Regression & Correlation

Cahit Karakuş

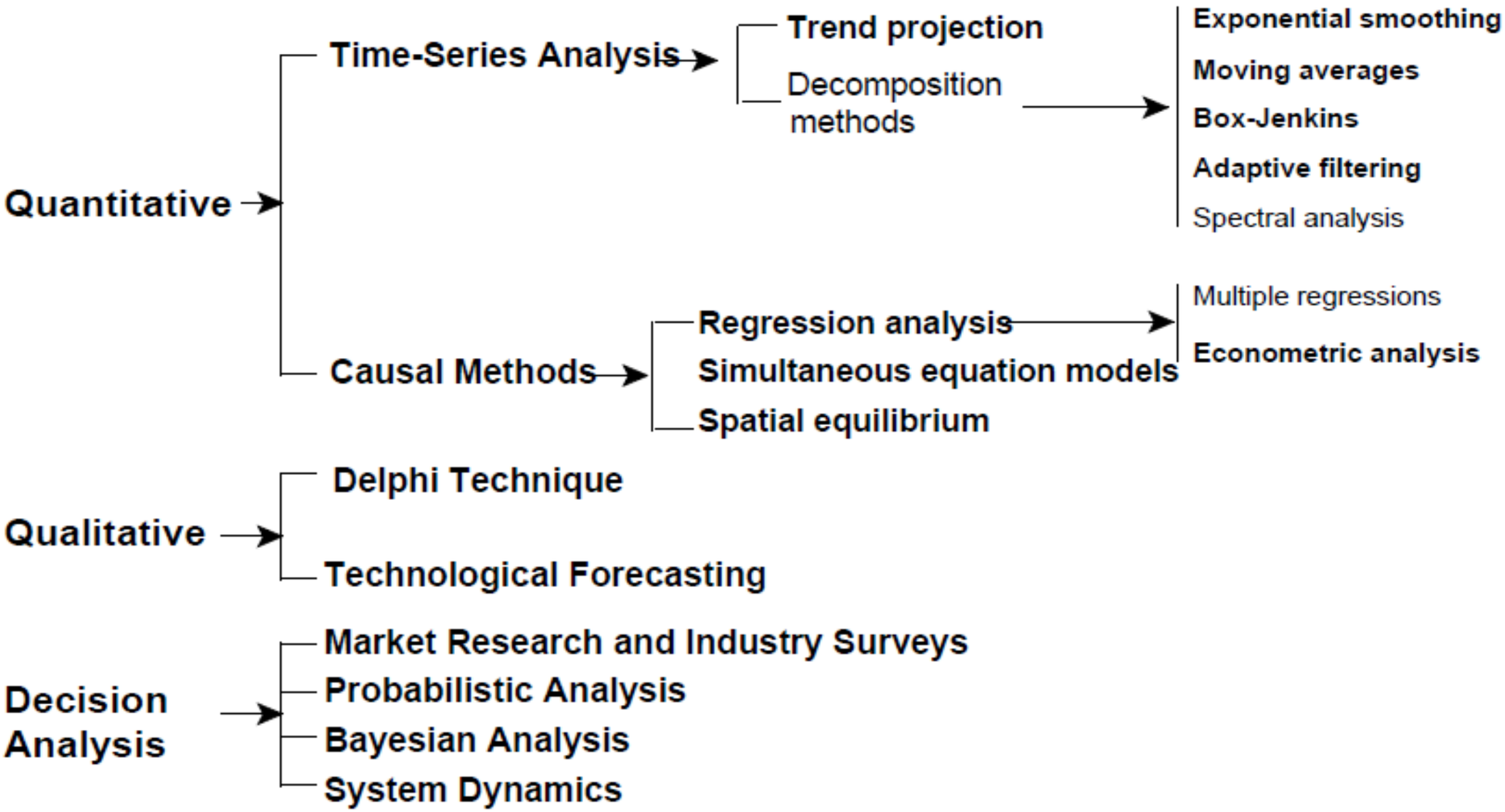
İstanbul, Turkey

Veri Türleri

Veriler genel olarak aşağıdaki üç türe ayrılabilir:

- **Zaman serisi verileri,** art arda zaman aralıkları boyunca toplanan, kaydedilen veya gözlemlenen verilerden oluşur.
- **Kesitsel veriler,** zaman içinde tek bir noktada toplanan gözlemlerdir.
- **Panel verileri,** bir havayolu örneği için taşınan yıllık yolcular gibi zaman içinde tekrarlanan kesitsel ölçümlerdir.

Forecasting Techniques



- Regression Analysis
  - Linear Regression:
- **<u>Simple Linear Regression</u>**: {y, x}
- Multiple Linear Regression: {y; x_1, … , x_p}
- Multivariate Linear Regression: {y_1, … , y_n; x_1, … , x_p}
- Correlation Analysis

Formulation of the Model

i) Linear: $Y = a + bX_1 + cX_2 + \ldots zX_n$

ii) Multiplicative or log-log: $Y = aX_1^{\,b}\, X_2^{\,c} \ldots X_n^{\,z}$

$$\log Y = \log(a) + b \log X_1 + c \log X_2 + \ldots z \log X_n$$

iii) Linear-log: $e^Y = aX_1^{\,b}\, X_2^{\,c} \ldots X_n^{\,z}$

$$Y = \log(a) + b \log X_1 + c \log X_2 + \ldots z \log X_n$$

iv) Log-linear: $\log Y = a + bX_1 + cX_2 + \ldots zX_n$

Matematik / İstatistik Modeli nedir?

- Genellikle değişkenler arasındaki ilişkiyi tanımlar

Türler:

- Deterministik Modeller (rastgelelik yok)
- Olasılık Modelleri (rastgelelikle)

Deterministik Modeller:

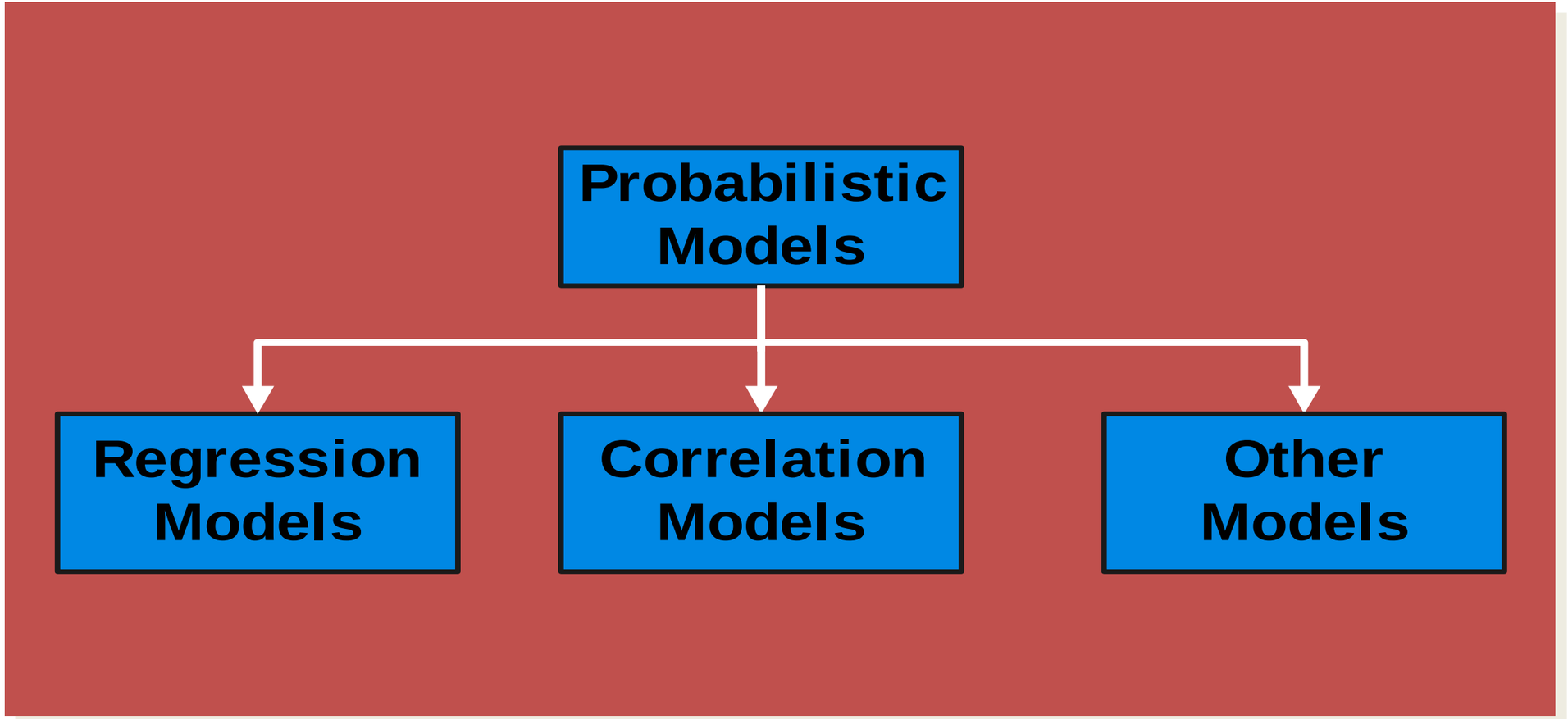
- Kesin ilişkileri varsayar
- Tahmin hatası önemsiz olduğunda uygundur.

Probabilistic Models

- Hypothesize two components
  - Deterministic
  - Random error

- Example: sales volume (y) is 10 times advertising spending (x) + random error
  - $y = 10x + \varepsilon$
  - Random error may be due to factors other than advertising

Types of Probabilistic Models

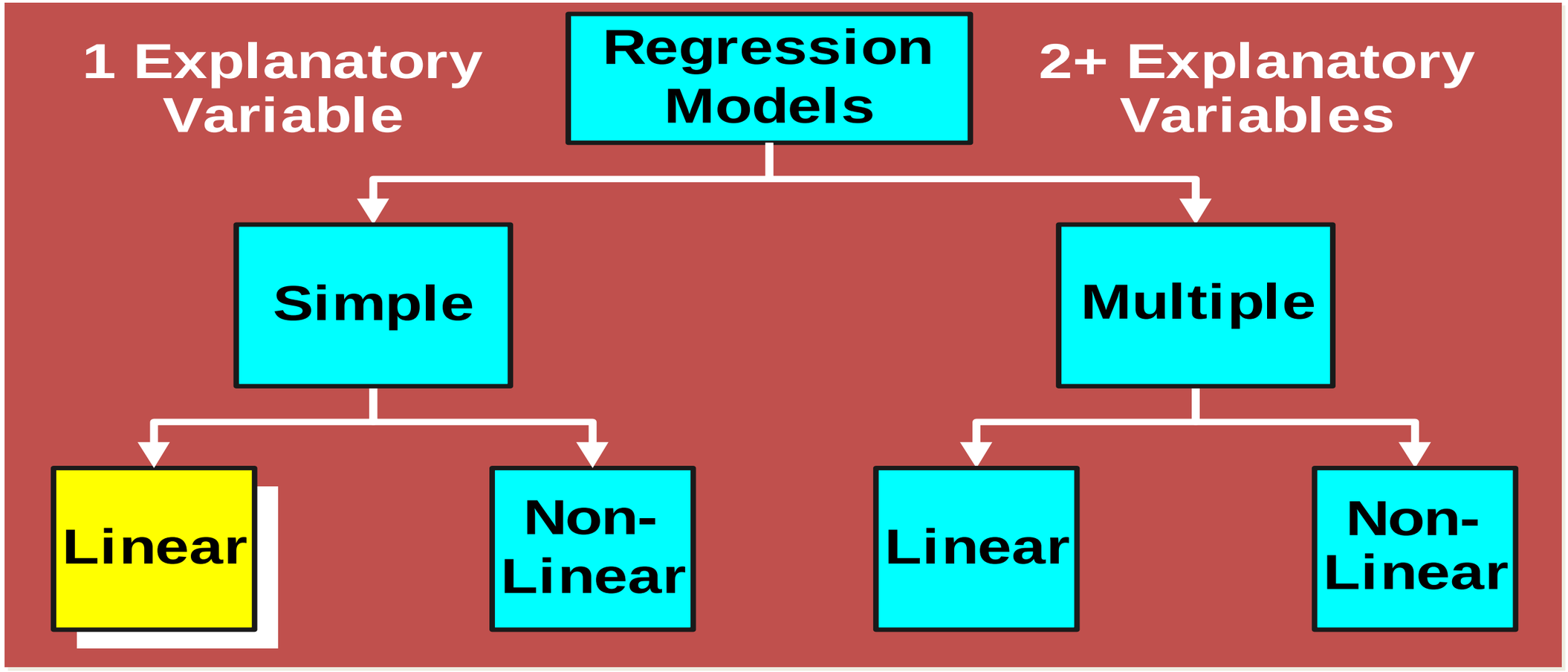


Regression Models (Forecasting and Planning)

Regresyon analizi

- Bir bağımlı değişken ile açıklayıcı değişken (ler) arasındaki ilişki, ilişki kurmak için denklem ve sayısal bağımlı (Yanıt) değişken kullanır.
- Esas olarak kestirim ve tahmin için kullanılır.

Types of Regression Models



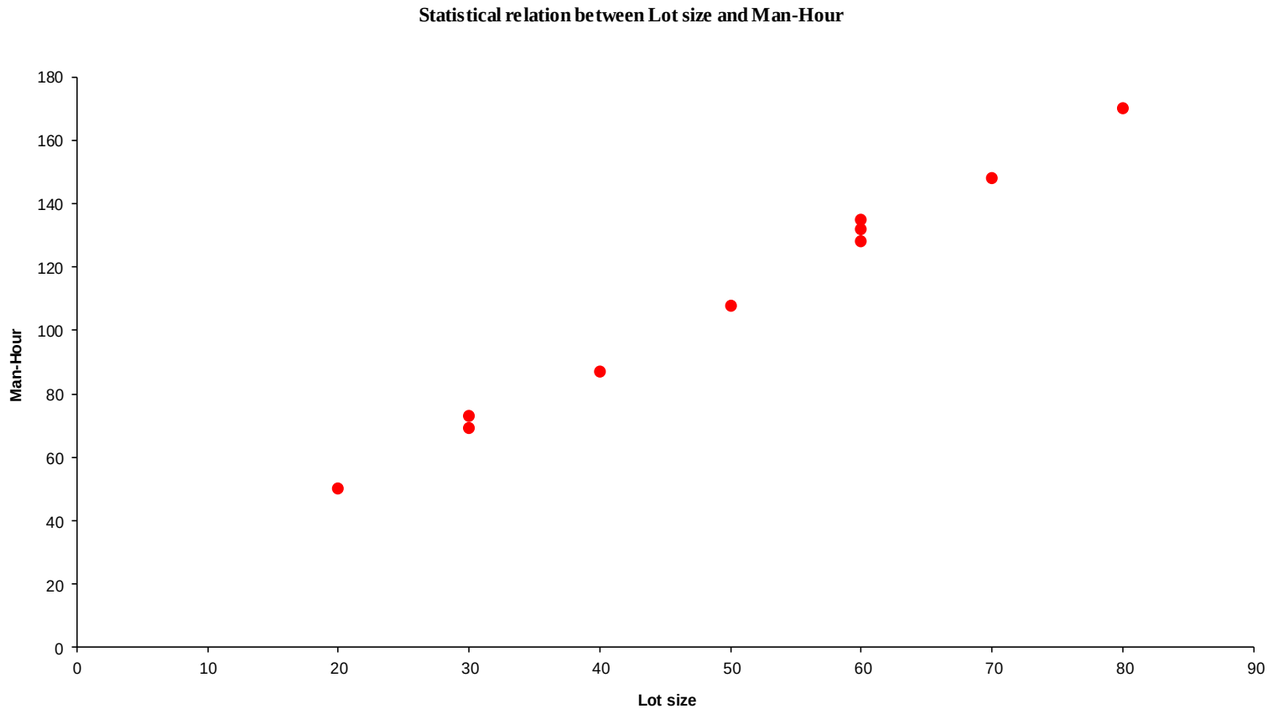
Regression Models

- Answers ‘What is the relationship between the variables?’
- Equation used
  - One numerical dependent (response) variable
    - What is to be predicted
  - One or more numerical or categorical independent (explanatory) variables
- Used mainly for prediction and estimation

- The goal of the analyst who studies the data is to find a functional relation

  between the response variable y and the predictor variable x.

$$y = f(x)$$



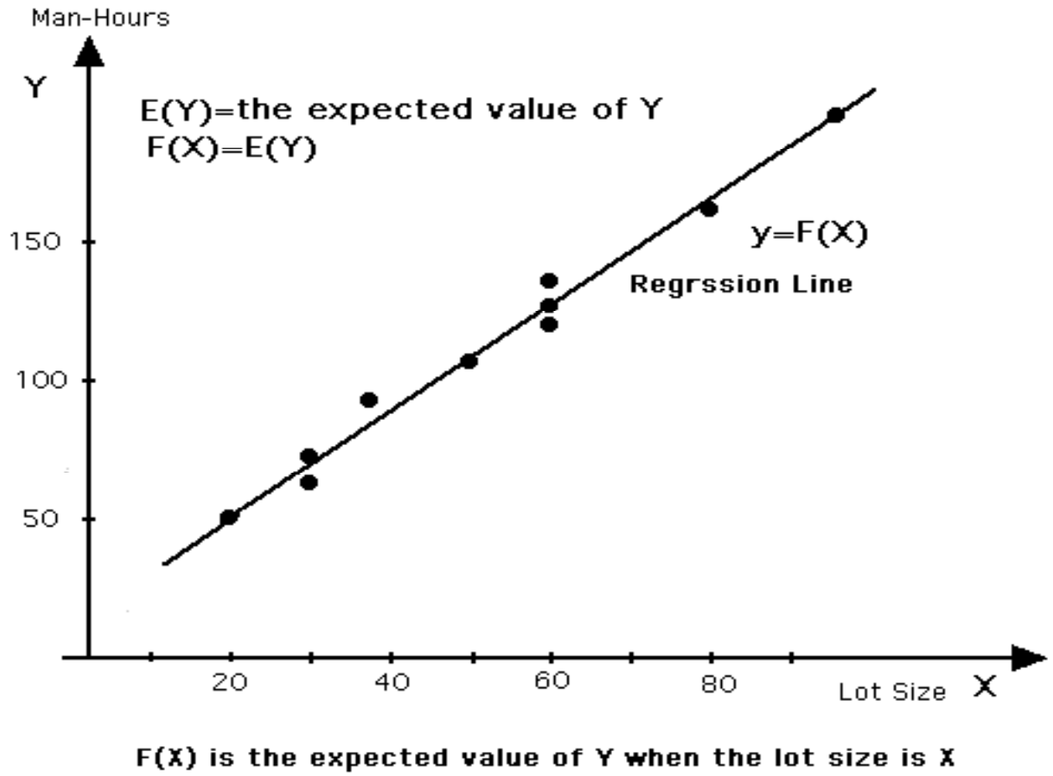
Regression Function

- The statement that the relation between X and Y is statistical should be interpreted as providing the following guidelines:

1. Regard Y as a random variable.
2. For each X, take $f(x)$ to be the expected value (i.e., mean value) of y.
3. Given that $E(Y)$ denotes the expected value of Y, call the equation

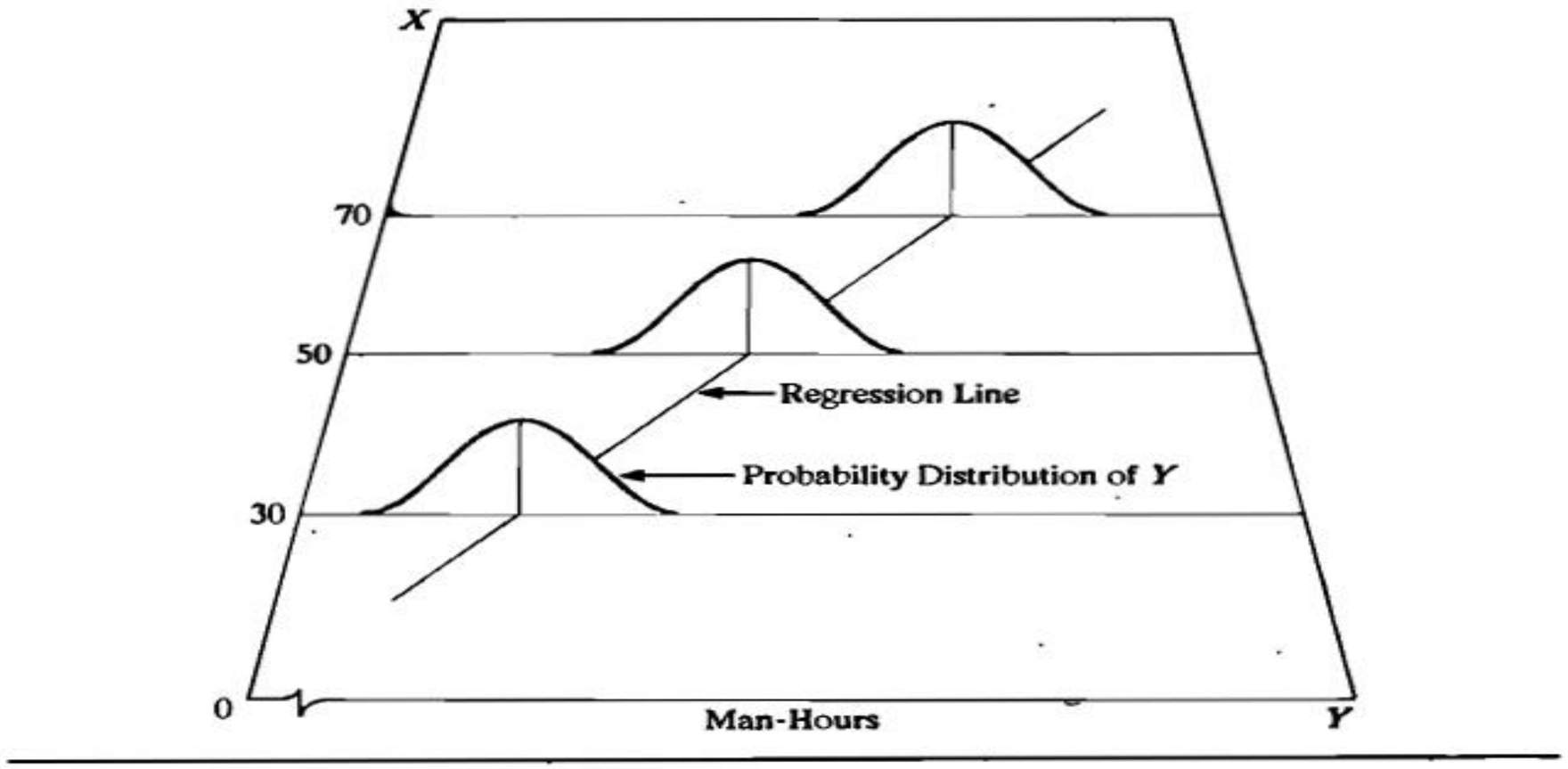
$$E(Y) = f(x)$$

the regression function.



F(X) is the expected value of Y when the lot size is X

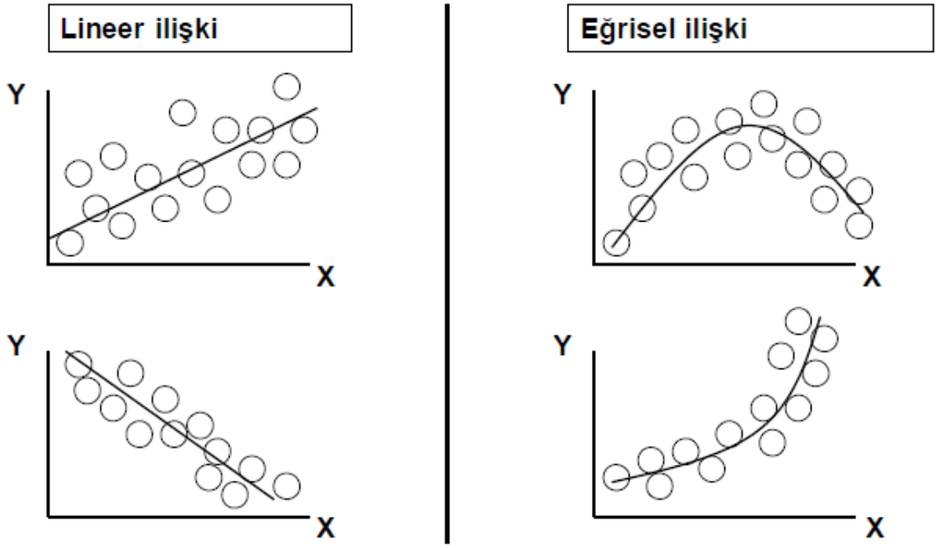
Pictorial Presentation of Linear Regression Model

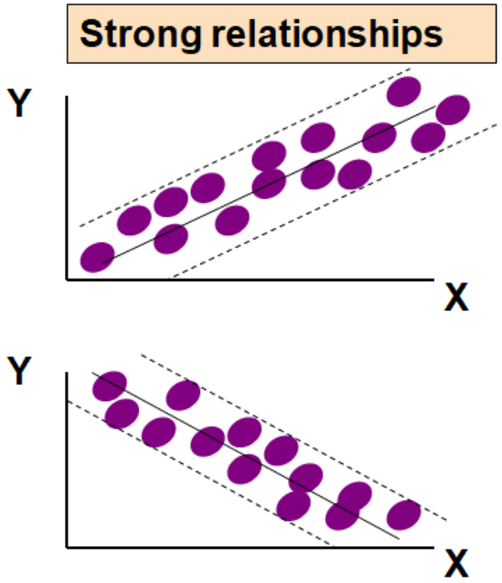


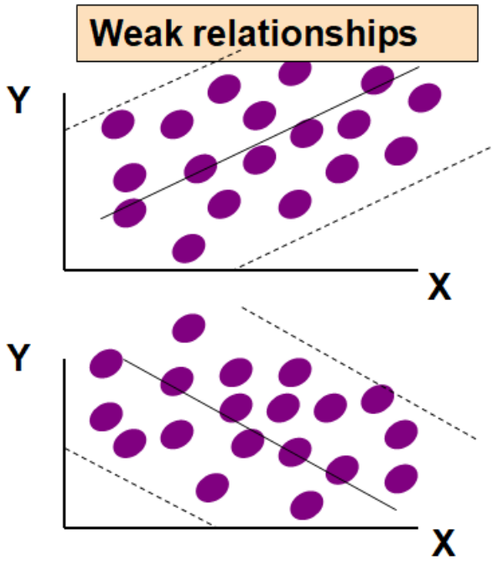
Specifying the Model

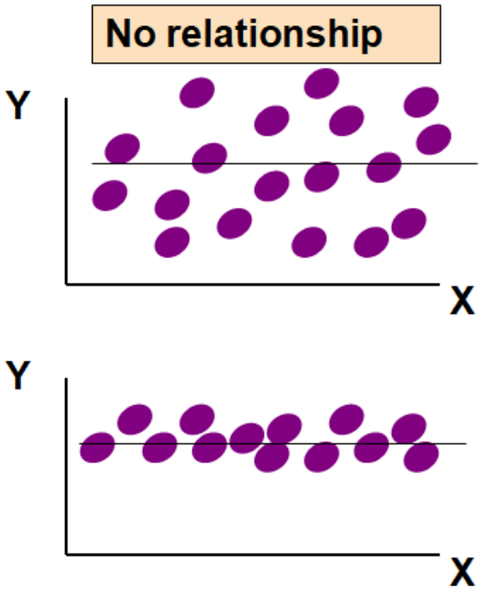
1. Define variables
    - Conceptual (e.g., Advertising, price)
    - Empirical (e.g., List price, regular price)
    - Measurement (e.g., $, Units)
2. Hypothesize nature of relationship
    - Expected effects (i.e., Coefficients' signs)
    - Functional form (linear or non-linear)
    - Interactions

Types of Relationships





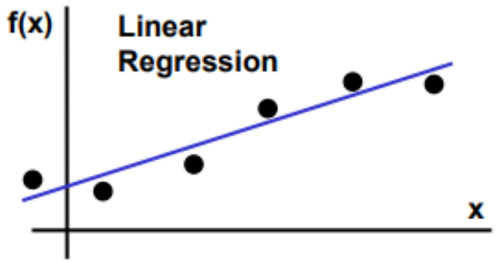


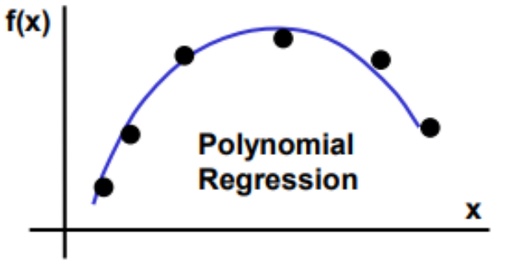


About Curve Fitting

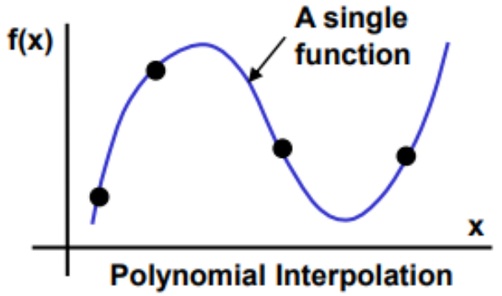
- Curve fitting is expressing a discrete set of data points as a continuous function.
- It is frequently used in engineering. For example the emprical relations that we use in heat transfer and fluid mechanics are functions fitted to experimental data.

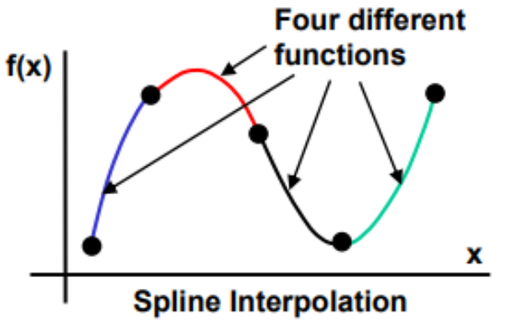
- **Regression:** Mainly used with experimental data, which might have significant amount of error (noise). No need to find a function that passes through all discrete points.





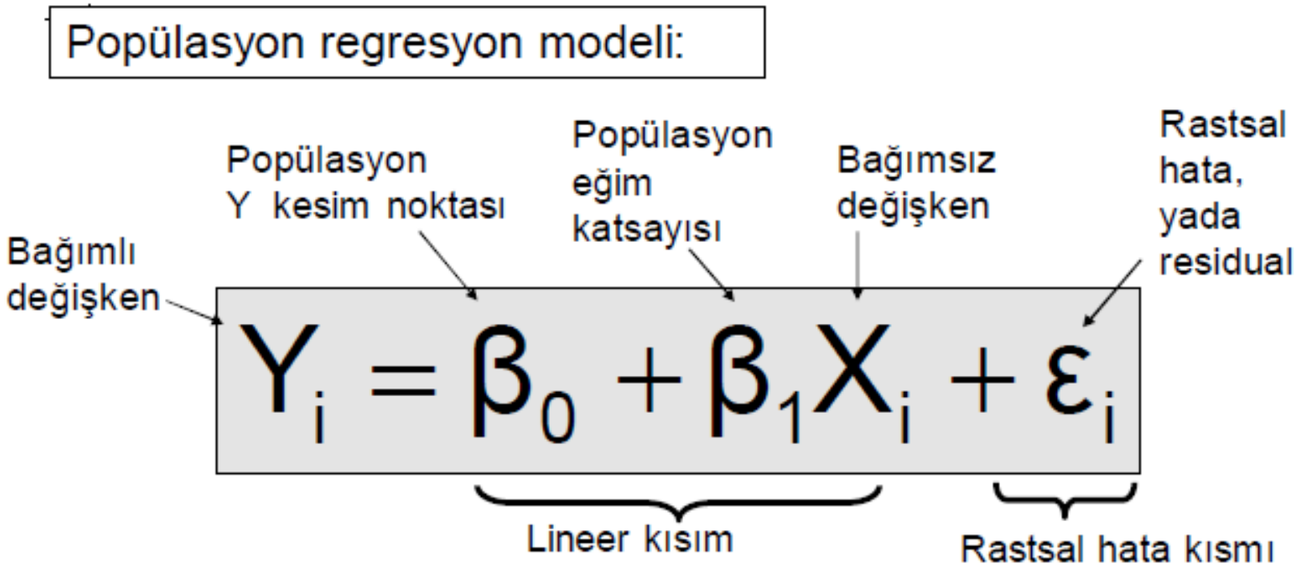
- **Interpolation:** Used if the data is known to be very precise. Find a function (or a series of functions) that passes through all discrete points.





Doğrusal Regresyon Modeli

Değişkenler arasındaki ilişki doğrusal bir fonksiyondur



Basit Doğrusal Regresyon Modeli

- Basit Doğrusal Regresyon Modeli

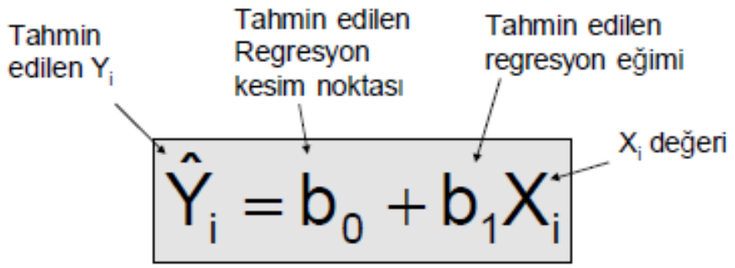
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Basit Doğrusal Regresyon Denklemi

$$E(y) = \beta_0 + \beta_1 x$$

- Tahmini Basit Doğrusal Regresyon Denklemi

$$\hat{y} = b_0 + b_1 x$$



e_i hata terimi sıfır ortalamaya sahip

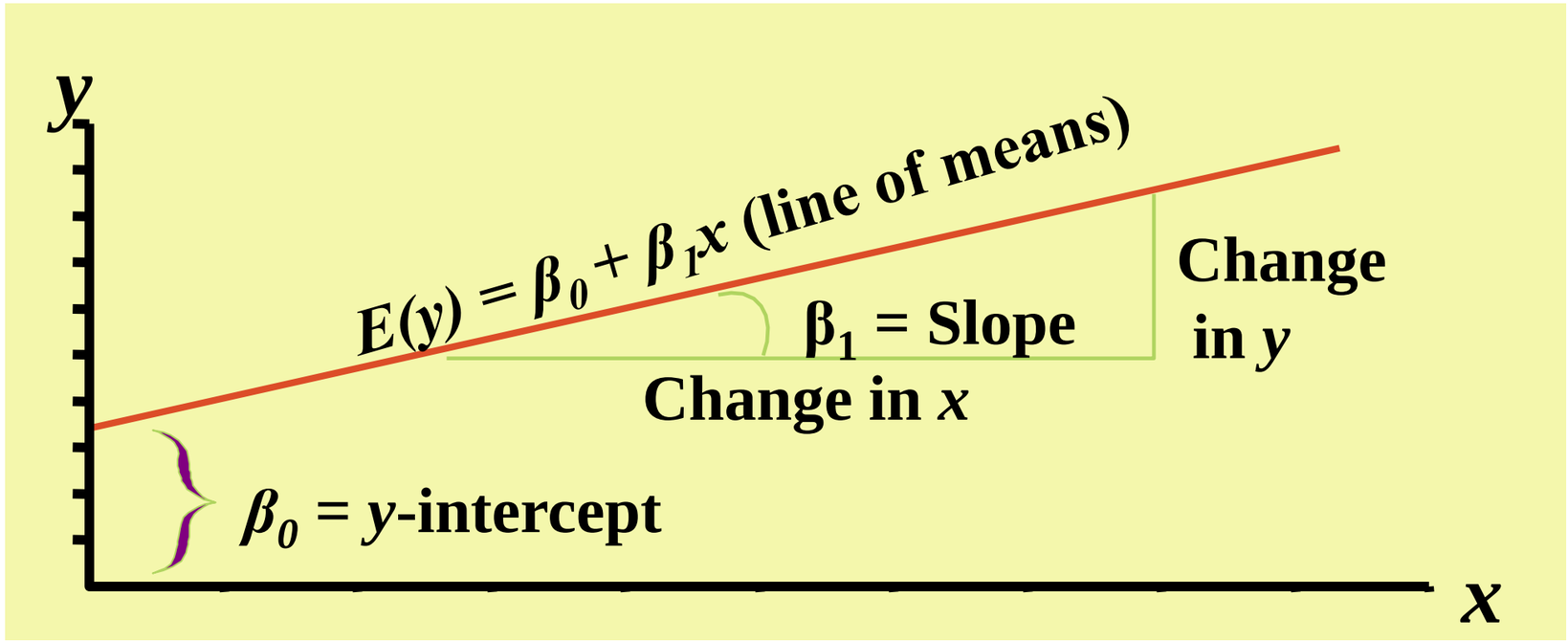
- General regression model
  1. β_0, and β_1 are parameters
  2. X is a known constant
  3. Deviations ε are independent N(o, σ^2)
- The values of the regression parameters β_0, and β_1 are not known.We estimate them from data.
- β_1 indicates the change in the mean response per unit increase in X.

- If the scatter plot of our sample data suggests a linear relationship between two variables i.e.

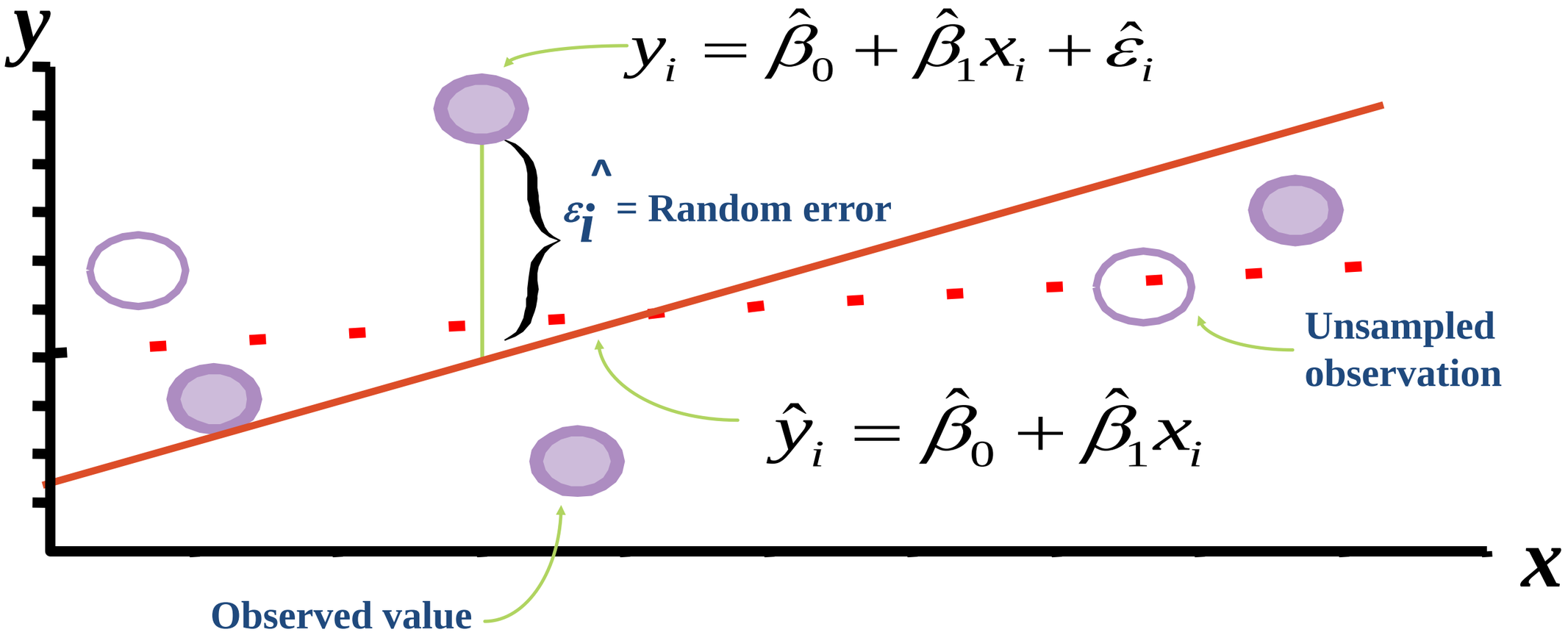
$$y = \beta_0 + \beta_1 x$$

  we can summarize the relationship by drawing a straight line on the plot.
- Least squares method give us the “best” estimated line for our set of sample data.

Line of Means



Sample Linear Regression Model



Estimating Parameters: Least Squares Method

En küçük kareler yöntemi

- Gerçek yaşamın çeşitli alanlarında herhangi bir uygulama ile toplanan veriler tablo şekline getirilerek incelenir ve toplanan veriyi modelleyen bir fonksiyon bulunmaya çalışılır. Çoğu zaman bu veri tablosuna tam olarak uyan bir fonksiyon bulmak mümkün olmaz; veri tablosuna en iyi uyan fonksiyon belirlenmeye çalışılır. Bir veri tablosuna en iyi uyan fonksiyonu bulma sürecine regresyon analizi denir.
- Regresyon analizi yaparken en çok kullanılan yöntemlerden biri en küçük kareler yöntemidir. Büyük matematikçi C. F. Gauss'un 18 yaşındayken (1795) geliştirdiği bu yöntem, ilk kez 1801 de Cres astroidinin yörüngesinin belirlenmesinde kullanılmış.
- Belli ölçümler sonucunda i = 1, 2, . . . , n için (xi , yi) verileri elde edilmiş olsun. Burada, her bir yi nin xi ye bağlı olarak değiştiği varsayılmaktadır. (xi , yi) düzlemde noktalar olarak düşünüldüğünde, pratikte bu noktalar düzgün bir eğri üzerinde, başka bir deyimle, bilinen bir fonksiyonun grafiği üzerinde bulunmazlar. Hatta bazı durumlarda, (xi , yi) ler arasında ne tür bir bağıntı bulunduğu dahi bilinmeyebilir.
- Yapılan ölçümlerin doğası gereği, her i = 1, 2, . . . , n için yi = f (xi) olacak biçimde bir fonksiyonun var olduğu, ölçümlerde yapılan hata nedeniyle bu eşitliklerin bazıları veya hepsinin sağlanmadığı kabul edilebilir. Bu düşünceyle, ölçülen yi değeri f (xi) için yaklaşık değer kabul edilerek bu yaklaşımdaki hatanın minimum olduğu f fonksiyonu belirlenmeye çalışılır. Bu amacı gerçekleştirmek için f fonksiyonunun bir takım parametrelere bağlı bir ifadesi bulunduğu varsayılıp eldeki veriler yardımıyla bu parametreler belirlenmeye çalışılır.

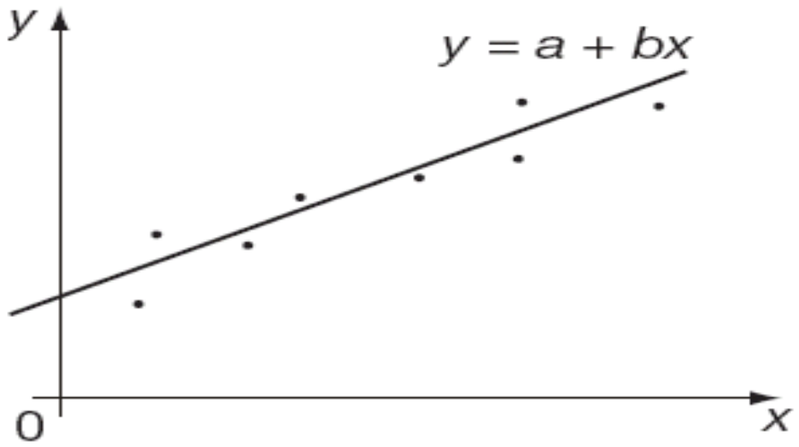
Method of least squares

Fitting a straight-line graph

Drawing a straight line of best fit through a set of plotted points by eye introduces unnecessary errors. To minimise errors the method of least squares is used where the sum of the squares of the vertical distances from the straight line is minimised.

Assume that the equation of the line of best fit is given by:

$$y = a + bx$$



The Line

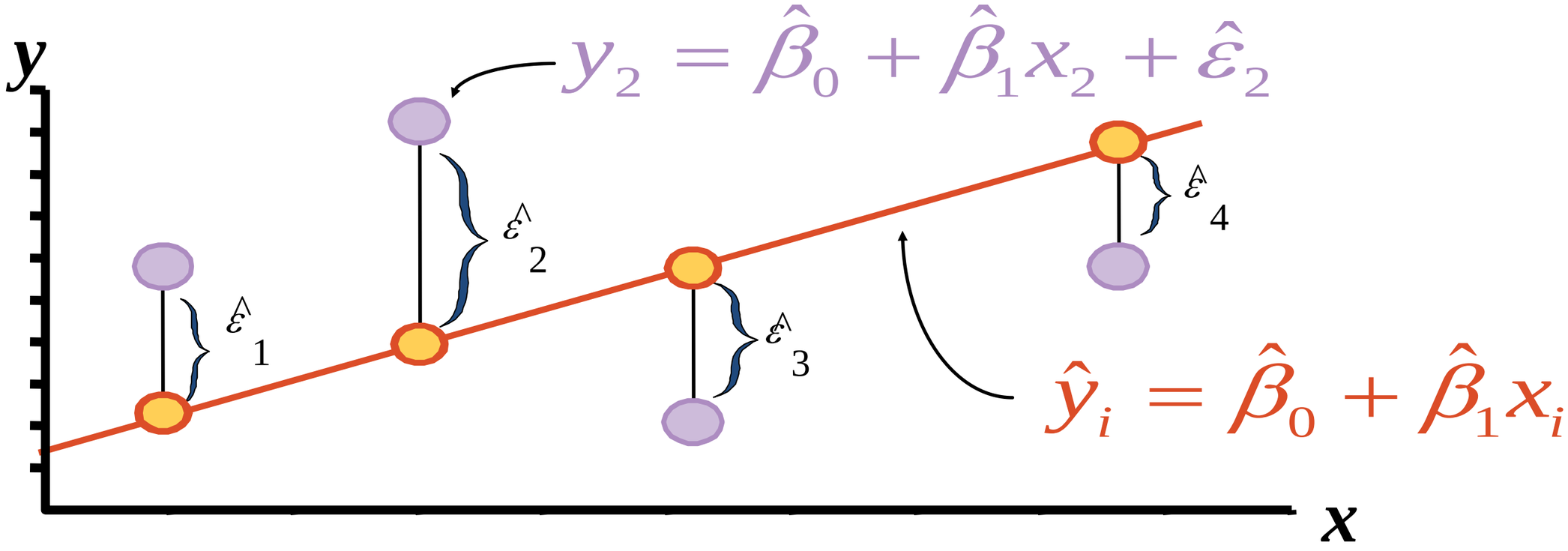
- Our aim is to calculate the values **m** (slope) and **b** (y-intercept) in the equation of a line :
- $y = mx + b$

Where:

- **y** = how far up
- **x** = how far along
- **m** = Slope or Gradient (how steep the line is)
- **b** = the Y Intercept (where the line crosses the Y axis)

Least Squares Graphically

$$\text{LS minimizes } \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$$



Tanım 1. $y_i - f(x_i)$ farklarından her birine bir **artık** denir.

En küçük kareler yönteminde aranan fonksiyon, ya da onun parametreleri, tüm artıkların kareleri toplamı olan

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 = (y_1 - f(x_1))^2 + \cdots + (y_n - f(x_n))^2$$

ifadesini minimum yapacak şekilde belirlenir. Bu, yönteme neden en küçük kareler yöntemi dendiğini açıklar. Sözü edilen kareler toplamının minimum olması için her bir hatanın küçük olması gerektiğine dikkat ediniz.

Least Squares

- ‘Best fit’ means difference between actual *y* values and predicted *y* values are a minimum
  - *But* positive differences off-set negative

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- Least Squares minimizes the Sum of the Squared Differences (SSE)

Minimizing the Square of Individual errors

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$ Sum of squares of the residuals

- Determine the unknowns a_0 and a_1 by minimizing S_r.
- To do this set the derivatives of S_r wrt a_0 and a_1 to zero.

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i) = 0 \quad \rightarrow \quad n\,a_0 + \left(\sum x_i\right)a_1 = \sum y_i$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n}[(y_i - a_0 - a_1 x_i)x_i] \quad \rightarrow \quad \left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 = \sum x_i y_i$$

or $$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_i y_i \end{Bmatrix}$$ These are called the normal equations.

- Solve these for a_0 and a_1. The results are

$$a_1 = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad a_0 = \bar{y} - a_1 \bar{x}$$

where y-bar and x-bar are the means of y and x, respectively.

En küçük kareler yöntemi

En küçük kareler yöntemi bu *E* toplamının en küçük değerinin bulunmasını hedefler ki bunun için problemin iki parametresi olan *a0* ve *a1* büyüklüklerinin en uygun değerinin bulunması gerekir.
E büyüklüğünün minimum olması için bu büyüklüğün *a0* ve *a1* parametrelerine göre türevlerinin sıfır olması gerekir:

$$\frac{\partial E}{\partial a_0} = \sum_{k=1}^{N} 2(y_k - a_0 - a_1 x_k)(-1) = 0$$

$$\frac{\partial E}{\partial a_1} = \sum_{k=1}^{N} 2(y_k - a_0 - a_1 x_k)(-x_k) = 0$$

Bu eşitlikler a_0 ve a_1 için yazılmış birer denklem olup düzenlenerek matris formda

$$\begin{bmatrix} N & \sum_{k=1}^{N} x_k \\ \sum_{k=1}^{N} x_k & \sum_{k=1}^{N} x_k^2 \end{bmatrix} \cdot \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{Bmatrix} \sum_{k=1}^{N} y_k \\ \sum_{k=1}^{N} x_k y_k \end{Bmatrix}$$

- We will write an estimated regression line based on sample data as

$$\hat{y} = b_0 + b_1 x$$

- The method of least squares chooses the values for b_0, and b_1 to minimize the sum of squared errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y - b_0 - b_1 x\right)^2$$

- Using calculus, we obtain estimating formulas:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b_1 = r \frac{S_y}{S_x}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- veri dağılımını (x_k, y_k) $k=1,2,\ldots,N$

- buna uydurulacak doğruyu $y(x) = a_0 + a_1 x$

- Hatalar $e_k = y_k - y(x_k)$

- Hataların kareleri $e_k^2 = [y_k - y(x_k)]^2 = (y_k - a_0 - a_1 x_k)^2$

- Hataların kareleri toplamı $E = \sum_{k=1}^{N} (y_k - a_0 - a_1 x_k)^2$

Computation Table

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
x_1	y_1	x_1^2	y_1^2	$x_1 y_1$
x_2	y_2	x_2^2	y_2^2	$x_2 y_2$
:	:	:	:	:
x_n	y_n	x_n^2	y_n^2	$x_n y_n$
Σx_i	Σy_i	Σx_i^2	Σy_i^2	$\Sigma x_i y_i$

- We have:

$n = 10 \qquad \sum x = 564 \qquad \sum x^2 = 32604$

$\sum y = 14365 \qquad \sum xy = 818755$

- The least squares estimates of the regression coefficients are:

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10(818755) - (564)(14365)}{10(32604) - (564)^2} = 10.8$$

$$b_0 = 1436.5 - 10.8(56.4) = 828$$

- The estimated regression function is:

$$\hat{y} = 828 + 10.8x$$

$$\text{Sales} = 828 + 10.8\ \text{Expenditure}$$

- This means that if the weekly advertising expenditure is increased by \$1 we would expect the weekly sales to increase by \$10.8.

Method of least squares

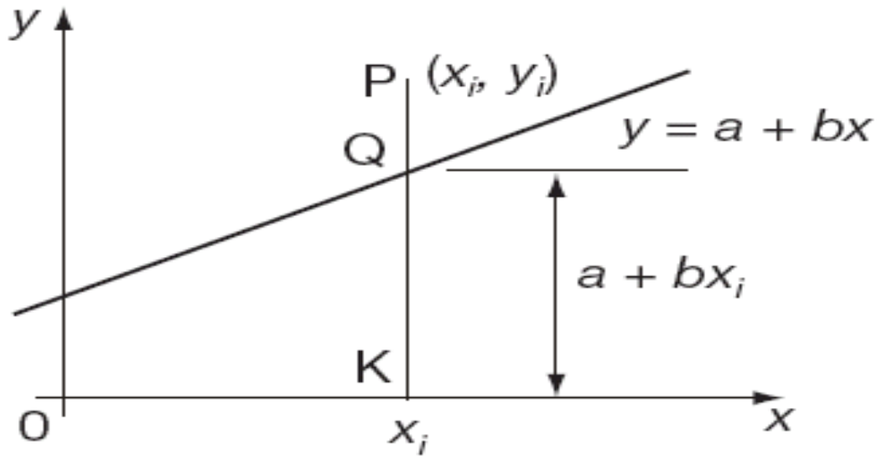
Fitting a straight-line graph

The i^{th} point plotted, (x_i, y_i), is a vertical distance from the line:

$$y_i - a - bx_i$$

The sum of the squares of these differences for all n points plotted is then:

$$S = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$



Method of least squares

Fitting a straight-line graph

The values of a and b must now be determined that gives S its minimum value. For S to be a minimum:

$$\frac{\partial S}{\partial a} = 0 \quad \text{and} \quad \frac{\partial S}{\partial b} = 0$$

This yields the two simultaneous equations from which the values of a and b can be found:

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Steps

To find the line of best fit for **N** points:

Step 1: For each (x,y) point calculate x^2 and xy

Step 2: Sum all x, y, x^2 and xy, which gives us Σx, Σy, Σx^2 and Σxy

Step 3: Calculate Slope **m**:

$$\mathbf{m} = \frac{N \Sigma(xy) - \Sigma x \, \Sigma y}{N \Sigma(x^2) - (\Sigma x)^2}$$

(N is the number of points.)

Step 4: Calculate Intercept **b**:

$$\mathbf{b} = \frac{\Sigma y - m \, \Sigma x}{N}$$

Step 5: Assemble the equation of a line

$$y = mx + b$$

Step 3: Calculate Slope **m**:

$$\mathbf{m} = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

$$= \frac{5 \times 263 - 26 \times 41}{5 \times 168 - 26^2}$$

$$= \frac{1315 - 1066}{840 - 676}$$

$$= \frac{249}{164} = 1{,}5183\ldots$$

Step 4: Calculate Intercept **b**:

$$\mathbf{b} = \frac{\Sigma y - m\,\Sigma x}{N}$$

$$= \frac{41 - 1{,}5183 \times 26}{5}$$

$$= 0{,}3049\ldots$$

Step 5: Assemble the equation of a line:

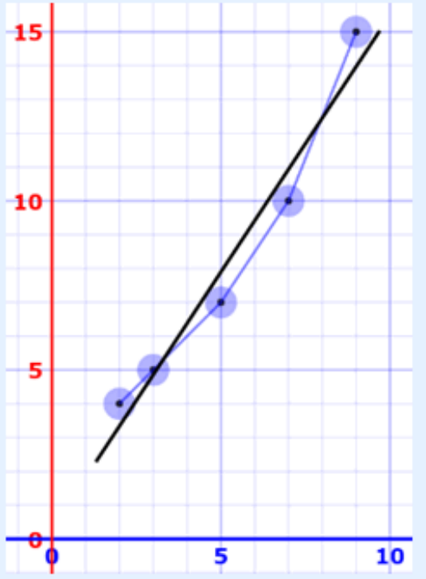
$$y = mx + b$$

$$y = 1{,}518x + 0{,}305$$

Let's see how it works out:

x	y	y = 1,518x + 0,305	error
2	4	3,34	−0,66
3	5	4,86	−0,14
5	7	7,89	0,89
7	10	10,93	0,93
9	15	13,97	−1,03

Here are the (x,y) points and the line **y = 1,518x + 0,305** on a graph:



Nice fit!

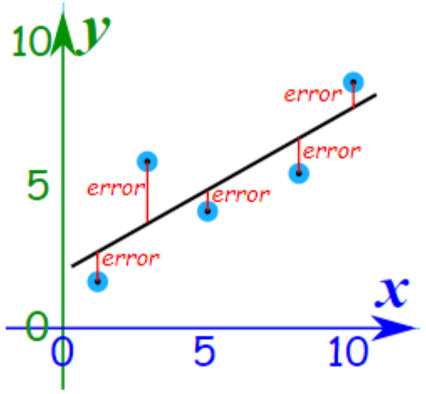
Sam hears the weather forecast which says "we expect 8 hours of sun tomorrow", so he uses the above equation to estimate that he will sell

$$y = 1{,}518 \times 8 + 0{,}305 = 12{,}45 \text{ Ice Creams}$$

Sam makes fresh waffle cone mixture for 14 ice creams just in case. Yum.

How does it work?

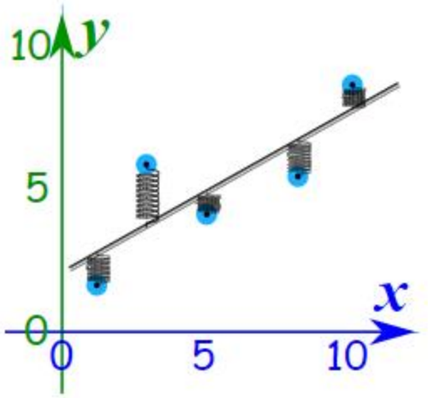
It works by making the total of the **square of the errors** as small as possible (that is why it is called "least squares"):



The straight line minimizes the sum of squared errors

So, when we square each of those errors and add them all up, the total is as small as possible.

You can **imagine** (but not accurately) each data point connected to a straight bar by springs:



Example: Sam found how many **hours of sunshine** vs how many **ice creams** were sold at the shop from Monday to Friday:

"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

Let us find the best **m** (slope) and **b** (y-intercept) that suits that data

$$y = mx + b$$

Step 1: For each (x,y) calculate x^2 and xy:

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 2: Sum x, y, x^2 and xy (gives us Σx, Σy, Σx^2 and Σxy):

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	**Σy: 41**	**Σx^2: 168**	**Σxy: 263**

Also **N** (number of data values) = **5**

Use least-squares regression to fit a straight line to

x	1	3	5	7	10	12	13	16	18	20
y	4	5	6	5	8	7	6	9	12	11

$n = 10$

$\sum x_i = 105$

$\sum y_i = 73$

$\bar{x} = 10.5$

$\bar{y} = 7.3$

$\sum x_i^2 = 1477$

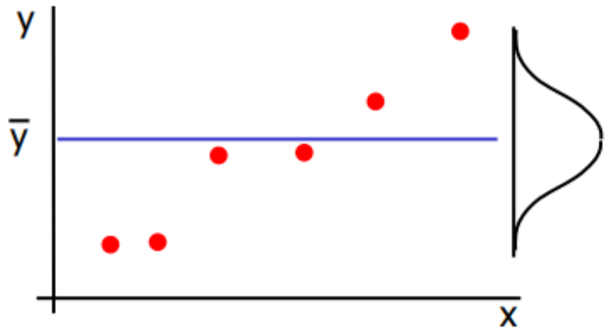
$\sum x_i y_i = 906$

$$a_1 = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{10*906 - 105*73}{10*1477 - 105^2} = 0.3725$$

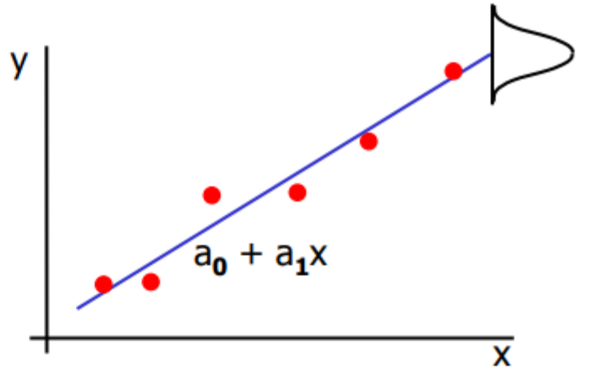
$$a_0 = 7.3 - 0.3725*10.5 = 3.3888$$

Error of Linear Regression (How good is the best line?)

Spread of data around the mean



Spread of data around the regression line



$$S_t = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s_y = \sqrt{\frac{S_t}{n-1}}$$ std. deviation

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - a_0 - a_1x_i)^2$$

$$s_{y/x} = \sqrt{\frac{S_r}{n-2}}$$ std. error of estimate

• The improvement obtained by using a regression line instead of the mean gives a maesure of how good the regression fit is.

coefficient of determination $r^2 = \dfrac{S_t - S_r}{S_t}$

correlation coefficient $r = \dfrac{n\sum(x_iy_i) - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$

How to interpret the correlation coefficient?

- Two extreme cases are
  - $S_r = 0 \rightarrow r=1$ describes a perfect fit (straight line passing through all points).
  - $S_r = S_t \rightarrow r=0$ describes a case with no improvement.
- Usually an r value close to 1 represents a good fit. But be careful and always plot the data points and the regression line together to see what is going on.

Example (cont'd): Calculate the correlation coefficient.

$$n = 10$$
$$\sum x_i = 105$$
$$\sum y_i = 73$$
$$\bar{x} = 10.5$$
$$\bar{y} = 7.3$$
$$\sum x_i^2 = 1477$$
$$\sum x_i y_i = 906$$

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 64.1$$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 = 12.14$$

$$r^2 = \frac{S_t - S_r}{S_t} = 0.8107$$

$$r = 0.9$$

Lineer olmayan veriler

veri seti non-lineer bir dağılım gösterebilir.

$y = a \cdot x^{b}$, $y = a \cdot e^{bx}$

Bu iki fonksiyon artık lineerdir.

Log(y)=log(a)+blog(x)
v= αu + β
v=log(y)
u=logx
α=b
β=log(a)

Ln(y)=ln(a)+bln(x)
v= αu + β
v=ln(y)
u=lnx
α=b
β=ln(a)

En küçük kareler polinomları

yakınlaşım polinomu $$y = \sum_{j=0}^{n} a_j x^j = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$$

hatalar $$e_k = y_k - \sum_{j=0}^{n} a_j x_k^j$$

hataların kareleri $$e_k^2 = \left[y_k - \sum_{j=0}^{n} a_j x_k^j \right]^2$$

kareler toplamı $$E = \sum_{k=1}^{N} \left[y_k - \sum_{j=0}^{n} a_j x_k^j \right]^2$$

En küçük kareler polinomları

kareler toplamının en küçük değerini elde etmek için E toplamının a_i ($i=0,1,2,\ldots,n$) paremetrelerine göre türevlerinin sıfır olması gerekmektedir:

$$\frac{\partial E}{\partial a_i} = \sum_{k=1}^{N} 2\left[y_i - \sum_{j=0}^{n} a_j x_k^j \right] \cdot \left[-x_k^i \right] = 0 \; ; \qquad i = 0,1,2,\cdots,n$$

Bu şekilde elde edilen $n+1$ adet lineer denklem eş-zamanlı olarak çözülerek a_j katsayıları elde edilebilir. Bunun için yukarıdaki denklem sistemi düzenlenerek

$$\sum_{j=0}^{n} \left(\sum_{k=1}^{N} a_j x_k^j x_k^i \right) = \sum_{k=1}^{N} x_k^i y_k \; ; \qquad i = 0,1,2,\cdots,n$$

En küçük kareler polinomları

$$\begin{bmatrix} N & \sum_{k=1}^{N} x_k^1 & \sum_{k=1}^{N} x_k^2 & \sum_{k=1}^{N} x_k^3 & \cdots & \sum_{k=1}^{N} x_k^n \\ \sum_{k=1}^{N} x_k^1 & \sum_{k=1}^{N} x_k^2 & \sum_{k=1}^{N} x_k^3 & \sum_{k=1}^{N} x_k^4 & \cdots & \sum_{k=1}^{N} x_k^{n+1} \\ \sum_{k=1}^{N} x_k^2 & \sum_{k=1}^{N} x_k^3 & \sum_{k=1}^{N} x_k^4 & \sum_{k=1}^{N} x_k^5 & \cdots & \sum_{k=1}^{N} x_k^{n+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{k=1}^{N} x_k^n & \sum_{k=1}^{N} x_k^{n+1} & \sum_{k=1}^{N} x_k^{n+2} & \sum_{k=1}^{N} x_k^{n+3} & \cdots & \sum_{k=1}^{N} x_k^{n+n} \end{bmatrix} \cdot \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ \cdots \\ a_n \end{Bmatrix} = \begin{Bmatrix} \sum_{k=1}^{N} y_k \\ \sum_{k=1}^{N} x_k y_k \\ \sum_{k=1}^{N} x_k^2 y_k \\ \cdots \\ \sum_{k=1}^{N} x_k^n y_k \end{Bmatrix}$$

şeklinde yazılabilir. Bu denklem sistemi Gauss eliminasyon yöntemi ile çözülebilir.

Aşağıdaki veri noktalarına en küçük kareler yaklaşımı ile kuadratik bir eğri uydurunuz

x_k	0.050	0.110	0.150	0.310	0.460	0.520	0.700	0.740	0.820	0.980	1.171
y_k	0.956	0.890	0.832	0.717	0.571	0.539	0.378	0.370	0.306	0.242	0.104

Probability Distribution of Random Error

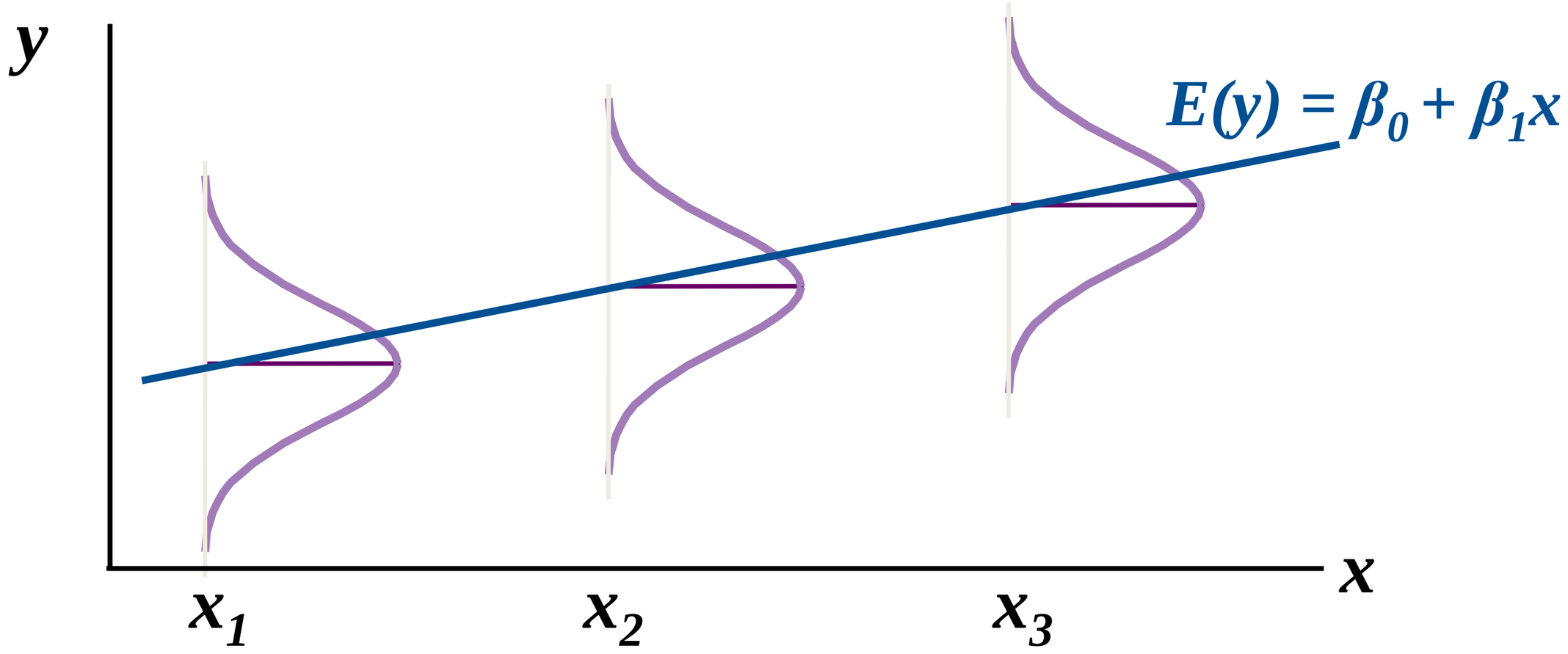
Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. **Specify probability distribution of random error term**
    - Estimate standard deviation of error
4. Evaluate model
5. Use model for prediction and estimation

Linear Regression Assumptions

1. Mean of probability distribution of error, ε, is 0
2. Probability distribution of error has constant variance
3. Probability distribution of error, ε, is normal
4. Errors are independent

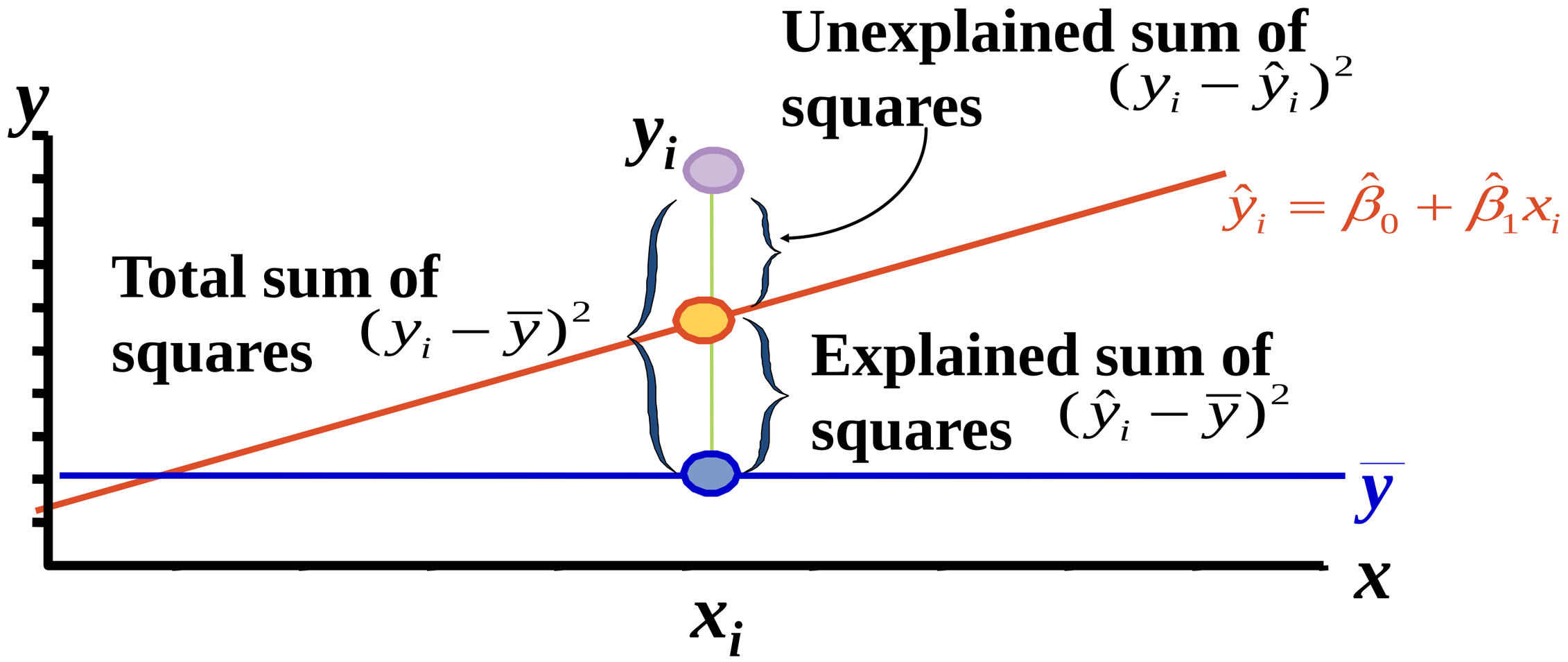
Error Probability Distribution



Random Error Variation

- Variation of actual y from predicted y, $\hat{y}$
- Measured by standard error of regression model
  - Sample standard deviation of ε : s
- Affects several factors
  - Parameter significance
  - Prediction accuracy

Variation Measures



Estimation of σ^2

$$s^2 = \frac{SSE}{n-2} \quad where \quad SSE = \sum (y_i - \hat{y}_i)^2$$

$$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$$

Calculating SSE, s^2, s Example

You're a marketing analyst for Hasbro Toys. You gather the following data:

Ad $	Sales (Units)
1	1
2	1
3	2
4	2
5	4

Find **SSE**, s^2, and **s**.

Calculating SSE Solution

x_i	y_i	$\hat{y} = -.1 + .7x$	$y - \hat{y}$	$(y - \hat{y})^2$
1	1	.6	.4	.16
2	1	1.3	-.3	.09
3	2	2	0	0
4	2	2.7	-.7	.49
5	4	3.4	.6	.36
				SSE=1.1

Calculating s^2 and s Solution

$$s^2 = \frac{SSE}{n-2} = \frac{1.1}{5-2} = .36667$$

$$s = \sqrt{.36667} = .6055$$

Residual Analysis

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)

Checking for Normality

- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Evaluating the Model

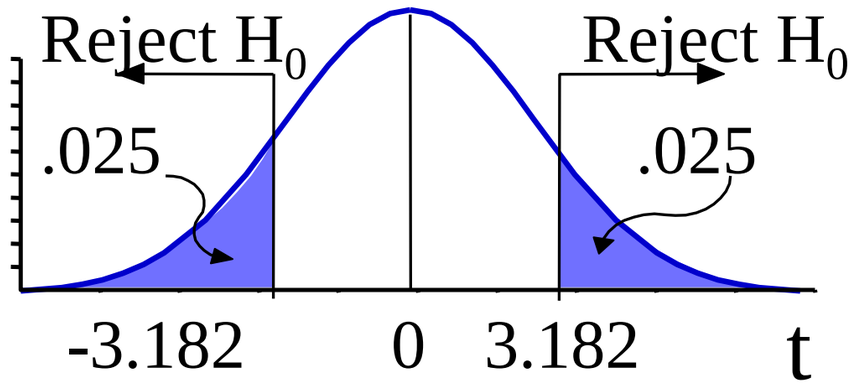
Testing for Significance

Test of Slope Coefficient

- Shows if there is a linear relationship between *x* and *y*
- Involves population slope β_1
- Hypotheses
  - H_0: $\beta_1 = 0$ (No Linear Relationship)
  - H_a: $\beta_1 \neq 0$ (Linear Relationship)
- Theoretical basis is sampling distribution of slope

Test of Slope Coefficient Solution

- **H0: $\beta_1 = 0$**
- **Ha: $\beta_1 \neq 0$**
- **$\alpha = .05$**
- **df = 5 - 2 = 3**
- **Critical Value(s):**



Test of Slope Coefficient Solution

Test Statistic:

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{.70}{.1914} = 3.657$$

Decision:

Reject at α = .05

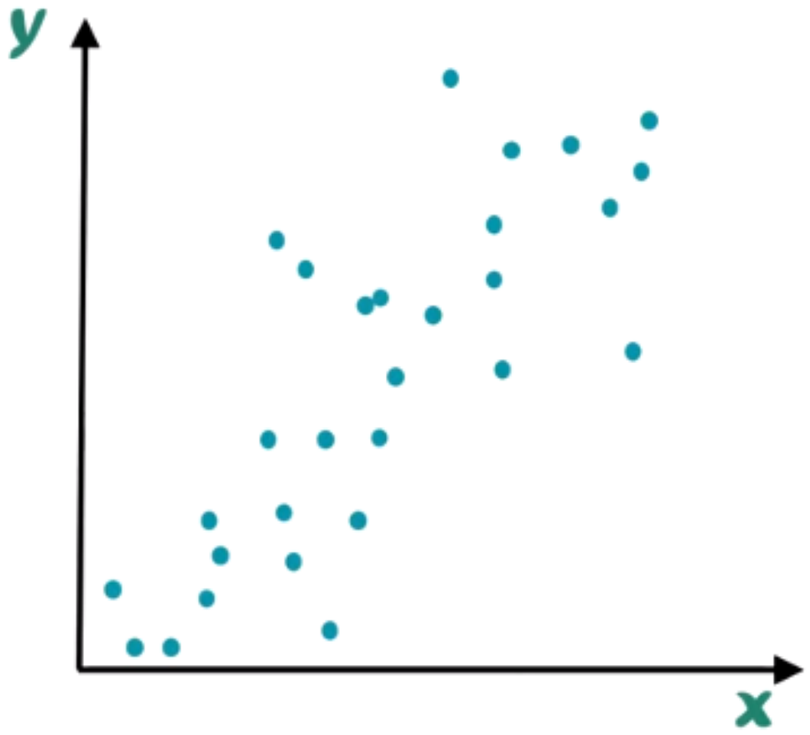
Conclusion:

There is evidence of a relationship

Correlatin and Regression

Korelasyon katsayısı

- Korelasyon, iki değişken arasındaki ilişkinin istatistiksel olarak ölçülmesini ifade eder. İki değişkenin ilişkili olma derecesini gösterir.
- Popülasyon korelasyon katsayısı (rho), değişkenler arasındaki ilişkinin gücünü ölçer.
- Örnek korelasyon katsayısı r, rho'nun bir tahminidir ve örnek gözlemlerdeki doğrusal ilişkinin gücünü ölçmek için kullanılır.



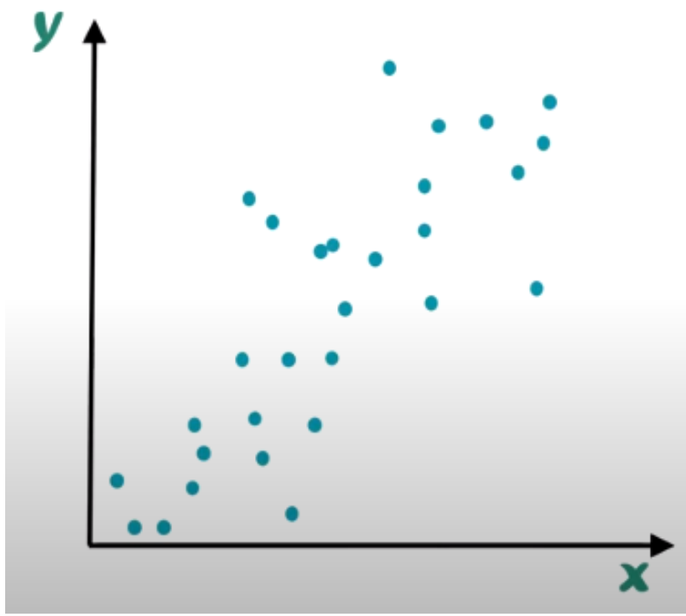
Korelasyon Katsayısı Nedir?

- Regresyon modelleri ile değişkenler arasındaki ilişki matematiksel bir fonksiyon ile belirlenir.
- Korelasyon katsayısı iktisat teorisi ile birlikte ekonometrik modeldeki değişkenlerin birbirlerini açıklamakta yeterli olup olmadığını ortaya koyar.
- Regresyon analizinde bağımsız değişkenlerin katsayılarının işareti değişkenler arasındaki ilişkinin yönünü belirtir. Ancak ilişkinin derecesi sadece bu işaret ile bulunmaz.
- Bu durumda kovaryans kavramı devreye girer.
- Varyans bir serinin aritmetik ortalaması etrafındaki dağılımının ölçüşüyken, kovaryans iki serinin karşılıklı ortalamaları etrafında dağılımlarının ölçüsüdür.

Korelasyon Katsayısı Nedir?

- Kovaryans da varyans gibi tek başına bir anlam ifade etmez.
- Kovaryans varyanstan farklı olarak pozitif veya negatif işaretli olabilir ve kovaryansın işareti, regresyon modelinin bağımsız değişkeninin katsayısının işareti gibi değişkenler arasındaki ilişkinin yönünü gösterir.
- Kovaryans standart bir ölçü olmadığından değişkenler arasındaki ilişkinin derecesi ile ilgili bilgi vermese de, standart ölçü haline getirilerek ilişkinin derecesi belirlenebilir.
- Kovaryans standart ölçü şekline getirilmesi ile elde edilen katsayıya korelasyon katsayısı adı verilmektedir.
- Değişkenler arasında ilişki olmadığında kovaryans sıfır olacağından r =0 olur.
- Değişkenler arasında doğru yönlü ilişki olduğunda kovaryans pozitif işaretli olacağından korelasyon katsayısı da pozitif işaretli; değişkenler arasında ters yönlü ilişki olduğunda kovaryans negatif işaretli olacağından korelasyon katsayısı da negatif işaretli olacaktır.
- Korelasyon katsayısı l'e yaklaştıkça ilişki kuvvetlenirken, 0'a yaklaştıkça ilişki zayıflamaktadır.
- Örneğin, korelasyon katsayısı -0,95 ise değişkenler arasında ters yönlü kuvvetli ilişki, +0,20 ise değişkenler arasında doğru yönlü zayıf ilişki olduğu şeklinde açıklanabilir.

Correlation



Correlation Coefficient is a statistical measure which determines how strongly the pair of variables are correlated or connected. It is denoted by **"r"** and it ranges from **-1** to **+1**.

Correlation coefficient

- Korelasyon katsayısı + 1 ila 0 ila - 1 arasında değişen bir ölçekte ölçülür. İki değişken arasındaki tam korelasyon + 1 veya -1 ile ifade edilir. Diğeri arttıkça bir değişken arttığında korelasyon pozitiftir; Diğeri arttıkça azaldığında negatiftir. Korelasyonun tamamen yokluğu, 0 ile temsil edilir. Şekilde, korelasyonun bazı grafiksel temsillerini verir.
- Bir korelasyon katsayısı, iki değişken arasındaki istatistiksel ilişki anlamına gelen bir tür korelasyonun sayısal bir ölçümüdür.
- Değişkenler, genellikle örnek olarak adlandırılan belirli bir veri gözlem kümesinin iki sütunu veya bilinen bir dağılıma sahip çok değişkenli rasgele değişkenin iki bileşeni olabilir.

Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Coefficient of Correlation

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$SS_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

Correlation

The Pearson product-moment correlation coefficient

The Pearson product-moment correlation coefficient r gives the strength of a linear relationship between the n values of two variables x_i and y_i for $i = 1 \ldots n$, where r is given by the equation:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Correlation

Spearman's rank correlation coefficient

Another method of measuring correlation that does not use the actual values of the data but rather the rankings of the data values is Spearman's rank correlation coefficient where d_i is the difference in ranking between the n values of two variables x_i and y_i for $i = 1 \ldots n$ and where r_s is given by the equation:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Correlation

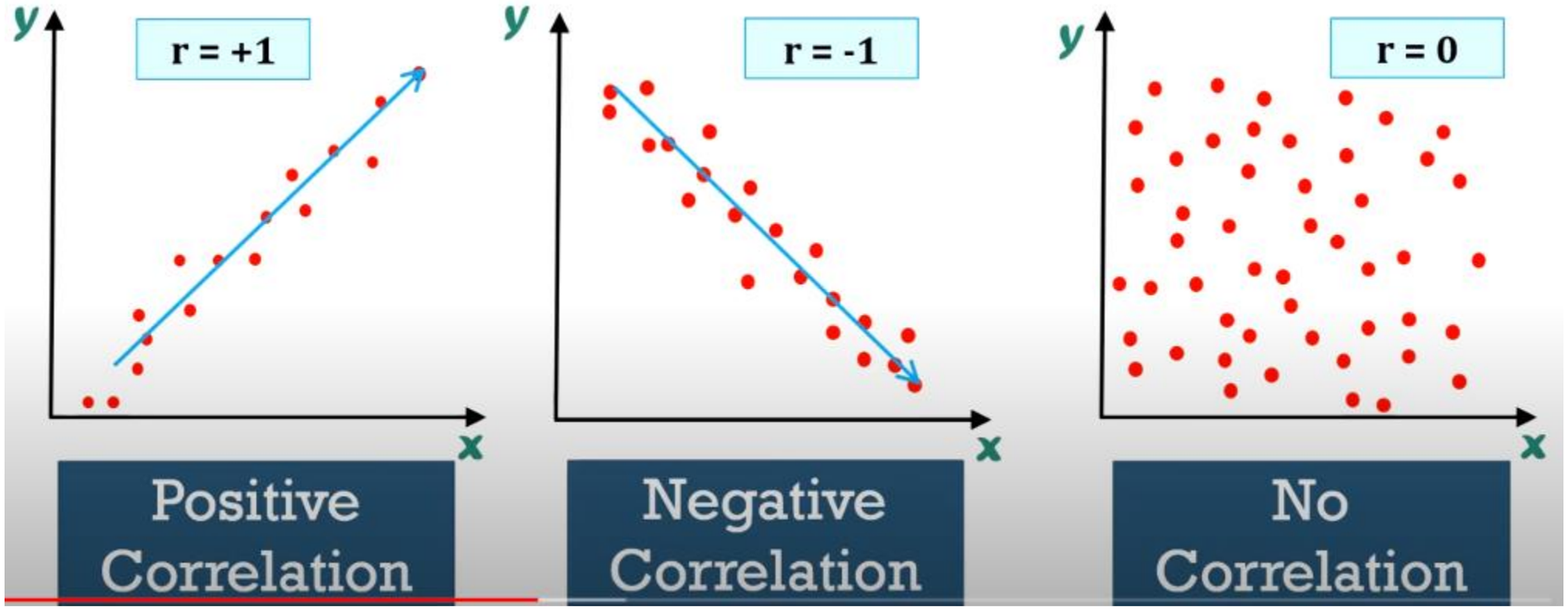
Measures of correlation

- İki değişken arasındaki korelasyonun gücü bir korelasyon katsayısı ile verilir - değeri -1 ile +1 arasında değişen bir sayı. En güçlü pozitif korelasyon, korelasyon katsayısının +1 olduğu zamandır - bu durumda iki değişken mükemmel bir uyum içinde birlikte artar ve azalır.

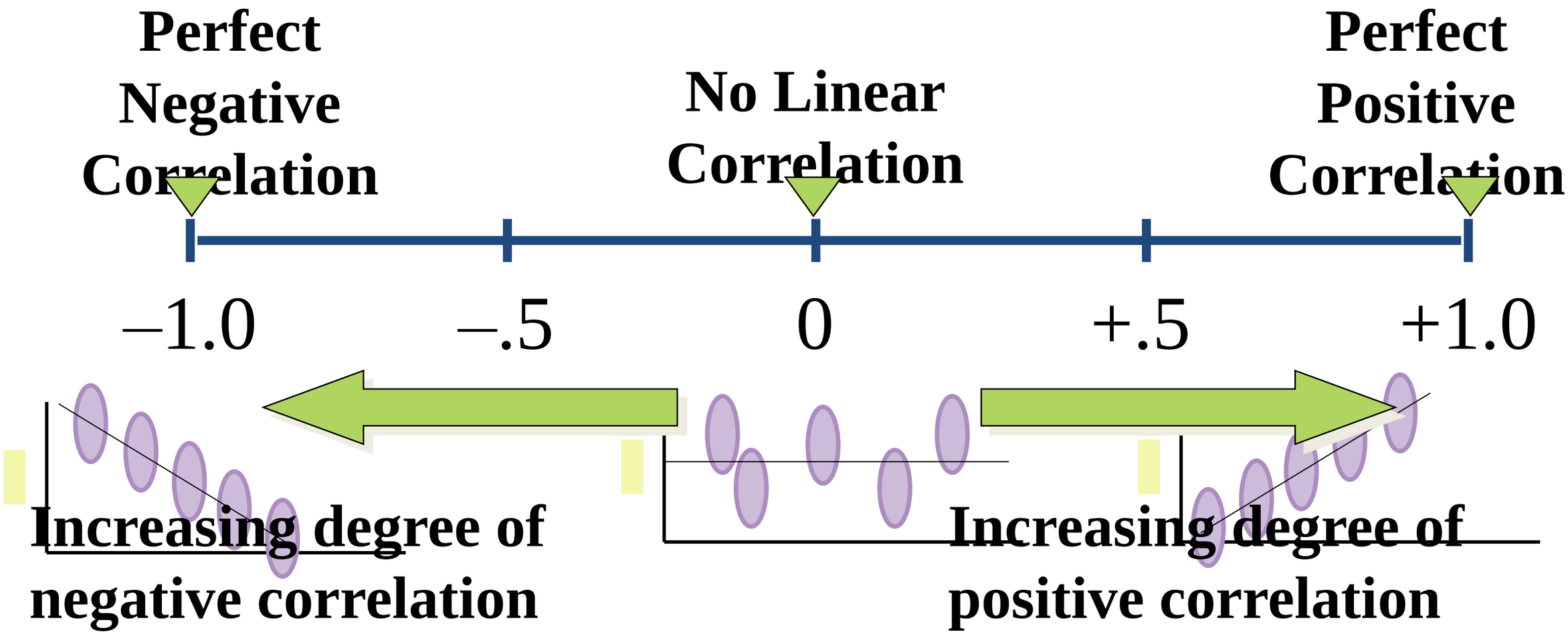
- En güçlü negatif korelasyon, korelasyon katsayısının −1 olduğu zamandır - bu durumda, bir değişken, diğeri azalırken veya arttıkça, yine mükemmel bir uyum içinde artacak veya azalacaktır.

- Korelasyon katsayısı sıfır olduğunda, iki değişken arasında herhangi bir korelasyon yoktur.

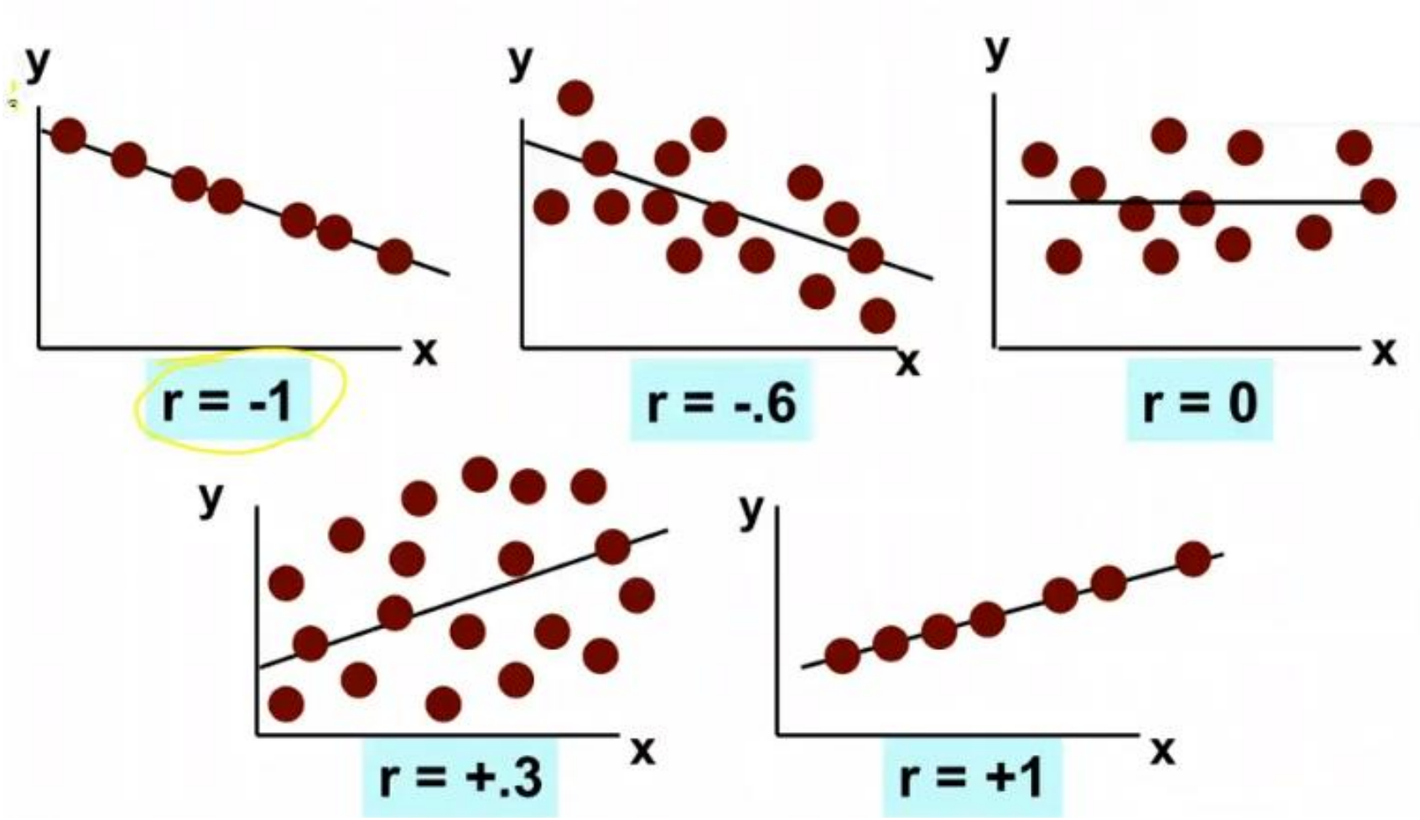
Correlation



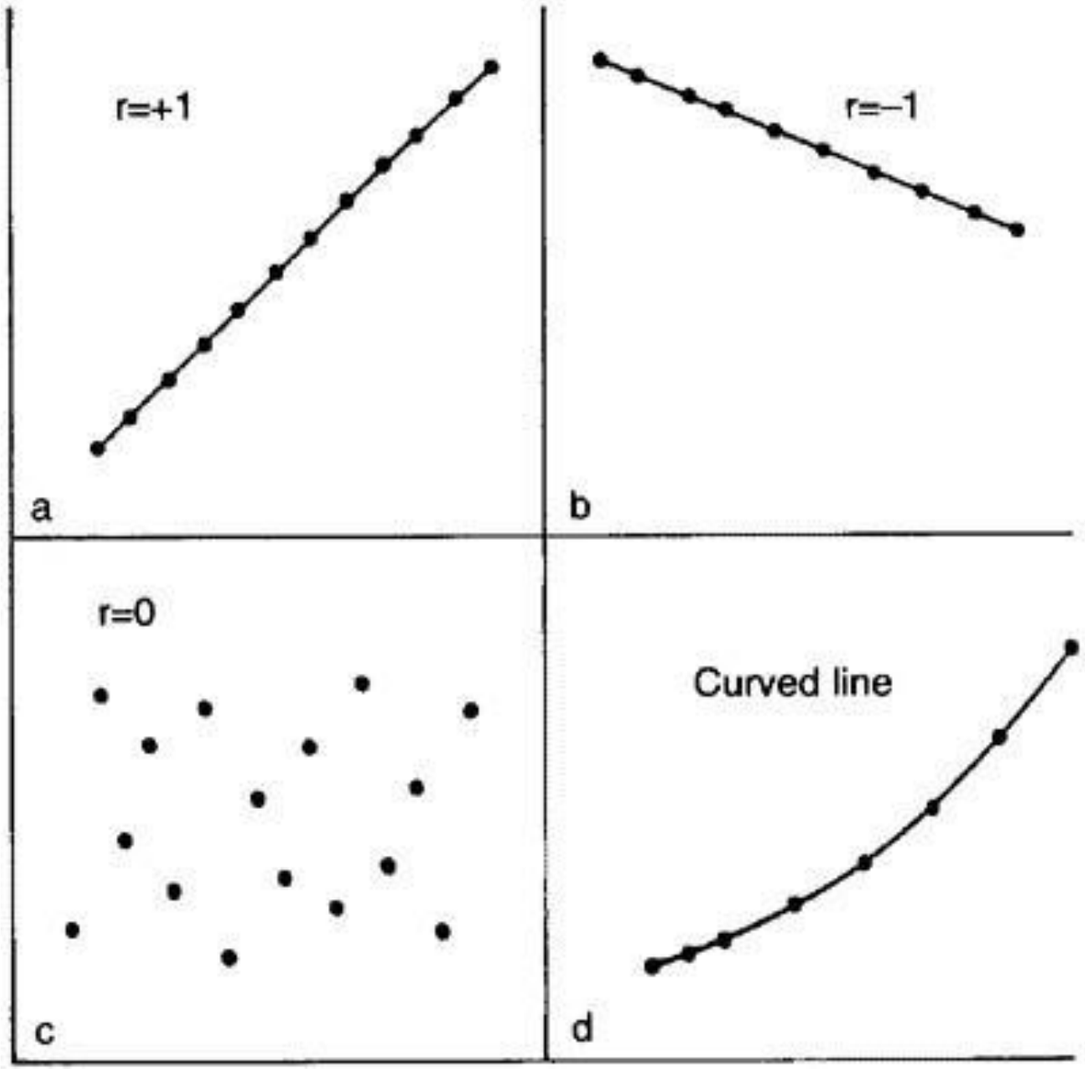
Coefficient of Correlation Values



Correlation Coefficient

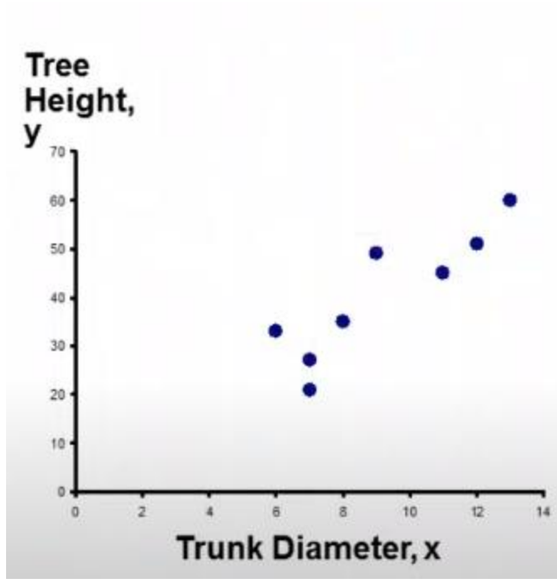


Korelasyon gösterimleri



Calculation Example

Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	Σ=73	Σ=3142	Σ=14111	Σ=713



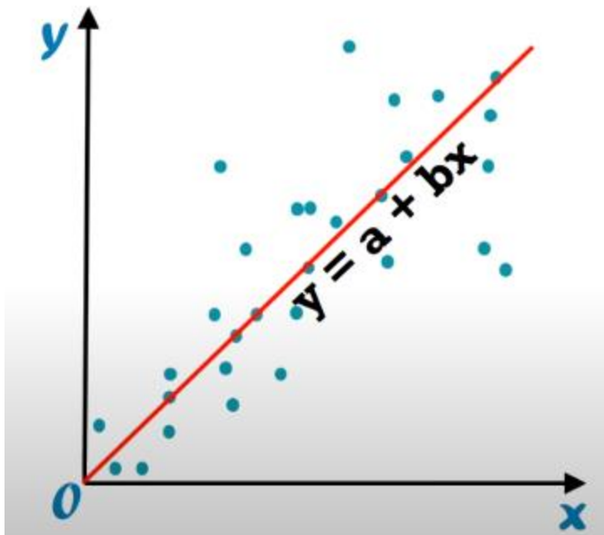
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

$$= 0.886$$

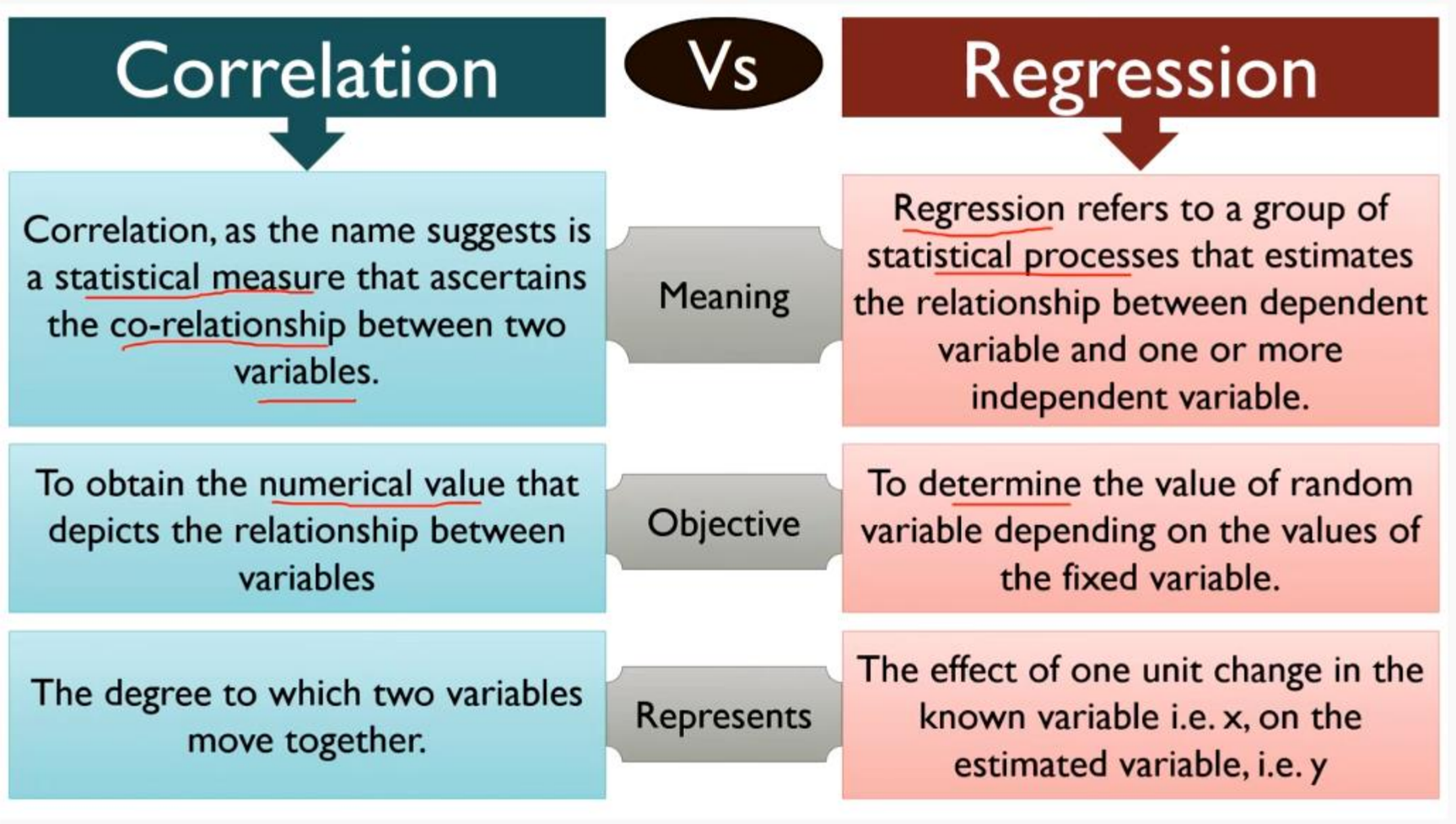
r = 0.886 → relatively strong positive linear association between x and y

Regression

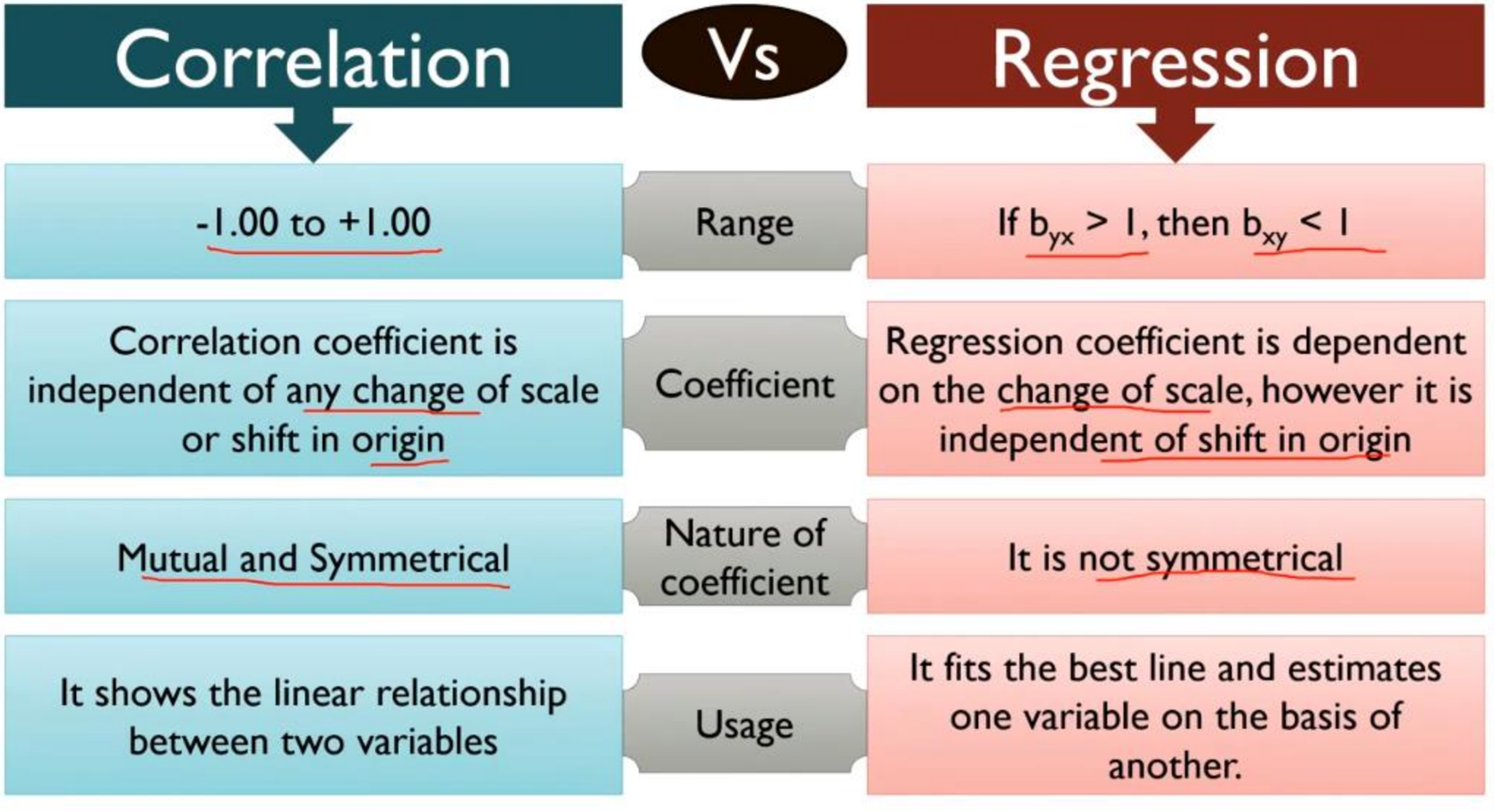


Regression implies the statistical tool used to identify the **nature of relationship** existing between a **dependent variable** and a set of **independent variables**.

Correlation and Regression



Correlation and Regression



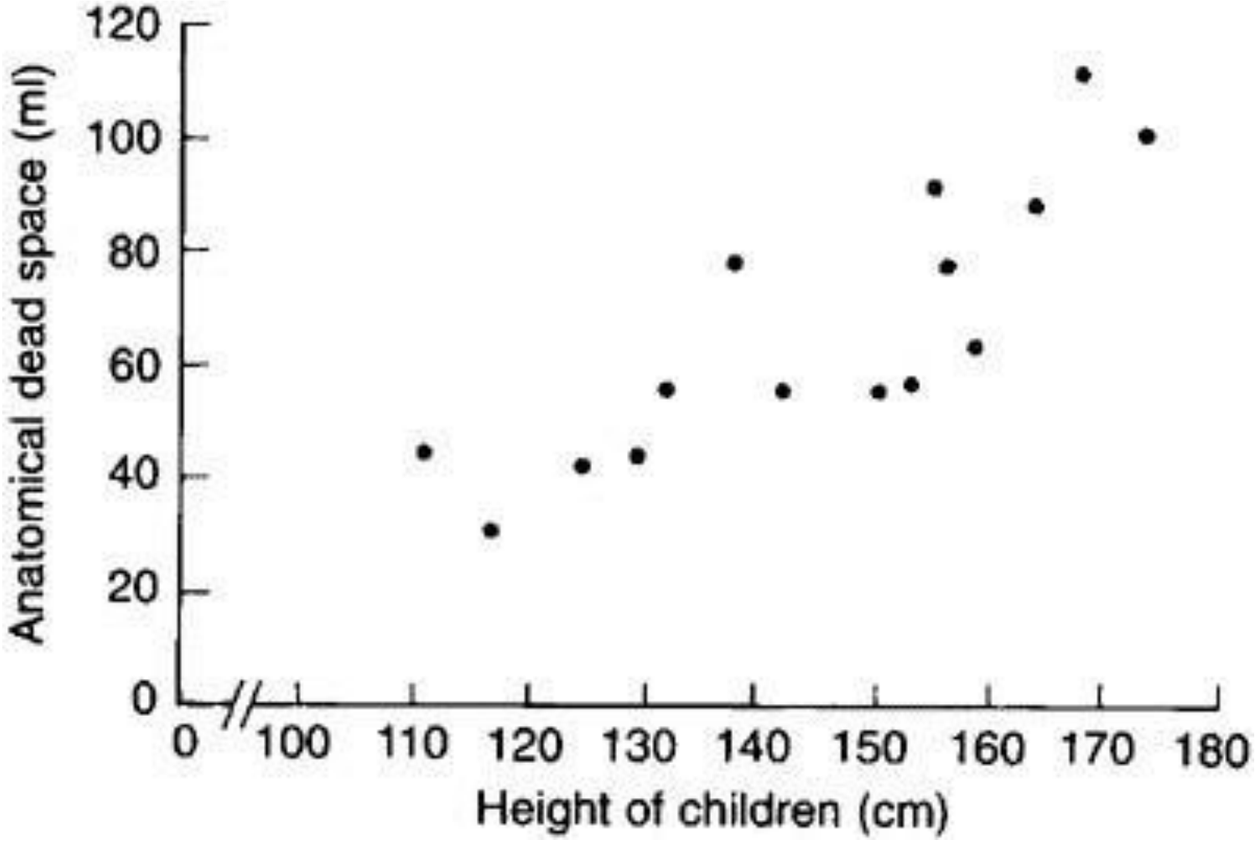
Örnek

15 çocuktaki boyları ve pulmoner anatomik ölü sayılarını göstermek için dağılım diyagramı yapılrken, çocuk doktoru tablodaki (1), (2) ve (3) sütunlarında olduğu gibi rakamları ortaya koydu. İki değişkenden biri açıkça bağımsız olarak tanımlanabilirse, gözlemleri bağımsız değişkenin seri sırasına göre düzenlemek yararlıdır. Bağımlı değişken için karşılık gelen rakamlar daha sonra bağımsız değişken için artan serilerle ilişkili olarak incelenebilir. Bu şekilde, dağılım diyagramında göründüğü gibi aynı resmi elde ederiz, ancak sayısal formda.

Table 11.1 Correlation between height and pulmonary anatomical dead space in 15 children

Child number	Height (cm)	Dead space (ml), y
1	110	44
2	116	31
3	124	43
4	129	45
5	131	56
6	138	79
7	142	57
8	150	56
9	153	58
10	155	92
11	156	78
12	159	64
13	164	88
14	168	112
15	174	101
Total	**2169**	**1004**
Mean	**144.6**	**66.933**

Örnek



$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2 (y - \bar{y})^2]}}$$

Matlab:
R = corrcoef(A,B) returns coefficients between two random variables A and B.

Coefficient of Determination

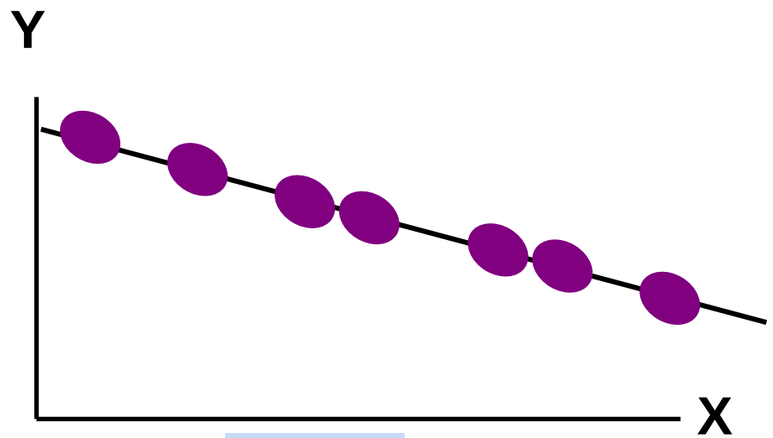
Proportion of variation ‘explained’ by relationship between *x* and *y*

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SS_{yy} - SSE}{SS_{yy}}$$

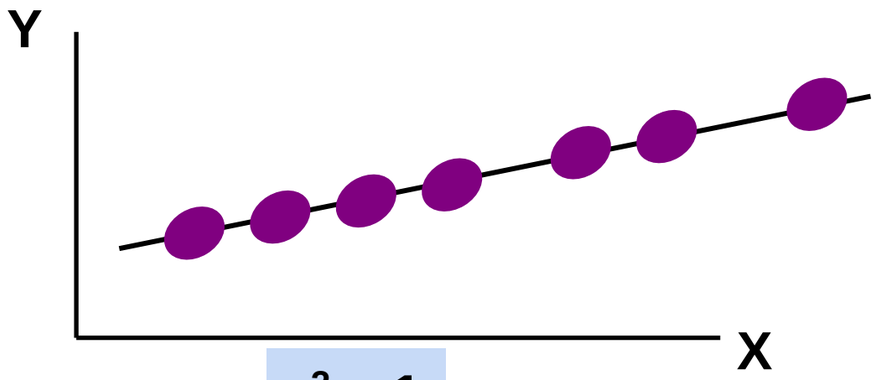
$0 \le r^2 \le 1$

$r^2 = (\text{coefficient of correlation})^2$

Examples of Approximate r^2 Values



$r^2 = 1$



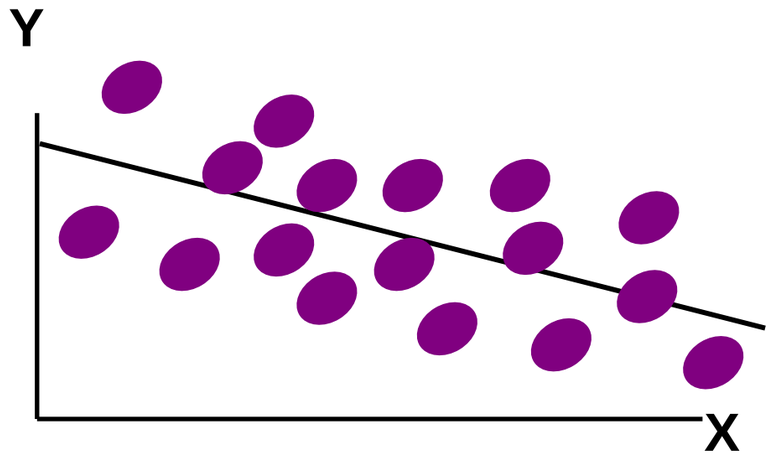
$r^2 = 1$

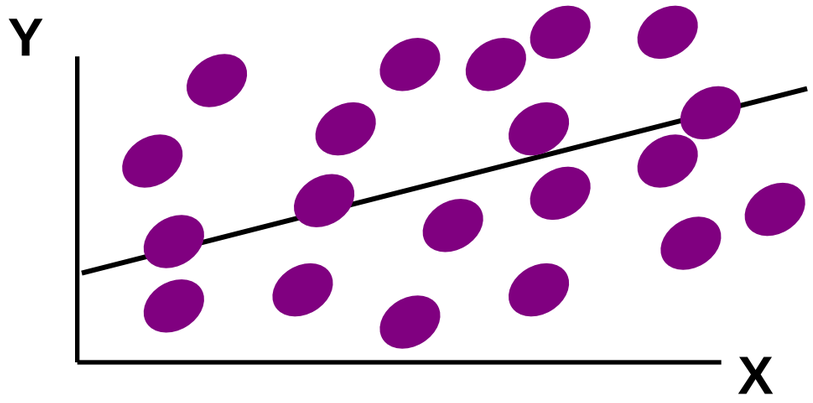
$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Examples of Approximate r^2 Values



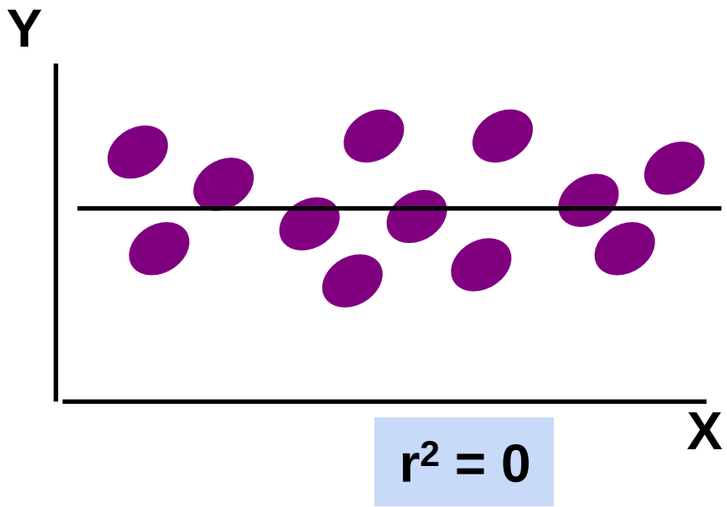


$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

Examples of Approximate r^2 Values



$r^2 = 0$

$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Coefficient of Determination Example

r = .904.

Ad $	Sales (Units)
1	1
2	1
3	2
4	2
5	4

Calculate and interpret the **coefficient of determination.**

Coefficient of Determination Solution

r^2 = (coefficient of correlation)^2

$r^2 = (.904)^2$

$r^2 = .817$

Interpretation: About 81.7% of the sample variation in Sales (y) can be explained by using Ad \$ (x) to predict Sales (y) in the linear model.

Using the Model for Prediction & Estimation

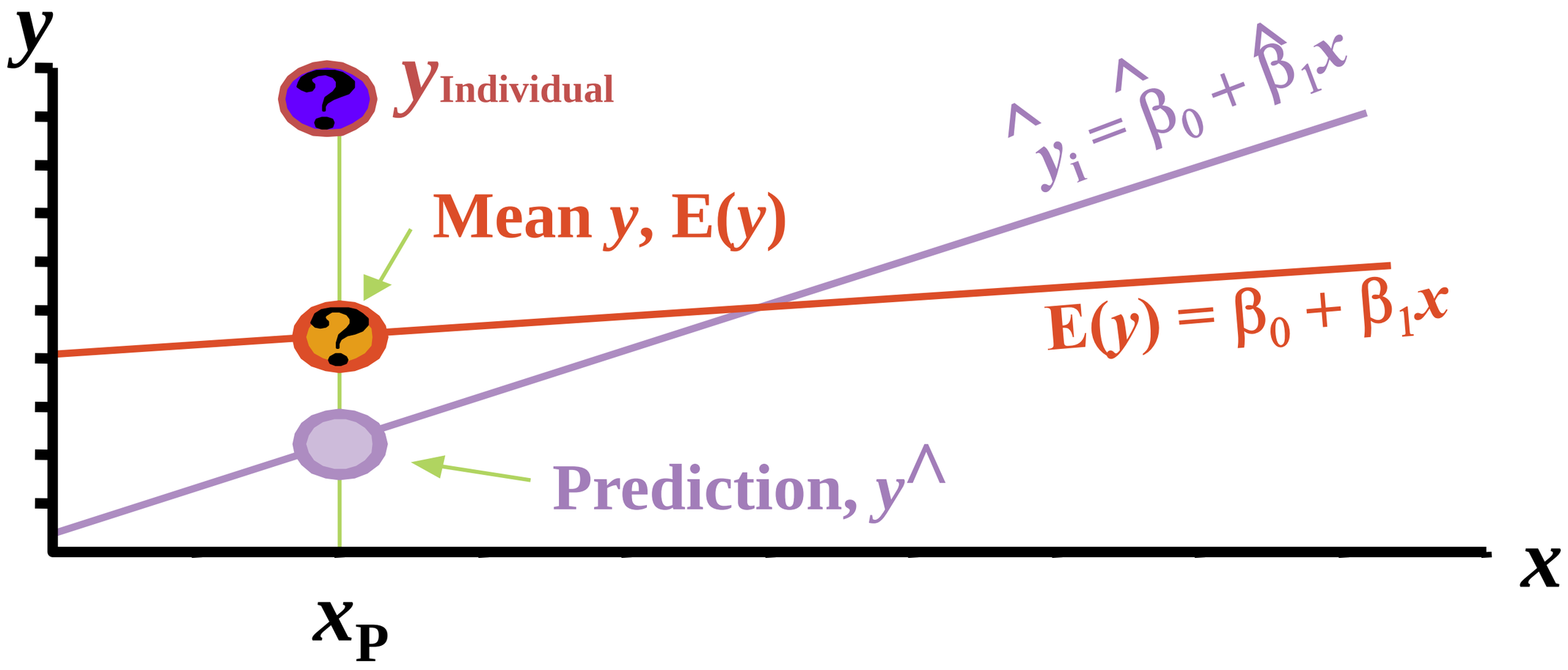
Regression Modeling Steps

1. Hypothesize deterministic component
2. Estimate unknown model parameters
3. Specify probability distribution of random error term
   - Estimate standard deviation of error
4. Evaluate model
5. **Use model for prediction and estimation**

Prediction With Regression Models

- Types of predictions
  - Point estimates
  - Interval estimates
- What is predicted
  - Population mean response $E(y)$ for given x
    - Point on population regression line
  - Individual response (y_i) for given x

What Is Predicted



Confidence Interval Estimate for Mean Value of y at $x = x_p$

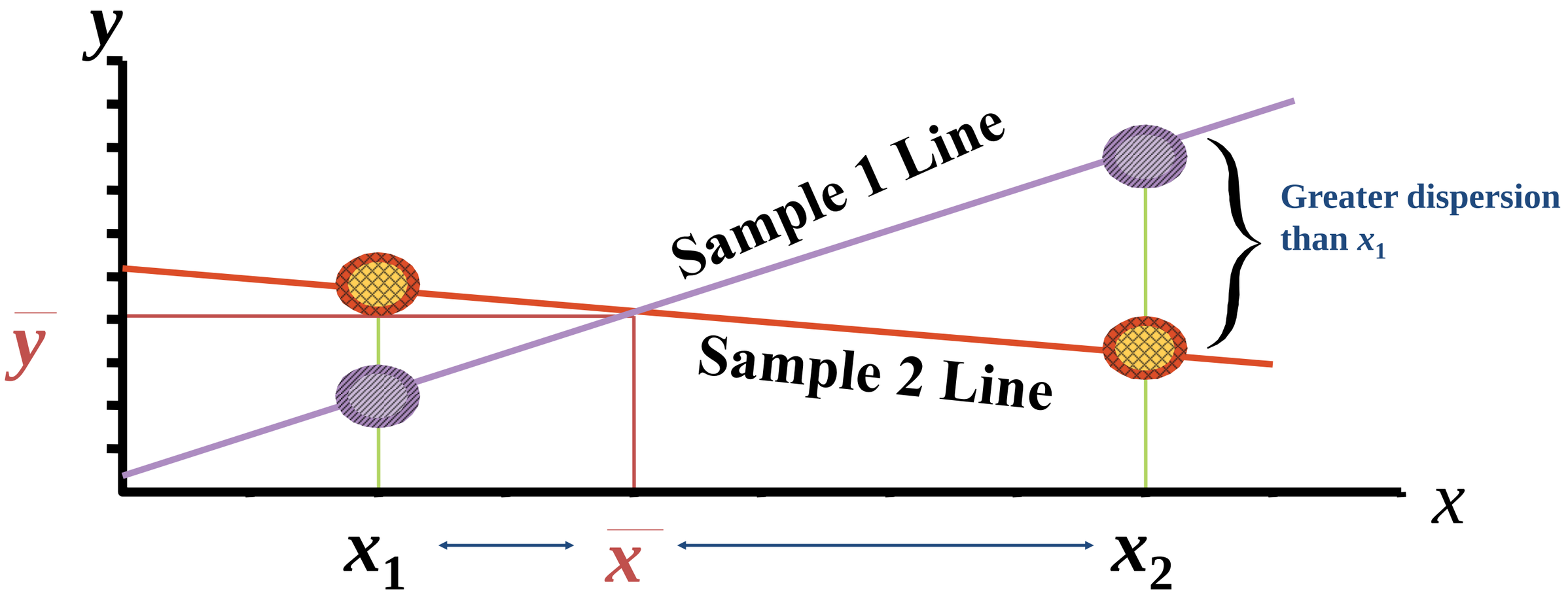
$$\hat{y} \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

$$\text{df} = n - 2$$

Factors Affecting Interval Width

1. Level of confidence $(1 - \alpha)$
   - Width increases as confidence increases
2. Data dispersion (s)
   - Width increases as variation increases
3. Sample size
   - Width decreases as sample size increases
4. Distance of x_p from mean $\bar{x}$
   - Width increases as distance increases

Why Distance from Mean?



Confidence Interval Estimate Example

You find $\hat{\beta}_0 = -.1$, $\hat{\beta}_1 = .7$ and $s = .6055$.

Ad $	Sales (Units)
1	1
2	1
3	2
4	2
5	4

Find a **95%** confidence interval for the **mean** sales when advertising is **$4**.

Solution Table

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	**10**	**55**	**26**	**37**

Confidence Interval Estimate Solution

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

x to be predicted

$$\hat{y} = -.1 + (.7)(4) = 2.7$$

$$2.7 \pm (3.182)(.6055)\sqrt{\frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$1.645 \leq E(Y) \leq 3.755$$

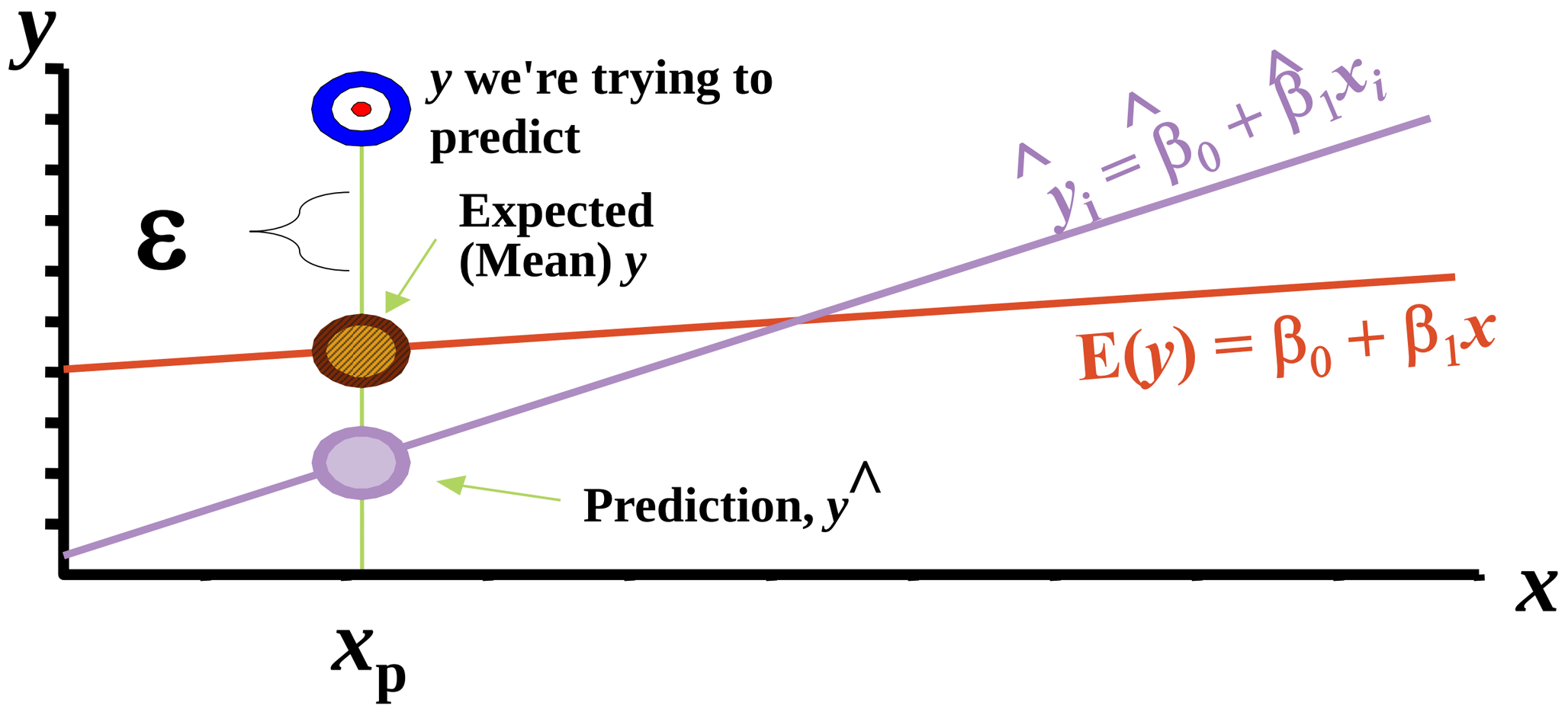
Prediction Interval of Individual Value of y at $x = x_p$

$$\hat{y} \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

Note!

$$\text{df} = n - 2$$

Why the Extra ‘S’?



Prediction Interval Solution

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

x to be predicted

$$\hat{y} = -.1 + (.7)(4) = 2.7$$

$$2.7 \pm (3.182)(.6055)\sqrt{1 + \frac{1}{5} + \frac{(4-3)^2}{10}}$$

$$.503 \leq y_4 \leq 4.897$$

Confidence Intervals v. Prediction Intervals