7973      470.96
## 3       22.10  75.38 452.7621 470.9168 457.4221      442.35
## 4       31.25  36.83 443.9494 462.1044 448.6097      428.77
## 5       22.99  49.42 435.1368 453.2921 439.7973      451.41
## 6       29.30  61.23 452.7621 470.9168 457.4221      426.25
## 7       22.72  60.34 443.9494 462.1044 448.6097      453.13
```

-Sıcaklık 13.97 ve nem 84.60 iken %95 güvenle net saatlik elektrik enerjisi çıkışı 435.14 MW ile 453.29 MW arasında değişir.