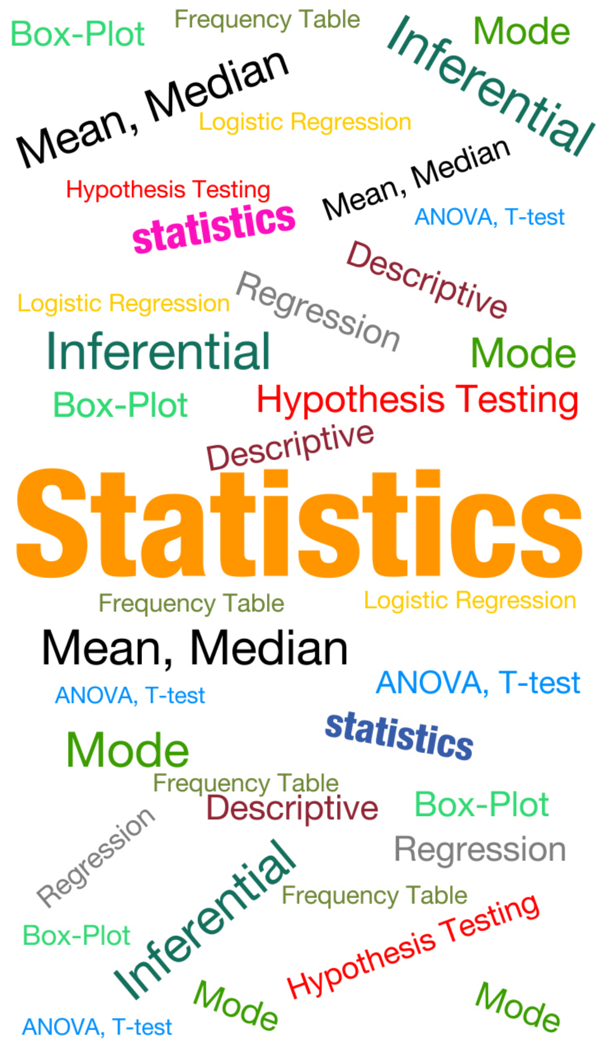


R ile Temel İstatistik

Muhammed Fatih TÜZEN

11 01 2022

İÇİNDEKİLER

1 Giriş

İstatistik; amacın belirlenmesi, çalışmanın planlanması, verilerin toplanması, değerlendirilmesi ve karara varılması sürecini içeren bir bilim dalıdır. İstatistik bilimi içinde örneklemden elde edilen bilgileri kitlelere genelleme, tahminler yapma, değişkenler arasındaki ilişkileri ortaya çıkarma gibi konular yer almaktadır.

Uygulamalı istatistikler iki alana ayrılabilir: tanımlayıcı istatistikler ve çıkarımsal istatistikler. Tanımlayıcı istatistikler, tabloları, grafikleri ve özet ölçüleri kullanarak verileri düzenleme, görüntüleme ve tanımlama yöntemlerinden oluşur. Buna karşılık çıkarımsal istatistikler, bir popülasyon hakkında kararlar veya tahminler yapmak için örnek sonuçlarını kullanan yöntemlerden oluşur.

Tanımlayıcı istatistik, bir dizi değeri veya bir veri kümesini özetlemeyi, tanımlamayı ve sunmayı amaçlayan bir istatistik dalıdır. Tanımlayıcı istatistikler genellikle herhangi bir istatistiksel analizin ilk adımı ve önemli bir parçasıdır. Verilerin kalitesini kontrol etmeyi sağlar ve net bir genel bakışa sahip olarak verileri anlamaya yardımcı olur. Tanımlayıcı istatistikler, merkezi eğilim ölçüleri ve dağılım ölçüleri olmak üzere ikiye ayrılır.

2 Merkezi Eğilim Ölçüleri

Dağılımın konumu hakkında bilgi veren ölçümlerdir. Aritmetik ortalama, geometrik ortalama, harmonik ortalama, düzeltilmiş ortalama, ortanca, çeyrekler, yüzdelikler konum ölçülerine örnek olarak verilebilir.

2.1 Aritmetik Ortalama

- Günlük hayatta en sık kullanılan merkezi eğilim ölçüsüdür.
- Üzerinde inceleme yapılan veri setindeki elemanların toplanıp incelenen eleman sayısına bölünmesiyle elde edilir.
- Konum olarak verilerin en çok hangi değer etrafında toplandığının ya da yoğunlaştığının sayısal bir ölçüsüdür.
- Hem kitle hem de örneklem için hesaplanır.
- Dağılışların yerinin belirlenmesinde en çok kullanılan yer ölçüsü aritmetik ortalamadır; ve tek başına ortalama sözcüğünden aritmetik ortalama anlaşılır.
- Aritmetik ortalama bütün değerlerin ağırlığını eşit kabul ettiğinden dağılımı her zaman en iyi şekilde temsil etmeyebilir. Ayrıca aritmetik ortalama, veri kümesindeki aşırı değerlerden çok kolay etkilenir.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

```
mean(airquality$Wind)
```

```
## [1] 9.957516
```

```
mean(airquality$Ozone, na.rm = TRUE) # NA'ler kaldırılarak ortalama hesaplanır
```

```
## [1] 42.12931
```

2.2 Geometrik Ortalama

- Periyodik artışlar veya azalmalar (değişim oranları) içeren enflasyon veya nüfus değişiklikleri gibi konuları incelerken, geometrik ortalama, incelenen tüm dönem boyunca ortalama değişikliği bulmak için daha uygundur.
- Eğer veriler sıfır ya da negatif değerler içeriyorsa geometrik ortalama hesaplanamaz.
- Geometrik ortalama, uç değerlerden aritmetik ortalamaya göre daha az etkilenmektedir.
- Geometrik Ortalama <= Aritmetik Ortalama

$$G.O. =^n \sqrt{\prod_{i=1}^{n} X_i}$$

```
# R programında hazır geometrik ortalama fonksiyonu yoktur.
# 1. yol
geo_mean <- function(x){
  x <- na.omit(x)
  (prod(x))^(1/length(x))
}

round(geo_mean(airquality$Wind),3)
```

```
## [1] 9.273
```

```
round(geo_mean(airquality$Ozone),3)
```

```
## [1] 30.524
```

```
# 2. yol
library(psych)
round(geometric.mean(airquality$Wind),3)
```

```
## [1] 9.273
```

```
round(geometric.mean(airquality$Ozone),3)
```

```
## [1] 30.524
```

2.3 Medyan (Ortanca)

- Gözlem değerleri küçükten büyüğe sıralandığında ortada kalan gözlem değeridir.
- Bir seride yer alan gözlemlerin tümünün hesaba katılmadığı ortalamalardan biridir.
- Basit serilerde seri tek sayıda gözlemden oluşuyorsa serinin gözlem değerleri küçükten büyüğe sıralandığında tam ortada yer alan gözlem değeridir.
- Seri çift sayıda gözlemden oluşuyorsa ortada kalan iki gözlem değerinin aritmetik ortalaması medyandır.
- Medyan, ölçümlerin %50'sinin üzerinde, %50'sinin aşağısında yer aldığı merkezi değerdir.
- Dağılımdaki aşırı değerlerden etkilenmez.
- Aritmetik ortalamaya kıyasla daha tutarlı bir sonuç elde edilir.
- Her bir veri seti için bir tek medyan söz konusudur.
- Medyanın zayıf tarafı serideki bütün değerleri dikkate almaması sebebi ile matematik işlemlere elverişli değildir.
- Gözlem sayısı (n) tek ise , $\widetilde{X} = X_{\frac{n+1}{2}}$
- Gözlem sayısı (n) çift ise , $\widetilde{X} = \frac{X_{\frac{n}{2}} + X_{\frac{n+1}{2}}}{2}$

```
median(airquality$Wind)
```

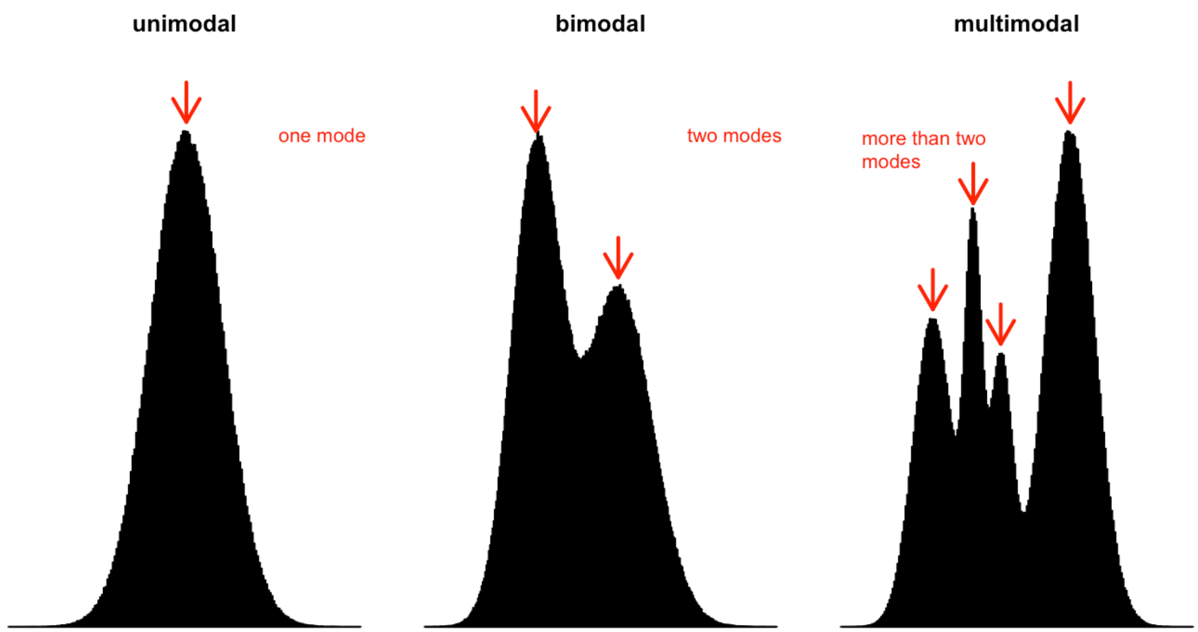
```
## [1] 9.7
```

```
median(airquality$Ozone,na.rm = TRUE)
```

```
## [1] 31.5
```

2.4 Mod (Tepe değeri)

- En sık ortaya çıkan (en yüksek frekanslı) ölçümdür.
- Dağılımdaki aşırı değerlerden etkilenmez
- Her dağılımda tepe değeri bulunmayabilir.
- Bazı dağılımlarda birden fazla tepe değeri bulunabilir.
- Tepe değeri aritmetik işlemler için elverişli değildir.
- Tüm veri değerlerini göz önünde bulundurmadığı için tutarlı olmayan bir merkezi eğilim ölçüsüdür.
- Gözlem sayısı az olduğunda tepe değer güvenilir bir ölçü değildir.




```r
# R programında hazır mod fonksiyonu yoktur.

library(DescTools)
Mode(airquality$Wind)
```

```
## [1] 11.5
## attr(,"freq")
## [1] 15
```

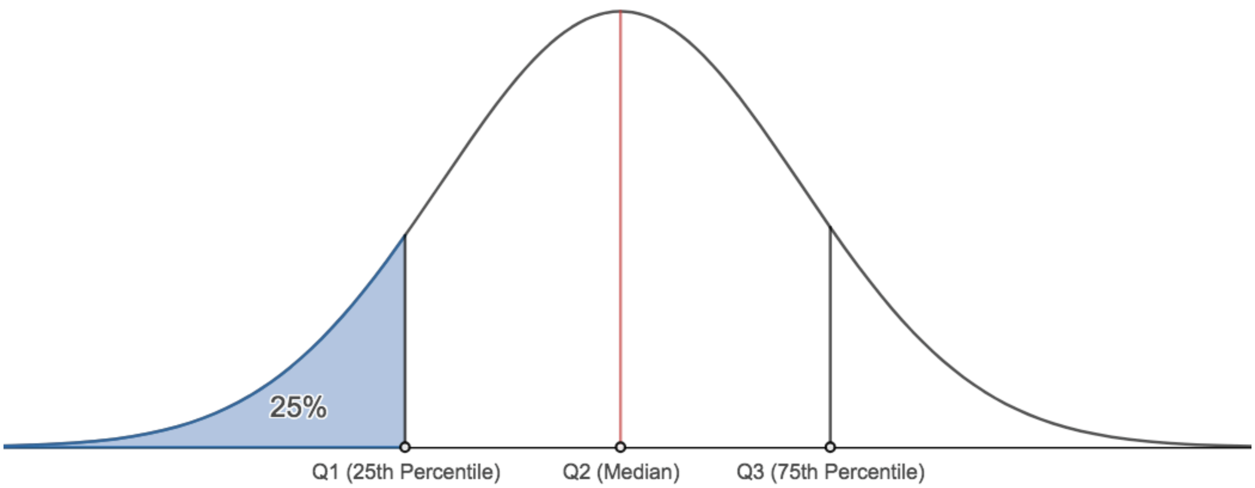
```r
Mode(airquality$Solar.R,na.rm = TRUE)
```

```
## [1] 238 259
## attr(,"freq")
## [1] 4
```

2.5 Çeyreklikler

- Birinci Bölen ilk yüzde 25 nci noktadır ve verinin ¼ kadarı birinci bölen içerisinde kalır.

- İkinci Bölen ilk yüzde 50 nci noktadır ve verinin yarısı bu noktanın altında kalır(½) aynı zamanda ikinci bölen medyan olarak ta bilinir.
- Üçüncü Bölen ilk yüzde 75 nci veri kümesidir ve bütün verinin ¾ kadarı bu noktanın altında kalır.
- Gözlem sayısı (n) tek ise , $Q_1 = X_{\frac{n+1}{4}}$
- Gözlem sayısı (n) çift ise , $Q_1 = \frac{X_{\frac{n}{4}} + X_{\frac{n}{4}+1}}{2}$
- Gözlem sayısı (n) tek ise , $Q_3 = X_{\frac{3(n+1)}{4}}$
- Gözlem sayısı (n) çift ise , $Q_3 = \frac{X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1}}{2}$




```
quantile(airquality$Wind,na.rm = TRUE)
```

```
##    0%   25%   50%   75%  100%
##   1.7   7.4   9.7  11.5  20.7
```

```
median(airquality$Wind,na.rm = TRUE)
```

```
## [1] 9.7
```

```
quantile(airquality$Wind,na.rm = TRUE,probs = 0.75) #Q3
```

```
##  75%
## 11.5
```

```
quantile(airquality$Wind,na.rm = TRUE,probs = 0.25) #Q1
```

```
## 25%
## 7.4
```

```
quantile(airquality$Wind,na.rm = TRUE,probs = c(0.20,0.50,0.80)) # %20,%50,%80
```

```
##   20%   50%   80%
##  6.90  9.70 12.96
```

```
quantile(airquality$Solar.R,na.rm = TRUE)
```

```
##      0%    25%    50%    75%   100%
##   7.00 115.75 205.00 258.75 334.00
```

```
median(airquality$Solar.R,na.rm = TRUE)
```

```
## [1] 205
```

3 Dağılım Ölçüleri

Ortalama, medyan ve mod gibi merkezi eğilim ölçüleri, bir veri setinin dağılımının bütün resmini ortaya koymaz. Aynı ortalamaya sahip iki veri seti tamamen farklı yayılımlara sahip olabilir. Bir veri seti için gözlem değerleri arasındaki farklılık, diğer veri seti için olduğundan çok daha büyük veya daha küçük olabilir. Bu nedenle, ortalama, medyan veya mod tek başına genellikle bir veri kümesinin dağılımının şeklini ortaya çıkarmak için yeterli bir ölçü değildir. Bu yüzden veri değerleri arasındaki varyasyon hakkında bazı bilgiler sağlayabilecek bir ölçülere de ihtiyaç vardır. Bu ölçülere dağılım (yayılım) ölçüleri denir. Birlikte ele alınan merkezi eğilim ve dağılım ölçüleri, tek başına merkezi eğilim ölçülerinden ziyade bir veri setinin daha iyi bir resmini verir. Değişim aralığı, çeyrekler arası genişlik, varyans, standart sapma, basıklık, çarpıklık, min, max başlıca dağılım ölçüleri arasındadır.

3.1 Değişim Aralığı (Açıklık)

- Veri setindeki en büyük değer ile en küçük değer arasındaki farktır.
- En basit dağılım ölçüsü olmakla birlikte uç ve aykırı değerlerden etkilenmesi olumsuz yönüdür.

- Serinin sadece 2 gözlemine bağlı olarak hesaplanan bu ölçü değişkenliğin şekli hakkında çok fazla bilgi vermediğinden diğer değişkenlik ölçüleri kadar sık kullanılmaz.

$$D.A = max(X) - min(X)$$

```r
# 1. yol
max(airquality$Ozone,na.rm = TRUE)-min(airquality$Ozone,na.rm = TRUE)
```

```
## [1] 167
```

```r
# 2. yol
range(airquality$Ozone,na.rm = TRUE)
```

```
## [1]   1 168
```

```r
range(airquality$Ozone,na.rm = TRUE)[2]-range(airquality$Ozone,na.rm = TRUE)[1]
```

```
## [1] 167
```

3.2 Çeyrekler Arası Genişlik

- Dağılımdaki verilerin ortadaki % 50'sinin yer aldığı aralığı belirlemek için kullanılır.
- Aşırı uç değerlerden etkilenmez. Çünkü çeyreklikler arası genişlik dağılımdaki değerlerin merkezdeki %50'si ile ilgilenir.
- Çeyrekler arası bir genişlik, değerlerin büyük kısmının nerede olduğunu gösteren bir ölçüdür.
- Çeyrek Sapma 3. çeyrek ile 1. çeyrek arasındaki farktır.
- IQR (Interquartile Range) olarak ifade edilir.

$$IQR = Q_3 - Q1$$

```r
# 1.yol
q3 <- quantile(airquality$Wind,na.rm = TRUE,probs = 0.75) #Q3
q1 <- quantile(airquality$Wind,na.rm = TRUE,probs = 0.25) #Q1
q3-q1
```

```
## 75%
## 4.1
```

```
# 2. yol
IQR(airquality$Wind,na.rm = TRUE)
```

```
## [1] 4.1
```

3.3 Varyans ve Standart Sapma

Gözlem değerlerinin aritmetik ortalamadan sapmaları dikkate alınarak farklı değişkenlik ölçüleri geliştirilebilir. Ancak gözlemlerin aritmetik ortalamadan sapmalarının her zaman sıfıra eşittir. Bu sorunu ortadan kaldırmak için gözlemlerin aritmetik ortalamadan olan sapmalarının karelerinin toplamının gözlem sayısına oranı değişkenlik ölçüsü olarak yorumlanabilir. Bu ölçü varyans olarak adlandırılır.

- Bir dağılımda değerler aritmetik ortalamadan uzaklaştıkça dağılımın yaygınlığı artar.
- Varyansın karekökü standart sapmadır. Genel olarak, bir veri kümesi için standart sapmanın daha düşük bir değeri, o veri kümesinin değerlerinin ortalama etrafında nispeten daha küçük bir aralığa yayıldığını gösterir. Buna karşılık, bir veri kümesi için standart sapmanın daha büyük bir değeri, o veri kümesinin değerlerinin, ortalama etrafında nispeten daha geniş bir aralığa yayıldığını gösterir.
- Kitle varyansı σ^2 ile standart sapma ise σ ile gösterilmektedir. Örneklem standart sapması ise s ile ifade edilir.

$$s = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{n-1}}$$

```
var(airquality$Wind,na.rm=TRUE)
```

```
## [1] 12.41154
```

```
sd(airquality$Wind,na.rm=TRUE)
```

```
## [1] 3.523001
```

```
var(airquality$Solar.R,na.rm=TRUE)
```

```
## [1] 8110.519
```

```
sd(airquality$Solar.R,na.rm=TRUE)
```

```
## [1] 90.05842
```

3.4 Değişim Katsayısı

- Farklı serilerin değişkenliklerinin karşılaştırılmasında, farklı birimlerle ölçülmüş veri setleri söz konusu olduğundan standart sapma kullanışlı değildir.
- Bunun yerine ilgili serilerin standart sapmaları serilerin ortalama değerinin yüzdesi olarak ifade edilir ve gözlem değerlerinin büyüklüklerinden kaynaklanan farklılık ortadan kalkmış olur.
- Elde edilen bu yeni değişkenlik ölçüsü kullanılarak serilerin birbirlerine göre daha değişken ya da daha homojen oldukları konusunda yorum yapılabilir.
- Bu değer ne kadar küçükse dağılım o kadar homojendir, değişkenlik azdır. Yüzdesel olarak ifade edilir.
- Değişim Katsayısı standart sapmanın aritmetik ortalamaya bölünüp 100 ile çarpılmasıyla elde edilir.

$$D.K. = \frac{S}{\bar{X}} \times 100$$

```
dk_wind <- sd(airquality$Wind,na.rm=TRUE)/mean(airquality$Wind,na.rm=TRUE)
dk_wind
```

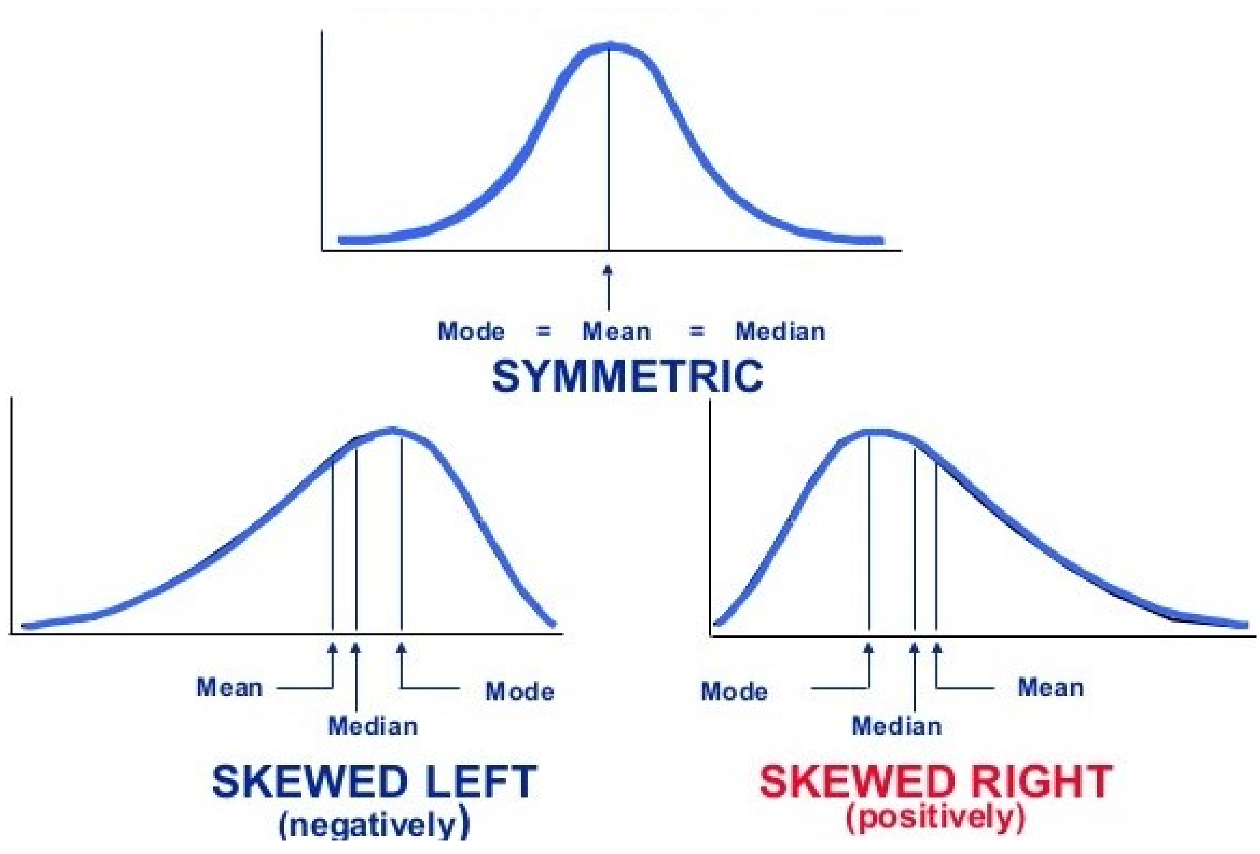
```
## [1] 0.3538032
```

```
dk_solar <- sd(airquality$Solar.R,na.rm=TRUE)/mean(airquality$Solar.R,na.rm=TRUE)
dk_solar
```

```
## [1] 0.4843634
```

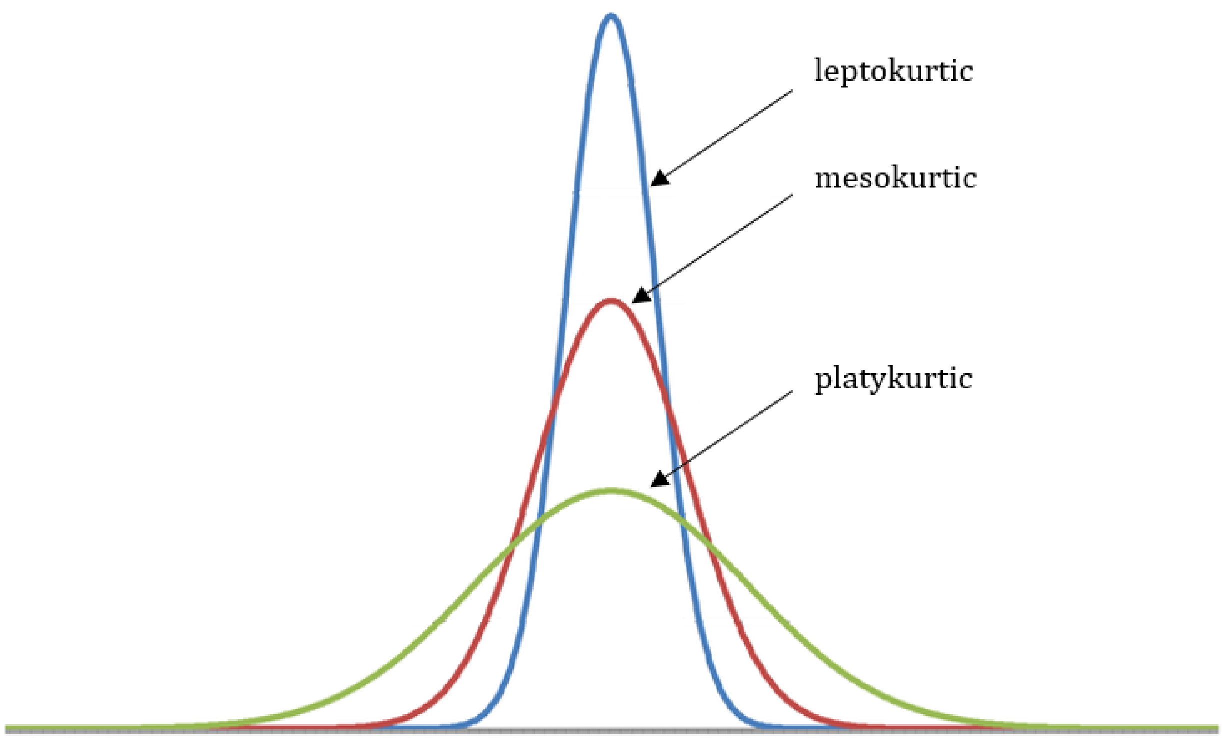
3.5 Çarpıklık ve Basıklık

- Bir dağılımın normal dağılıma göre çarpık olup olmadığını belirlemede kullanılır. Simetrik dağılımlarda ortalama, ortanca ve tepe değeri birbirine eşittir.
- Çarpıklık katsayısı 0 ise dağılım simetriktir, 0'dan küçük ise sola çarpıktır (negatif çarpıklık), 0'dan büyük ise sağa çarpıktır (pozitif çarpıklık).
- Pozitif çarpıklıkta sağ kuyruk daha uzun iken negatif çarpıklıkta sol kuyruk daha uzundur.
- Aritmetik Ortalama, Medyan ve Mod arasındaki ilişkilere göre de çarpıklık belirlenebilir.
  - Mod < Medyan < Ortalama ise, dağılım sağa-çarpık yani (+) yöne eğilimli dağılımdır.
  - Ortalama < Medyan < Mod ise, dağılım sola-çarpık yani (-) yöne eğilimli dağılımdır.
  - Ortalama = Mod = Medyan ise, dağılım simetrik dağılımdır.



- Bir dağılımın normal dağılıma göre basık olup olmadığını belirlemede kullanılır.

- Basıklık katsayısı sıfırdan büyükse normal dağılıma göre daha sivri, küçük ise daha basıktır.
- Basıklık katsayısı 3'e eşit ise seri normal dağılıma (mesokurtic) sahiptir. Eğer3'ten küçük ise, bir platykurtik dağılımı gösterir (daha kısa kuyruklu normal dağılımdan daha düz). Eğer 3'ten büyük ise, bir leptokurtik dağılımı gösterir (daha uzun kuyruklu normal dağılımdan daha doruğa).
- İki veya daha fazla simetrik dağılım karşılaştırıldığında aralarındaki fark basıklık ile incelenir.



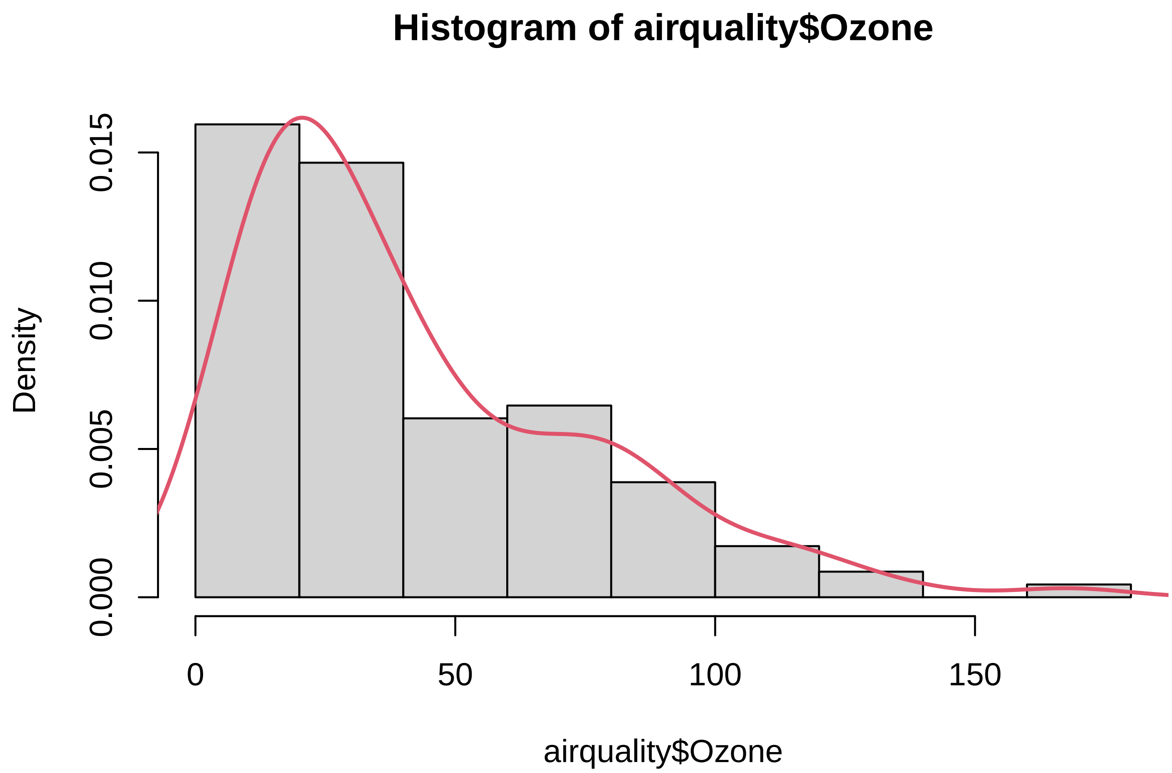

```
library(moments)
skewness(airquality$Ozone,na.rm = TRUE) # sağa çarpık
```

```
## [1] 1.225681
```

```
kurtosis(airquality$Ozone,na.rm = TRUE) # sivri
```

```
## [1] 4.184071
```

```
hist(airquality$Ozone,freq = FALSE)
lines(density(airquality$Ozone,na.rm = TRUE),col = 2, lwd = 2)
```



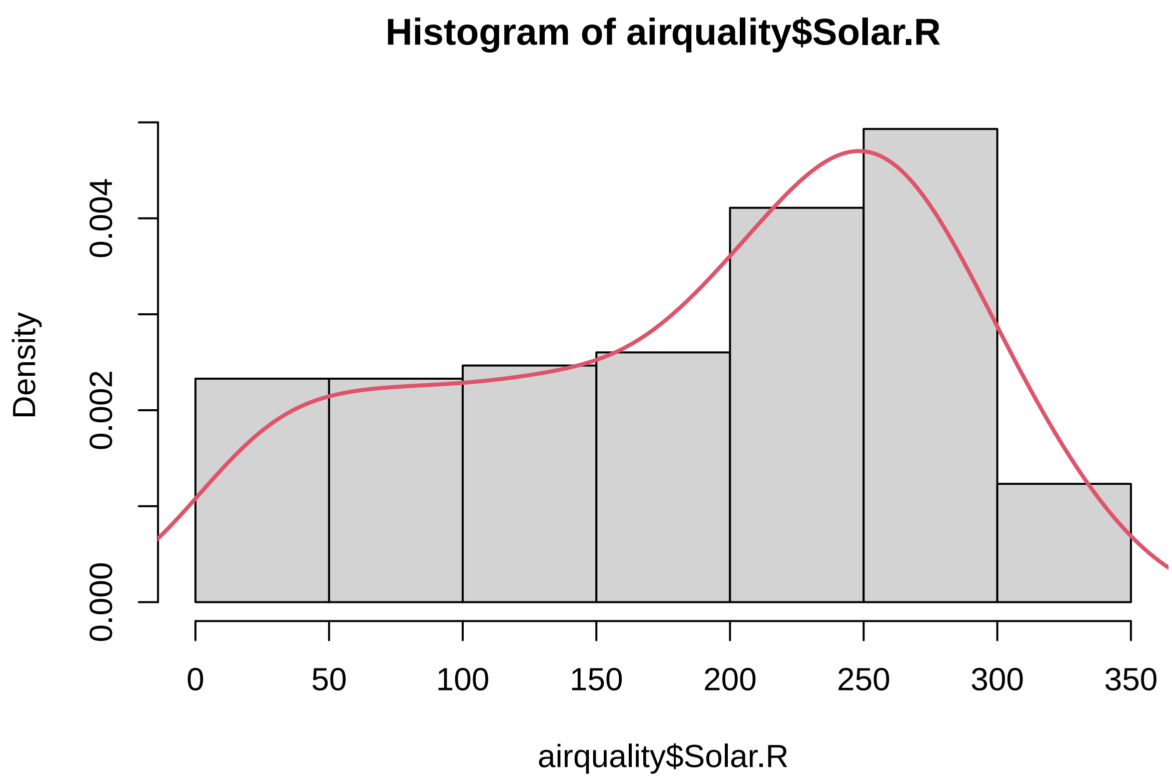

```
skewness(airquality$Solar.R,na.rm = TRUE) # sola çarpık
```

```
## [1] -0.4236342
```

```
kurtosis(airquality$Solar.R,na.rm = TRUE) # sivri
```

```
## [1] 2.023567
```

```
hist(airquality$Solar.R,freq = FALSE)
lines(density(airquality$Solar.R,na.rm = TRUE),col = 2, lwd = 2)
```



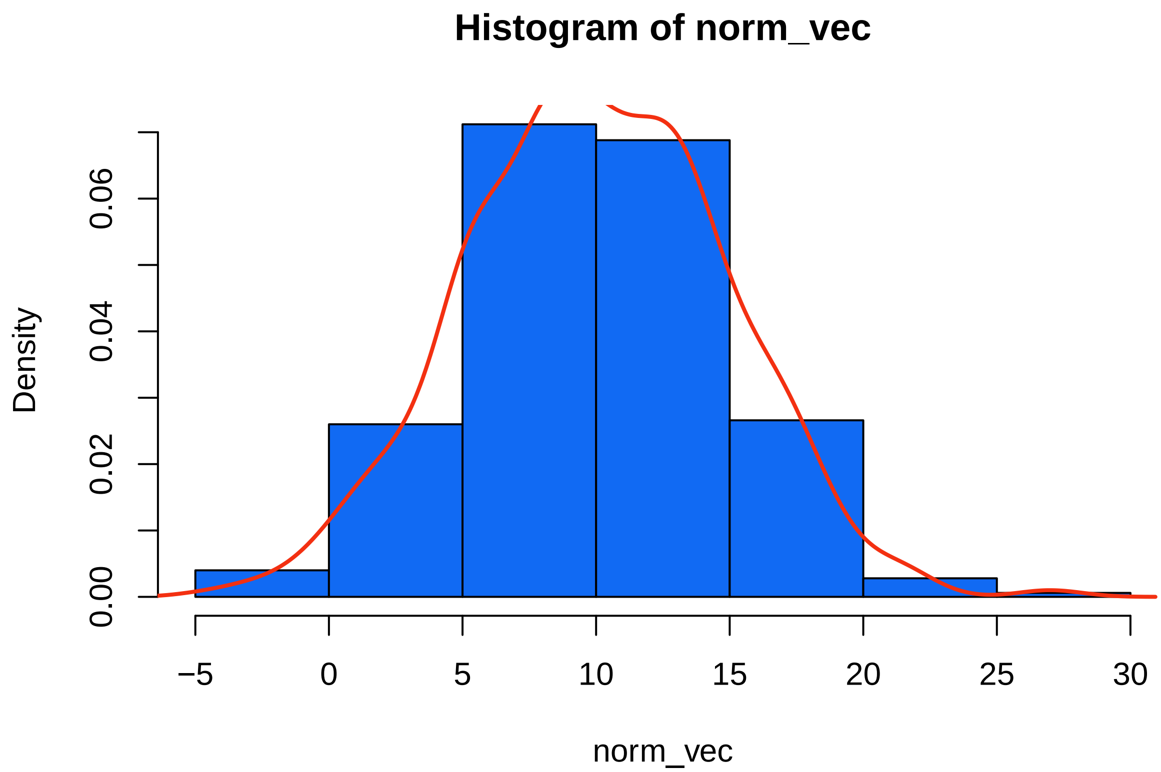

```
# normal dağılımdan veri üretelim
norm_vec <- rnorm(1000,10,5)
skewness(norm_vec) # sola çarpık
```

```
## [1] 0.03981022
```

```
kurtosis(norm_vec) # sivri
```

```
## [1] 3.00519
```

```
hist(norm_vec,freq = FALSE,col="#116AF3") # renk kodları da kullanılabilir.
lines(density(norm_vec),col = "#F33011", lwd = 2)
```



4 Aykırı ve Uç Değerler

Uç Değer; bireysel farklılıklardan dolayı ortaya çıkan, diğer bireylerden farklılık gösteren değerlerdir. **Aykırı Değer** ise ölçüm ya da kayıt hataları, farklı bir popülasyondan gelen bir gözlemin veya olağandışı aşırı bir gözlemin sonucu gibi nedenlerden dolayı ortaya çıkabilir. Uç ve aykırı değerler diğer değerlerden çok büyük olabileceği gibi çok küçük de olabilir.

Bir aykırı değer gözlemlersek, nedenini belirlemeye çalışmak önemlidir. Bir aykırı değer, bir ölçüm veya kayıt hatasından kaynaklanıyorsa veya başka bir nedenle açıkça veri kümesine ait değilse, aykırı değer basitçe kaldırılabilir. Bununla birlikte, bir aykırı değer için herhangi bir açıklama mevcut değilse, onu veri setinde tutup tutmama kararını vermek zordur.

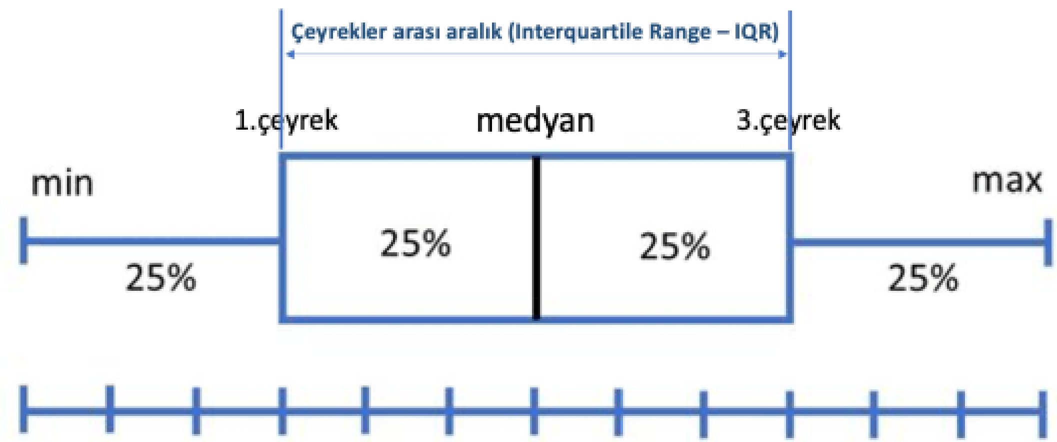
Aykırı değerler olabilecek gözlemleri tespit etmek için bir teşhis aracı olarak çeyreklikleri ve IQR'yi kullanabiliriz. Bu nedenle, bir veri setinin alt limitini ve üst limitini tanımlarız. Alt sınır, ilk çeyreğin 1.5×IQR altında kalan sayıdır; üst sınır, üçüncü çeyreğin 1,5×IQR üzerinde kalan sayıdır. Alt sınırın altında veya üst sınırın üzerinde olan gözlemler potansiyel aykırı değerlerdir.

$$AltSınır = Q_1 - 1.5 \times IQR$$

$$stSınır = Q_3 + 1.5 \times IQR$$

Ayrıca aykırı değerlerin tespi için görsel bir araç olarak boxplot grafikleri de kullanılabilir. Kutu ve bıyık diyagramı olarak da adlandırılan bir boxplot grafiği, beş sayılı özete dayanır ve bir veri kümesinin merkezinin ve varyasyonunun grafiksel bir görüntüsünü sağlamak için kullanılabilir.

Boxplot, beş ölçü kullanarak verilerin grafiksel bir sunumunu verir: en küçük değer (min), birinci çeyreklik (Q_1) , medyan, üçüncü çeyreklik (Q_3) en büyük değer. Kutunun farklı bölümleri arasındaki boşluk, verilerdeki dağılım (yayılma) ve çarpıklık derecesini gösterir.




```
summary(airquality$Ozone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   18.00   31.50   42.13   63.25  168.00      37
```

```
fivenum(airquality$Ozone)
```

```
## [1]   1.0  18.0  31.5  63.5 168.0
```

```
out <- boxplot.stats(airquality$Ozone)$out
out
```

```
## [1] 135 168
```

```
out_ind <- which(airquality$Ozone %in% c(out)) # outlier indeksleri
out_ind
```

```
## [1]  62 117
```

```
# outlier olarak görülen satırlar
airquality[out_ind, ]
```

```
##     Ozone Solar.R Wind Temp Month Day
## 62    135     269  4.1   84     7   1
## 117   168     238  3.4   81     8  25
```

```
# el ile hesaplama
q1 <- quantile(airquality$Ozone, 0.25,na.rm = TRUE)
q1
```

```
## 25%
##  18
```

```
q3 <- quantile(airquality$Ozone, 0.75,na.rm = TRUE)
q3
```

```
##   75%
## 63.25
```

```
altsinir <- q1 - 1.5 * IQR(airquality$Ozone,na.rm = TRUE)
altsinir
```

```
##     25%
## -49.875
```

```
ustsinir <- q3 + 1.5 * IQR(airquality$Ozone,na.rm = TRUE)
ustsinir
```

```
##     75%
## 131.125
```

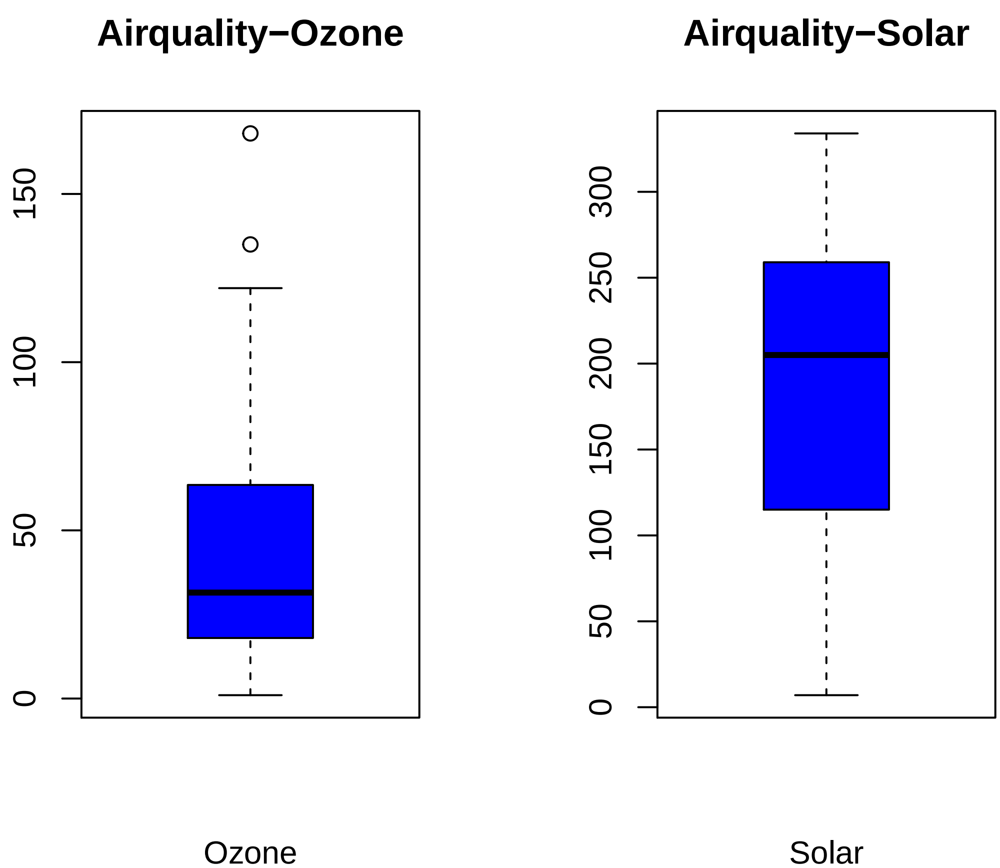
```
# Bu yönteme göre, -49.875'in altındaki ve 131.125'in üzerindeki tüm gözlemler,
# potansiyel aykırı değerler olarak kabul edilecektir.

outlier_sira <- which(airquality$Ozone < altsinir | airquality$Ozone > ustsinir)
outlier_sira
```

```
## [1]  62 117
```

```
airquality[outlier_sira,]
```

```
##     Ozone Solar.R Wind Temp Month Day
## 62    135     269  4.1   84     7   1
## 117   168     238  3.4   81     8  25
```

```
# boxplot

par(mfrow=c(1,2))
boxplot(airquality$Ozone,
        col = 'blue',
        xlab = 'Ozone',
        main = 'Airquality-Ozone')

boxplot(airquality$Solar.R,
        col = 'blue',
        xlab = 'Solar',
        main = 'Airquality-Solar')
```



5 İlişki Ölçüleri

Önceki bölümlerde, bir dağılımı tanımlayan ve özet istatistikleri hesaplayan tek bir değişkene odaklanmıştık. Tek bir değişkeni tanımlayan istatistiklere tek değişkenli istatistikler denir. İki değişken arasındaki ilişkiyi incelersek, iki değişkenli istatistiklere atıfta bulunuruz. Birkaç değişken arasındaki ilişkiler aynı anda incelenirse, çok değişkenli istatistiklere atıfta bulunuruz. İlişki ölçüleri, iki değişken arasındaki ilişkinin boyutunu özetlemek için araçlar sağlar.

İlişkiyi ölçmek için birçok araç türü olmasına rağmen, kovaryans ve Pearson korelasyon katsayıları "sayısal" veri türü için en bilinen ve yaygın araçlardır. Kovaryans ve korelasyon arasındaki temel fark, kovaryans, değerin işaretine (+'ve veya -'ve) bağlı olarak ilişkinin yönünü gösterir. Ancak korelasyon, değişkenler arasındaki "**doğrusal**" ilişkinin gücünü gösterir.

Kategorik veriler için ki-kare testi kullanılmkatadır. Spearman rho ve Kendall Tau korelasyon katsayıları da vardır ancak bunlar parametrik olmayan testlerdir ve yaygın olarak kullanılmazlar.

Değişkenler arasındaki ilişkiyi çizgi veya saçılım grafiği çizerek de incelenebilir. Ancak, bu grafiklere bakarak ilişkiden emin olmak her zaman mümkün olmayabilir. İstatistikte testler her zaman görsel araçlardan daha güçlüdür. Görsel araçlar fikir verir, testler ise fikirleri doğrular.

5.1 Kovaryans

Kovaryans, iki değişkenin ortak değişkenliğinin bir ölçüsüdür. Kovaryans $(-\infty,\infty)$ aralığında herhangi bir değer alabilir. Bir değişkenin büyük/küçük değerleri esas olarak diğer değişkenin daha büyük/küçük değerlerine karşılık geliyorsa kovaryans pozitiftir. Değişkenler zıt davranış gösterme eğilimindeyse kovaryans negatiftir. Kovaryans s_{xy} ile gösterilir ve aşağıdaki şekilde hesaplanır.

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
cov(iris$Sepal.Length,iris$Petal.Length) # pozitif ilişki var
```

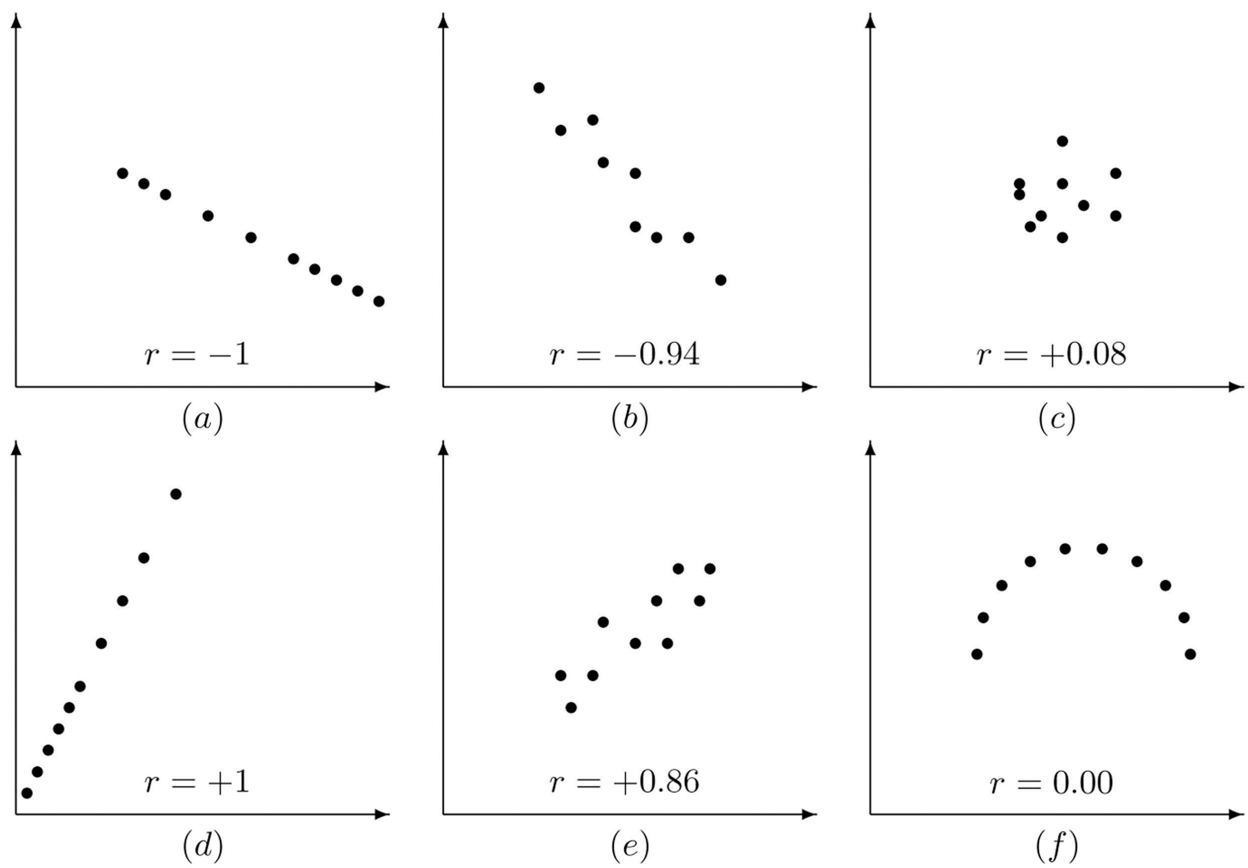
```
## [1] 1.274315
```

```
cov(iris$Sepal.Length,iris$Sepal.Length)
```

```
## [1] 0.6856935
```

5.2 Korelasyon

Korelasyon, nicel değişkenler arasındaki ilişkiyi incelemek için yaygın olarak kullanılan bir yöntemdir. **Karl Pearson'ın** Pearson moment korelasyon katsayısı olarak da bilinen doğrusal korelasyon katsayısı **r**'dir. Doğrusal korelasyon katsayısı, iki değişken arasındaki doğrusal ilişkinin gücünü ölçer.



- Korelasyon, kovaryansın standartlaştırılmış halidir.
- Standartlaştırmadan kaynaklanan bilgi kaybı vardır.
- Standartlaştırılmış olduğu için korelasyonun birimi yoktur, birimsizdir.

- Korelasyon -1 ve +1 arasında değer alır.
- Korelasyon , ±1'e yakınsa, iki değişken yüksek oranda ilişkilidir ve bir saçılım grafiği üzerinde çizilirse, veri noktaları bir çizgi etrafında kümelenir.
- Korelasyon , ±1'den uzaksa, veri noktaları daha geniş bir alana dağılır.
- Korelasyon 0'a yakınsa, veri noktaları esasen yatay bir çizgi etrafında dağılır ve bu, değişkenler arasında neredeyse hiçbir doğrusal ilişki olmadığını gösterir.
- r=1 ise değişkenler arasında pozitif yönlü tam bir doğrusal ilişki vardır.
- r=-1 ise değişkenler arasında negatif (ters) yönlü tam bir doğrusal ilişki vardır.
- r=0 ise değişkenler arasında doğrusal ilişki yoktur.
- Korelasyon nedensel ilişki değildir.
- Korelasyon değişkenler arasındaki sebep sonuç ilişkilerini açıklamaz.
- Korelasyon matematiksel ilişkidir.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

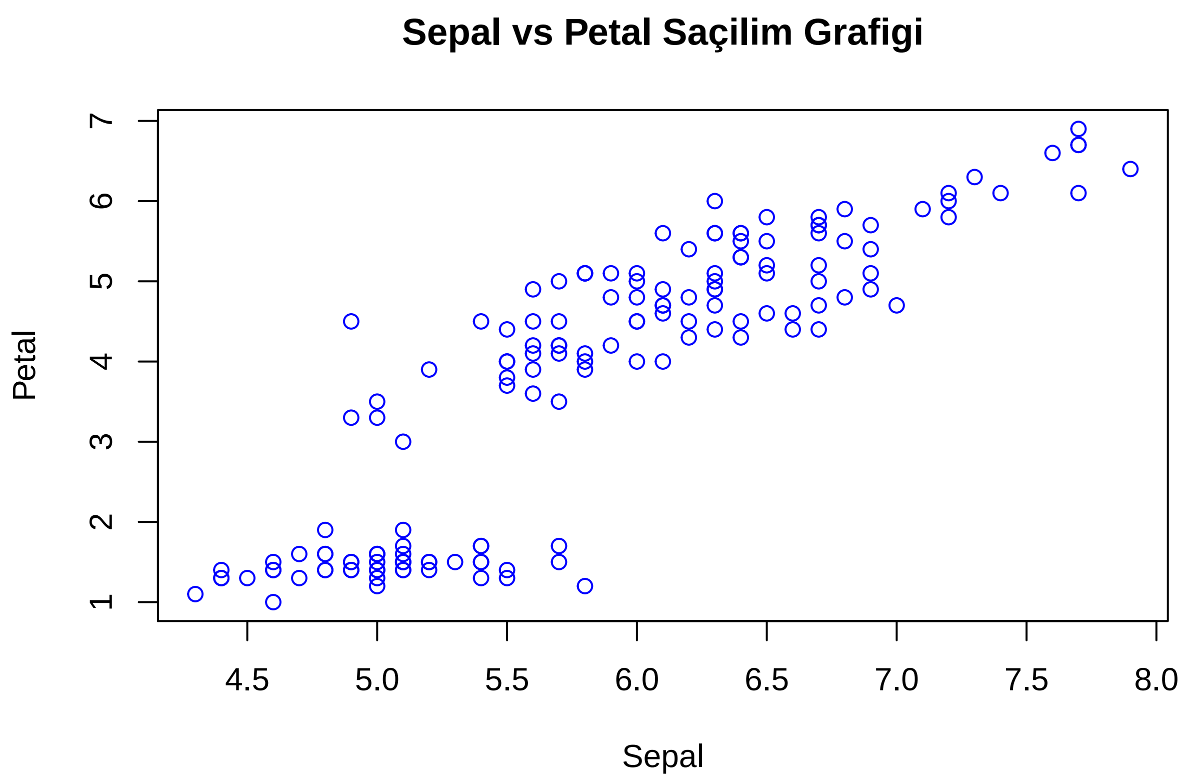
İki değişken arasındaki doğrusal ilişkinin miktarı için açık bir sınıflandırma kuralı yoktur. Bununla birlikte, aşağıdaki tablo, Pearson çarpım momenti korelasyon katsayısının sayısal değerlerinin nasıl ele alınacağı konusunda temel bir fikir verebilir.

Korelasyon Katsayısı (r)	İlişkinin Derecesi
$r > 0.90$	Çok kuvvetli
$0.70 < r \leq 0.90$	Kuvvetli
$0.50 < r \leq 0.70$	Orta
$0.30 < r \leq 0.50$	Düşük
$r < 0.30$	Zayıf

```
cor(iris$Sepal.Length,iris$Petal.Length) # kuvvetli ilişki vardır.
```

```
## [1] 0.8717538
```

```
plot(iris$Sepal.Length,iris$Petal.Length,
     col="blue",
     xlab = "Sepal",
     ylab = "Petal",
     main = "Sepal vs Petal Saçılım Grafiği")
```



5.3 Kontenjans Katsayısı

Kontenjans katsayısı C, kategorik veriler için $\chi 2$ tabanlı bir ilişki ölçüsüdür. Bağımsızlık için $\chi 2$ testine dayanır. $\chi 2$ istatistiği, kontenjans durum tablolarındaki (iki yönlü tablo, çapraz tablo tablosu veya çapraz tablolar olarak da bilinir) değişkenler arasında istatistiksel bir ilişki olup olmadığını değerlendirmeyi sağlar. Bu tür tablolarda değişkenlerin dağılımı matris formatında gösterilir. İki nominal (kategorik) değişken arasında anlamlı bir ilişki olup olmadığını belirlemek için kullanılır.

$$\chi 2 = \sum \frac{(G-B)^2}{B}$$

Burada G gözlemlenen frekansı ve B ise beklenen frekansı temsil eder . Ki-kare test istatistiği ile iki kategorik değişken arasında ilişki olup olmadığı araştırılır. Hipotez aşağıdaki gibi kurulur:

H_0 : Değişkenler arasında ilişki yoktur.

H_1 : Değişkenler arasında ilişki vardır.

Kontenjans katsayısı ise şu şekilde elde edilir:

$$C = \sqrt{\frac{\chi 2}{n + \chi 2}}$$

Burada n satır ve sütun toplamlarını ifade eder. C katsayısı 0 ile 1 arasında bir değer alır. C=0 olması iki değişken arasında ilişki olmadığına, C=1 olması ile tam ilişkili olduğu anlamına gelir.

```
# öğrencilerin sigara içme alışkanlığının egzersiz düzeyi ile ilişkili
# olup olmadığını inceleyelim.

library(MASS)
head(survey)
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height      M.I
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00   Metric
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80 Imperial
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA     <NA>
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00   Metric
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00   Metric
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72 Imperial
##      Age
## 1 18.250
## 2 17.583
## 3 16.917
## 4 20.333
## 5 23.667
## 6 21.000
```

```
nrow(survey)
```

```
## [1] 237
```

```
tbl <-  table(survey$Smoke, survey$Exer)
tbl
```

```
##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

```
# 1.yol
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

```
# 0.4828 p değeri .05 anlamlılık düzeyinden büyük olduğu için sigara
# içme alışkanlığının öğrencilerin egzersiz düzeyinden bağımsız olduğu
# sıfır hipotezini reddedemeyiz.

# 2.yol
summary(tbl)
```

```
## Number of cases in table: 236
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.489, df = 6, p-value = 0.4828
##  Chi-squared approximation may be incorrect
```

```
# 3. yol
library(vcd)
assocstats(tbl)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 5.8015  6  0.44579
## Pearson          5.4885  6  0.48284
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.151
## Cramer's V        : 0.108
```

```
# C Katsayısı
chi_squ_val <- chisq.test(tbl)$statistic
sqrt(chi_squ_val/(sum(tbl)+chi_squ_val))
```

```
## X-squared
##  0.150758
```

6 Doğrusal Regresyon

Basit doğrusal regresyon, iki nicel değişken arasındaki doğrusal ilişkiyi değerlendirmeye izin veren istatistiksel bir yaklaşımdır. Daha doğrusu, ilişkinin nicelleştirilmesini ve öneminin değerlendirilmesini sağlar. Çoklu doğrusal regresyon, bu yaklaşımın bir yanıt değişkeni (nicel) ile birkaç açıklayıcı değişken (nicel veya nitel) arasındaki doğrusal ilişkileri değerlendirmeyi mümkün kılması anlamında, basit doğrusal regresyonun bir genellemesidir.

Gerçek dünyada, çoklu doğrusal regresyon, basit doğrusal regresyondan daha sık kullanılır. Bu çoğunlukla böyledir çünkü, Çoklu doğrusal regresyon, diğer değişkenlerin etkisini kontrol ederken (yani etkiyi ortadan kaldırırken) iki değişken arasındaki ilişkiyi değerlendirmeye izin verir. Veri toplamanın da kolaylaşmasıyla, veriler analiz edilirken daha fazla değişken dahil edilebilir ve dikkate alınabilir.

Basit doğrusal regresyonda, değişkenlerden biri yanıt veya bağımlı değişken olarak kabul edilir ve y ekseninde temsil edilir. Diğer değişken ise açıklayıcı veya bağımsız değişken olarak da adlandırılır ve x ekseninde temsil edilir.

Basit doğrusal regresyon, iki değişken arasında doğrusal bir ilişkinin varlığını değerlendirmeye ve bu bağlantıyı nicelleştirmeye izin verir. Doğrusallığın, iki değişkenin doğrusal olarak bağımlı olup olmadığını test etmesi ve ölçmesi anlamında doğrusal regresyonda güçlü bir varsayım olduğuna dikkat etmek gerekmektedir.

Doğrusal regresyonu güçlü bir istatistiksel araç yapan şey, açıklayıcı/bağımsız değişken bir birim arttığında yanıtın/bağımlı değişkenin hangi nicelikle değiştiğini ölçmeye izin vermesidir. Bu kavram lineer regresyonda anahtardır ve aşağıdaki soruları yanıtlamaya yardımcı olur:

- Reklama harcanan miktar ile belirli bir dönemdeki satışlar arasında bir bağlantı var mı?
- Tütün vergilerindeki artış tüketimini azaltır mı?
- Bölgeye bağlı olarak bir konutun en olası fiyatı nedir?
- Bir kişinin bir uyarana tepki verme süresi cinsiyete bağlı mıdır?

Basit doğrusal regresyon analizinde, bağımlı değişken y ile bağımsız değişken x arasındaki ilişki doğrusal bir denklem şeklinde verilir.

$$y = \beta_0 + \beta_1 x$$

Burada, β_0 sayısına kesme noktası denir ve regresyon doğrusu ile y ekseninin (x=0) kesişme noktasını tanımlar. β_1 sayısına regresyon katsayısı denir. Regresyon doğrusu eğiminin bir ölçüsüdür. Böylece β_1, x değeri 1 birim arttığında y değerinin ne kadar değiştiğini gösterir. Model, x ve y arasında kesin bir ilişki verdiği için deterministik bir model olarak kabul edilir.

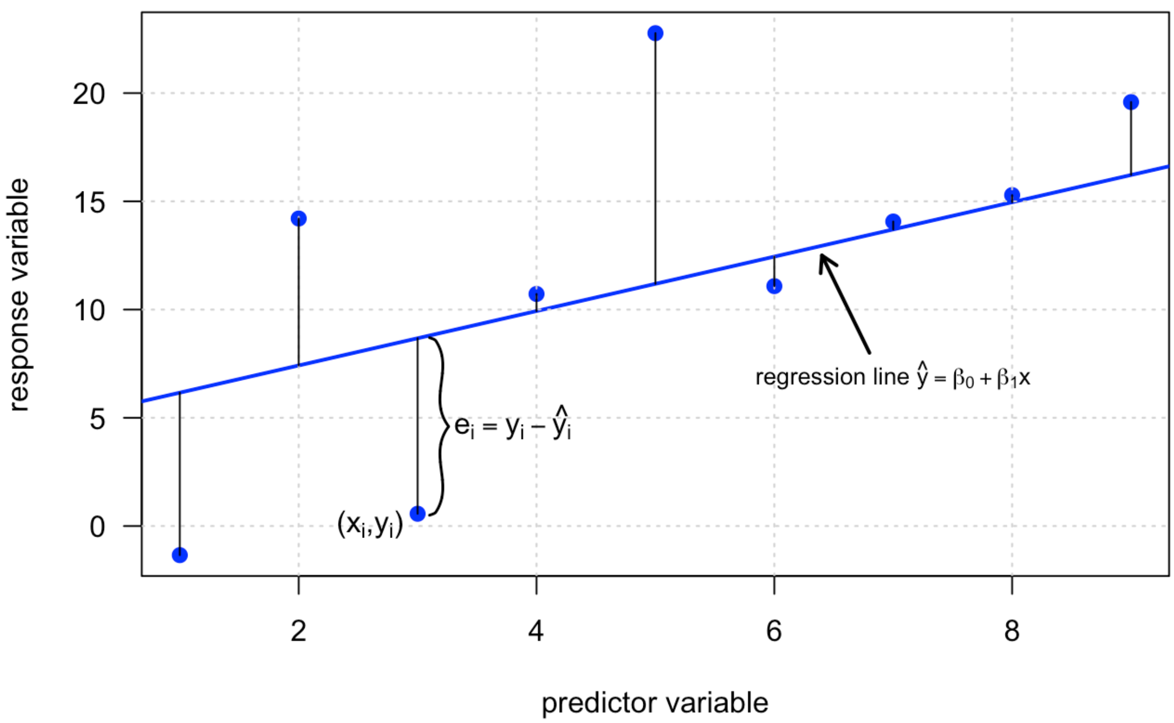
Ancak birçok durumda, iki değişken x ve y arasındaki ilişki kesin değildir. Bunun nedeni, bağımlı değişken y'nin, tahmin değişkeni x tarafından tam olarak yakalanmayan diğer bilinmeyen ve/veya rastgele süreçlerden etkilenmesidir. Böyle bir durumda veri noktaları düz bir çizgi üzerinde sıralanmaz. Bununla birlikte, veriler hala temeldeki doğrusal bir ilişkiyi takip edebilir. Bu bilinmeyenleri dikkate almak için lineer model denklemine ε ile gösterilen rastgele bir hata terimi eklenir, böylece yukarıdaki deterministik modelin aksine olasılıklı bir model elde edilir.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Burada hata terimi ε_i'nin bağımsız normal dağılımlı değerlerden oluştuğu varsayılır, $e_i \sim N(0, \sigma^2)$.

Doğrusal regresyon modeli hakkında aşağıdaki varsayımlar yapılır:

- Bağımlı değişken tesadüfi bir değişkendir ve normal dağılım göstermektedir.
- Tahmin hataları tesadüfidir ve normal dağılım gösterirler.
- Hatalar birbirinden bağımsızdır (otokorelasyon yoktur).
- Hata varyansı sabittir ve veriler arasında hiç değişmediği varsayılır (eşit varyanslılık-homoscedasticity).
- Eğer çoklu regresyon analizi yapılıyorsa, bağımsız değişkenlerin birbirleri ile bağlantısının olmaması gereklidir. Buna çoklu bağlantı (multicollinearity) olmaması varsayımı adı verilir.
- Bağımlı değişken ile bağımsız değişkenler arasında doğrusal bir ilişki olmalıdır.
- Gözlem sayısı parametre sayısından büyük olmalıdır.



Kalıntı olarak da adlandırılan her bir belirli değer çifti (x_i, y_i) için hata e_i, gözlemlenen y_i değeri ile $\hat{y_i}$ tahmin değerinin farkıyla hesaplanır. En iyi modele karşılık gelen regresyon eğrisini elde etmek için en En Küçük Kareler yöntemi (EKK) kullanılır. Kullanılan veriler ile en uygun doğruyu elde etmek için hata karelerinin hata toplamı en aza indirilir. Bir başka deyişle bu yöntem, ölçüm sonucu elde edilmiş veri noktalarına mümkün olduğu kadar yakın geçecek bir eğri bulmaya yarar.

$$min(EKK) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Eşitlik β_0 ve β_1 parametrelerine göre kısmi türevleri alınarak sıfıra eşitlendiğinde β_0 ve β_1 parametrelerinin EKK tahminleri elde edilir.

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

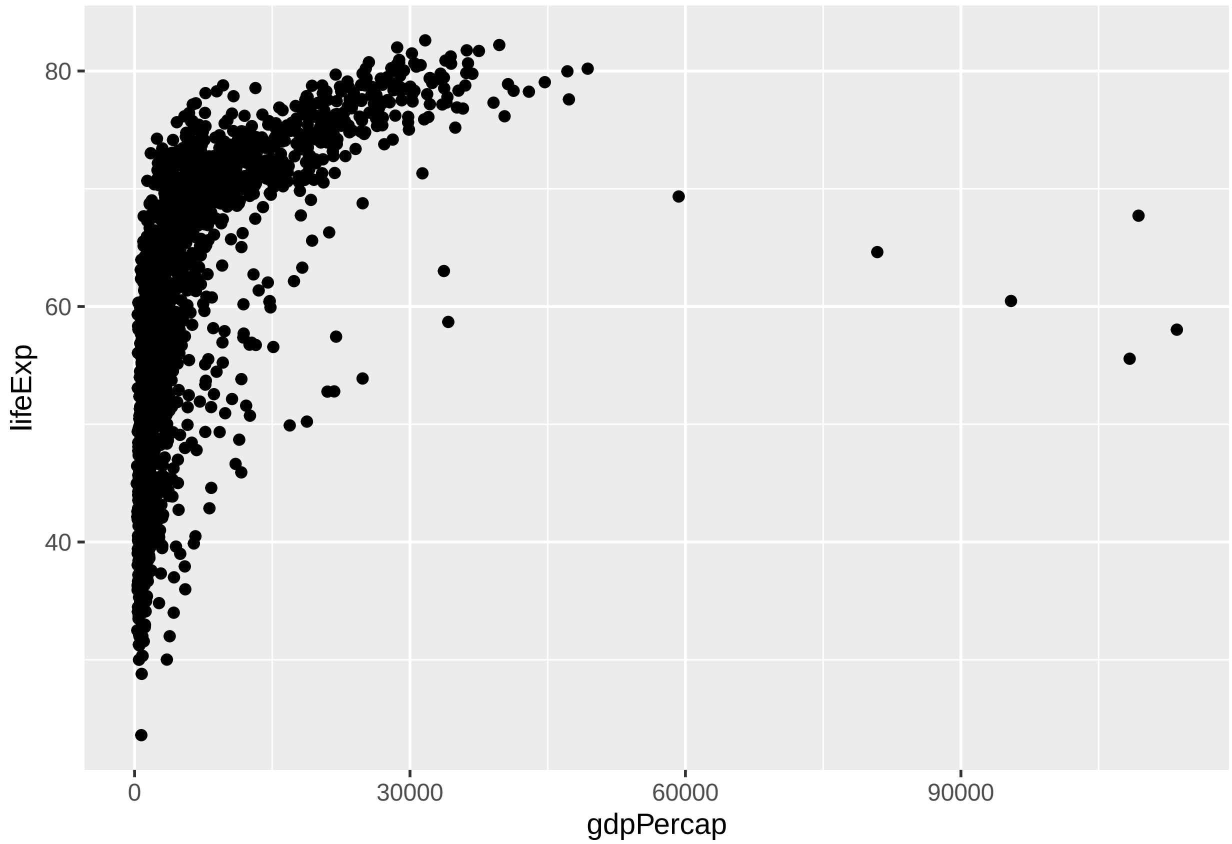
```
library(gapminder)
library(dplyr)
library(ggplot2)

# gapminder veri setine bakalım

glimpse(gapminder)
```

```
## Rows: 1,704
## Columns: 6
## $ country   <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", ~
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, ~
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, ~
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8~
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12~
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, ~
```

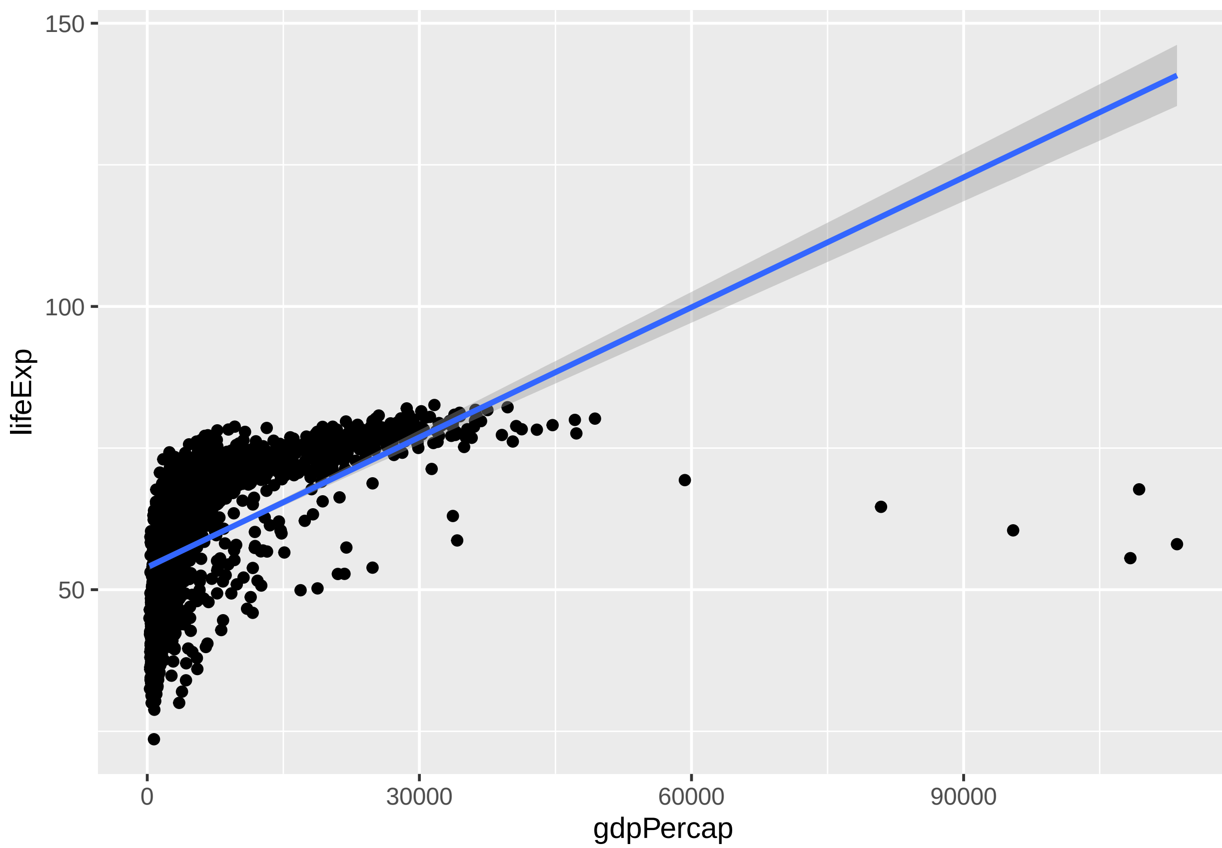
```
summary(gapminder)
```

```
##         country        continent        year         lifeExp
##  Afghanistan:  12   Africa  :624   Min.   :1952   Min.   :23.60
##  Albania    :  12   Americas:300   1st Qu.:1966   1st Qu.:48.20
##  Algeria    :  12   Asia    :396   Median :1980   Median :60.71
##  Angola     :  12   Europe  :360   Mean   :1980   Mean   :59.47
##  Argentina  :  12   Oceania : 24   3rd Qu.:1993   3rd Qu.:70.85
##  Australia  :  12                  Max.   :2007   Max.   :82.60
##  (Other)    :1632
##       pop               gdpPercap
##  Min.   :     60011   Min.   :   241.2
##  1st Qu.:   2793664   1st Qu.:  1202.1
##  Median :   7023596   Median :  3531.8
##  Mean   :  29601212   Mean   :  7215.3
##  3rd Qu.:  19585222   3rd Qu.:  9325.5
##  Max.   :1318683096   Max.   :113523.1
##
```

```
# kişi başına milli gelir ile yaşam beklentisi değişkenlerini görselleştirelim.

ggplot(gapminder, aes(gdpPercap, lifeExp)) +
  geom_point()
```




```
ggplot(gapminder, aes(gdpPercap, lifeExp)) +
  geom_point() +
  geom_smooth(method = "lm",se=TRUE)
```




```
# regresyon modeli kuralım

model1 <- lm(lifeExp ~ gdpPercap, data = gapminder)
model1
```

```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = gapminder)
##
## Coefficients:
## (Intercept)    gdpPercap
##  53.9555609    0.0007649
```

Yani burada söyleyebileceğimiz şey, GSYİH'daki her 1 artış için, yaşam beklentisinde 0.0007649 yıllık bir artış görmeyi bekleyebiliriz. Bu özellikle büyük değil - ama o zaman, GSYİH'de tek bir dolarlık artış da çok fazla değil! Modelimizi daha iyi anlayabilmek için model üzerinde summary() fonksiyonunu kullanabiliriz.

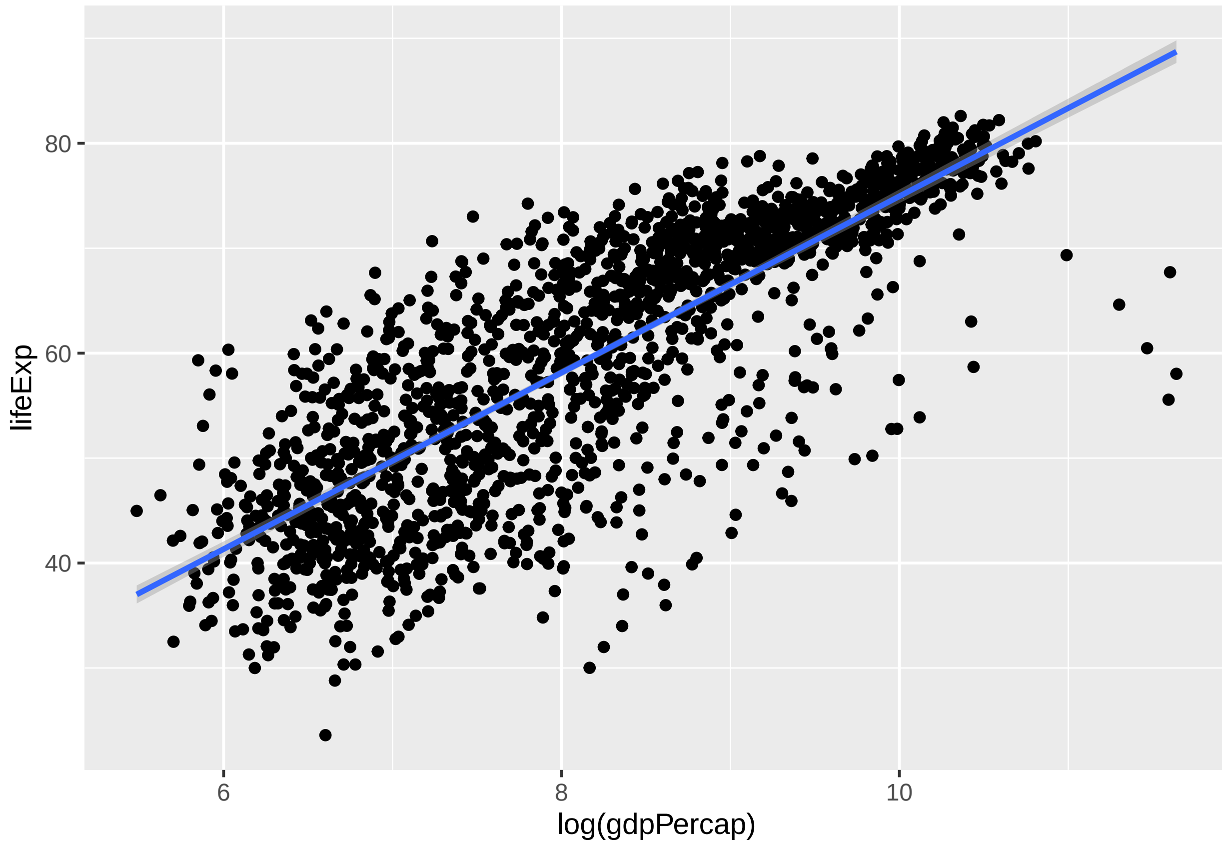
```
summary(model1)
```

```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.754  -7.758   2.176   8.225  18.426
##
## Coefficients:
##                Estimate  Std. Error t value            Pr(>|t|)
## (Intercept) 53.95556088  0.31499494  171.29 <0.0000000000000002 ***
## gdpPercap    0.00076488  0.00002579   29.66 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.49 on 1702 degrees of freedom
## Multiple R-squared:  0.3407, Adjusted R-squared:  0.3403
## F-statistic: 879.6 on 1 and 1702 DF,  p-value: < 0.00000000000000022
```

Burada modelimizin verilere ne kadar iyi uyduğu hakkında biraz daha bilgi alıyoruz. Genel modelimiz ve her değişken için p-değerlerini görebiliriz. R^2 değeri, veri kümenizdeki varyansın ne kadarının modeliniz tarafından açıklanabileceğini - temel olarak, modelinizin verilere ne kadar iyi uyduğunu gösterir. Bu değer 0 ile 1 arasında değişir ve büyük olması beklenir. Genel olarak, modelinizde kaç değişken kullandığınızı telafi eden düzeltilmiş R^2'yi kullanırız - aksi halde başka bir değişken eklemek her zaman R^2'yi artırır.

Ancak GSYİH'nın logaritması alındığında değişkenlerimiz arasında çok daha normal bir doğrusal ilişki görebiliriz.

```
ggplot(gapminder, aes(log(gdpPercap), lifeExp)) +
  geom_point() +
  geom_smooth(method = "lm",se=TRUE)
```




```
model2 <- lm(lifeExp ~ log(gdpPercap), data = gapminder)
summary(model2)
```

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap), data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.778  -4.204   1.212   4.658  19.285
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     -9.1009     1.2277  -7.413    0.000000000000193 ***
## log(gdpPercap)   8.4051     0.1488  56.500 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.62 on 1702 degrees of freedom
## Multiple R-squared:  0.6522, Adjusted R-squared:  0.652
## F-statistic:  3192 on 1 and 1702 DF,  p-value: < 0.00000000000000022
```

R^2 değerimizin arttığını görebiliyoruz. İlk modelde bu değer 0,34 iken ikinci modelde 0,652 olarak bulunmuştur Bu nedenle, verilerimizi log-dönüştürmek, modelimizin verilere daha iyi uymasına yardımcı oluyor gibi görünüyor. Veri setimizdeki continent (kıta) ve year (yıl) değişkenlerini de modele ekleyerek çoklu regresyon analizi sonuçlarına bakalım.

```
model3 <- lm(lifeExp ~ log(gdpPercap) + continent + year, data = gapminder)
summary(model3)
```

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap) + continent + year, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.0433  -3.2175   0.3482   3.6657  15.1321
##
## Coefficients:
##                      Estimate  Std. Error t value            Pr(>|t|)
## (Intercept)       -465.869597   16.674319  -27.94 <0.0000000000000002 ***
## log(gdpPercap)       5.023835    0.159473   31.50 <0.0000000000000002 ***
## continentAmericas    8.925906    0.462954   19.28 <0.0000000000000002 ***
## continentAsia        7.062939    0.395901   17.84 <0.0000000000000002 ***
## continentEurope     12.507788    0.509676   24.54 <0.0000000000000002 ***
## continentOceania    12.750719    1.274763   10.00 <0.0000000000000002 ***
## year                 0.241637    0.008586   28.14 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.813 on 1697 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.7975
## F-statistic:  1119 on 6 and 1697 DF,  p-value: < 0.00000000000000022
```

Bu sonuçlara göre R^2 değeri 0.79'a yükselmiştir. değişken sayısını artırmak model başarısını artırmış görünüyor. Ayrıca katsayıların hepsinin de anlamlı çıktığı göz ardı edilmemelidir.

Afrika kıtası haricinde, veri kümemizdeki kıtaların her biri için bir satır var. Bunun sebebi Afrika kıtası referans kıta olarak burada belirlenmesinden kaynaklanmaktadır. Yani kıtalara göre verileri yorumlarken Afirika kıtasına göre değerlendirme yapılacaktır. Örneğin Avrupa'da olmak ortalama olarak, Afrika'da olmaktan 12.27 yıl daha fazla yaşam beklentisine sahip olmak anlamına gelmektedir.

Model her bağımsız değişkenin birbirinden bağımsız olduğunu varsaymasıdır. Bununla birlikte, bunun GSYİH ve kıta için doğru olmadığından oldukça emin olabiliriz - genellikle Okyanusya'daki çoğu ülkenin, örneğin Afrika'daki çoğu ülkeden daha yüksek kişi başına GSYİH'ya sahip olduğunu varsayabiliriz. Bu nedenle, bu iki değişken arasında bir etkileşim terimi eklemeliyi düşünebiliriz. Bunu, model ifademizde bu terimler arasındaki + yerine * ile değiştirerek yapabiliriz.

```r
model4 <- lm(lifeExp ~ log(gdpPercap) * continent + year, data = gapminder)
summary(model4)
```

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap) * continent + year, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.2340  -2.9548   0.1681   3.3382  14.9649
##
## Coefficients:
##                                   Estimate  Std. Error t value
## (Intercept)                    -476.845698   17.080099 -27.918
## log(gdpPercap)                    4.848360    0.268597  18.051
## continentAmericas               -13.205551    4.646858  -2.842
## continentAsia                     4.405673    2.655124   1.659
## continentEurope                  30.952598    4.415078   7.011
## continentOceania                 76.626813   35.240823   2.174
## year                              0.247825    0.008681  28.550
## log(gdpPercap):continentAmericas  2.596704    0.555943   4.671
## log(gdpPercap):continentAsia      0.347108    0.346221   1.003
## log(gdpPercap):continentEurope   -1.934524    0.498751  -3.879
## log(gdpPercap):continentOceania  -6.487055    3.605512  -1.799
##                                              Pr(>|t|)
## (Intercept)                    < 0.0000000000000002 ***
## log(gdpPercap)                 < 0.0000000000000002 ***
## continentAmericas                          0.004539 **
## continentAsia                              0.097239 .
## continentEurope                    0.00000000000341 ***
## continentOceania                           0.029815 *
## year                           < 0.0000000000000002 ***
## log(gdpPercap):continentAmericas   0.00000323702307 ***
## log(gdpPercap):continentAsia               0.316216
## log(gdpPercap):continentEurope             0.000109 ***
## log(gdpPercap):continentOceania            0.072164 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.728 on 1693 degrees of freedom
## Multiple R-squared:  0.8045, Adjusted R-squared:  0.8034
## F-statistic: 696.8 on 10 and 1693 DF,  p-value: < 0.00000000000000022
```

Sonuçlar R^2 değerinin 0.80'e yükseldiğini gösteriyor. Şimdi bu modeli kullanarak yeni bir gözlem ile kestirim yapalım.

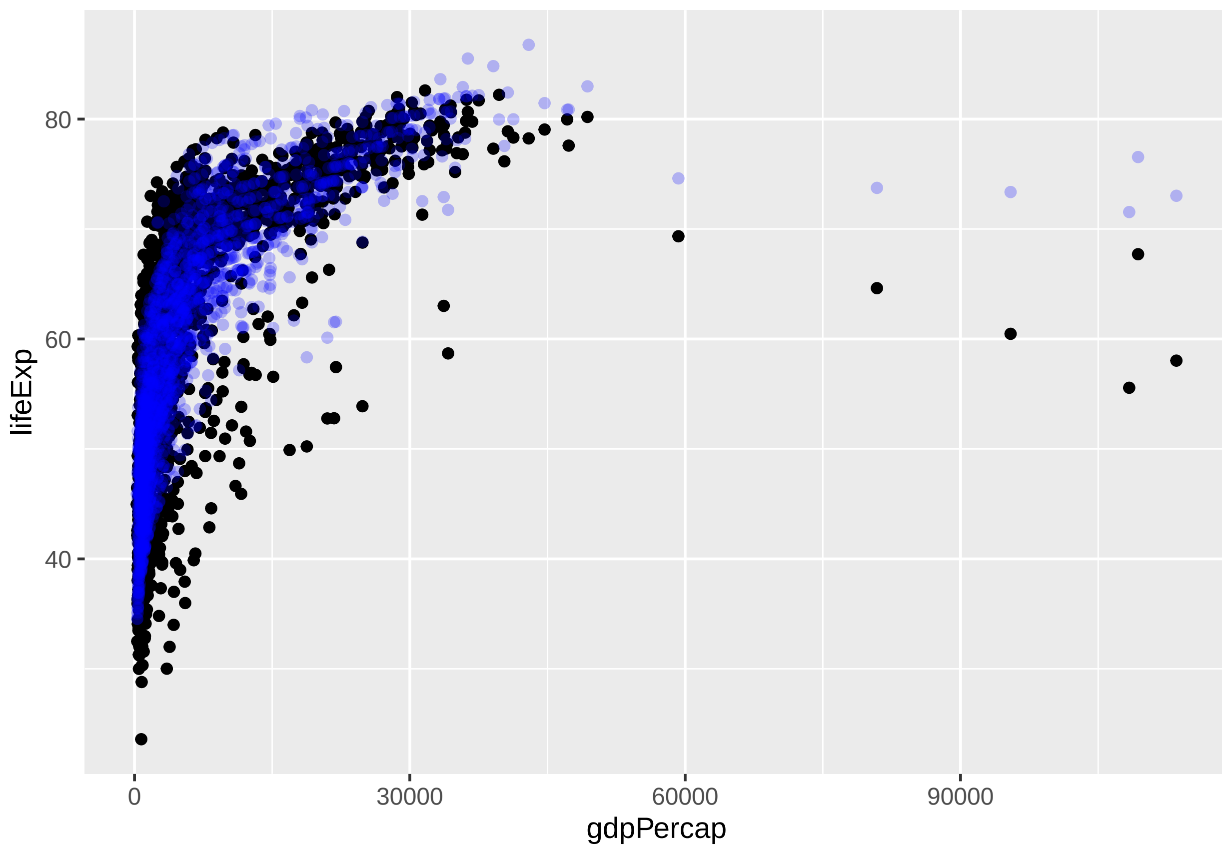
```
# yeni verilerler kestirim
gap_pred <- data.frame(lifeExp=c(70,75,80),
                       gdpPercap=c(9000,12000,15000),
                       continent=c("Asia","Americas","Europe"),
                       year=c(2012,2012,2012))

predict(model4, newdata = gap_pred,interval = "confidence", level = 0.99)
```

```
##        fit      lwr      upr
## 1 73.48759 72.35145 74.62374
## 2 78.50070 77.08816 79.91325
## 3 80.74877 79.69798 81.79956
```

```
# model tahminlerine ulaşmak
fitted <- data.frame(predict=model4$fitted.values)
head(fitted,10)
```

```
##     predict
## 1  45.90794
## 2  47.41599
## 3  48.85531
## 4  49.99046
## 5  50.59449
## 6  52.14781
## 7  54.52173
## 8  55.04663
## 9  54.87211
## 10 55.99799
```

```
# original ve kestirim sonuçlarını görselleştirelim
ggplot(gapminder, aes(gdpPercap)) +
  geom_point(aes(y = lifeExp)) +
  geom_point(aes(y = fitted$predict), color = "blue", alpha = 0.25)
```



Model sonuçları içerisinde bakılması gereken en önemli kısımlardan birisi de artıklardır. Artıklar kullanılarak modellerin başarılarını ölçen metrikler bulunmaktadır. Artıkların ortalaması ya da **RMSE**(Root mean square error) bunlardan bazılarıdır. Ayrıca **AIC**, **BIC**gibi bilgi kriterleri de model başarımlarını ölçmede yardımcı metriklerdir.

```
# Modellerin metriklerini bir araya getirelim

ME <- function(model){
  mean(residuals(model))

}

RMSE <- function(model){
  sqrt(sum(residuals(model)^2) / df.residual(model))
}

adj.R2 <- function(model){
   summary(model)$adj.r.squared
}

metrics <-
  data.frame(
```

```
    model = c("model1", "model2", "model3", "model4"),
    ME = c(ME(model1), ME(model2), ME(model3), ME(model4)),
    AIC = c(AIC(model1), AIC(model2), AIC(model3), AIC(model4)),
    adj.R2 = c(
      adj.R2(model1),
      adj.R2(model2),
      adj.R2(model3),
      adj.R2(model4)
    ),
    RMSE = c(RMSE(model1), RMSE(model2), RMSE(model3), RMSE(model4))
  )

metrics
```

```
##    model                          ME      AIC    adj.R2      RMSE
## 1 model1 -0.000000000000000073877129 12850.41 0.3403256 10.491319
## 2 model2  0.000000000000000005144947 11760.42 0.6520423  7.619535
## 3 model3  0.000000000000000055878558 10843.29 0.7974611  5.813257
## 4 model4  0.000000000000000009246565 10796.71 0.8033820  5.727655
```

Model sonuçlarının daha güzel ve temiz (tidy) bir formatta görünmesi için **broom** paketi kullanılabilir.

```
library(broom)

# gözlem düzeyinde sonuçlar
augment(model4)
```

```
## # A tibble: 1,704 x 9
##    lifeExp `log(gdpPercap)` continent  year .fitted    .hat .sigma .cooksd
##      <dbl>            <dbl> <fct>     <int>   <dbl>   <dbl>  <dbl>   <dbl>
##  1    28.8             6.66 Asia       1952    45.9 0.00654   5.71 0.00537
##  2    30.3             6.71 Asia       1957    47.4 0.00591   5.71 0.00483
##  3    32.0             6.75 Asia       1962    48.9 0.00544   5.71 0.00433
##  4    34.0             6.73 Asia       1967    50.0 0.00530   5.72 0.00378
##  5    36.1             6.61 Asia       1972    50.6 0.00570   5.72 0.00337
##  6    38.4             6.67 Asia       1977    52.1 0.00547   5.72 0.00288
##  7    39.9             6.89 Asia       1982    54.5 0.00474   5.72 0.00285
##  8    40.8             6.75 Asia       1987    55.0 0.00552   5.72 0.00313
##  9    41.7             6.48 Asia       1992    54.9 0.00715   5.72 0.00350
## 10    41.8             6.45 Asia       1997    56.0 0.00777   5.72 0.00443
## # ... with 1,694 more rows, and 1 more variable: .std.resid <dbl>
```

```
#model düzeyinde sonuçlar
glance(model4)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik    AIC    BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
## 1     0.805         0.803  5.73      697.       0    10 -5386. 10797. 10862.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```