

İST359 REGRESYON ÇÖZÜMLEMESİ ÖDEVİ

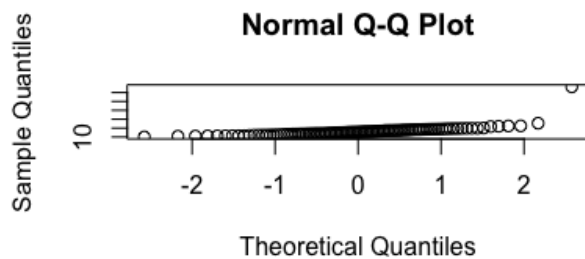
Şeydanur Kayır
21723933

Senaryo;

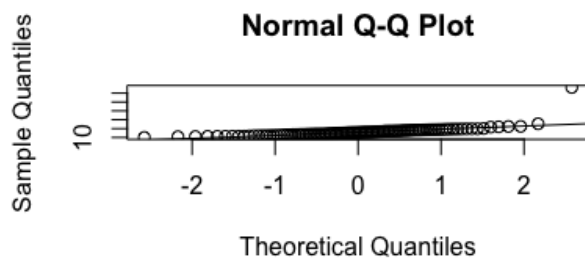
Bir ankette trafik kazalarının oluş nedenleri incelenmek isteniyor. Bu amaçla kazaların hız, yol koşulları ve hava durumuna göre üç farklı durumla 100 kişi rasgele seçiliyor.

```
data <- read_table2("Desktop/data.txt")  
names(data)<-c("y","x1","x2","x3")  
attach(data)
```

qqnorm(y)



qqline(y)



H0: Verilerin dağılışı ile normal dağılım arasında fark yoktur.

shapiro.test(y)

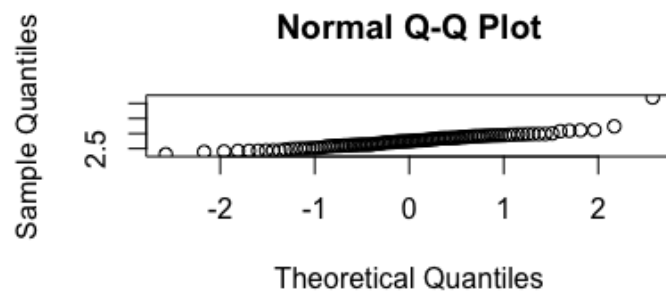
Shapiro-Wilk normality test

data: y

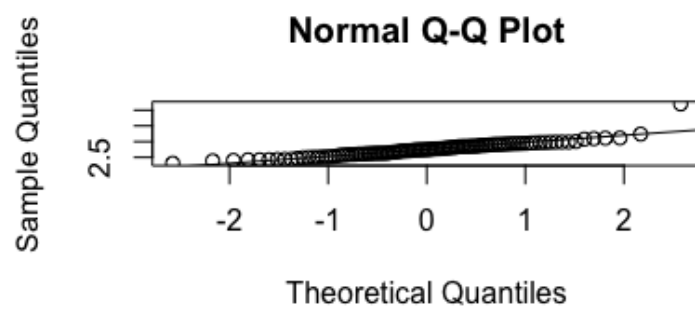
W = 0.52422, p-value < 2.2e-16

lny<-log(y)

```
qqnorm(lny)
```



```
qqline(lny)
```



```
shapiro.test(lny)
```

Shapiro-Wilk normality test

```
data: lny
```

```
W = 0.86066, p-value = 2.997e-08
```

```
boxplot(y)$out
```

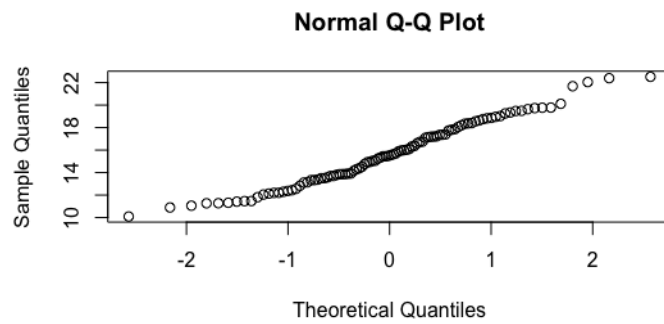
```
[1] 25.47472 66.37160
```

```
which(y %in% boxplot(y)$out)
```

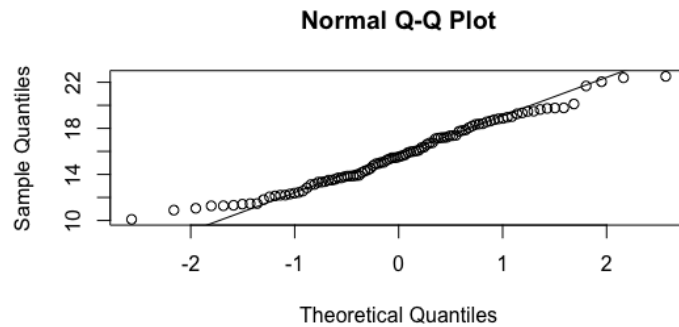
```
[1] 21 60
```

```
data1<-data[-c(21,60),]
```

```
qqnorm(data1$y)
```



qqline(data1\$y)



shapiro.test(data1\$y)

Shapiro-Wilk normality test

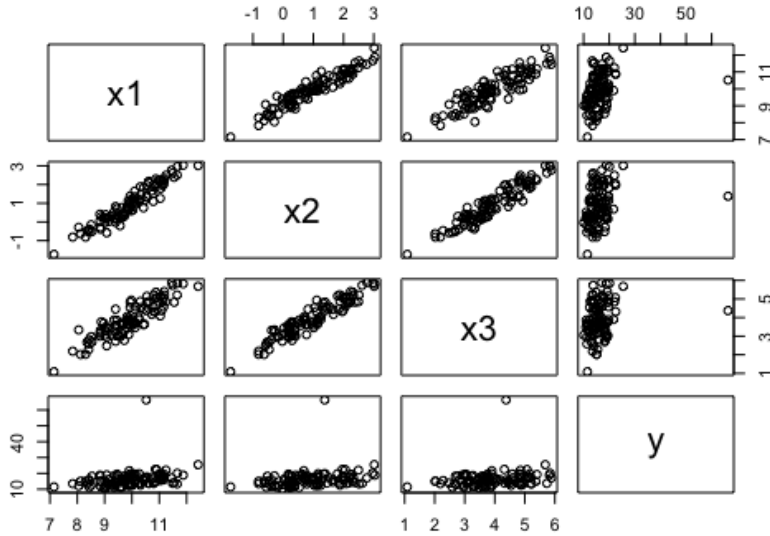
data: data1\$y

W = 0.97619, p-value = 0.07212

H_0 : Verilerin dağılışı ile normal dağılım arasında fark yoktur.

$p = 0,07212 > \alpha = 0,05$ olduğundan H_0 red edilemez. Verilerin %95 güvenle normal dağılıma sahip olduğu söylenebilir.

```
datay<- cbind(x1,x2,x3,y)
pairs(data1)
```



Bağımlı değişken ile bağımsız değişkenler arasında ilişki var ve doğrusal model oluşturulabilir. Ancak bağımsız değişkenlerde birbirleriyle ilişkili durumdadır.

```
sonuc<-lm(y~x1+x2+x3)
summary(sonuc)
```

Call:

```
lm(formula = y ~ x1 + x2 + x3)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.059	-2.724	-0.602	1.370	49.060

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.8435	18.6210	-0.153	0.879
x1	2.4156	1.9084	1.266	0.209
x2	0.3981	2.5950	0.153	0.878
x3	-1.3281	1.8081	-0.735	0.464

Residual standard error: 5.74 on 96 degrees of freedom

Multiple R-squared: 0.08548, Adjusted R-squared: 0.0569

F-statistic: 2.991 on 3 and 96 DF, p-value: 0.03476

Multiple R-squared değerin 1'e yakın olmaması bağımsız değişkenin bağımlı değişkeni iyi bir şekilde açıklamadığı anlamına gelir . Yani açıklanamayan kısım %92 dir. Açıklanamayan kısım için farklı değişkenler olabilir. Kurulan regresyon modelinin anlamlı olduğu %95 güvenilirlikle söylenebilir.

$H_0 : \beta = 0$ (Model anlamlı değildir)

$H_0 : \beta \neq 0$ (Model anlamlıdır)

$P = 0,03 < \alpha = 0,05$ H_0 red.

$$\hat{y}_i = -2.8435 + 2.4156X_{i1} + 0.3981X_{i2} - 1.3281 X_{i3} \pm 5.74$$

(18.621) (1.9084) (2.5950) (1.8081)

Katsayı Yorumları:

$$b_0 = -2.8435$$

Trafik kazalarının etkisi – olması gibi bir durum söz konusu olmayacağından denklem gereği bir katsayıdır ve yorumlanmasına gerek yoktur.

$$b_1 = 2.4156$$

x_2 sabit tutulduğunda, hız yapıldığı durumunun 1 olay daha artması yol koşullarının ortalama olarak 2.4156 gün arttırmaktadır. Güven değeri $p=0.209 > \alpha = 0.05$ olduğundan katsayı istatistiksel olarak anlamsızdır.

$$b_2 = 0.3981$$

x_1 sabit tutulduğunda, yol koşullarının 1 olay daha artması yol koşullarının ortalama olarak 0.3981 etkinlikte arttırmaktadır. Güven değeri $p=0.878 > \alpha = 0.05$ olduğundan katsayı istatistiksel olarak anlamsızdır.

$$b_3 = -1.3281$$

x_1 ve x_2 sabit tutulduğunda , hava durumunun 1 olay daha artması yol koşullarının ve hız durumunun ortalama olarak -1.3281 etkinlikte azaltmaktadır. Güven değeri $p=0.464 > \alpha = 0.05$ olduğundan katsayı istatistiksel olarak anlamsızdır.

ARTIK İNCELEMESİ

```
influence.measures(sonuc)
inf<-ls.diag(sonuc)
inf
```

Aykırı Değer (r_i)

Standartlaştırılmış hataların $(-2,+2)$ aralığında olması istenir. Student tipi artıkların ise $(-3,+3)$ arasında olması istenir. Bu aralığın dışında kalan değerlerin aykırı değer olduğu söylenmektedir. Bu durumda 60. Gözlem hem standartlaştırılmış hem de student tipi artıklar için aykırı değer olarak belirlenmiştir.

Cook Uzaklığı(D_i)

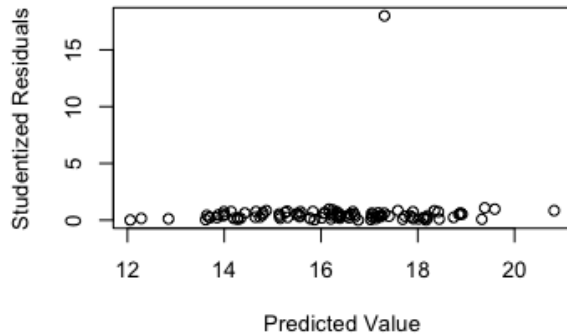
Etkin değer olup olmadığı incelendiğinde 60. Gözlemin aynı zamanda etkin bir aykırı değer olduğu görülmektedir.

$D_{60}=0,001 < 0,04$ olduğundan 60. gözlem etkili gözlem değildir.

Gözlem Uzaklığı (h_{ii})

37. gözlem ve 96. Gözlem sırasıyla 0.10345524,0,10795833<0,06 olduğundan verilerde uç değer yoktur.

```
par(mfrow=c(2,2))
plot(predict(sonuc), inf$stud.res, ylab="Studentized Residuals", xlab="Predicted Value")
```



AYKIRI DEĞER İNCELEMELERİ

COOK UZAKLIĞI

```
library(zoo)
```

```
n<-98
```

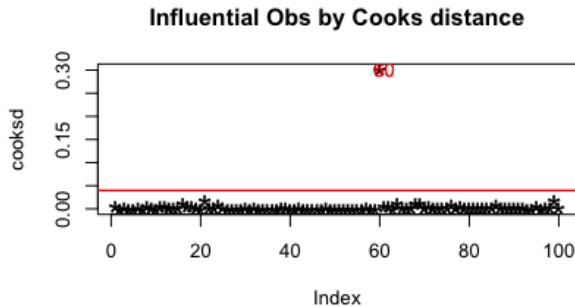
```
k<-4
```

```
cooksds<-cooks.distance(sonuc)
```

```
plot(cooksds, pch="*", cex=2, main="Influential Obs by Cooks distance")
```

```
abline(h = if (length(data$y)>50) 4/length(data$y) else 4/(length(data$y)-(length(data)-1)-1)  
, col="red")
```

```
text(x=1:length(cooksds)+1, y=cooksds, labels=ifelse(cooksds>if (length(data$y)>50)  
4/length(data$y) else 4/(length(data$y)-(length(data)-1)-1),names(cooksds),""), col="red")
```



```
cooksdd<- cooksds[-c(16,21,24,60,64,68,69,76,86,99 )]
```

```
plot(cooksdd, pch="*", cex=2, main="Influential Obs by Cooks distance")
```

```
abline(h = if (n>50) 4/n else 4/(n-k-1) , col="red")
```

```
text(x=1:length(cooksdd)+1, y=cooksdd, labels=ifelse(cooksdd>if (n>50) 4/n else 4/(n-k-  
1),names(cooksdd),""), col="red")
```

```
boxplot(cooksdd)$out
```

```
shapiro.test(cooksdd)
```

```
lncooksdd<-log(cooksdd)
```

```
shapiro.test(lncooksdd)
```

```
boxplot(cooksdd)$out
```

```
cooksdd2<-cooksdd[-c(12,18)]
```

```
shapiro.test(cooksdd2)
```

Shapiro-Wilk normality test

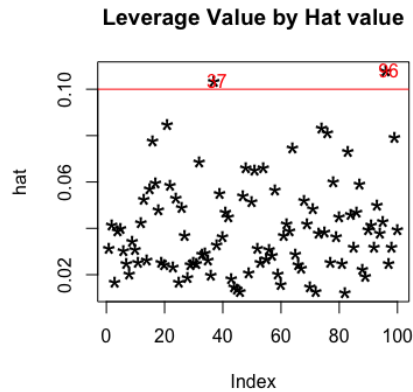
data: cooksdd2

W = 0.8525, p-value = 6.736e-08

Normallik varsayımını sağlamıyor

UÇ DEĞER İLE AYKIRILIK İNCELEMESİ

```
hat<-inf$hat
plot(hat, pch="*", cex=2, main="Leverage Value by Hat value")
abline(h = 2*length(data)/length(y) , col="red")
text(x=1:length(hat)+1, y=hat, labels=ifelse(hat>2*length(data)/length(y),index(hat),""),
col="red")
```



```
boxplot(hat)$out
which(hat%in%boxplot(hat)$out)
hatt<- hat[-c(37,96)]
```

```
plot(hatt, pch="*", cex=2, main="Leverage Value by Hat value")
abline(h = 2*(k+1)/n , col="red")
text(x=1:length(hat)+1, y=hat, labels=ifelse(hat>2*(k+1)/n,index(hat),""), col="red")
```

```
shapiro.test(hatt)
Shapiro-Wilk normality test
```

data: hatt
W = 0.94032, p-value = 0.0002355

```
lnhatt1<-log(hatt)
shapiro.test(lnhatt1)
Shapiro-Wilk normality test
```

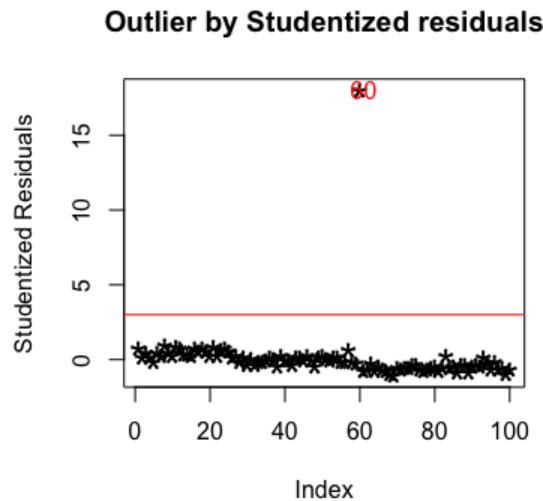
data: lnhatt1
W = 0.9838, p-value = 0.2718

p değeri 0.05 den büyük çıktı ve artık değer bulunmadığı için kabul ediyoruz.

```
plot(lnhatt1, pch="*", cex=2, main="Leverage Value by Hat value")
abline(h = 2*(k+1)/n , col="red")
text(x=1:length(lnhatt1)+1, y=lnhatt1, labels=ifelse(lnhatt1>2*(k+1)/n,index(lnhatt1),""),
col="red")
```

STUDENT TÜRÜ ARTIKLAR

```
stud<-inf$stud.res
plot(stud, pch="*", cex=2, main="Outlier by Studentized residuals",ylab="Studentized
Residuals", xlab="Index")
abline(h = c(-3,3) ,col="red")
text(x=1:length(stud)+1, y=stud, labels=ifelse(stud<-3 & stud>3,index(stud),""), col="red")
```



```
boxplot(stud)$out
which(stud%in%boxplot(std)$out)
studd<- stud[-c(60)]
shapiro.test(studd)
```

Shapiro-Wilk normality test

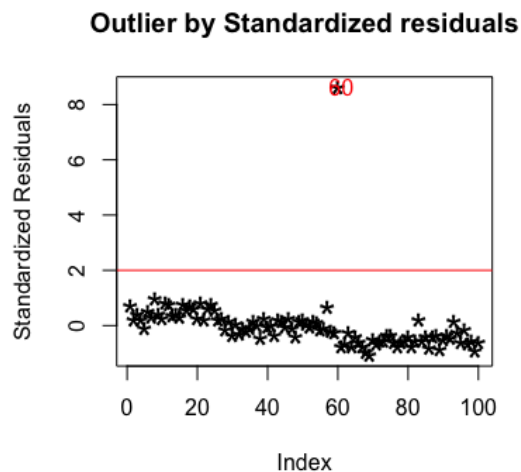
data: studd
W = 0.97671, p-value = 0.07619

```
plot(studd, pch="*", cex=2, main="Outlier by Studentized residuals",ylab="Studentized
Residuals", xlab="Index")
abline(h = c(-3,3) ,col="red")
text(x=1:length(studd)+1, y=studd, labels=ifelse(studd<-3 & studd>3,index(studd),""),
col="red")
```

p değeri 0.05 den büyük çıktı ve artık değer bulunmadığı için kabul ediyoruz.

STANDARTLAŞTIRILMIŞ ARTIKLAR

```
std<-inf$std.res
plot(std, pch="*", cex=2, main="Outlier by Standardized residuals",ylab="Standardized
Residuals", xlab="Index")
abline(h = c(-2,2) ,col="red")
text(x=1:length(std)+1, y=std, labels=ifelse(std<-2 & std>2,index(std),""), col="red")
```



```
boxplot(std)$out
which(std%in%boxplot(std)$out)
stddd<- std[-c(60)]
shapiro.test(stddd)
Shapiro-Wilk normality test
```

data: stddd
W = 0.97633, p-value = 0.07113

```
plot(stddd, pch="*", cex=2, main="Outlier by Standardized residuals",ylab="Standardized
Residuals", xlab="Index")
abline(h = c(-2,2) ,col="red")
text(x=1:length(stddd)+1, y=stddd, labels=ifelse(stddd<-2 & stddd>2,index(std),""), col="red")
shapiro.test(stddd)
```

p değeri 0.05 den büyük çıktı ve artık değer bulunmadığı için kabul ediyoruz

```
confint(sonuc,level = .99)
      0.5 %   99.5 %
(Intercept) -51.779667 46.092682
x1          -2.599735  7.430866
x2          -6.421626  7.217853
x3          -6.079822  3.423625
```

üç bağımsız değişkenin katsayı kestirimleri de sıfır değerini içermediğinden anlamlıdır.

x2 sabit tutulduğunda, hız durumu 1 durum daha arttığında yol koşullarının -2.599735 ile 7.430866 arasında arttığı %99 güven düzeyinde söylenebilir.

x1 sabit tutulduğunda, yol koşulları durumunun 1 durum daha arttığında hız durumunun -6.421626 ile 7.217853 arasında arttığı %99 güven düzeyinde söylenebilir.

x3 sabit tutulduğunda, yol koşulları 1 durum daha arttığında hava durumunun -6.079822 ile 3.423625 arasında arttığı %99 güven düzeyinde söylenebilir.

DEĞİŞEN VARYANSLILIK SORUNU

```
library(lmtest)
bptest(sonuc)
```

studentized Breusch-Pagan test

data: sonuc
BP = 0.72056, df = 3, p-value = 0.8684

Bu durumda değişen varyans sorunu bulunmamaktadır ($p=0.8684 > \alpha=0.05$).

```
summary(lm(abs(residuals(sonuc)) ~ fitted(sonuc)))
```

Call:
lm(formula = abs(residuals(sonuc)) ~ fitted(sonuc))

Residuals:

Min	1Q	Median	3Q	Max
-3.290	-1.444	-0.658	0.615	45.935

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.2454	4.6630	-0.482	0.631
fitted(sonuc)	0.3102	0.2844	1.091	0.278

Residual standard error: 4.89 on 98 degrees of freedom

Multiple R-squared: 0.012, Adjusted R-squared: 0.001918

F-statistic: 1.19 on 1 and 98 DF, p-value: 0.278

Ho: Model anlamsızdır.

$p=0.278 > \alpha=0.05$ ise artıklarla kestirim değerleri arasında oluşturulan modelin anlamsız olduğu söylenebilir. Değişen varyanslılık yoktur.

ÖZ İLİŞKİ SORUNU

dwtest(sonuc)

Durbin-Watson test

data: sonuc

DW = 1.8281, p-value = 0.1859

alternative hypothesis: true autocorrelation is greater than 0

ÇOKLU BAĞLANTI SORUNU

library(car)

vif(sonuc)

	x1	x2	x3
	10.867649	20.546306	9.114318

Koşul sayısı 30 dan küçük olduğu için x1,x2,x3 çoklu bağlantıdan etkilenmemektedir.

```

ort1<-mean(x1)
kt1<-sum((x1-ort1)^2)
skx1<-(x1-ort1)/(kt1^0.5)
ort2<-mean(x2)
kt2<-sum((x2-ort2)^2)
skx2<-(x2-ort2)/(kt2^0.5)
ort3<-mean(x3)
kt3<-sum((x3-ort3)^2)
skx3<-(x3-ort3)/(kt3^0.5)

```

```

x<-cbind(skx1,skx2,skx3)
sm<-eigen(t(x) %*%x)
signif(sm$values,3)
[1] 2.8600 0.1110 0.0321

```

```

signif(sm$vectors,3)
      [,1] [,2] [,3]
[1,] -0.574 0.6760 0.461
[2,] -0.585 0.0547 -0.809
[3,] -0.572 -0.7350 0.364

```

```

t(V)%*%V
      [,1] [,2] [,3]
[1,] 1.0000000e+00 -1.665335e-16 -5.551115e-17
[2,] -1.665335e-16 1.0000000e+00 2.220446e-16
[3,] -5.551115e-17 2.220446e-16 1.0000000e+00

```

```

V%*%diag(sm$values) %*% t(V)
      [,1] [,2] [,3]
[1,] 1.0000000 0.9525017 0.8893460
[2,] 0.9525017 1.0000000 0.9430901
[3,] 0.8893460 0.9430901 1.0000000

```

İLERİYE DOĞRU SEÇİM YÖNTEMİ

```
library(stats)
lm.null<-lm(y~1)
forward<- step(lm.null,y~x1+x2+x3, direction = "forward")
```

Start: AIC=356.36

y ~ 1

	Df	Sum of Sq	RSS	AIC
+ x1	1	270.09	3189.0	350.23
+ x2	1	219.10	3240.0	351.82
+ x3	1	152.22	3306.9	353.86
<none>			3459.1	356.36

Step: AIC=350.23

y ~ x1

	Df	Sum of Sq	RSS	AIC
<none>			3189.0	350.23
+ x3	1	24.8239	3164.2	351.45
+ x2	1	7.8209	3181.2	351.98

Bilgi kriteri değeri =350.23

forward

Call:

```
lm(formula = y ~ x1)
```

Coefficients:

(Intercept)	x1
-0.1403	1.6573

X1 değişkeniyle kurduğu model de diğer değişkenler modele alındığında anlamlı çıkmadığı için modelde yalnızca x1 yer almıştır.

summary(forward)

Call:

lm(formula = y ~ x1)

Residuals:

Min	1Q	Median	3Q	Max
-5.850	-2.681	-0.530	1.595	49.074

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.1403	5.7376	-0.024	0.98054
x1	1.6573	0.5753	2.881	0.00487 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.704 on 98 degrees of freedom

Multiple R-squared: 0.07808, Adjusted R-squared: 0.06867

F-statistic: 8.3 on 1 and 98 DF, p-value: 0.004871

En iyi model: $y_i = -0.1403 + 1.6573x_{1i} \pm 5.704$

(5.7376) (0.001)

Modelin de anlamlı olduğu görülmektedir (p=0.00)

X1 tarafından %0.07 de açıklanabilmektedir.

GERİYE DOĞRU SEÇİM YÖNTEMİ

```
backward<-step(sonuc,direction="backward")
```

Start: AIC=353.42

$y \sim x1 + x2 + x3$

	Df	Sum of Sq	RSS	AIC
- x2	1	0.776	3164.2	351.45
- x3	1	17.779	3181.2	351.98
- x1	1	52.794	3216.2	353.08
<none>			3163.4	353.42

Step: AIC=351.45

$y \sim x1 + x3$

	Df	Sum of Sq	RSS	AIC
- x3	1	24.824	3189.0	350.23
<none>			3164.2	351.45
- x1	1	142.692	3306.9	353.86

Step: AIC=350.23

$y \sim x1$

	Df	Sum of Sq	RSS	AIC
<none>			3189.0	350.23
- x1	1	270.09	3459.1	356.36

En büyük aic değeri x1 dir.

Geriye doğru seçim yönteminin özelliğinden tüm değişkenler modelde olarak başlıyor. İlk model tüm bağımsız değişkenlerin modelde bulunduğu durumdur. İkinci modelde $x2$ değişkeni modelden çıkmış, üçüncü modelde ise $x3$ değişkeni modelden çıkmış, dördüncü modelde $x3$ model dışında kalmıştır. Son modelde yalnızca $x1$ modelde kalmıştır.

```
summary(backward)
```

Call:

`lm(formula = y ~ x1)`

Residuals:

Min 1Q Median 3Q Max

-5.850 -2.681 -0.530 1.595 49.074

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.1403	5.7376	-0.024	0.98054
x1	1.6573	0.5753	2.881	0.00487 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.704 on 98 degrees of freedom

Multiple R-squared: 0.07808, Adjusted R-squared: 0.06867

F-statistic: 8.3 on 1 and 98 DF, p-value: 0.004871

Burada da yalnızca x1 değişkeninin olduğu model en iyi model olarak elde edilmiştir. Modelin anlamlı olduğu söylenebilir (p=0.004). X1 değişkeninin anlamlı olduğu da görülmektedir (P=0.004).

ADIMSAL SEÇİM YÖNTEMİ

library(MASS)

step.model <- stepAIC(sonuc, direction = "both", trace = FALSE)

summary(step.model)

Call:

lm(formula = y ~ x1)

Residuals:

Min	1Q	Median	3Q	Max
-5.850	-2.681	-0.530	1.595	49.074

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.1403	5.7376	-0.024	0.98054
x1	1.6573	0.5753	2.881	0.00487 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.704 on 98 degrees of freedom

Multiple R-squared: 0.07808, Adjusted R-squared: 0.06867

F-statistic: 8.3 on 1 and 98 DF, p-value: 0.004871

İlk modele x_1 değişkeni alınarak başlamış. Diğer bağımsız değişkenler anlamlı olmadığından yöntem tamamen ileri doğru seçim yöntemi gibi sonra ermiştir. En iyi model yalnızca x_1 'in olduğu modeldir.

En iyi model geriye doğru seçim yöntemidir. Çünkü her değişkeni alıp sırayla elemiştir. Hepsini değerlendirdiği için daha kesin sonuç olacağını düşünüyorum.

UYUM KESTİRİMİ

```
predict(sonuc, data.frame(x1=8.536118, x2=-0.52000588, x3=2.691658 ), interval =  
'prediction')  
      fit      lwr      upr  
1 13.99425 2.420786 25.56771
```

ÖN KESTİRİM

```
predict(sonuc, data.frame(x1=1.912, x2=5.684, x3=9.325 ), interval = 'prediction')  
      fit      lwr      upr  
1 -8.346576 -57.15641 40.46326
```

UYUM KESTİRİMİ GÜVEN ARALIĞI

```
predict(sonuc, data.frame(x1=8.536118, x2=-0.52000588, x3=2.691658 ), interval =  
'confidence', level = .95)  
      fit      lwr      upr  
1 13.99425 11.96733 16.02117
```

ÖN KESTİRİM GÜVEN ARALIĞI

```
predict(sonuc, data.frame(x1=1.912, x2=5.684, x3=9.325 ), interval = 'confidence', level=  
.95)  
      fit      lwr      upr  
1 -8.346576 -55.80776 39.1146
```