**Scatter Plot & Box Plot**

**Scatter Plot**

Before we take up the discussion of correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a scatter plot.

A scatter plot shows the direction of a relationship between the variables. A clear direction happens when there is either:

* High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
* High values of one variable occurring with low values of the other variable.

Korelasyon tartışmasına başlamadan önce, iki değişken x ve y arasındaki ilişkiyi göstermenin bir yolunu incelememiz gerekiyor. En yaygın ve en kolay yol bir dağılım grafiğidir.

Bir dağılım grafiği, değişkenler arasındaki ilişkinin yönünü gösterir. Aşağıdakilerden biri olduğunda net bir yön olur:

• Bir değişkenin yüksek değerlerinin diğer değişkenin yüksek değerleriyle ortaya çıkması veya bir değişkenin düşük değerlerinin diğer değişkenin düşük değerleriyle ortaya çıkması.

• Bir değişkenin yüksek değerleri ile diğer değişkenin düşük değerlerinin ortaya çıkması.

**📝Scatter Plot**  
A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis.

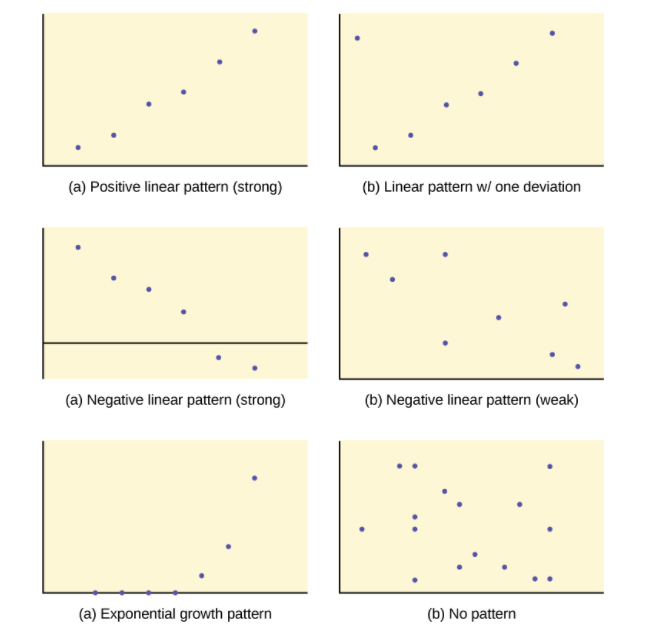
You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

When you look at a scatterplot, you want to notice the overall pattern and any deviations from the pattern. The following scatterplot examples illustrate these concepts.

Dağılım grafiğine bakarak ve noktaların bir çizgiye, bir güç işlevine, üstel bir işleve veya başka bir işlev türüne ne kadar yakın olduğunu görerek ilişkinin gücünü belirleyebilirsiniz. Doğrusal bir ilişki için bir istisna vardır. Tüm noktaların "mükemmel bir uyum" sağlayan yatay bir çizgi üzerine düştüğü bir dağılım grafiğini düşünün. Yatay çizgi aslında hiçbir ilişki göstermez.

Bir dağılım grafiğine baktığınızda, genel deseni ve desenden sapmaları fark etmek istersiniz. Aşağıdaki dağılım grafiği örnekleri bu kavramları göstermektedir.





**Scatter Plot & Box Plot**

**Box Plot**

**Box plots** (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: *the minimum value, the first quartile, the median, the third quartile,* and *the maximum value.* We use these values to compare how close other data values are to them.

(**Box plot,** veri konsantrasyonunun iyi bir grafik görüntüsünü verir. Ayrıca uç değerlerin çoğu veriden ne kadar uzak olduğunu da gösterirler. Beş değerden bir kutu grafiği oluşturulur: minimum değer, ilk çeyrek, medyan, üçüncü çeyrek ve maksimum değer. Bu değerleri, diğer veri değerlerinin onlara ne kadar yakın olduğunu karşılaştırmak için kullanırız.)

One of the more effective graphical summaries of a data set, the box plot generally shows mean, median, 25th and 75th percentiles, and outliers

Bir veri kümesinin en etkili grafik özetlerinden biri olan kutu grafiği genellikle ortalama, medyan, 25. ve 75. yüzdelik dilimleri ve aykırı değerleri gösterir

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

Bir box plot oluşturmak için yatay veya dikey bir sayı doğrusu ve dikdörtgen bir kutu kullanın. En küçük ve en büyük veri değerleri, eksenin uç noktalarını etiketler. İlk çeyrek kutunun bir ucunu ve üçüncü çeyrek kutunun diğer ucunu gösterir. Verilerin yaklaşık yüzde 50'si kutunun içine düşüyor. "Bıyıklar", kutunun uçlarından en küçük ve en büyük veri değerlerine kadar uzanır. Medyan veya ikinci çeyrek, birinci ve üçüncü çeyrekler arasında olabilir veya biri, diğeri veya her ikisi olabilir. Kutu grafiği, verilerin iyi ve hızlı bir resmini verir.

You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values

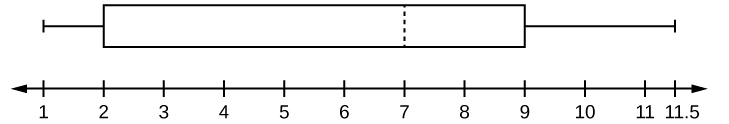
Aykırı değerleri işaretleyen noktalara sahip kutu ve bıyık grafikleriyle karşılaşabilirsiniz. Bu durumlarda, bıyıklar minimum ve maksimum değerlere uzanmıyor

Consider the following dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 2 | 2 | 4 | 6.8 | 7 | 8 | 8.3 | 9 | 10 | 10 | 11.5 |

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

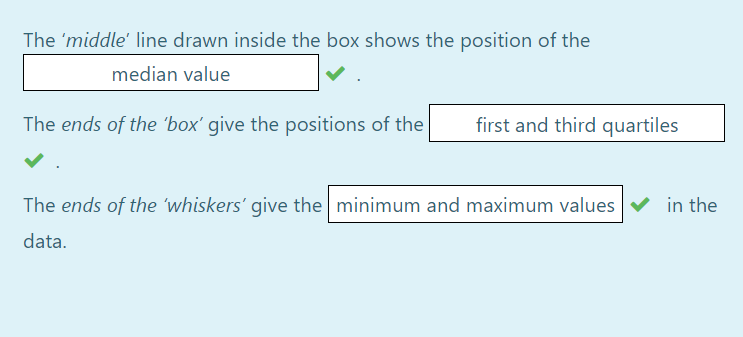
İlk çeyrek iki, medyan yedi ve üçüncü çeyrek dokuz. En küçük değer bir, en büyük değer 11.5'tir. Aşağıdaki görüntü, oluşturulmuş kutu grafiğini göstermektedir.



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

İki bıyık, birinci çeyrekten en küçük değere ve üçüncü çeyrekten en büyük değere kadar uzanır. Medyan kesikli bir çizgi ile gösterilir.





## Correlation and Covariance

### Correlation and Covariance

**Correlation** is one of the most common statistical concepts. It is a statistical technique that determines how one variable changes with another variable. It gives us the degree of the relationship between the two variables.

Korelasyon en yaygın istatistiksel kavramlardan biridir. Bir değişkenin başka bir değişkenle nasıl değiştiğini belirleyen istatistiksel bir tekniktir. Bize iki değişken arasındaki ilişkinin derecesini verir.

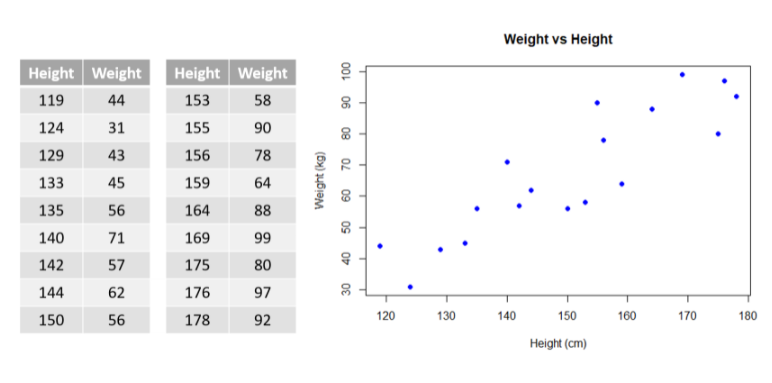
**Correlation**  
A relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone.

Olgular veya şeyler arasında veya yalnızca şansa dayalı olarak beklenmeyen bir şekilde değişme, ilişkilendirilme veya birlikte meydana gelme eğiliminde olan matematiksel veya istatistiksel değişkenler arasında var olan bir ilişki.

For example, consider the correlation between *height* and *weight* variables. If taller people also have more weight it can be said there is a correlation between height and weight.

Örneğin, boy ve kilo değişkenleri arasındaki ilişkiyi düşünün. Daha uzun boylu insanlar da daha fazla kiloya sahipse, boy ile kilo arasında bir ilişki olduğu söylenebilir.





**Covariance** provides similar information with correlation. However, correlation goes beyond covariance and gives information also about the strength of the relationship between two variables. Covariance does not provide information about the strength of the relationship.

Kovaryans, korelasyon ile benzer bilgiler sağlar. Ancak korelasyon kovaryansın ötesine geçer ve iki değişken arasındaki ilişkinin gücü hakkında da bilgi verir. Kovaryans ilişkinin gücü hakkında bilgi sağlamaz.

**Covariance**  
The expected value of the product of the deviations of two random variables from their respective means.

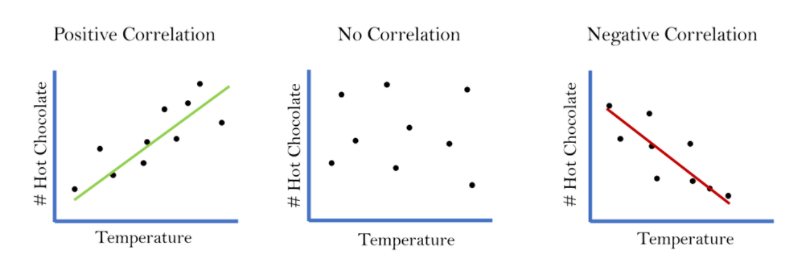
**Covaryans: İki rastgele değişkenin kendi ortalamalarından sapmalarının çarpımının beklenen değeri**.



Both can be positive or negative. Correlation or covariance is positive if one increases the other also increases and negative if one increases the other decreases. The figure below shows the positive and negative correlations between the two variables.

Her ikisi de olumlu veya olumsuz olabilir. Korelasyon veya kovaryansta biri arttıkça ​​diğeri de artarsa pozitiftir. biri artıp ​​diğeri azalırsa negatiftir.. Aşağıdaki şekil, iki değişken arasındaki pozitif ve negatif korelasyonları göstermektedir.

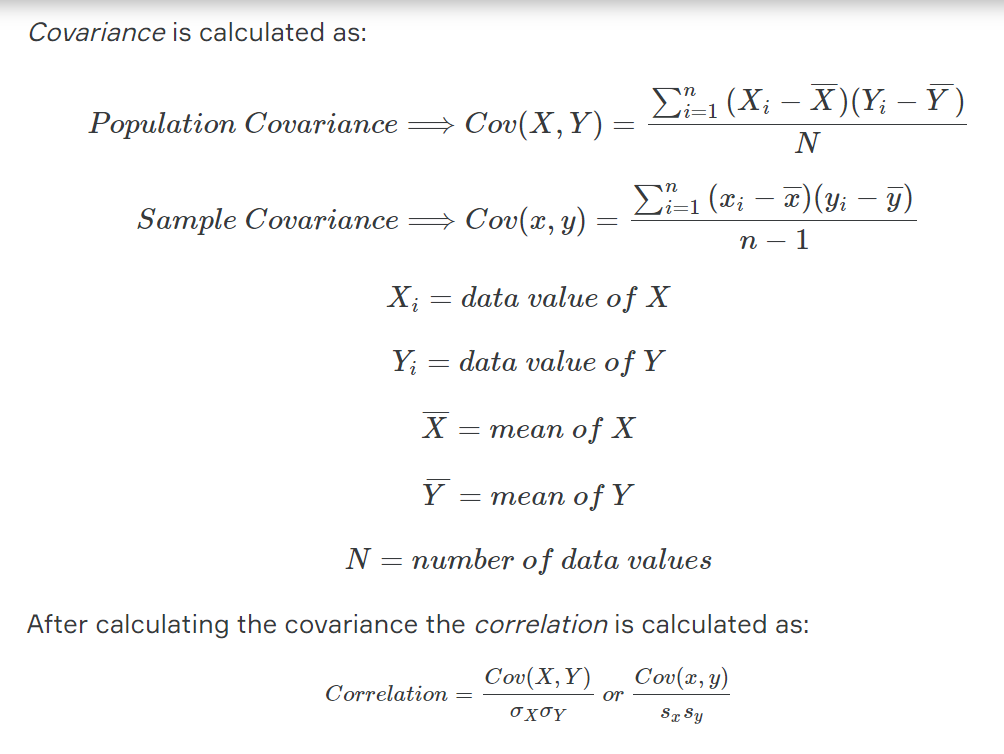




## Correlation and Covariance

### Formulation

*Covariance* is calculated as:



We can say correlation is calculated by the division of covariance by the standard deviation of variables. Because the correlation is a number between -1 and 1 it is often referred to as the **correlation coefficient**.

As we said earlier; covariance and correlation provide similar information while correlation goes beyond covariance and also gives us an idea about the **strength of the relationship**.

**Korelasyonun kovaryansın değişkenlerin standart sapmasına bölünmesiyle hesaplandığını söyleyebiliriz**. Korelasyon -1 ile 1 arasında bir sayı olduğu için genellikle korelasyon katsayısı olarak adlandırılır.

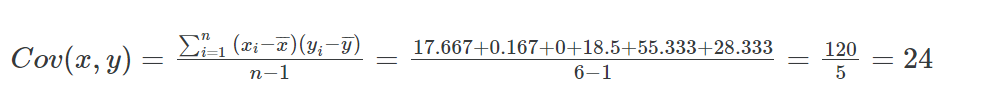
Daha önce de söylediğimiz gibi; **kovaryans ve korelasyon benzer bilgiler sağlarken, korelasyon kovaryansın ötesine geçer ve ayrıca bize ilişkinin gücü hakkında bir fikir verir.**

The following data shows the number of people with their corresponding temperature in a supermarket.

**Temperature (°F)** = 93, 84, 82, 78, 98, 70 (Variable x)  
**Number of People** = 13, 10, 11, 8, 15, 9 (Variable y)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| xixi | xi−x¯¯¯xi−x¯ | yiyi | yi−y¯¯¯yi−y¯ | (xi−x¯¯¯)(yi−y¯¯¯)(xi−x¯)(yi−y¯) |
| 93 | 8.8333 | 13 | 2 | 17.667 |
| 84 | -0.1667 | 10 | -1 | 0.167 |
| 82 | -2.1667 | 11 | 0 | 0 |
| 78 | -6.1667 | 8 | -3 | 18.5 |
| 98 | 13.8333 | 15 | 4 | 55.333 |
| 70 | -14.1667 | 9 | -2 | 28.333 |

Putting temperature and number of people values in the formula we calculate the covariance:



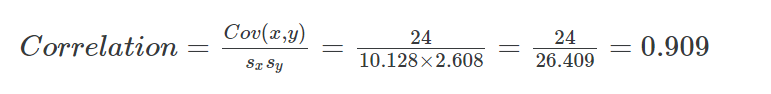
The covariance between the temperature and the number of people is 24. Since the covariance is positive, the temperature and number of people have a positive relationship. As temperature rises, so does the number of people.

Sıcaklık ile insan sayısı arasındaki kovaryans 24'tür. Kovaryans pozitif olduğu için sıcaklık ile insan sayısı arasında pozitif bir ilişki vardır. Sıcaklık arttıkça insan sayısı da artıyor.

However there is no information about how strong the relationship is, and the answer to this problem is the correlation:

Ancak ilişkinin ne kadar güçlü olduğuna dair bir bilgi yoktur ve bu sorunun cevabı korelasyondur.





The 0.909 which is called the correlation coefficient shows that there is a very strong correlation between temperature and number of people.

Korelasyon katsayısı olarak adlandırılan 0.909, sıcaklık ile insan sayısı arasında çok güçlü bir korelasyon olduğunu gösterir.



**💡Tips:**

* Correlation is calculated by the division of covariance by the standard deviation of variables.
* Correlation gives information about the strength of the relationship between two variables.
* • **Korelasyon, kovaryansın değişkenlerin standart sapmasına bölünmesiyle hesaplanır.**
* **• Korelasyon, iki değişken arasındaki ilişkinin gücü hakkında bilgi verir.**

## Correlation and Covariance

### Pearson Correlation Coefficient

There are different methods to calculate the correlation coefficient between two variables. The most famous one is the Pearson Correlation Coefficient. It is a number between -1 and 1 that indicates the strength of the relationship.

İki değişken arasındaki korelasyon katsayısını hesaplamak için farklı yöntemler vardır. En ünlüsü Pearson Korelasyon Katsayısı. İlişkinin gücünü gösteren -1 ile 1 arasında bir sayıdır.

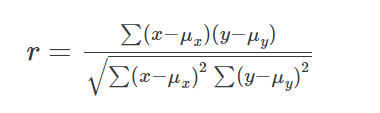


*-1 : Indicates complete negative correlation*

*+1 : Indicates complete correlation*

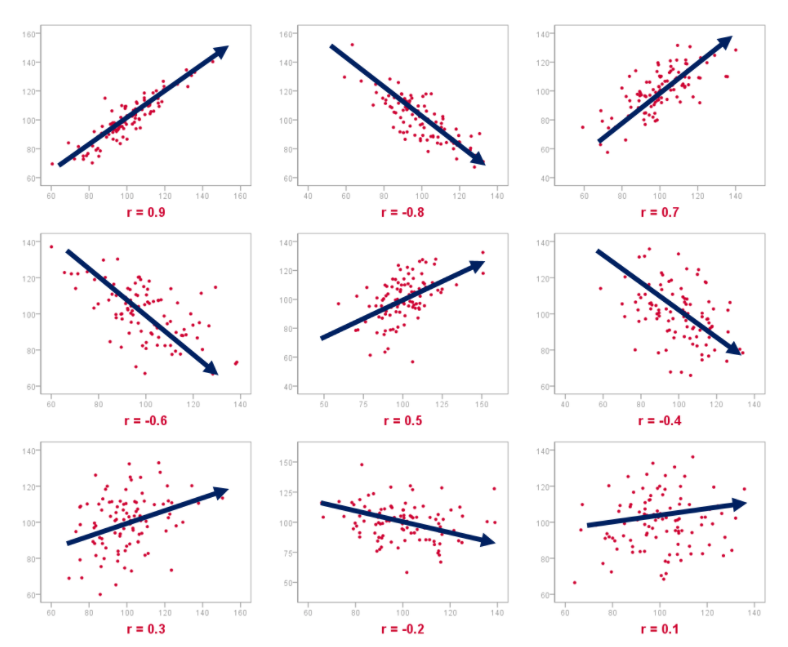
*0 : Indicates no correlation*

The Pearson correlation coefficient is calculated as follows:



It captures not only the strength but also the direction of the linear association between two continuous variables. And it tries to draw a line of best fit through the data points of two variables. The picture below, in which the correlation coefficient is denoted by “r” illustrates this point very well. Blue arrows are the lines of best fit through the data points of two variables.

İki sürekli değişken arasındaki doğrusal ilişkinin sadece gücünü değil aynı zamanda yönünü de yakalar. Ve iki değişkenin veri noktalarından en uygun çizgiyi çizmeye çalışır. Korelasyon katsayısının “r” ile gösterildiği aşağıdaki resim bu noktayı çok iyi göstermektedir. Mavi oklar, iki değişkenin veri noktalarından geçen en uygun çizgilerdir.



**Tips:**

* Pearson Correlation Coefficient tries to draw a line of best fit through the data points of two variables.
* Linear transformations have no effect on Pearson's correlation coefficient.
* Pearson's correlation is symmetric in the sense that the correlation of X with Y is the same as the correlation of Y with X.
* • **Pearson Korelasyon Katsayısı, iki değişkenin veri noktaları aracılığıyla en uygun çizgiyi çizmeye çalışır.**
* **• Doğrusal dönüşümlerin Pearson korelasyon katsayısı üzerinde etkisi yoktur.**
* **• Pearson korelasyonu, X'in Y ile korelasyonunun, Y'nin X ile korelasyonu ile aynı olması anlamında simetriktir.**

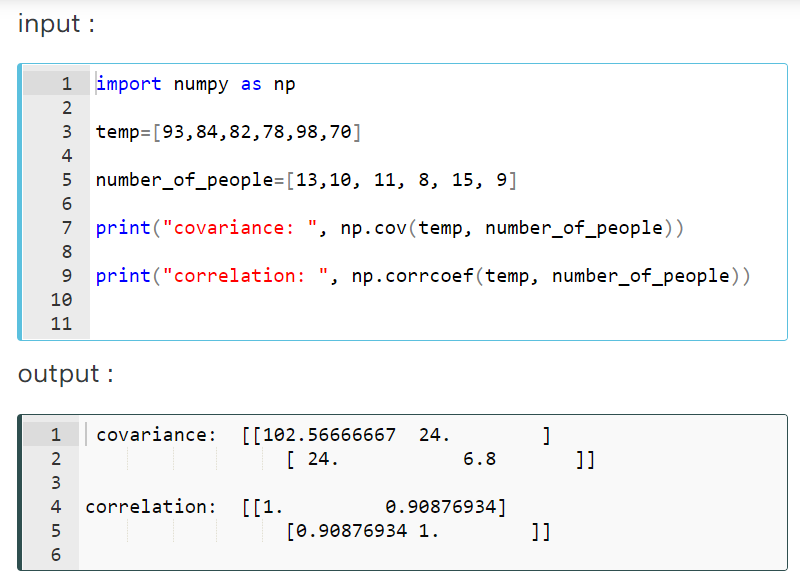
## Correlation and Covariance

### Calculate Correlation and Covariance with Python

We can easily calculate covariance and correlation with numpy. The results of **np.cov()** and **np.corrcoef()** commands are displayed in matrix form. You remember the covariance between temperature and number of people was 24, and the correlation was 0,9. Try to find these values in the result matrices and ignore the other values. You can compare the values we get by numpy with the values we calculated manually.

Numpy ile kovaryans ve korelasyonu kolayca hesaplayabiliriz. np.cov() ve np.corrcoef() komutlarının sonuçları matris biçiminde görüntülenir. Sıcaklık ve insan sayısı arasındaki kovaryansın 24 olduğunu ve korelasyonun 0,9 olduğunu hatırlıyorsunuz. Sonuç matrislerinde bu değerleri bulmaya çalışın ve diğer değerleri yok sayın. Numpy ile aldığımız değerleri manuel olarak hesapladığımız değerlerle karşılaştırabilirsiniz.





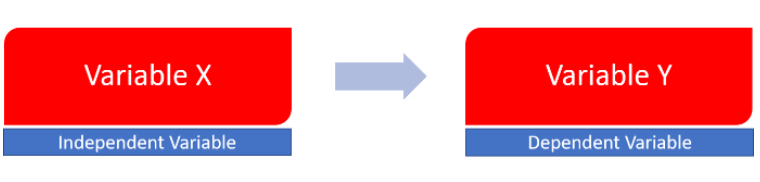
## Linear Regression

### What is Linear Regression?

The relationship between two continuous variables can be explained with a straight line in the simplest way. In regression analysis, the relationship between a *dependent variable* and an *independent variable* can be explained with a simple linear regression model.

**İki sürekli değişken arasındaki ilişki en basit şekilde düz bir çizgi ile açıklanabilir. Regresyon analizinde bağımlı değişken ile bağımsız değişken arasındaki ilişki basit bir doğrusal (lineer) regresyon modeli ile açıklanabilir**.





**Regression**  
A functional relationship between two or more correlated variables that is often empirically determined from data and is used especially to predict values of one variable when given values of the others.

Genellikle verilerden ampirik olarak belirlenen ve özellikle diğerlerinin değerleri verildiğinde bir değişkenin değerlerini tahmin etmek için kullanılan iki veya daha fazla ilişkili değişken arasındaki işlevsel bir ilişki.



In this lesson, we will address cases where there is only one independent variable. Cases with more than one independent variable will be examined in our Machine Learning course.

Bu derste, yalnızca bir bağımsız değişkenin olduğu durumları ele alacağız. Machine Learning kursumuzda birden fazla bağımsız değişkeni olan durumlar incelenecektir.



**Tips:**

* With the correlation coefficient, the variables X and Y are interchangeable.
* With regression, we are trying to predict the Y variable from X using a linear relationship.
* • Korelasyon katsayısı ile X ve Y değişkenleri birbirinin yerine kullanılabilir.
* • Regresyonla, doğrusal bir ilişki kullanarak X'ten Y değişkenini tahmin etmeye çalışıyoruz.

## Linear Regression

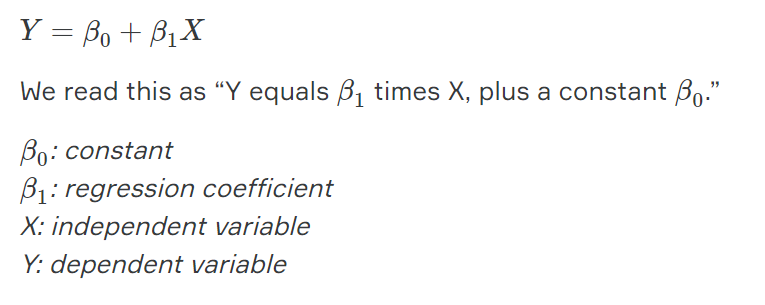
### Regression Equation

The least squares regression method is a statistical method used to write the mathematical relationship between two variables that change together.

En küçük kareler regresyon yöntemi, birlikte değişen iki değişken arasındaki matematiksel ilişkiyi yazmak için kullanılan istatistiksel bir yöntemdir.



Suppose Y is a dependent variable, and X is an independent variable.



In this formula, Y is the value of the dependent variable that we are trying to predict. β0 is the Y -intercept, the value of Y when X is equal to 0. β1 is the slope, for each increase in unit on the X -axis is a corresponding increase or decrease on the Y -axis.

Bu formülde Y, tahmin etmeye çalıştığımız bağımlı değişkenin değeridir. β0, Y -kesme noktasıdır, X 0'a eşit olduğunda Y'nin değeridir. β1 eğimdir, X eksenindeki birimdeki her artış, Y ekseninde karşılık gelen bir artış veya azalmadır.



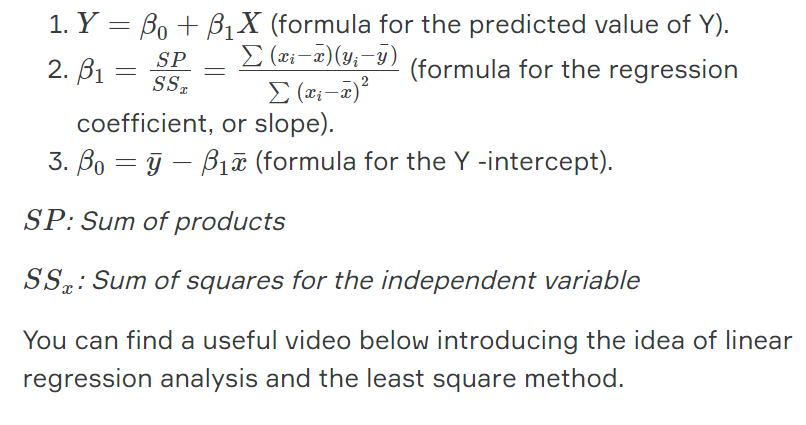
Basically, we call Y as dependent variable, X as independent variable, β0 and β1 as coefficients.

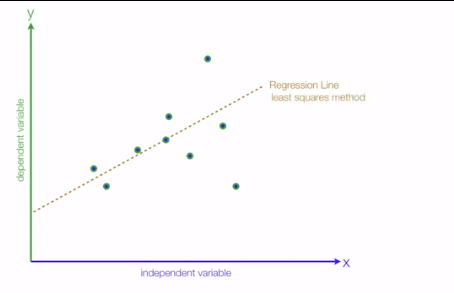
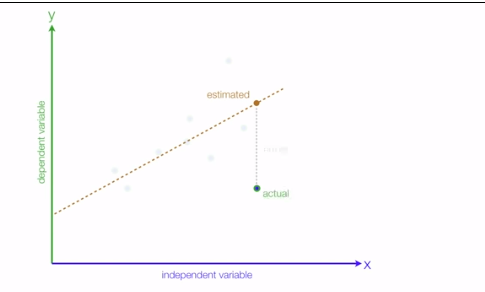
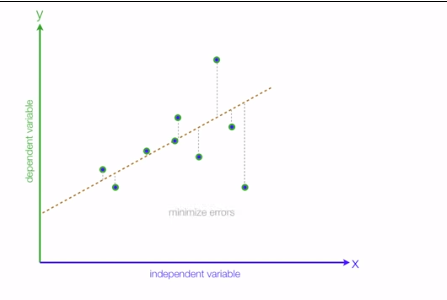
Temel olarak Y'yi bağımlı değişken, X'i bağımsız değişken, β0 ve β1'i katsayı olarak adlandırıyoruz.

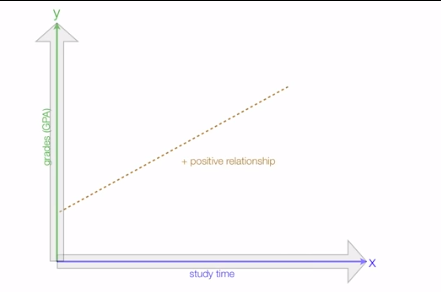
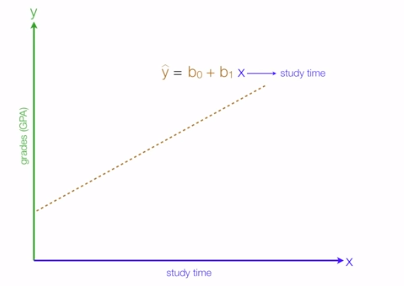
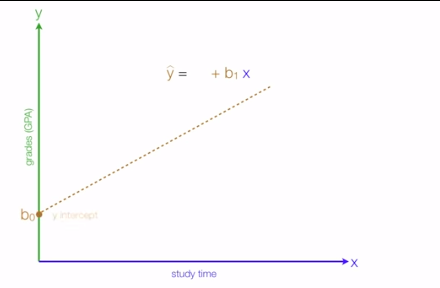
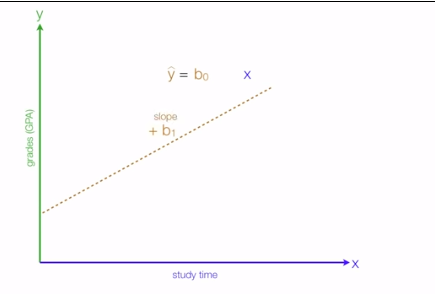


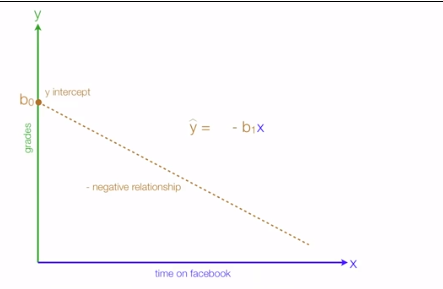
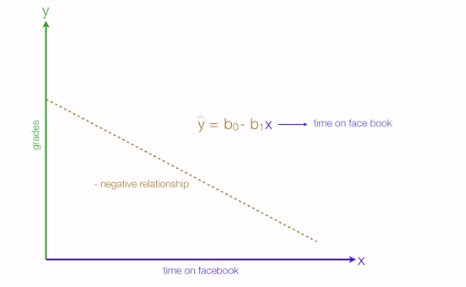
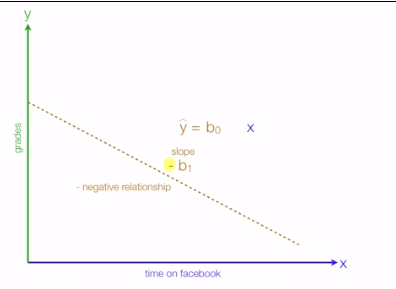
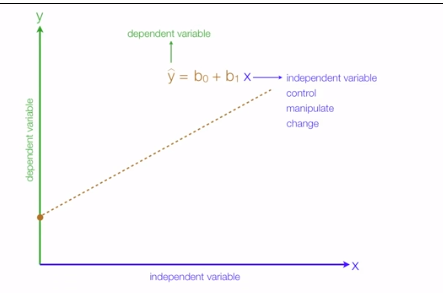
The values must be calculated on the basis of means and variances, specifically sums of squares. Therefore, the three formulas we are working with to draw a regression line are as follows:

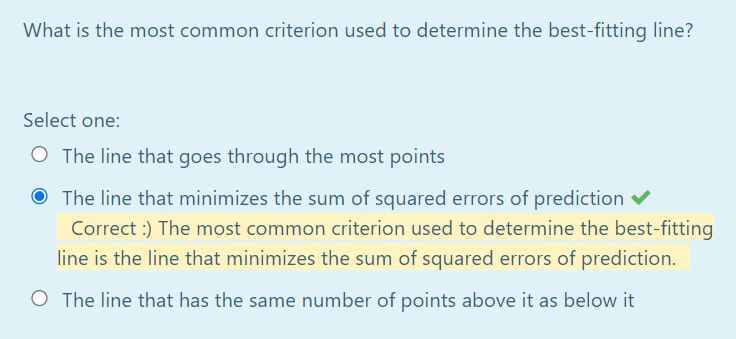
Değerler, ortalamalar ve varyanslar, özellikle kareler toplamı temelinde hesaplanmalıdır. Bu nedenle, bir regresyon çizgisi çizmek için üzerinde çalıştığımız üç formül aşağıdaki gibidir:





## Linear Regression

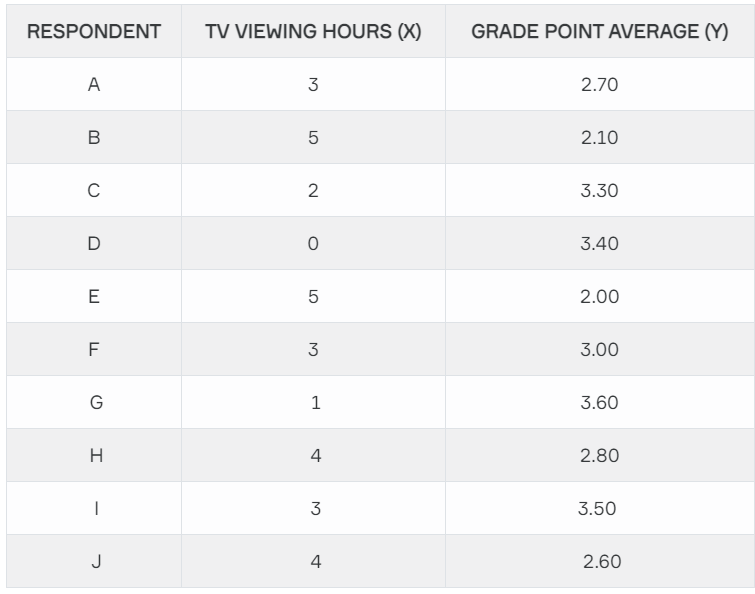
### Regression Example

#### **Grade Point Average and Television Viewing**

Suppose we want to learn more about the relationship between students’ GPAs and the time they spend watching television. Because watching television takes time away from more academic activities, we hypothesize that the two are negatively associated: As television viewing hours ( X ) increase, GPA ( Y ) decreases. To test our hypothesis, we sample 10 students and gather the following data.

Not Ortalaması ve Televizyon İzleme

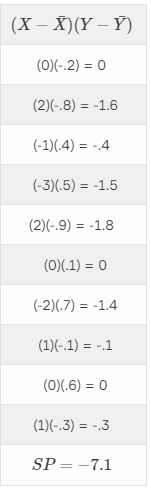
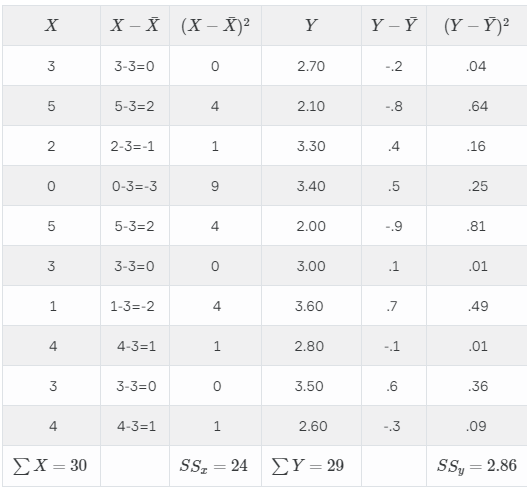
Öğrencilerin genel not ortalamaları ile televizyon izleyerek geçirdikleri zaman arasındaki ilişki hakkında daha fazla şey öğrenmek istediğimizi varsayalım. Televizyon izlemek daha fazla akademik aktiviteden zaman aldığından, ikisinin olumsuz bir şekilde ilişkili olduğunu varsayıyoruz: Televizyon izleme saatleri ( X ) arttıkça, not ortalaması ( Y ) düşüyor. Hipotezimizi test etmek için 10 öğrenciyi örnekliyoruz ve aşağıdaki verileri topluyoruz.



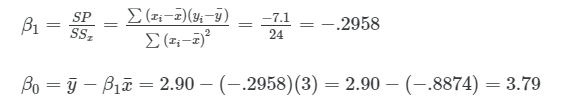
Before we can begin to solve for our predicted values of Y (Y=β0+β1XY=β0+β1X), we must first determine the regression coefficient ( β1  ) and the Y-intercept ( β0 ). To determine these values, we must first calculate the sum of products (SP), the sum of squares for the independent variable (SSx), the mean for the dependent variable (Y), and the mean for the independent variable (X). To do this, we set up a table similar to the one used to calculate a Pearson’s correlation since all of the values in the table will be of use to us at some point.

Öngörülen Y (Y=β0+β1X) değerlerimizi çözmeye başlamadan önce, önce regresyon katsayısını ( β1 ) ve Y kesişimini ( β0 ) belirlememiz gerekir. Bu değerleri belirlemek için önce çarpımların toplamını (SP), bağımsız değişkenin karelerinin toplamını (SSx), bağımlı değişkenin ortalamasını (Y) ve bağımsız değişkenin ortalamasını (X) hesaplamamız gerekir. Bunu yapmak için, tablodaki tüm değerler bir noktada bizim için yararlı olacağından, Pearson korelasyonunu hesaplamak için kullanılana benzer bir tablo oluşturduk.





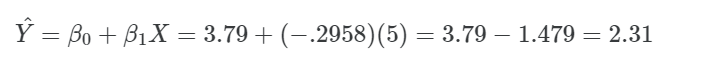
We can now calculate the regression coefficient ( β1 ) and the Y-intercept ( β0 ) as follows:



We can now use the values of the regression coefficient and the Y-intercept to calculate predicted values of y (Y) using given values of X to draw two points through which the fit line will pass. It is always a good idea to draw the fit line using the y -intercept as one point and the second point on the basis of using a large value of X in the formula for a line (Y=β0+β1XY=β0+β1X). This will produce two points that are far apart on the scatterplot, making the line easier to draw. First, we draw one point on the Y-axis at a GPA of 3.79 (the Y-intercept). Second, using an X value of 5, our highest value of X, we calculate the second point using the formula for a line as follows:

Artık, uygun doğrunun içinden geçeceği iki nokta çizmek için verilen X değerlerini kullanarak y (Y)'nin tahmin edilen değerlerini hesaplamak için regresyon katsayısı ve Y-kesişim değerlerini kullanabiliriz. Bir doğru (Y=β0+β1X) formülünde büyük bir X değeri kullanarak, bir nokta olarak y -kesme noktasını ve ikinci noktayı kullanarak uygun doğruyu çizmek her zaman iyi bir fikirdir. Bu, dağılım grafiğinde birbirinden uzak iki nokta üretecek ve çizginin çizilmesini kolaylaştıracaktır. İlk olarak, Y ekseni üzerinde 3.79 GPA'da (Y-kesişim noktası) bir nokta çiziyoruz. İkinci olarak, en yüksek X değerimiz olan 5'lik bir X değeri kullanarak, ikinci noktayı bir çizgi için aşağıdaki formülü kullanarak hesaplıyoruz:





This tells us that a student with an X value of 5 will have a Y value of 2.31.

Bu bize X değeri 5 olan bir öğrencinin Y değerinin 2.31 olacağını söyler.



## Linear Regression

### Explaining Variance in the Dependent Variable (R-squared)

From our previous example we first calculate Pearson's r as follows,



Pearson’s r indicates a strong negative association between GPA and television viewing. As the number of hours spent watching television each day increases, GPA tends to decrease. R-squared is nothing more than the square of Pearson’s r, but;it offers us unique insight on the relationship between the two variables in our analysis. When we square Pearson’s r we get:



The R2R2 value tells us what percent of the total variance in the dependent variable is explained by the independent variable. In other words, 74% of all the variance in GPA can be explained by how many hours of television students watch each day. Obviously these results are based on fabricated data because GPA can be a factor of so many other behaviors and conditions, however, R2R2 is a powerful tool that can be used to describe relationships between variables in a concise and easily understood manner.

R2R2 değeri bize bağımlı değişkendeki toplam varyansın yüzde kaçının bağımsız değişken tarafından açıklandığını söyler. Diğer bir deyişle, GNO'daki tüm varyansın %74'ü, öğrencilerin her gün kaç saat televizyon izlediği ile açıklanabilir. Açıkça bu sonuçlar uydurma verilere dayanmaktadır, çünkü GPA diğer birçok davranış ve koşulun bir faktörü olabilir, ancak R2R2, değişkenler arasındaki ilişkileri kısa ve kolay anlaşılır bir şekilde tanımlamak için kullanılabilecek güçlü bir araçtır.