Robust(Dayanıklı) Regression Uygulama

ELİF EKMEKCİ

2023-06-02

Veri Seti Açıklaması

Aşağıdaki veri analizimiz için Alan Agresti ve Barbara Finlay tarafından yayınlanan Sosyal Bilimler için İstatistiksel Yöntemler, Üçüncü Baskı'da yer alan suç veri kümesini kullanacağız (Prentice Hall, 1997). Değişkenler eyalet kimliği (sid), eyalet adı (state), 100.000 kişi başına şiddet suçları (crime), 1.000.000 kişi başına cinayet (murder), metropol alanlarda yaşayan nüfusun yüzdesi (pctmetro), nüfusun yüzdesi beyaz (pctwhite), lise veya üzeri eğitim almış nüfusun yüzdesi (pcths), yoksulluk sınırı altında yaşayan nüfusun yüzdesi (poverty) ve tek ebeveynli (single) nüfusun yüzdesidir. Veri setinde 51 gözlem mevcuttur. Bu çalışmada suçu tahmin etmek için poverty ve single değişkenlerini kullanacağız.

```
set.seed(300)
library(foreign)
cdata <- read.dta("https://stats.idre.ucla.edu/stat/data/crime.dta")
summary(cdata)
```

```
##       sid          state               crime            murder
##  Min.   : 1.0   Length:51          Min.   :  82.0   Min.   : 1.600
##  1st Qu.:13.5   Class :character   1st Qu.: 326.5   1st Qu.: 3.900
##  Median :26.0   Mode  :character   Median : 515.0   Median : 6.800
##  Mean   :26.0                      Mean   : 612.8   Mean   : 8.727
##  3rd Qu.:38.5                      3rd Qu.: 773.0   3rd Qu.:10.350
##  Max.   :51.0                      Max.   :2922.0   Max.   :78.500
##     pctmetro         pctwhite         pcths          poverty
##  Min.   : 24.00   Min.   :31.80   Min.   :64.30   Min.   : 8.00
##  1st Qu.: 49.55   1st Qu.:79.35   1st Qu.:73.50   1st Qu.:10.70
##  Median : 69.80   Median :87.60   Median :76.70   Median :13.10
##  Mean   : 67.39   Mean   :84.12   Mean   :76.22   Mean   :14.26
##  3rd Qu.: 83.95   3rd Qu.:92.60   3rd Qu.:80.10   3rd Qu.:17.40
##  Max.   :100.00   Max.   :98.50   Max.   :86.60   Max.   :26.40
##      single
##  Min.   : 8.40
##  1st Qu.:10.05
##  Median :10.90
##  Mean   :11.33
##  3rd Qu.:12.05
##  Max.   :22.10
```

En küçük kareler regresyonu yapalım

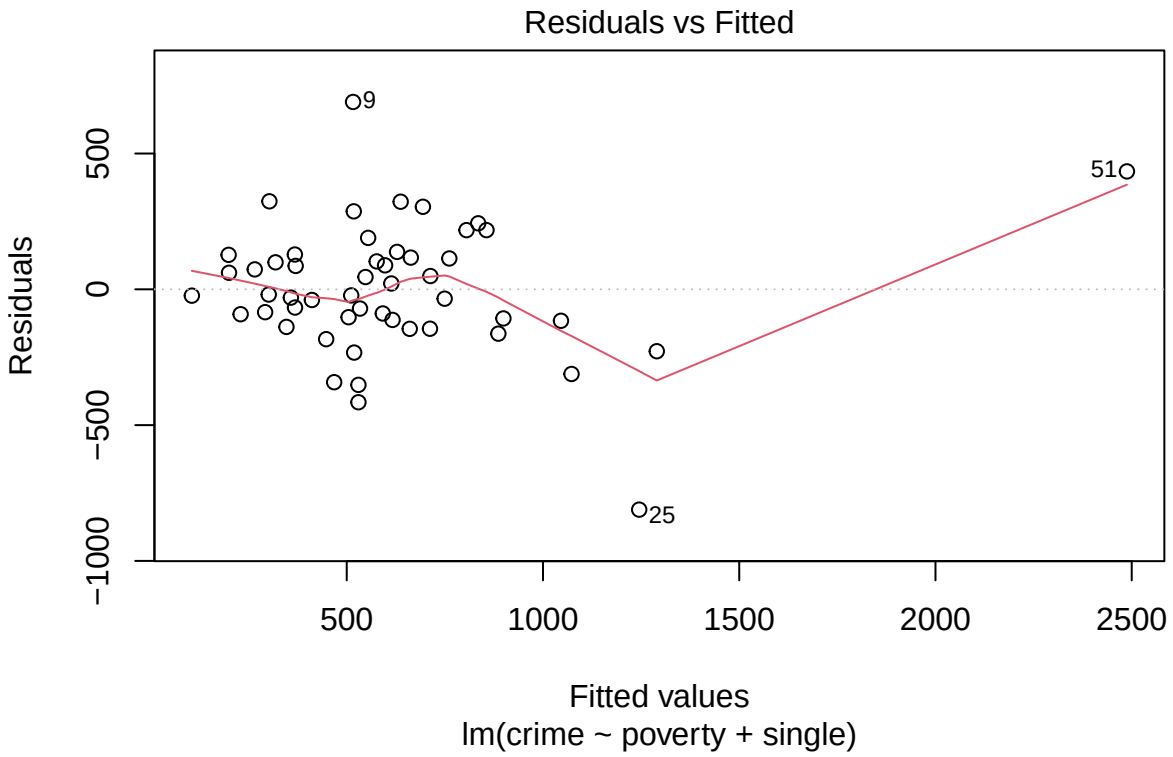
```r
summary(ols <- lm(crime~ poverty + single, data = cdata))
```

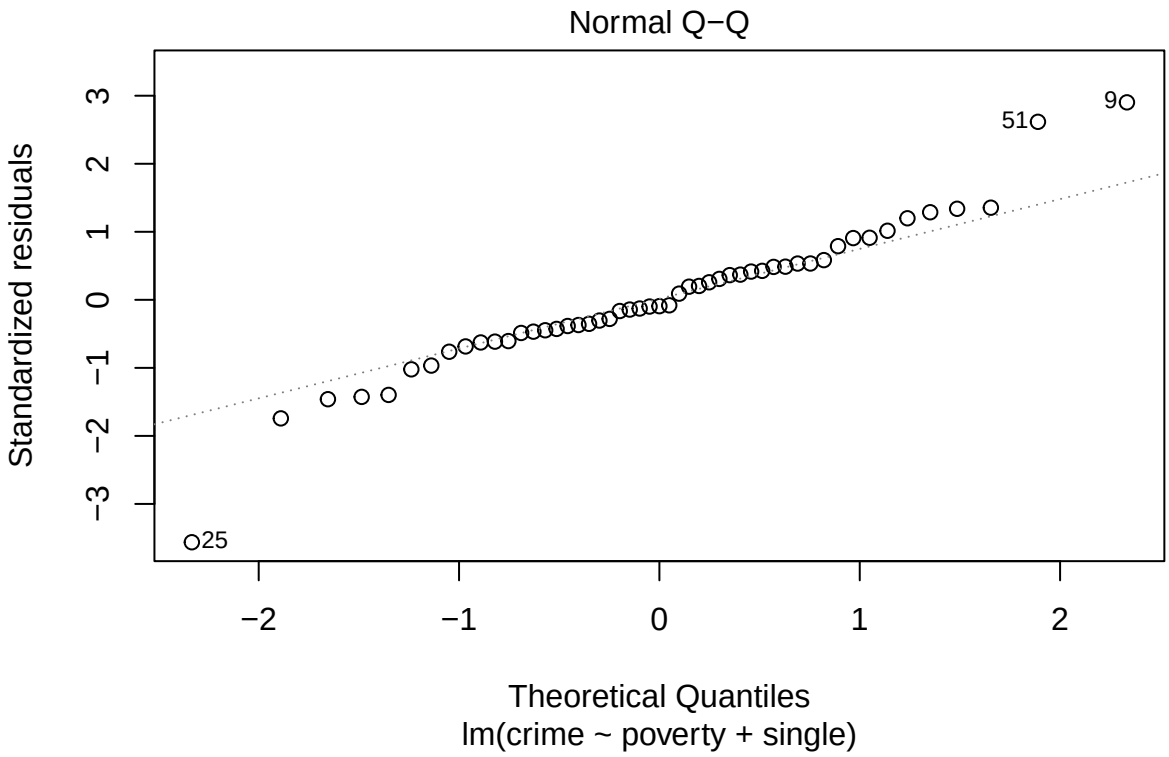
```
##
## Call:
## lm(formula = crime ~ poverty + single, data = cdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.14 -114.27  -22.44  121.86  689.82
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1368.189    187.205  -7.308 2.48e-09 ***
## poverty         6.787      8.989   0.755    0.454
## single        166.373     19.423   8.566 3.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 243.6 on 48 degrees of freedom
## Multiple R-squared:  0.7072, Adjusted R-squared:  0.695
## F-statistic: 57.96 on 2 and 48 DF,  p-value: 1.578e-13
```

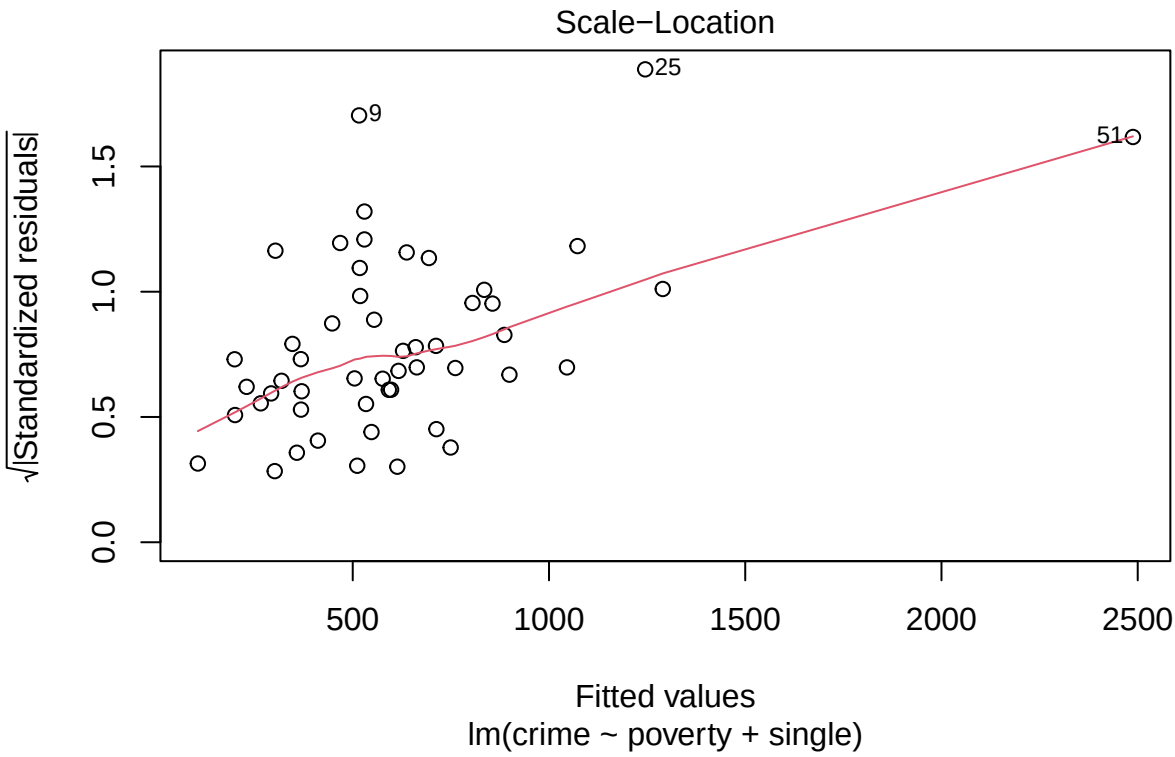
- **poverty** değişkeni anlamlı çıkmadı
- **NOT:** Bazen aykırı değer varlığı, bazı değişkenlerin kullanılmamasından kaynaklanabilir.

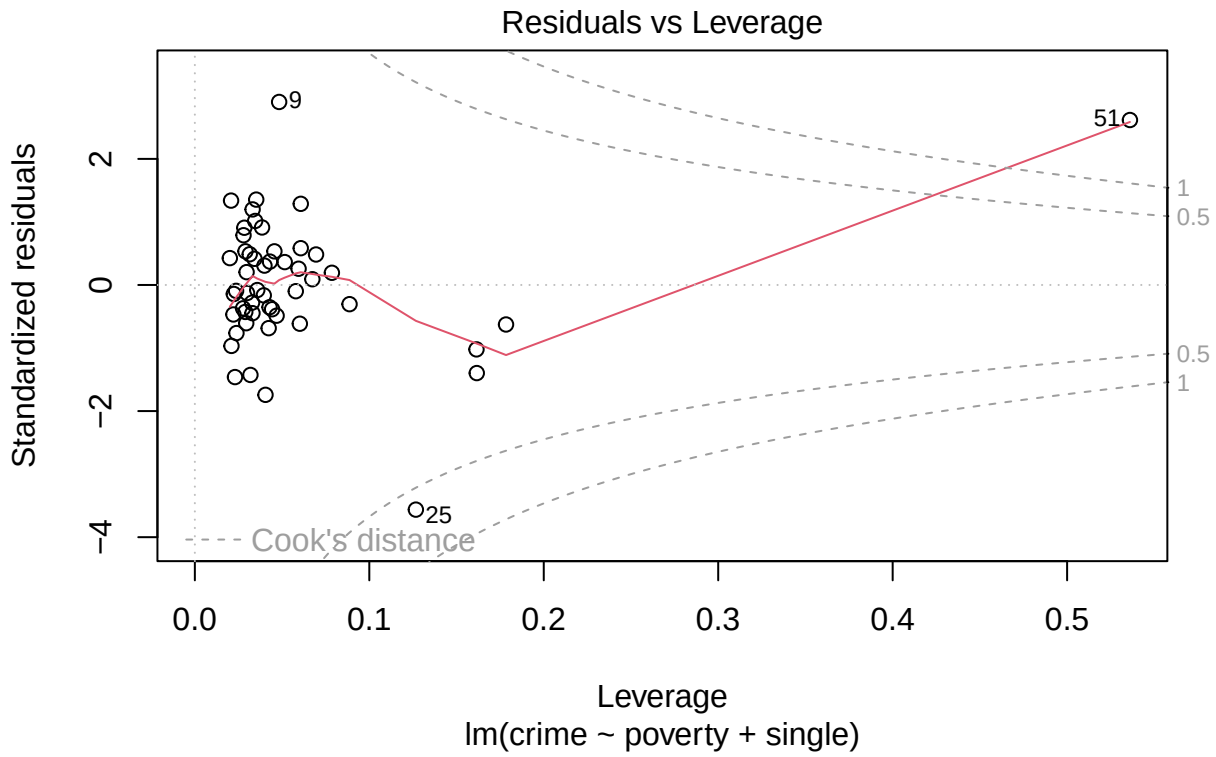
Grafik çizdirelim ve aykırı değerleri daha net görelim

```r
library(faraway)
plot(ols)
```


Residuals vs Fitted
Residuals
500
0
−500
−1000
9
51
25
500
1000
1500
2000
2500
Fitted values
lm(crime ~ poverty + single)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(crime ~ poverty + single)
9
51
25
−3
−2
−1
0
1
2
3


Scale−Location
25
9
51
√|Standardized residuals|
0.0
0.5
1.0
1.5
500
1000
1500
2000
2500
Fitted values
lm(crime ~ poverty + single)



Bu grafiklerden 9, 25 ve 51 gözlemlerini modelimiz için muhtemelen sorunlu olarak tanımlayabiliriz. Bu gözlemlerin hangi durumları temsil ettiklerine bakalım

```
cdata[c(9,25,51),]
```

```
##    sid state crime murder pctmetro pctwhite pcths poverty single
## 9    9    fl  1206    8.9     93.0     83.5  74.4    17.8   10.6
## 25  25    ms   434   13.5     30.7     63.3  64.3    24.7   14.7
## 51  51    dc  2922   78.5    100.0     31.8  73.1    26.4   22.1
```

```
# hangi eyaletlerin sorunlu oldugunu bulmak icin bu kodu calistirdik
```

Cook tdistance değeri **2p/n** den büyük olan gözlemleri ve bunlara karşılık gelen standartlaştırılmış artıkları inceleyelim.

```
library(MASS)
d1 <- cooks.distance(ols)
r <- stdres(ols) #stdress() uygun sekilde donusturulmus artiklarin vektoru
a <- cbind(cdata,d1,r)
a[d1>4/51,]
```

```
##    sid state crime murder pctmetro pctwhite pcths poverty single        d1
## 1    1    ak   761    9.0     41.8     75.2  86.6     9.1   14.3 0.1254750
## 9    9    fl  1206    8.9     93.0     83.5  74.4    17.8   10.6 0.1425891
## 25  25    ms   434   13.5     30.7     63.3  64.3    24.7   14.7 0.6138721
```

```
## 51  51    dc  2922   78.5    100.0     31.8  73.1    26.4   22.1 2.6362519
##            r
## 1  -1.397418
## 9   2.902663
## 25 -3.562990
## 51  2.616447
```

a[d1>4/51,] kodu ile cook distance değeri 2p/n'den büyük olan gözlemleri ve bu gözlemlere karşılık gelen standartlaştırılmış artıkları buluyoruz

Şimdi artıklara bakacağız. Artıkların mutlak değeri olan **rabs** adında yeni bir değişken üreteceğiz (çünkü artık işareti önemli değil). Daha sonra en yüksek mutlak artık değeri olan ilk 10 gözleme bakacağız.

```
rabs <- abs(r)
a <- cbind(cdata, d1, r, rabs)
asorted <- a[order(-rabs), ]
asorted[1:10, ]
```

```
##    sid state crime murder pctmetro pctwhite pcths poverty single         d1
## 25  25    ms   434   13.5     30.7     63.3  64.3    24.7   14.7 0.61387212
## 9    9    fl  1206    8.9     93.0     83.5  74.4    17.8   10.6 0.14258909
## 51  51    dc  2922   78.5    100.0     31.8  73.1    26.4   22.1 2.63625193
## 46  46    vt   114    3.6     27.0     98.4  80.8    10.0   11.0 0.04271548
## 26  26    mt   178    3.0     24.0     92.6  81.0    14.9   10.8 0.01675501
## 21  21    me   126    1.6     35.7     98.5  78.8    10.7   10.6 0.02233128
## 1    1    ak   761    9.0     41.8     75.2  86.6     9.1   14.3 0.12547500
## 31  31    nj   627    5.3    100.0     80.8  76.7    10.9    9.6 0.02229184
## 14  14    il   960   11.4     84.0     81.0  76.2    13.6   11.5 0.01265689
## 20  20    md   998   12.7     92.8     68.9  78.4     9.7   12.0 0.03569623
##            r     rabs
## 25 -3.562990 3.562990
## 9   2.902663 2.902663
## 51  2.616447 2.616447
## 46 -1.742409 1.742409
## 26 -1.460885 1.460885
## 21 -1.426741 1.426741
## 1  -1.397418 1.397418
## 31  1.354149 1.354149
## 14  1.338192 1.338192
## 20  1.287087 1.287087
```

```
# en yuksek mutlak artik degeri olan ilk 10 gozlem
```

NOT: Çıktıdan görüldüğü üzere en büyük artık değeri state = ms'de. Bu yüzden en küçük ağırlık bu eyalete verilecek.

NOT: Robust regresyon için MASS kütüphanesindeki **rlm()** fonksiyonunu kullanıyoruz.

Şimdi ilk sağlam regresyonumuzu gerçekleştirelim. Sağlam regresyon iteratif yeniden ağırlıklı en küçük kareler (IRLS) ile yapılır. Sağlam regresyon çalıştırma komutu MASS paketinde rlm'dir. IRLS için kullanılabilecek çeşitli ağırlık fonksiyonları vardır. Bu örnekte önce **Huber ağırlıklarını** kullanacağız. Daha sonra IRLS işlemi tarafından oluşturulan son ağırlıklara bakacağız.

```
summary(rr.huber <- rlm(crime ~ poverty+single, data = cdata))
```

```
##
## Call: rlm(formula = crime ~ poverty + single, data = cdata)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -846.09 -125.80  -16.49  119.15  679.94
##
## Coefficients:
##             Value      Std. Error t value
## (Intercept) -1423.0373   167.5899    -8.4912
## poverty         8.8677     8.0467     1.1020
## single        168.9858    17.3878     9.7186
##
## Residual standard error: 181.8 on 48 degrees of freedom
```

```
summary(rlm(crime ~ poverty+single, data = cdata, psi = psi.huber))
```

```
##
## Call: rlm(formula = crime ~ poverty + single, data = cdata, psi = psi.huber)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -846.09 -125.80  -16.49  119.15  679.94
##
## Coefficients:
##             Value      Std. Error t value
## (Intercept) -1423.0373   167.5899    -8.4912
## poverty         8.8677     8.0467     1.1020
## single        168.9858    17.3878     9.7186
##
## Residual standard error: 181.8 on 48 degrees of freedom
```

```
# default olarak huber agirliklandirilmasi yapiliyor
# psi bilesenini yazmazsak rlm fonksiyonu otomatik olarak huber agirliklarinin kullanir
```

Kabaca, mutlak artık azaldıkça, ağırlığın arttığını görebiliriz. Başka bir deyişle, büyük kalıntıları olan vakalar düşük ağırlıklı olma eğilimindedir. Bu çıktı bize Mississippi gözleminin en düşük ağırlıklı olacağını gösteriyor. Florida da önemli ölçüde düşük ağırlıklı olacaktır. Yukarıda gösterilmeyen tüm gözlemler 1 ağırlığa sahiptir. OLS regresyonunda, tüm vakalar 1 ağırlığa sahiptir. Bu nedenle, robust(sağlam) regresyonda bire yakın ağırlığa sahip vakalar ne kadar fazla olursa, OLS ve robust(sağlam) regresyonların sonuçları o kadar yakın olur.

ÖZETLE artığı fazla olanlara düşük ağırlık verilir. 53.satirdaki kod chunk çalıştırıldığında en fazla artık değerinin Mississippi'ye ait olduğu görülüyor. Dolayısıyla en küçük ağırlık bu gözleme verilecek.

Şimdi de **bisquare ağırlıklandırmasını** kullanarak regresyon modelimizi kuralım

```
rr.bisquare <- rlm(crime ~ poverty+single, data = cdata, psi = psi.bisquare)
summary(rr.bisquare)
```

```
##
## Call: rlm(formula = crime ~ poverty + single, data = cdata, psi = psi.bisquare)
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -905.59 -140.97  -14.98  114.65  668.38
##
## Coefficients:
##             Value      Std. Error t value
## (Intercept) -1535.3338   164.5062    -9.3330
## poverty        11.6903     7.8987     1.4800
## single        175.9303    17.0678    10.3077
##
## Residual standard error: 202.3 on 48 degrees of freedom
```

Tekrar ağırlıklara bakalım

```
biweights <- data.frame(state = cdata$state, resid = rr.bisquare$resid,weight = rr.bisquare$w )
biweights2 <- biweights[order(rr.bisquare$w),]
# order default olarak kucukten buyuge siraliyor
biweights2[1:15,]
```

```
##    state     resid      weight
## 25    ms -905.5931 0.007652565
## 9     fl  668.3844 0.252870542
## 46    vt -402.8031 0.671495418
## 26    mt -360.8997 0.731136908
## 31    nj  345.9780 0.751347695
## 18    la -332.6527 0.768938330
## 21    me -328.6143 0.774103322
## 1     ak -325.8519 0.777662383
## 14    il  313.1466 0.793658594
## 20    md  308.7737 0.799065530
## 19    ma  297.6068 0.812596833
## 51    dc  260.6489 0.854441716
## 50    wy -234.1952 0.881660897
## 5     ca  201.4407 0.911713981
## 10    ga -186.5799 0.924033113
```

Mississippi'ye verilen ağırlığın, bisquare ağırlıklandırması ile Huber ağırlıklandırmasına göre elde edilenden çok daha düşük olduğunu ve bu iki farklı ağırlıklandırma yönteminden parametre tahminlerinin farklı olduğunu görebiliriz.

-Sıradan en küçük kareler regresyonu ve robust(sağlam) regresyonun sonuçlarını karşılaştırırken, sonuçlar çok farklıysa, robust(sağlam) regresyondan gelen sonuçlar kullanılır.

-Büyük farklılıklar, model parametrelerinin aykırı değerlerden büyük oranda etkilendiğini göstermektedir.

-Farklı ağırlıklandırmaların avantajları ve dezavantajları vardır. Huber ağırlıkları şiddetli aykırı değerlerde zorluklar yaşayabilir ve bisquare ağırlıklar yakınsamada zorluk yaşayabilir veya birden fazla çözüm verebilir.

SONUÇ: İki modelin residual standart error'lerine bakıldığı zaman Huber yöntemi daha küçük residual standart error değerine sahiptir. Dolayısıyla Huber yöntemi ile kurulan model daha iyi performans gösterecektir.