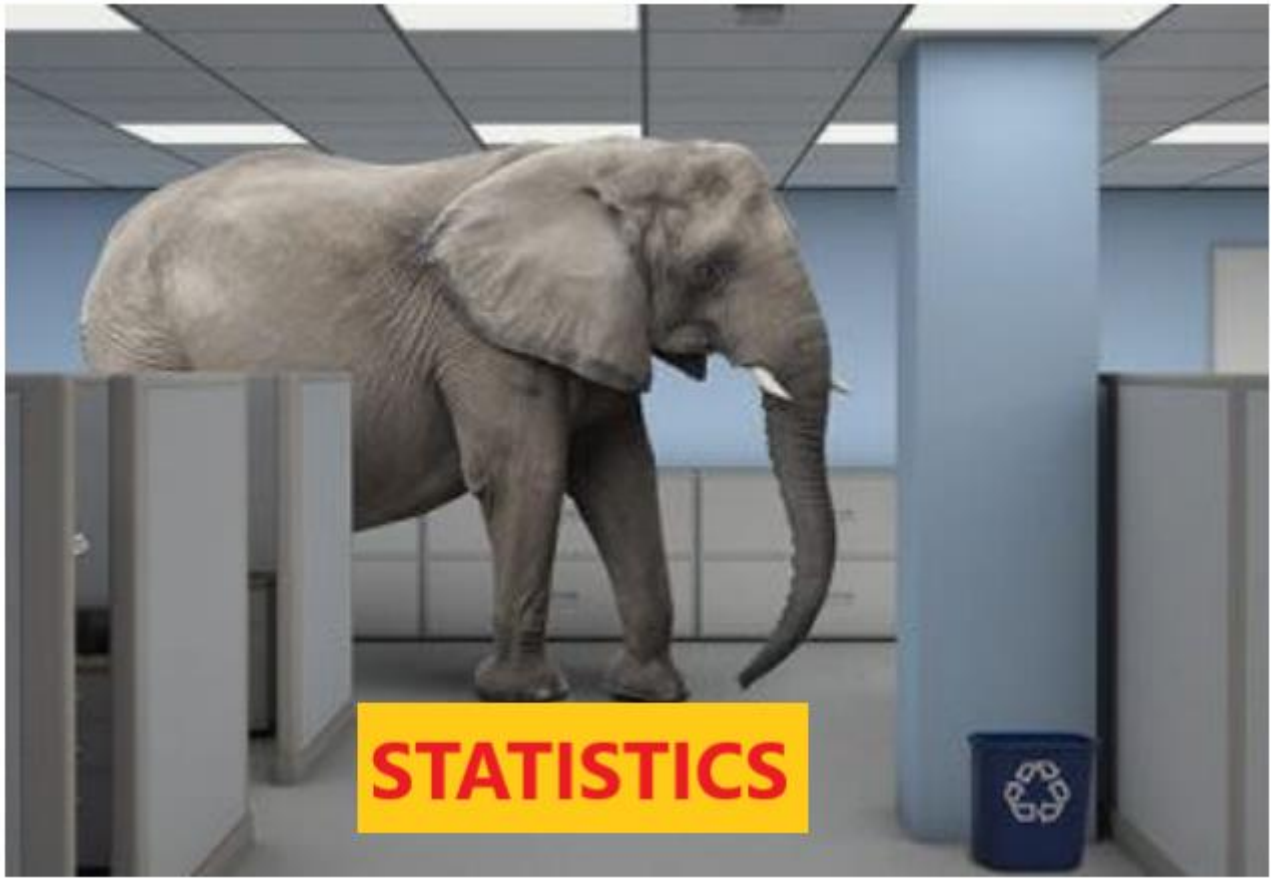
Predictive Analytics

Regression Modeling

Dr. Tanujit Chakraborty
@ Sorbonne
tanujitisi@gmail.com

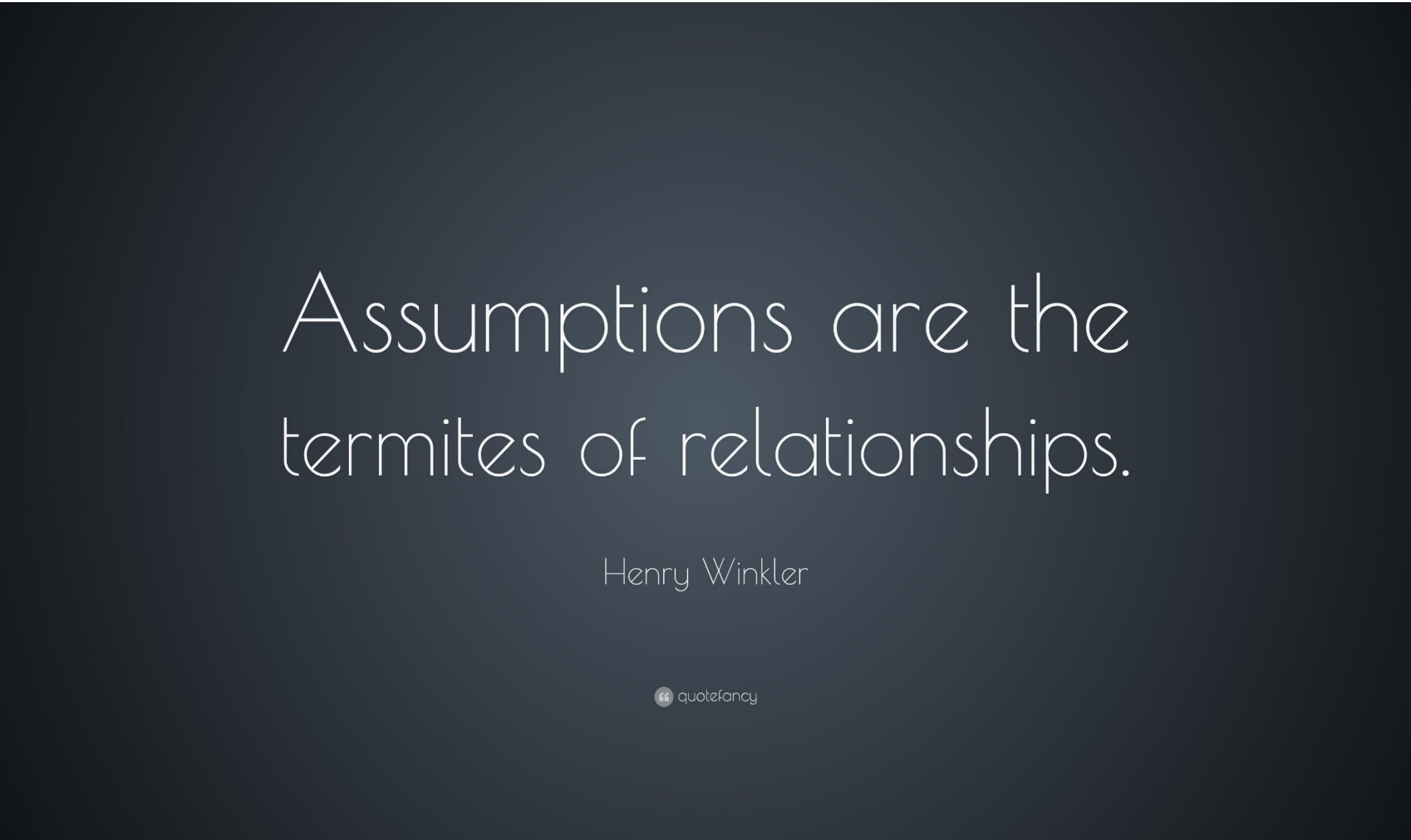
Why Statistics?



"Statistics is the universal tool of inductive inference, research in natural and social sciences, and technological applications. Statistics, therefore, must always have purpose, either in the pursuit of knowledge or in promotion of human welfare."

– Prof. Prasanta Chandra Mahalanobis, *Father of Statistics in India*

Quote of the day..



This presentation includes…

- Introduction
- Correlation Analyses
- Regression Analysis
  - Linear regression
  - Non-linear regression
  - Auto-regression
  - Demos

Data for Relationship Analysis

Univariate population: The population consisting of only one variable.

Example:

Temperatur	20	30	21	18	23	45	52

Here, statistical measures suffice to find a relationship.

Bivariate population: Here, the data happen to be with two variables.

Example:

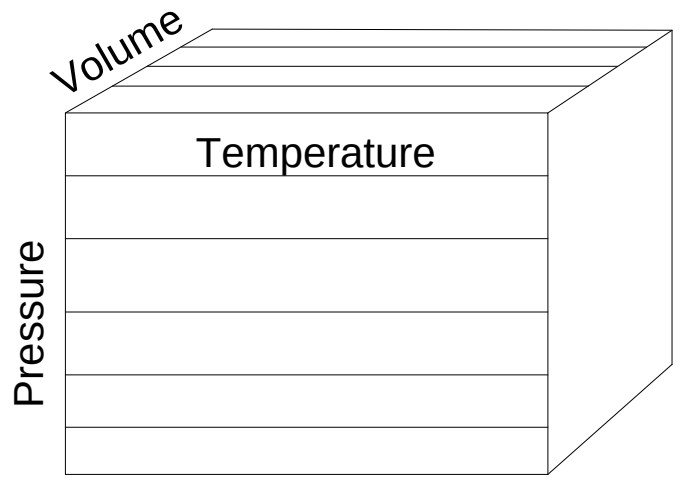
Pressure	1	1.1	... ...	0.8
Temperatı	35	41		29

Data for Relationship Analysis

Multivariate population: If the data happen to be one more than two variable.

.

Example:



If we add another variable say viscosity in addition to Pressure, Volume or Temperature?

Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population) ?

- If yes, of what degree?

Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

- If yes, of what degree and in which direction?

Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population) ?

Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

Solution

?

Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist relation between two variables (in case of bivariate population) ?

Q2: Is there any relationship between one variable in one side and two or more variables on the other side (in case of multivariate population)?

To find solutions to the above questions, two approaches are known.

- **Correlation Analysis**
- **Regression Analysis**

Correlation Analysis

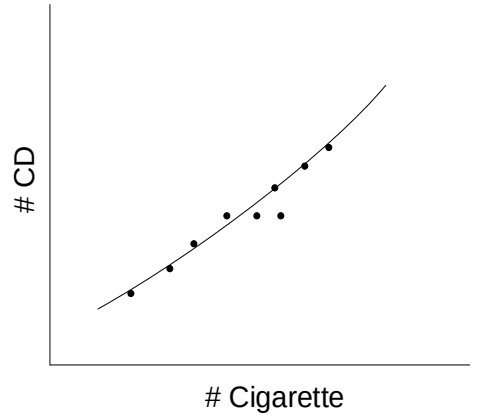
Correlation Analysis

In statistics, the word correlation is used to denote some form of association between two variables.

Example: Weight is correlated with height

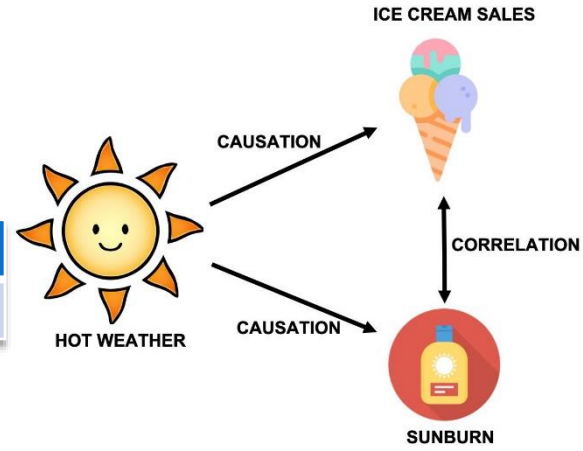
Correlation Analysis

Do you find any correlation between X and Y as shown in the table?



Look at the table given below

No. of CD's sold in shop X	25	30	35	42	48	52	56
No. of cigarette sold in Y	5	7	9	10	11	11	12

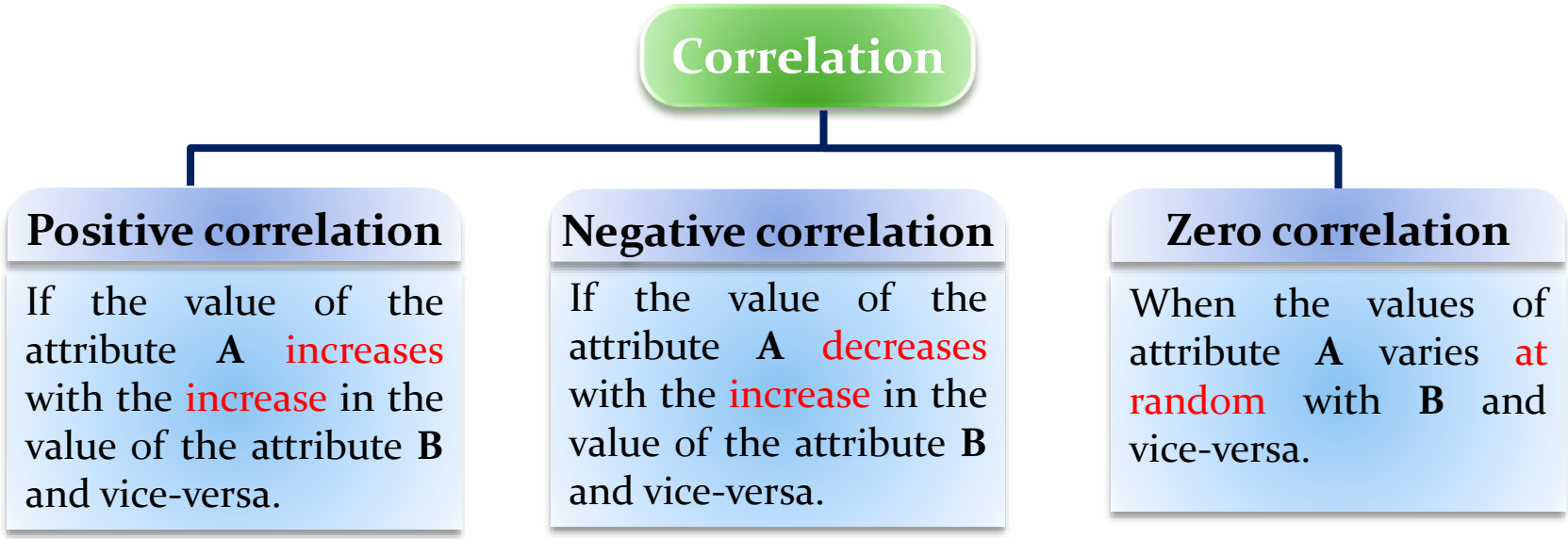


Note

- In data analytics, a correlation analysis makes sense only when a relationship makes sense.
- Correlation does NOT imply causation.

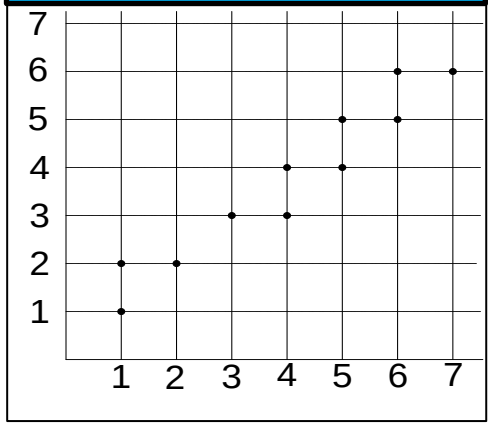
Correlation Analysis

A	a_1	a_2	a_3	a_4	a_5	a_6	… …
B	b_1	b_2	b_3	b_4	b_5	b_6	… …

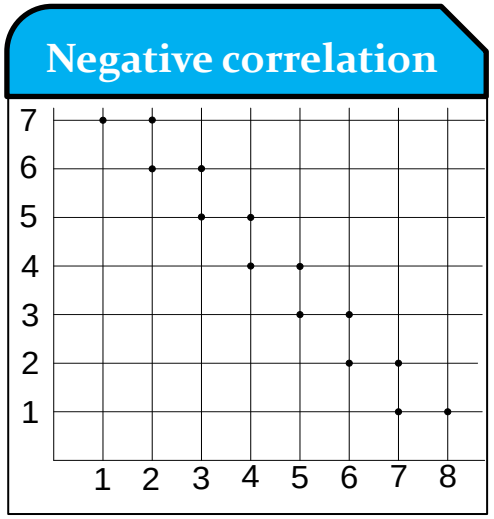


Correlation Analysis

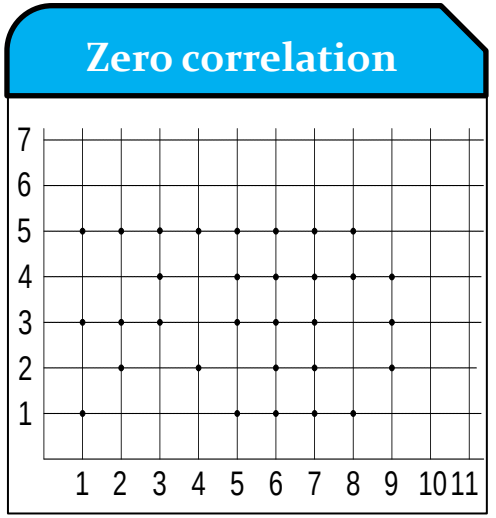
Positive correlation



Negative correlation



Zero correlation

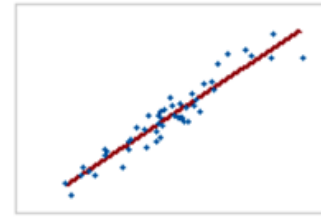


Form of Correlation

Concerning the form of a correlation, it could be linear, non-linear, or monotonic.

Linear Correlation

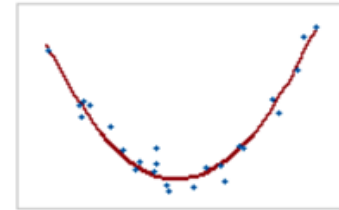
A correlation is linear when two variables change at constant rate.



Linear Correlation

Non-linear Correlation

In this case, the relationship between the variables graph as a curved pattern parabola, hyperbola ... etc).



Non-linear Correlation

Form of Correlation

Concerning the form of a correlation , it could be linear, non-linear, or **monotonic.**

Monotonicity of a function

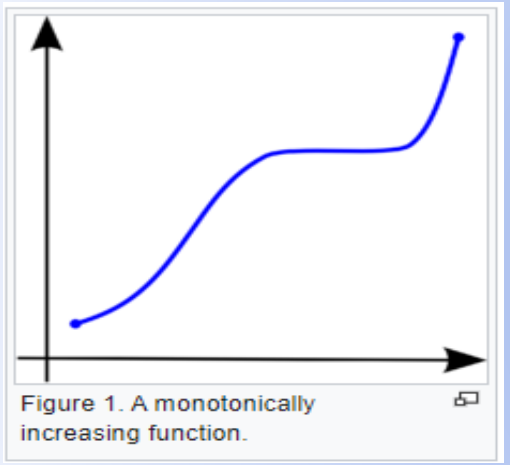


Figure 1. A monotonically increasing function.

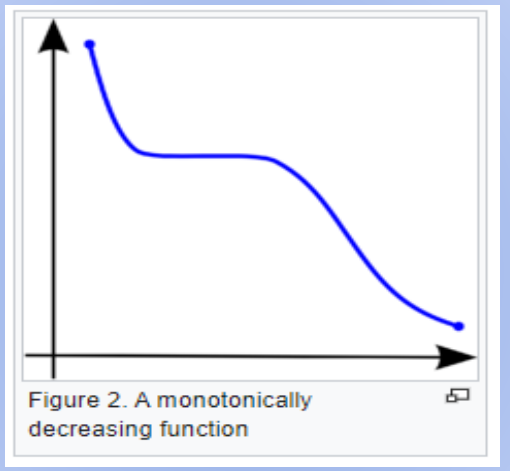


Figure 2. A monotonically decreasing function

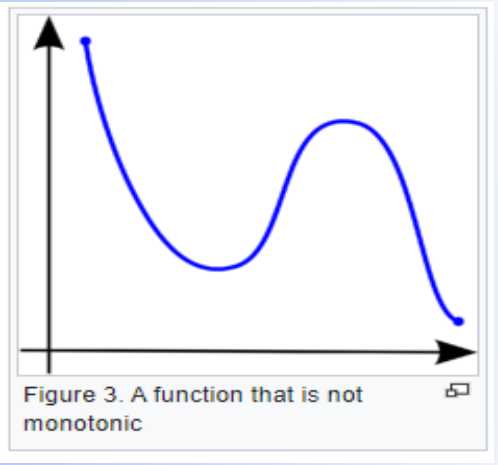


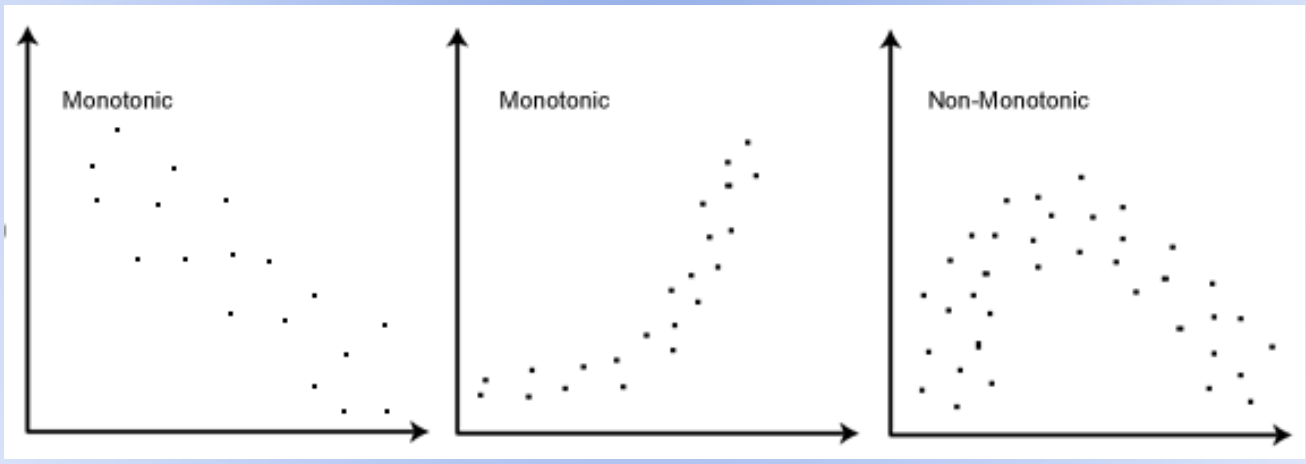
Figure 3. A function that is not monotonic

Form of Correlation

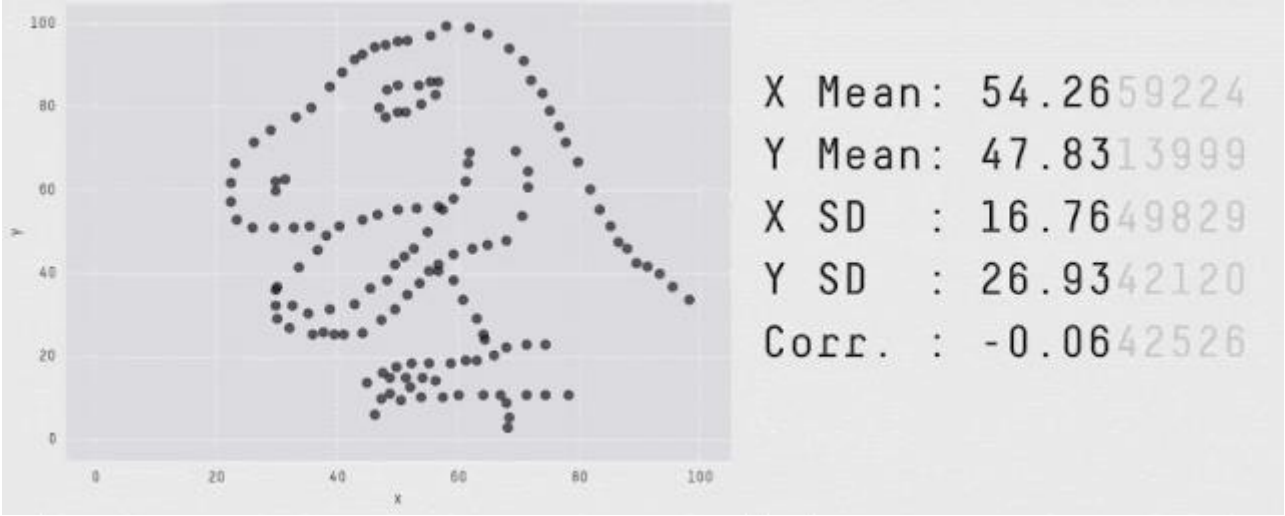
Concerning the form of a correlation , it could be linear, non-linear, or **monotonic** :

Monotonic and non-monotonic relations

Monotonic correlation: In a monotonic relationship, the variables tend to move in the same relative direction or opposite direction, but not necessarily at a constant rate.



Scatter Plot tells stories

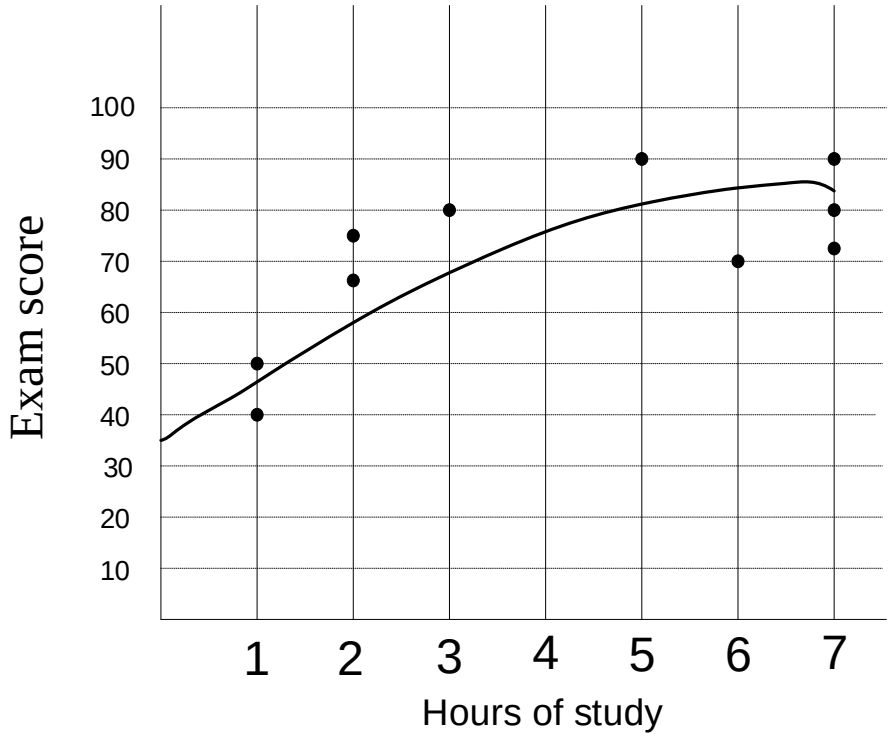


- In 1973, a famous statistician, Francis Anscombe, demonstrated how important it is to visualize the data. The concept got extended later to create Datasaurus Dozen.
- It is a collection of 12 scatterplots with the same means, standard deviations, and correlation coefficient for X and Y (up to 2 decimal places).
- However, the shape of the data is very different from each other. Therefore, the scatterplots tell very different stories about the behavior and interrelationships of X and Y.
- Data available at https://cran.r-project.org/web/packages/datasauRus/vignettes/Datasaurus.html

Correlation Analysis

We need to measure the degree of correlation between two attributes.

Hours Study	Exam Score
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100

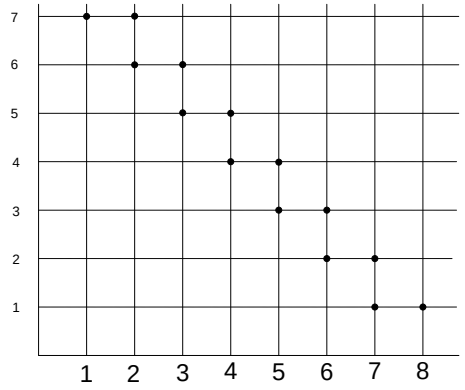


Correlation Coefficient

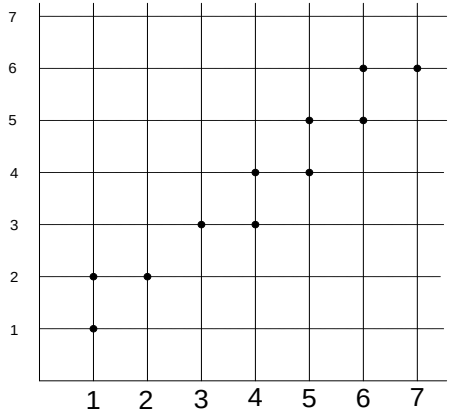
- **Correlation coefficient is used to measure the degree of association.**
- It is usually denoted by r.
- The value of r lies between +1 and -1.
- Positive values of r indicates positive correlation between two variables, whereas, negative values of r indicate negative correlation.
- $r = +1$ implies **perfect positive correlation**, and otherwise.
- The value of r nearer to +1 or -1 indicates **high degree of correlation** between the two variables.
- $r = 0$ implies, there is **no correlation**

Correlation Coefficient

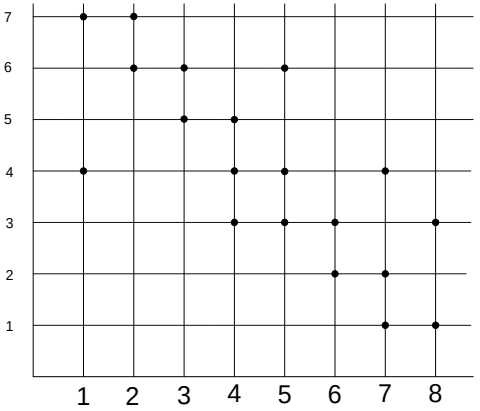
High Negative Correlation



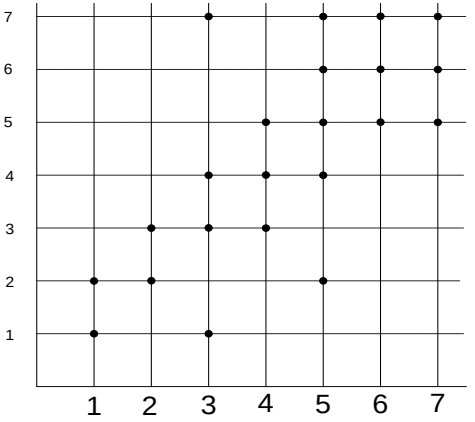
High Positive Correlation



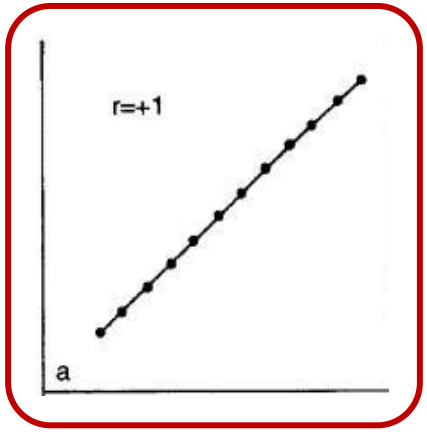
Low Negative Correlation

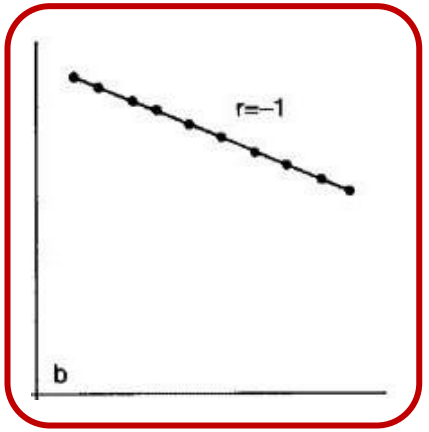


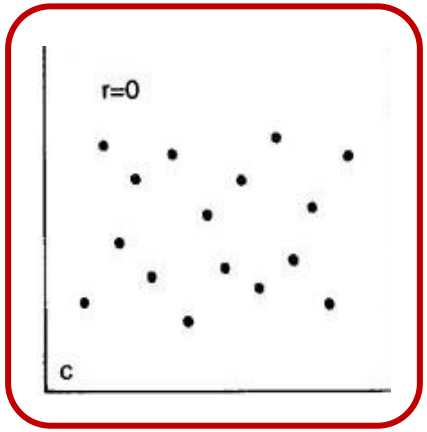
Low Positive Correlation

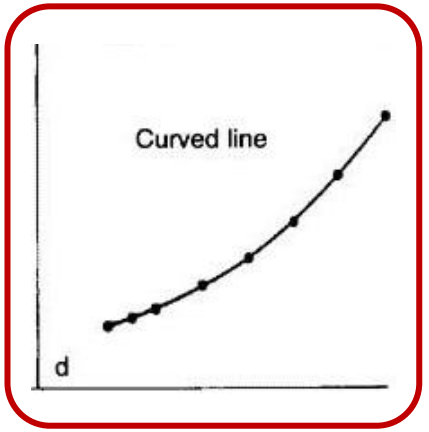


Correlation Coefficient

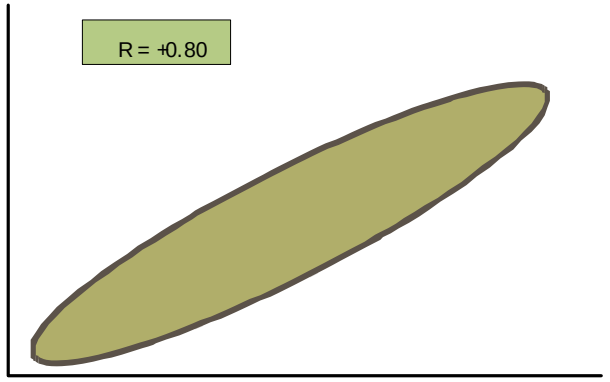


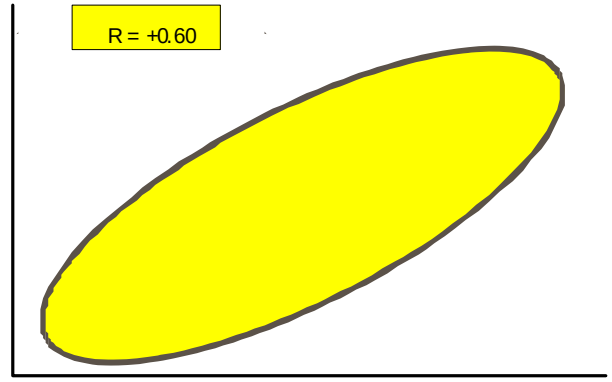


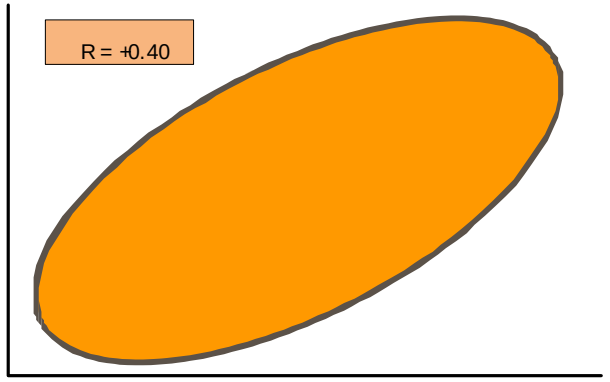


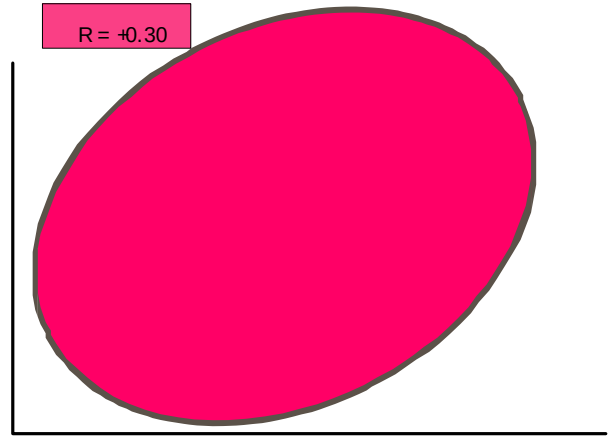


Correlation Coefficient

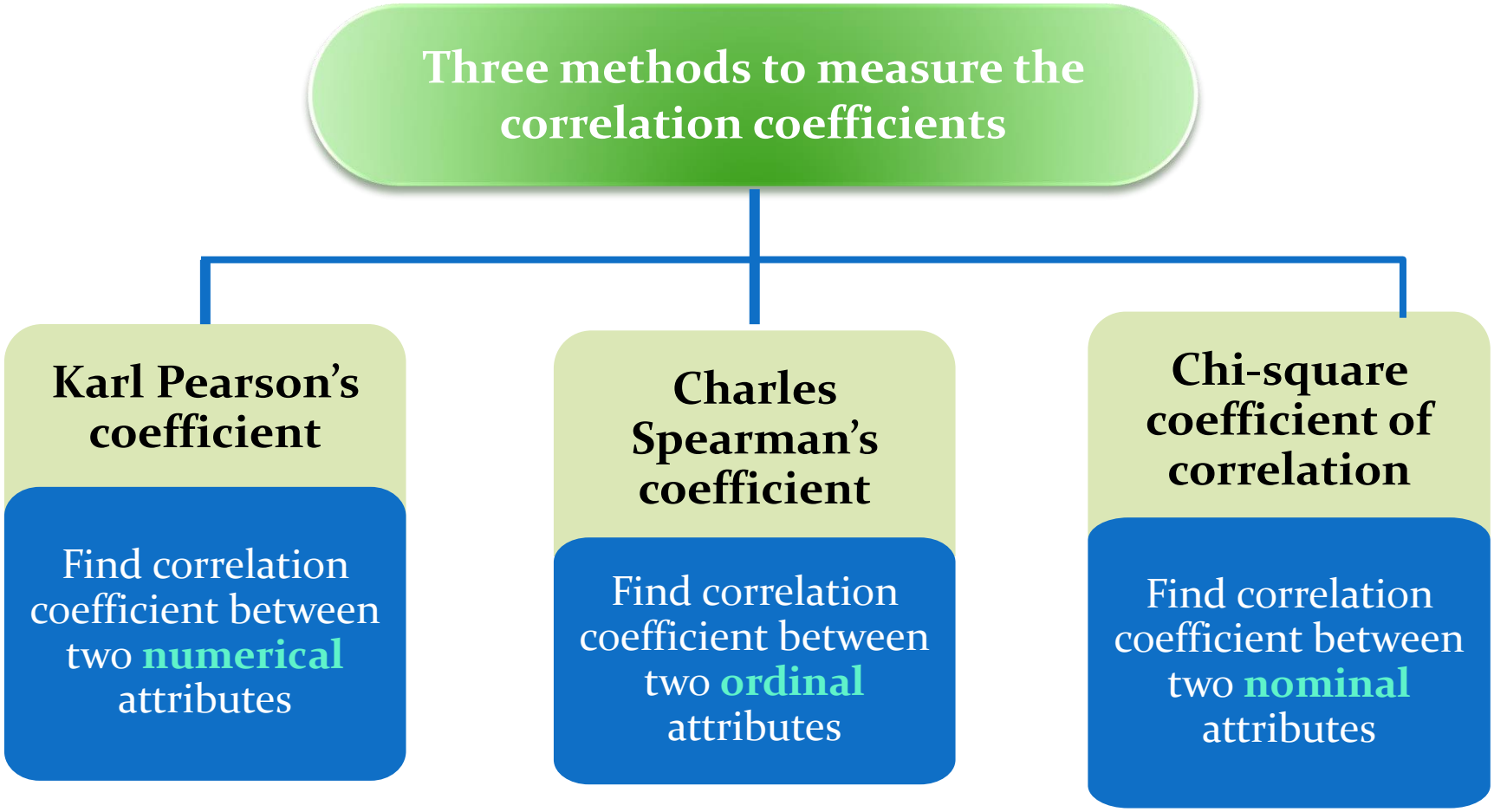








Measuring Correlation Coefficients



Pearson's Correlation Analysis

Karl Pearson's Correlation Analysis

This is also called Pearson's Product Moment Correlation

Definition : Karl Pearson's correlation coefficient

Let us consider two attributes are X and Y.

The Karl Pearson's coefficient of correlation is denoted by r^* and is defined as

$$r^* = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

where

$X_i = \text{i} - \text{th value of } X - \text{variable}, \quad \bar{X} = \text{mean of } X$

$Y_i = \text{i} - \text{th value of } Y - \text{variable}, \quad \bar{Y} = \text{mean of } Y$

$n = \text{number of pairs of observation of } X \text{ and } Y$

$\text{cov(X, Y)} = \text{covariance of X } and \text{ } Y, \sigma_X = \text{SD of } X, \quad \sigma_Y = \text{SD of } Y$

Karl Pearson's Coefficient of Correlation

Example : Correlation of Gestational Age and Birth Weight

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Karl Pearson's coefficient of Correlation

Example : Correlation of Gestational Age and Birth Weight

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

We wish to estimate the association between gestational age and infant birth weight.

Birth weight → Dependent variable

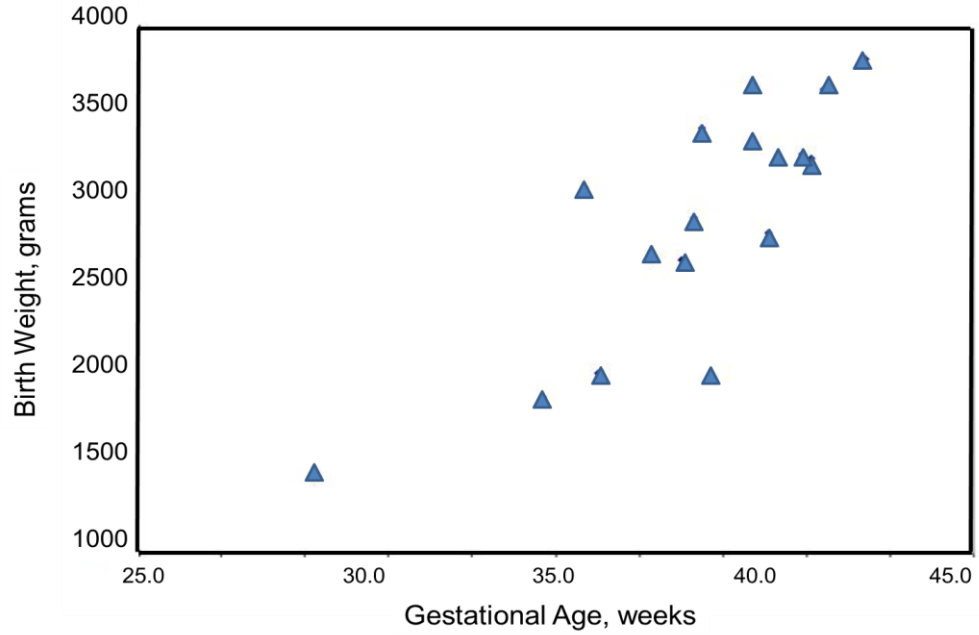
Gestational age → Independent variable

Thus

$$Y = birth\ weight \text{ and}$$

$$X = gestational\ age$$

The data are displayed in the scatter diagram.



Karl Pearson's coefficient of Correlation

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

For the given data

- $\bar{X} = \frac{\sum X}{n} = \frac{652.1}{17} = \mathbf{38.4}$
- $\bar{Y} = \frac{\sum Y}{n} = \frac{49334}{17} = \mathbf{2902}$
- $S_x^2 = \frac{\sum(X-\bar{X})^2}{n-1} = \frac{159.45}{16} = \mathbf{9.97}$
- $S_y^2 = \frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{7767660}{16} = \mathbf{485578.8}$
- $r^* = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\sqrt{\sum(Y_i-\bar{Y})^2}} = \mathbf{0.82}$

Conclusion: The sample's correlation coefficient indicates a strong positive correlation between Gestational Age and Birth Weight.

Significance Test

Definition : Karl Pearson's correlation coefficient

- Say we have an n sized sample data with two variables x and y.
- The sample correlation coefficient (r) between x and y is known
- The population correlation coefficient ρ between x and y is unknown
- **Goal:** We want to make an inference about the value of ρ based on r

Null hypothesis H_0: $\rho = r$

Alternative hypothesis H_1: $\rho \neq r$

Karl Pearson's Coefficient of Correlation

Significance Test

- To test whether the association is merely apparent, and might have arisen by chance use the *t* test in the following calculation

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- Here, the number of pair of observation is 17. Hence,

$$t = 0.82\sqrt{\frac{17-2}{1-0.82^2}} = 1.44$$

- Consulting the t-test table, at degrees of freedom 15 and for $\alpha = 0.05$, we find that t = 1.753.
- Thus, the value of Pearson's correlation coefficient in this case indicates that we fail to reject the null hypothesis.

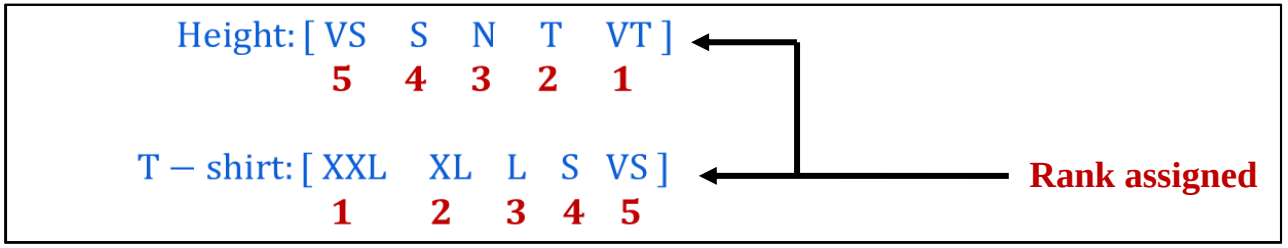
Rank Correlation Analysis

Charles Spearman's Correlation Coefficient

This correlation measurement is also called Rank correlation

- This technique is applicable to determine the degree of correlation between two variables in case of ordinal data.
- We can assign rank to the different values of a variable with ordinal data type.

Example



Charles Spearman's Correlation Coefficient

Definition: **Charles Spearman's correlation coefficient**

The rank correlation can be defined as

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

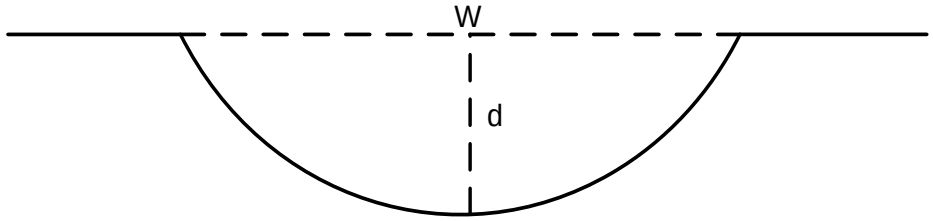
where d_i = Difference between ranks of i^{th} pair of the two variables

n = Number of pairs of observations

- The Spearman's coefficient is often used as a statistical methods to aid either providing or disproving a hypothesis.

Charles Spearman's Coefficient of Correlation

Example: The hypothesis that the depth of a river **does not progressively increase** further from the bank.



A sample of size 10 is collected to test the hypothesis, using Spearman's correlation coefficient.

Sample#	*Width in m*	*Depth in m*
1	0	0
2	50	10
3	150	28
4	200	42
5	250	59
6	300	51
7	350	73
8	400	85
9	450	104
10	500	96

Charles Spearman's Coefficient of Correlation

Step 1: Assign rank to each data. It is customary to assign rank 1 to the largest data, and 2 to next largest and so on.

Note: If there are two or more samples with the same value, the mean rank should be used.

Data	20	25	25	25	30
Assign rank	5	4	3	2	1
Final rank	5	3	3	3	1

Charles Spearman's Coefficient of Correlation

Step 2: The contingency table will look like

Sample	Width	Width r	Depth	Depth r	d	d^2
1	0	10	0	10	0	0
2	50	9	10	9	0	0
3	150	8	28	8	0	0
4	200	7	42	7	0	0
5	250	6	59	5	1	1
6	300	5	51	6	−1	1
7	350	4	73	4	0	0
8	400	3	85	3	0	0
9	450	2	104	1	1	1
10	500	1	96	2	−1	1

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 4}{10 \times 99}$$

$$\sum d^2 = 4$$

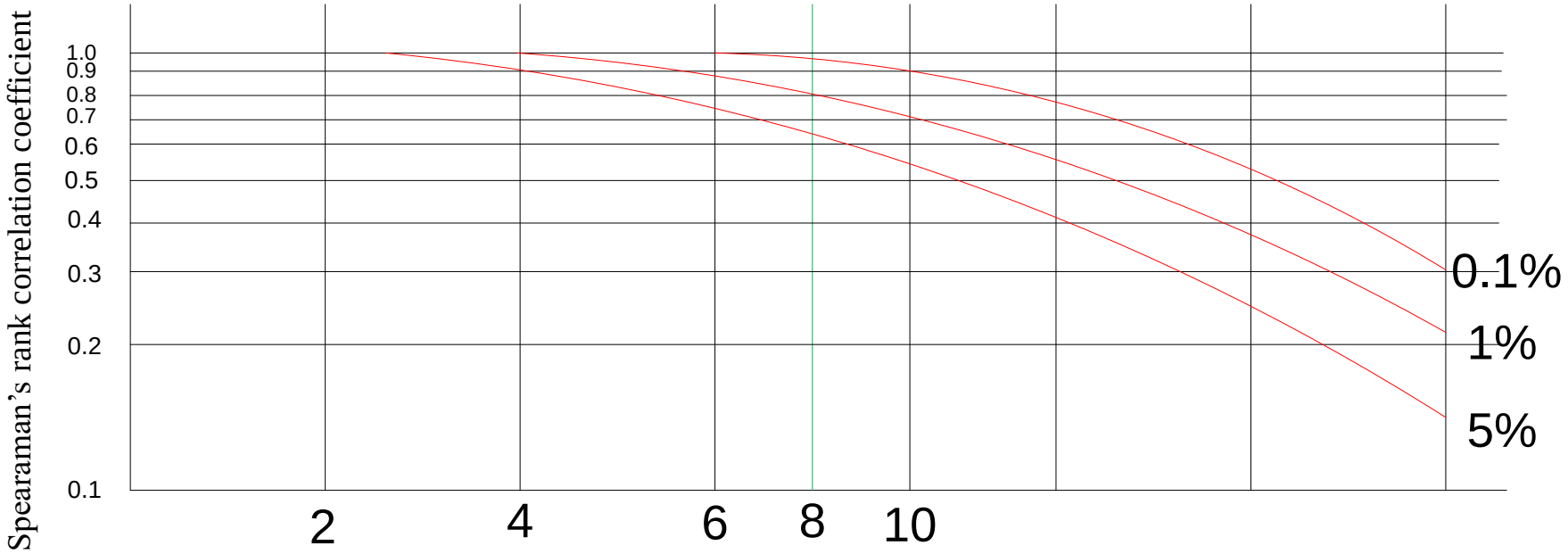
$$r_s = 0.9757$$

Charles Spearman's Coefficient of Correlation

Step 3: To see, if this r_s value is significant, the Spearman's rank significance table (or graph) must be consulted.

Note: The degrees of freedom for the sample $= n - 2 = 8$

Assume, the significance level $= 0.1\%$



Charles Spearman's Coefficient of Correlation

Step 4: Final conclusion

From the graph, we see that $r_s = 0.9757$ lies above the line at 8 and 0.1% significance level. Hence, there is a greater than 99% chance that the relationship is significant (i.e., not random) and hence the hypothesis should be rejected.

Thus, we can reject the hypothesis and conclude that in this case, depth of a river **progressively increases** the further the distance from the river bank.

χ^2 Correlation Analysis

Chi-Squared Test of Correlation

- This method is also alternatively termed as Pearson's χ^2–test or simply χ^2 -test
- This method is applicable to categorical (discrete) data only.

  - Suppose, two attributes *A* and *B* with categorical values

    $A = a_1, a_2, a_3, \ldots\ldots, a_m$ and

    $B = b_1, b_2, b_3, \ldots\ldots, b_n$

    having *m* and *n* distinct values.

A	a_1	a_2	a_3	a_1	a_5	a_1	… …
B	b_1	b_2	b_3	b_1	b_5	b_1	… …

Between whom we are to find the correlation relationship.

χ^2 –Test Methodology

Contingency Table

Given a data set, it is customary to draw a contingency table, whose structure is given below.

	b_1	b_2	-----	b_j	------	b_n	**Row Total**
a_1							
a_2							
⋮							
a_i							
⋮							
a_m							
Column Total							**Grand Total**

χ^2 –Test Methodology

Entry into Contingency Table: Observed Frequency

In contingency table, an entry O_{ij} denotes the event that attribute A takes on value a_i and attribute B takes on value b_j (i.e., $A = a_i$, $B = b_j$).

A	a_i	a_2	a_3	a_i	a_5	a_i	… …
B	b_j	b_2	b_3	b_j	b_5	b_j	… …

	b_1	b_2	-----	b_j	------	b_n	**Row Total**
a_1							
a_2							
⋮							
a_i				O_{ij}			
⋮							
a_m							
Column Total							**Grand Total**

χ^2 –Test Methodology

Entry into Contingency Table: Expected Frequency

In contingency table, an entry e_{ij} denotes the expected frequency, which can be calculated as

$$e_{ij} = \frac{Count(A = a_i) \times Count(B = b_j)}{Grand\ Total} = \frac{A_i \times B_j}{N}$$

	b_1	b_2	-----	b_j	------	b_n	Row Total
a_1							
a_2							
⋮							
a_i				e_{ij}			A_i
⋮							
a_m							
Column Total				B_j			N

A	B
…	…
ai	bj
…	…
ai	bj
…	…
…	…
ai	bj
…	…
…	…
…	…
…	…

χ^2 – Test

Definition: χ^2-Value

The $\boldsymbol{\chi}^2$ value (also known as the Pearson's $\boldsymbol{\chi}^2$ test) can be computes as

$$\boldsymbol{\chi}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the **o**bserved frequency

e_{ij} is the **e**xpected frequency

χ^2 – Test

- The cell that contribute the most to the χ^2 value are those whose actual count is very different from the expected.
- The χ^2 statistics tests the hypothesis that *A* and *B* are independent. The test is based on a significance level, with $(n\text{-}1) \times (m\text{-}1)$ degrees of freedom., with a contingency table of size $n \times m$
- If the hypothesis can be rejected, then we say that *A* and *B* are statistically related or associated.

χ^2 – Test

Example 3: Survey on Gender versus Hobby.

- Suppose, a survey was conducted among a population of size 1500. In this survey, gender of each person and their hobby as either “book” or “computer” was noted. The survey result obtained in a table like the following.

GENDER	HOBBY
.....	.....
.....	.....
M	Book
F	Computer
.....	.....
.....	.....
.....	.....

- We have to find if there is any association between Gender and Hobby of a people, that is, we are to test whether “gender” and “hobby” are correlated.

χ^2 –Test

Example : Survey on Gender versus Hobby.

From the survey table, the **observed frequency** are counted and entered into the contingency table, which is shown below.

GENDER	HOBBY
.....	.....
.....	.....
M	Book
F	Computer
.....	.....

HOBBY	GENDER			
		Male	Female	Total
	Book			
	Computer			
	Total			

χ^2 – Test

Example: Survey on Gender versus Hobby.

- From the survey table, the observed frequency are counted and entered into the contingency table, which is shown below.

		GENDER		
		Male	Female	Total
HOBBY	Book	250	200	450
	Computer	50	1000	1050
	Total	300	1200	**1500**

χ^2 – Test

Example: Survey on Gender versus Hobby.

- From the survey table, the expected frequency are counted and entered into the contingency table, which is shown below.

		GENDER		
HOBBY		Male	Female	Total
	Book	90	360	450
	Computer	210	840	1050
	Total	300	1200	**1500**

χ^2 – Test

- Using equation for χ^2 computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 507.93$$

- This value needs to be compared with the tabulated value of χ^2 (available in any standard book on statistics) with 1 degree of freedom (for a table of $m \times n$, the degrees of freedom is $(m-1)\times(n-1)$; here $m = 2$, $n = 2$).

- For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.01 significance level is 10.828. Since our computed value is above this, we reject the hypothesis that "Gender" and "Hobby" are independent and hence, conclude that the two attributes are *strongly correlated* for the given group of people.

Significance Test for χ^2 -Test

Cramer's V Test

- For χ^2 –test, the most commonly used test to measure the strength of the relation is Cramer's V test. The test takes the following form:

$$V = \sqrt{\frac{\chi^2/n}{(k-1)}}$$

- Here, n is the number of total observation, and k is the number of rows or columns, whichever is less.
- For the example case, n = 1500 and k = 2. Hence, V = 0.58.
- Thus, it is neither weak nor a strong correlation; this implies that **Gender** and **Hobby** are related with the degree of correlation 0.58

More on Correlation Analysis

Other Types of Correlation

- Binary variable to binary variable correlation
  - **Tetrachoric correlation**
- Nominal/ categorical valued variable to binary variable correlation
  - **Cramer's V correlation**
- Continuous variable to binary variable correlation
  - **Point-biserial correlation**

Tetrachoric correlation

- **Tetrachoric correlation** is a measure of the association between two binary variables, that is, variables that can only take on two values like "yes" and "no" or "good" and "bad."

- Suppose, we have the following 2×2 table with two variables, x and y, that both take on two values:

Here
a = Total count for $x = 0$ and $y = 0$
b = Total count for $x = 0$ and $y = 1$
c = Total count for $x = 1$ and $y = 0$
d = Total count for $x = 1$ and $y = 1$

	$y = 0$	$y = 1$
$x = 0$	a	b
$x = 1$	c	d

$$r_t = cos\left(\frac{180}{1+\sqrt{\frac{b*c}{a*d}}}\right)$$

Tetrachoric correlation: Example

- **Example:**

Suppose, we want to know whether or not gender is associated with political party preference so we take a simple random sample of 47 voters and survey them on their political party preference.

	$y =$ **party 1**	$y =$ **party 2**
$x = male$	9	15
$x = female$	13	10

Here
$a = 9$ for $x = male$ and $y = party\ 1$
$b = 15$ for $x = male$ and $y = party\ 2$
$c = 13$ for $x = female$ and $y = party\ 1$
$d = 10$ for $x = female$ and $y = party\ 2$

Tetrachoric correlation: Example

$$r_t = cos\left(\frac{180}{1+\sqrt{\frac{b*c}{a*d}}}\right)$$

Here
$a = 9$ for $x = male$ and $y = party\ 1$
$b = 15$ for $x = male$ and $y = party\ 2$
$c = 13$ for $x = female$ and $y = party\ 1$
$d = 10$ for $x = female$ and $y = party\ 2$

	$y =$ **party 1**	$y =$ **party 2**
$x = male$	9	15
$x = female$	13	10

- $r_t = cos\left(\frac{180}{1+\sqrt{\frac{b*c}{a*d}}}\right)$

  $= cos\left(\frac{180}{1+\sqrt{\frac{15*13}{9*10}}}\right)$

  $= cos\left(\frac{180}{1+1.471}\right)$

  $= cos\left(\frac{180}{2.471}\right) = cos(72.84) = 0.29$

- Here, the coefficient of correlation between gender and political party preference is 0.29.
- This correlation is significantly low, which indicates that **there is a weak correlation between gender and preference of political party .**

Cramer’s V correlation

Cramer's V correlation is used to measure the strength of association between two variables with nominal or categorical values.

Each variable can have two or more than two nominal or categorical values also.

Cramer's V correlation coefficient

$$r_{cv} = \sqrt{\frac{\frac{\chi 2}{n}}{\min(m-1, c-1)}}$$

$\chi 2$ = The Chi-square statistics
n = Total number of samples in the dataset
m = Number of classes of dependent variable
c= Number of columns in the dataset

Cramer's V correlation: Example

- **Example:**

Suppose, we want to know if there is any association between three different eye colors (blue, green and brown) and three regions (east, north and west). After surveying 50 random samples, the following data is obtained.

	Eye Color		
	Blue	Green	Brown
East	8	5	6
North	2	8	3
west	4	6	8

Cramer's V correlation: Example

	Eye Color			
	Blue	Green	Brown	**Row Total**
East	8	5	6	19
North	2	8	3	13
west	4	6	8	18
Column Total	14	19	17	**Grand total 50**

Step 1:

Here all the frequencies are called **observed frequency.**

Add all values row wise and column wise.
Here
row totals are 19,13,18
column totals are 14, 19, 17

Grand Total is 50

Cramer's V correlation: Example

	Eye Color			
	Blue	Green	Brown	**Row Total**
East	$\frac{19*14}{50} = 5.32$	$\frac{19*19}{50} = 7.22$	$\frac{19*17}{50} = 6.46$	19
North	$\frac{13*14}{50} = 3.64$	$\frac{13*19}{50} = 4.94$	$\frac{13*17}{50} = 4.42$	13
west	$\frac{18*14}{50} = 5.04$	$\frac{18*19}{50} = 6.84$	$\frac{18*17}{50} = 6.12$	18
Column Total	14	19	17	**Grand Total = 50**

Step 2:

Calculate expected frequencies for each cell.

Expected frequency

$$e_{ij} = \frac{(i^{th} \text{ row total}) * (j^{th} \text{ column total})}{Grand\ total}$$

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij} is the **o**bserved frequency

e_{ij} is the **e**xpected frequency

Cramer's V correlation: Example

Step 3:

Calculate $\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$

Observed values	Eye Color		
	Blue	Green	Brown
East	8	5	6
North	2	8	3
west	4	6	8

Expected values	Eye Color		
	Blue	Green	Brown
East	5.32	7.22	6.46
North	3.64	4.94	4.42
west	5.04	6.84	6.12

$$\chi^2 = \frac{(8-5.32)^2}{5.32} + \frac{(5-7.22)^2}{7.22} + \frac{(6-6.46)^2}{6.46} + \frac{(2-3.64)^2}{3.64} + \frac{(8-4.94)^2}{4.94} + \frac{(3-4.42)^2}{4.42} + \frac{(4-5.04)^2}{5.04} + \frac{(6-6.84)^2}{6.84} + \frac{(8-6.12)^2}{6.12} = 6.35$$

Cramer's V correlation: Example

Step 4:

Putting the below given values to the equation

$\chi^2 = 6.35$
$n = 50$
$m = 3$
$c = 3$

$$r_{cv} = \sqrt{\frac{\frac{\chi 2}{n}}{\min(m-1, c-1)}} = \sqrt{\frac{\frac{6.35}{50}}{\min(3-1, 3-1)}} = \sqrt{\frac{0.127}{2}} = 0.25$$

The correlation between three different eye colors (blue, green and brown) and three regions (east, north and west) is 0.25

It means **eye color is weakly associated with the regions.**

Point-biserial correlation

Point-biserial correlation is a measure of the association between a continuous valued and a binary valued variable.

Point-biserial correlation coefficient

$$r_{pb} = \left|\frac{M_1 - M_0}{S_n}\right| \sqrt{p * q}$$

where,
M_1= mean of values in x_i, when y= 1.
M_0 = mean of values in x_i , when y= 0.
S_n = standard deviation of the attribute values x_i with a sample of size n.
p = Proportion of cases for y = 0
q = Proportion of cases for y = 1

Point-biserial correlation: Example

Example:

Suppose we want to know whether or not gender is associated with weekly expenditure of the students, where we take a simple random sample of 7 students and survey on them.

$x = expanditure$	$y = gender$
12	**1**
8	**1**
7	1
22	0
18	0
16	0
20	0

Step 1:

Here, 1= male and 0 = female

$M_1 = \frac{(12+8+7)}{3} = 9$

$M_0 = \frac{(22+18+16+20)}{4} = 19$

$n = 7$

Point-biserial correlation: Example

$x =$ *expanditure*	$y =$ *gender*
12	**1**
8	**1**
7	1
22	0
18	0
16	0
20	0

- **Step 2:**
- $p = \frac{Total\ number\ of\ male}{n} = \frac{3}{7} = 0.43$
- $q = \frac{Total\ number\ of\ female}{n} = \frac{4}{7} = 0.47$

- **Step 3:**
- $S_n = \sqrt{\frac{(x_i - \bar{x})^2}{n}} = 5.85$ where, $\bar{x} = \frac{12+8+7+22+18+16+20}{7} = 14.71$
- So,
- $r_{pb} = \left|\frac{M_1 - M_0}{S_n}\right| \sqrt{p * q} = \left|\frac{9-19}{14.71}\right| \sqrt{0.43 * 0.47} = 0.85$

- Here, the coefficient of correlation between gender and weekly expenditure of the students is 0.85.
- It means **gender is strongly associated with weekly expenditure of the students**.

Galton Board: CLT

- Sir Francis Galton, Charles Darwin's half-cousin, invented the 'Galton Board' in 1874 to demonstrate that the normal distribution is a natural phenomenon.
- He is the founder of linear regression.
- Galton Board specifically shows that the binomial distribution approximates a normal distribution with a large enough sample size.

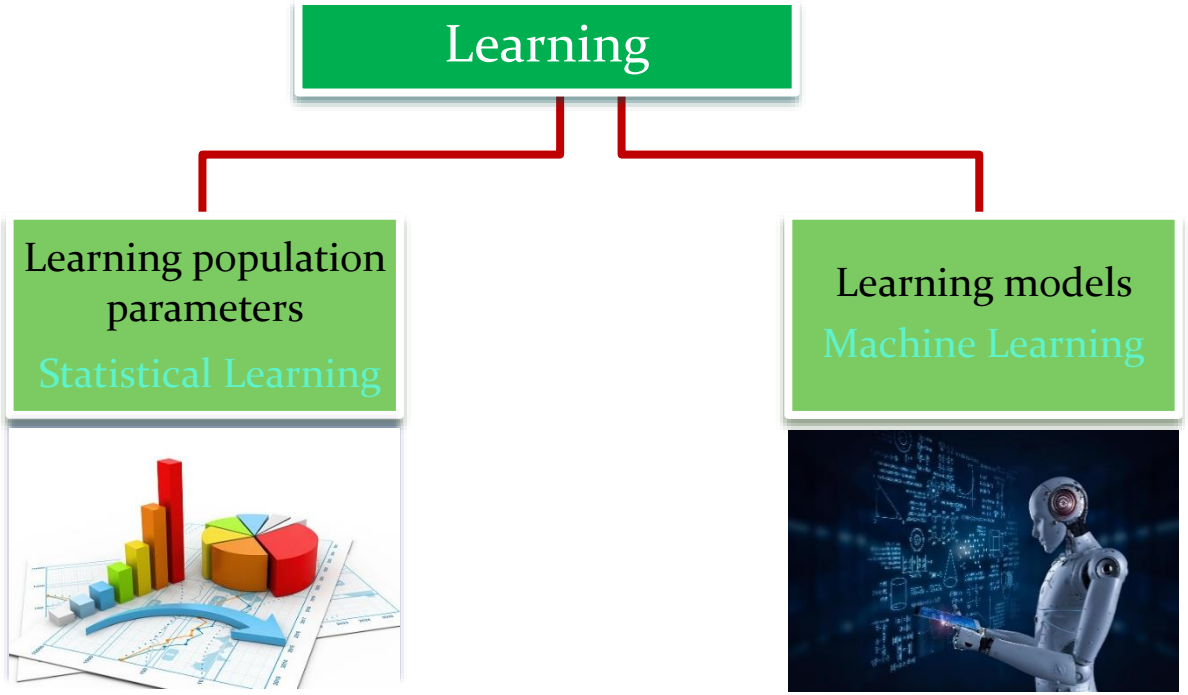
Do Remember!

Figure 14-3. Another example of association or causation. (DILBERT © 2011 Scott Adams. Used by permission of UNIVERSAL UCLICK. All rights reserved.)

Regression Analysis

Learning Strategies

There are two types of learning concepts:

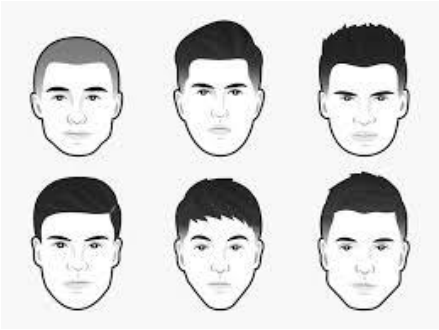


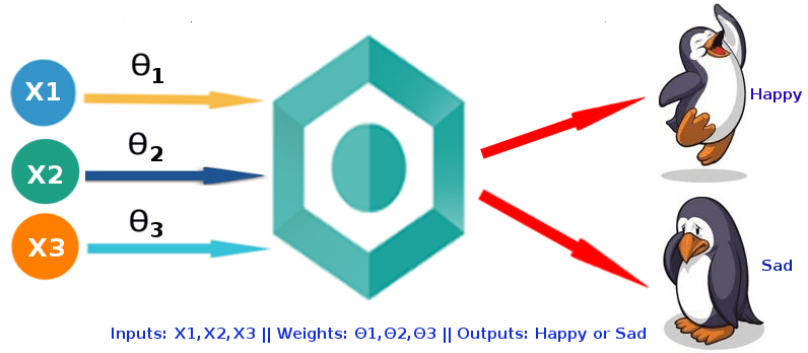
Statistical Learning

Usually assumes certain properties of the population from which we draw samples:

- Observation come from a normal population.
- Sample size is small.
- Population parameters like mean, variance, etc. are hold good.
- Requires measurement equivalent to interval scaled data.

Machine Learning





$y = f(x) = ax^2 + bx + c$

- Does not under any assumption
- Works well with high volume high dimensional data

Important Point

This learning strategy needs a very large sample data

Relationship Analysis

Relationship Analysis

- **Example: Wage Data**

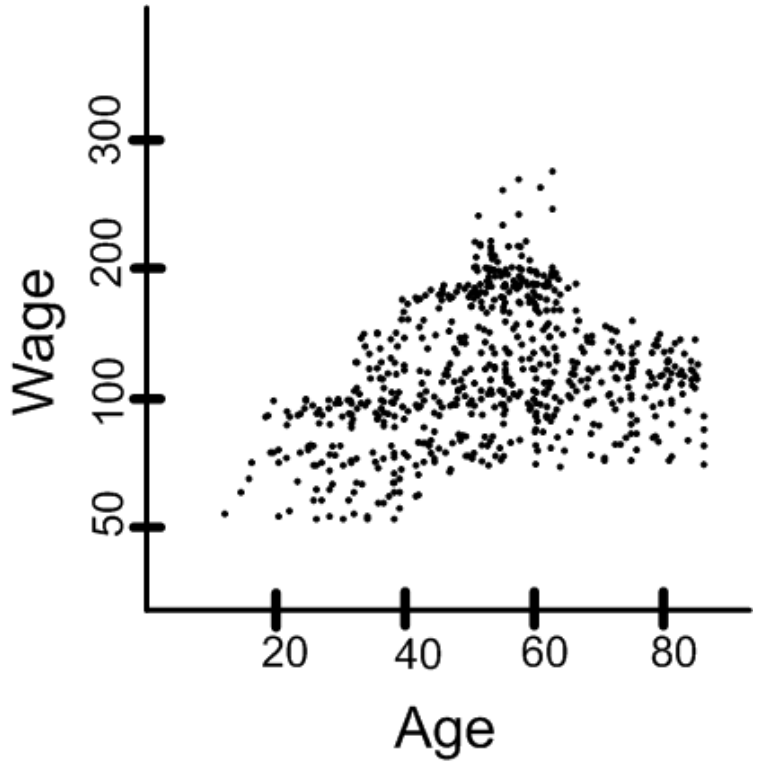
  A large data regarding the wages for a group of employees from the eastern region of India is given.

  In particular, we wish to understand the following relationships:

  - *Employee's age and wage:* How wages vary with ages?
  - *Calendar year and wage:* How wages vary with time?
  - *Employee's age and education:* Whether wages are anyway related with employees' education levels?

Relationship Analysis

- Example: Wage Data
  - Case I. Wage versus Age
    - From the data set, we have a graphical representations, which is as follows:

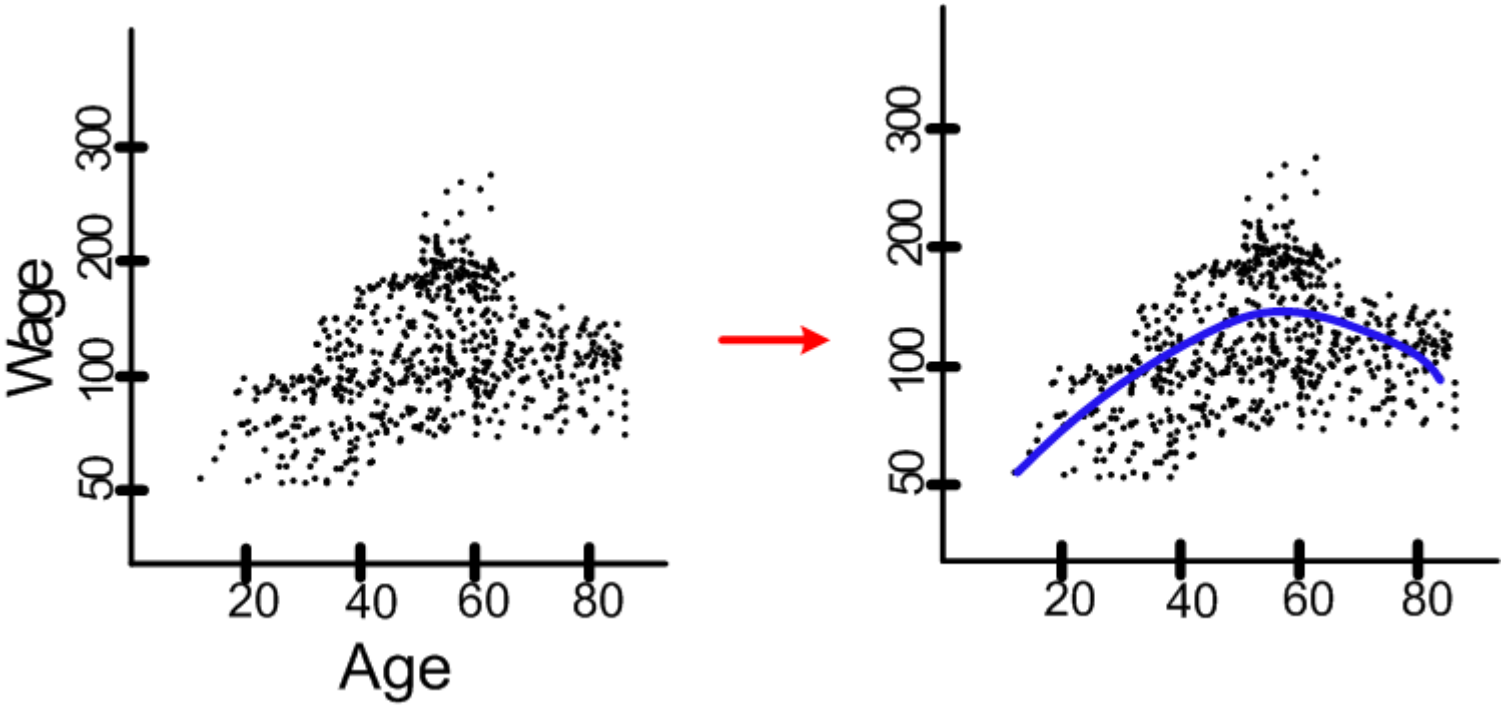


?

How wages vary with ages?

Relationship Analysis

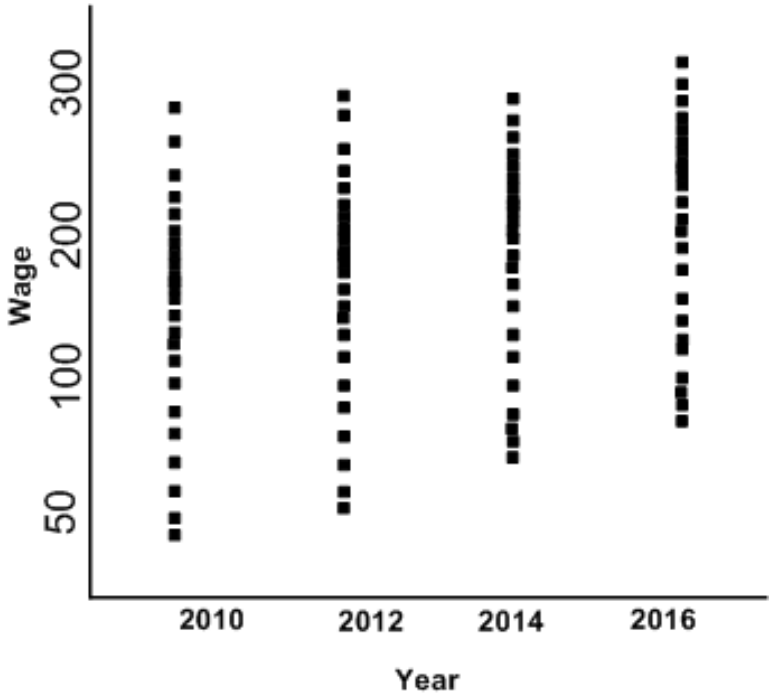
- Example: Wage Data
  - *Employee's age and wage:* How wages vary with ages?



Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

Relationship Analysis

- Example: Wage Data
  - Case II. Wage versus Year
    - From the data set, we have a graphical representations, which is as follows:

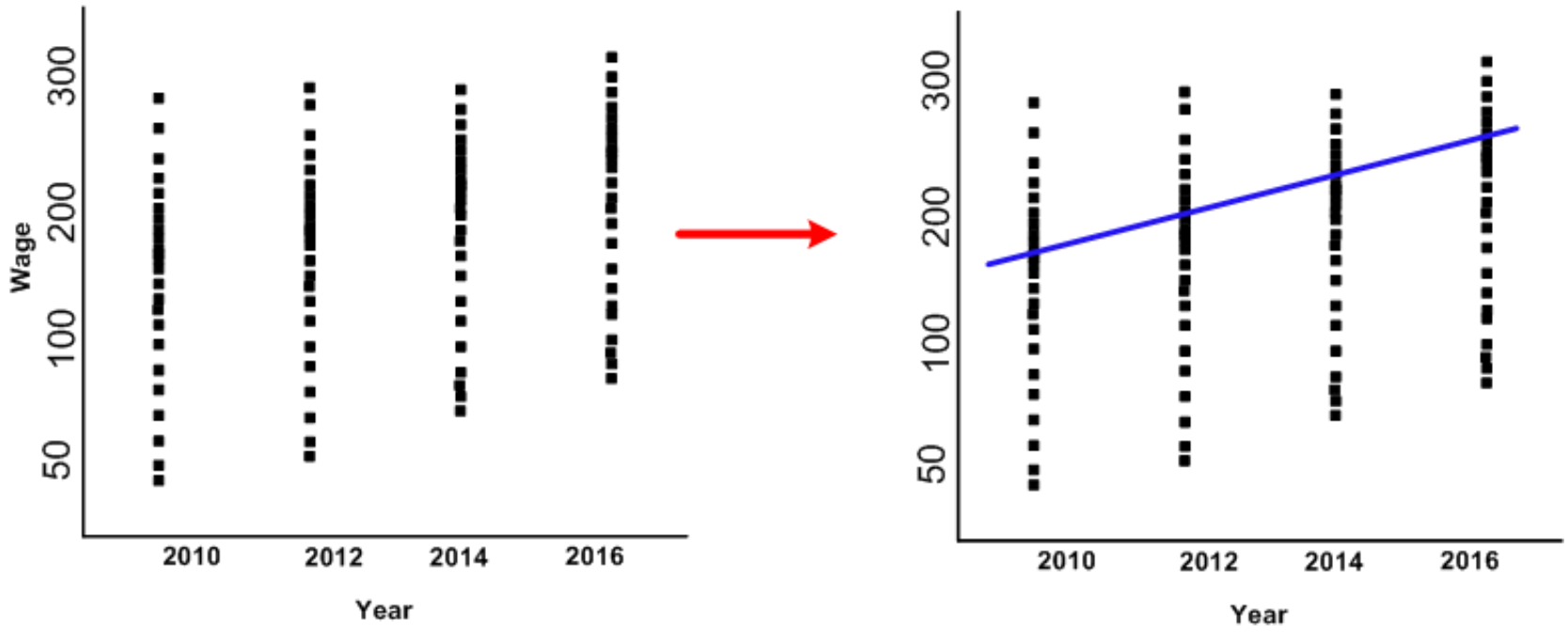


?

How wages vary with time?

Relationship Analysis

- Example: Wage Data
  - *Wage and calendar year:* How wages vary with years?

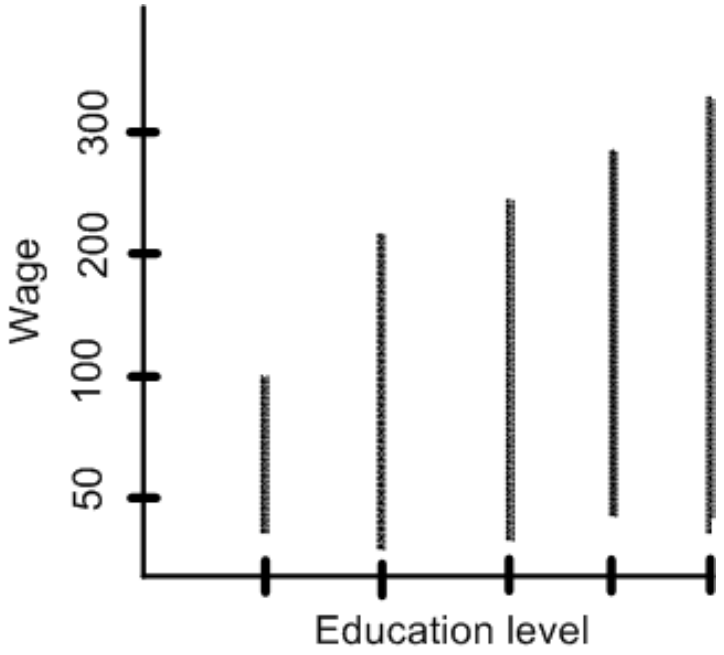


Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

.

Relationship Analysis

- Example: Wage Data
  - Case III. Wage versus Education
    - From the data set, we have a graphical representations, which is as follows:

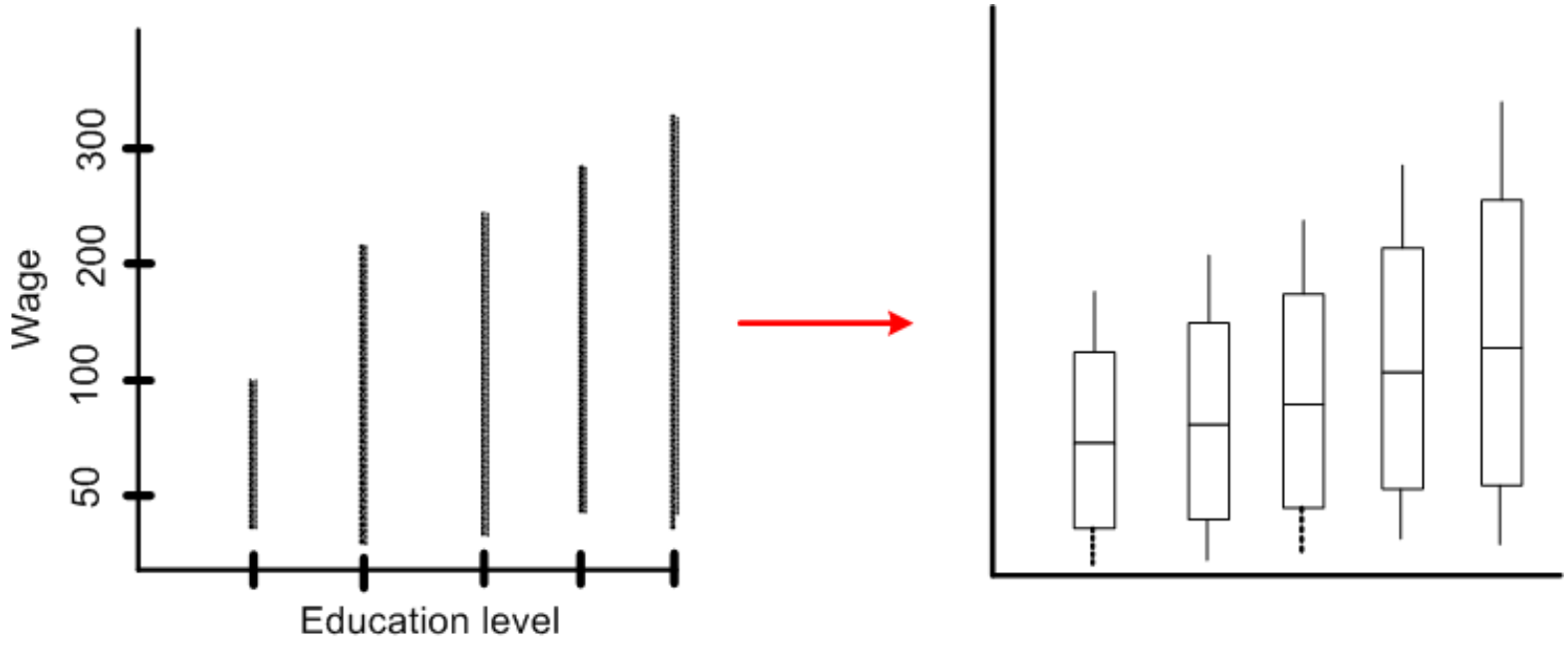


?

Whether wages are related with education?

Relationship Analysis

- Example: Wage Data
  - *Wage and education level:* Whether wages vary with employees' education levels?



Interpretation: On the average, wage increases with the level of education.

Relationship Analysis

What more information can we get?

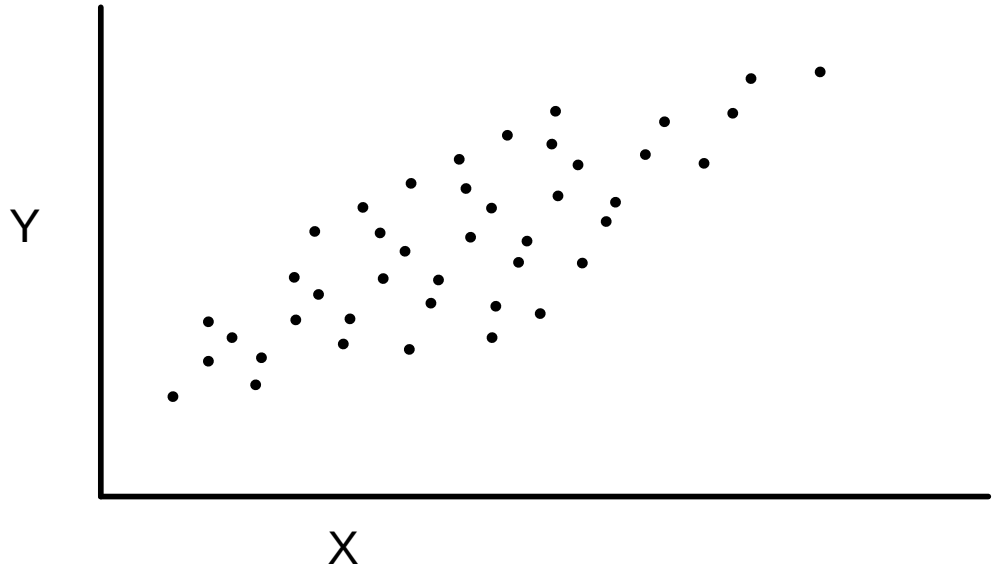
Whether wage has any association with both year and education level?

Given an employee's wage can we predict his age?

… and what's more?

Regression Analysis to find Relationships

A curious Question!

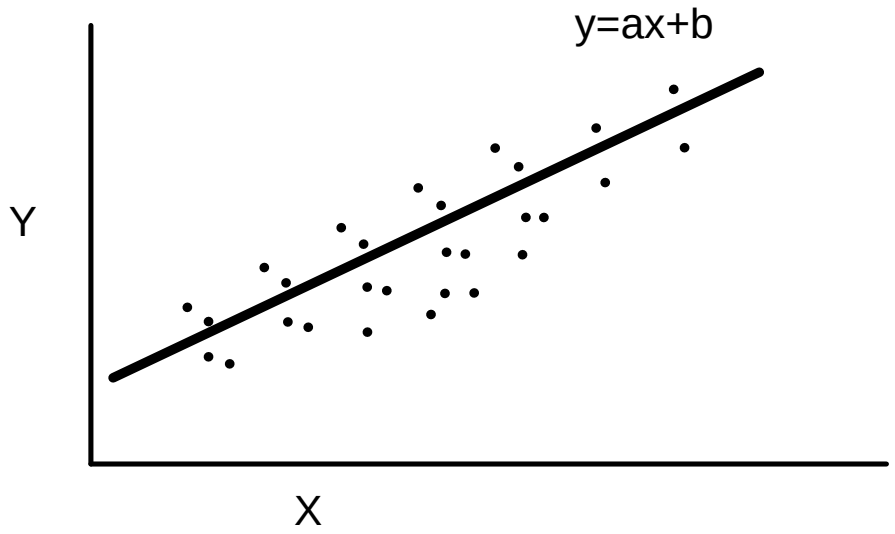


Suppose there are countably infinite points in the $XY\ plane$. We need a huge memory to store all such points.

Is there any way out to store this information with a least amount of memory?

Say, with two values only.

Yahoo!



Just decide the values of **a** and **b**
(as if storing one point's data only!)

Note: Here, tricks was to find a relationship among all the points.

Measures of Relationship

Univariate Population

Temperature	20	30	21	18	23	45	52

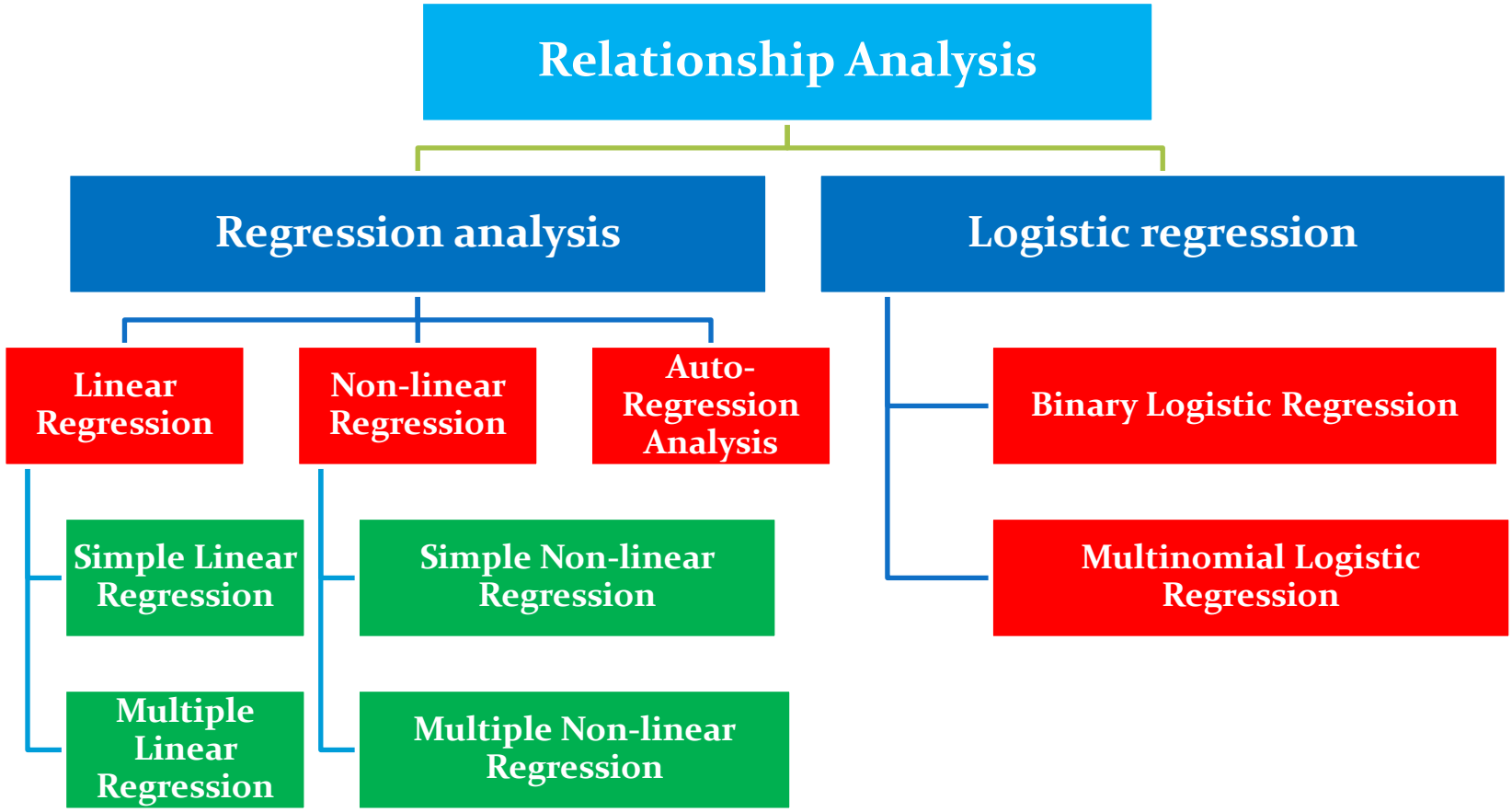
Bivariate Population

Temperature	20	30	21	18	23	45	52
Pressure	1	1.5	1.05	0.96	1.2	2.5	2.8

Multivariate Population

Temperature	20	30	21	18	23	45	52
Pressure	1	1.5	1.05	0.96	1.2	2.5	2.8
Volume	20	30	21	18	23	45	52

Measures of Relationship



Regression Analysis

Definition

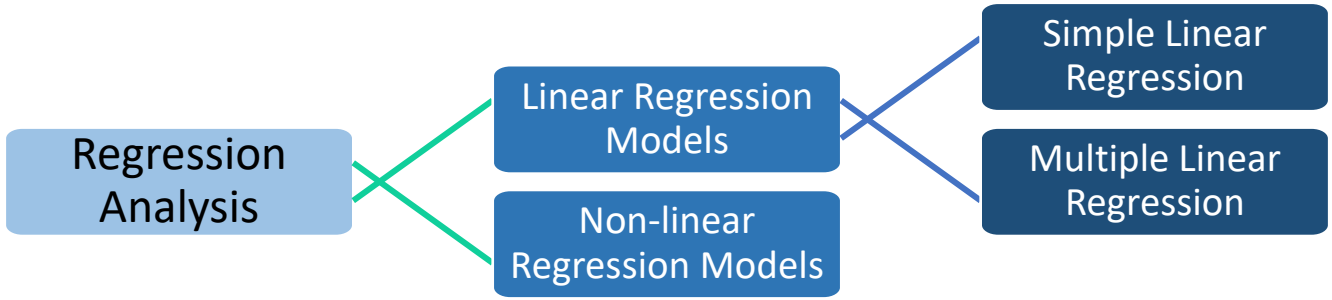
The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variable(s).

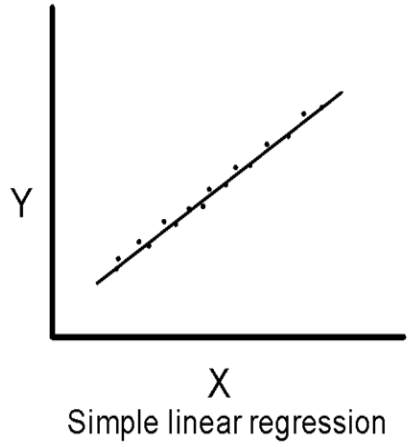
A Simple Example

How Exam Score is related to Hours of Study?

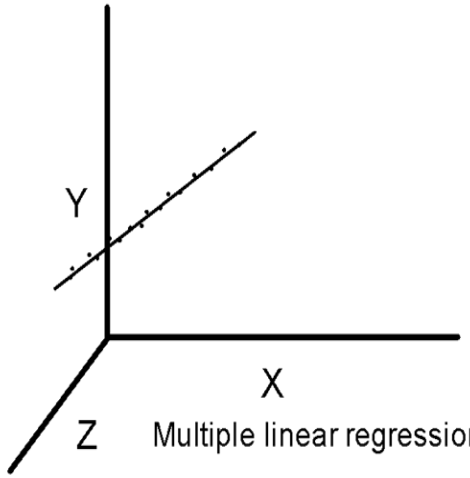
Hours Study	*Exam Score*
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100

Regression Analysis

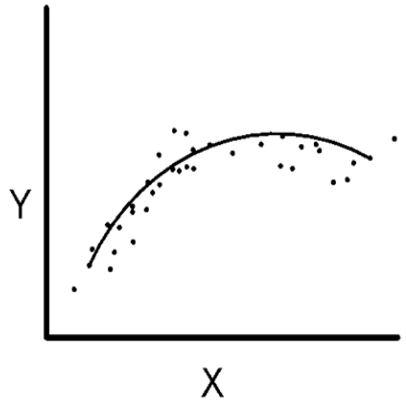




Simple linear regression



Multiple linear regression



Non-linear regression

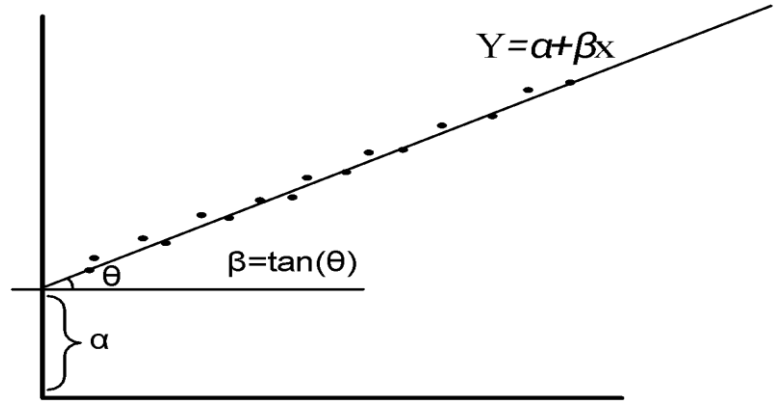
Simple Linear Regression

Simple Linear Regression Model

In simple linear regression, we have only two variables:

Dependent variable:
Also called *Response*, usually denoted as Y

Independent variable:
Also called *Regressor*, usually denoted as x

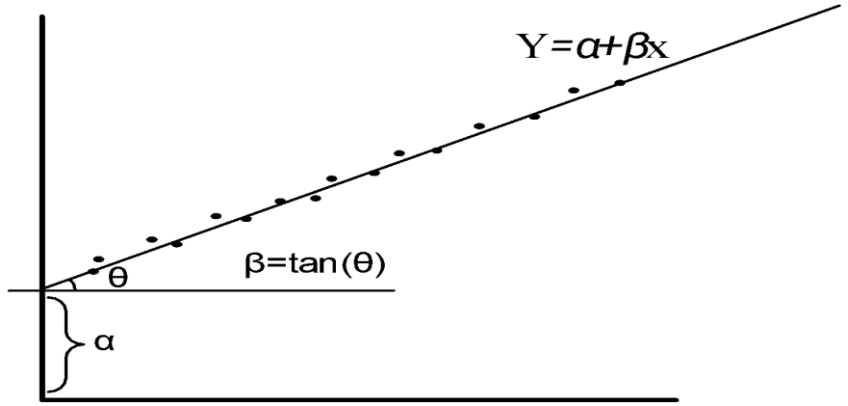


Linear regression

A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$

Simple Linear Regression Model

In simple linear regression, we have only two variables:



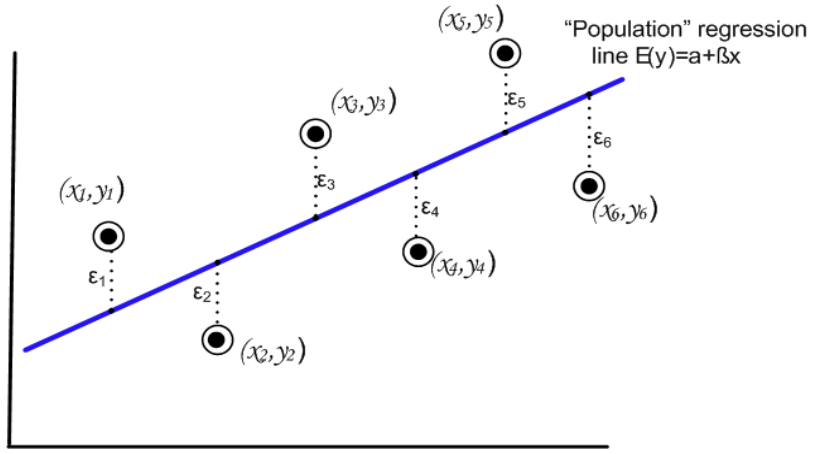
Note

- There are infinite number of lines (and hence α_s and β_s)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α and β) quantifying the strength of that relationship.

Regression Analysis

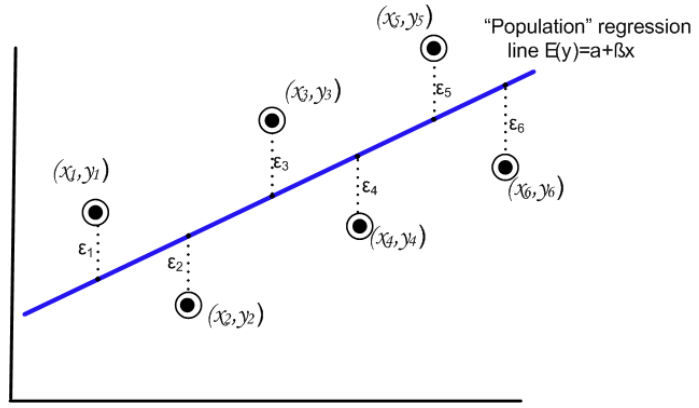
Given: The set $[(x_i, y_i), i = 1,2,3, \dots, n]$ of data involving n pairs of (x, y) values

Objective: To find "true" or population regression line, such that $Y = \alpha + \beta x + \in$



Here, $\in$ is a random variable with $E(\in) = 0$ and $var(\in) = \sigma^2$. The quantity σ^2 is often called the **error variance**.

Regression Analysis



Note

- $E(\in) = 0$ implies that at a specific x, the y values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- The values of the regression coefficients α and β to be estimated from data

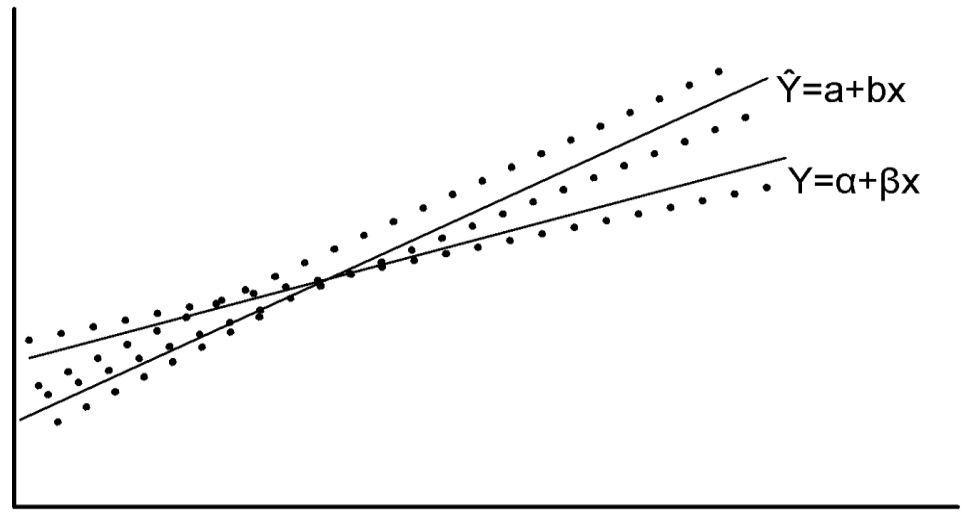
True versus Fitted Regression Line

- The task in regression analysis is to estimate the regression coefficients α and β.

- Suppose, we denote the estimates *a* for α and *b* for β. Then the fitted regression line is

$$\hat{Y} = a + bx$$

where, $\hat{Y}$ is the predicted or fitted value

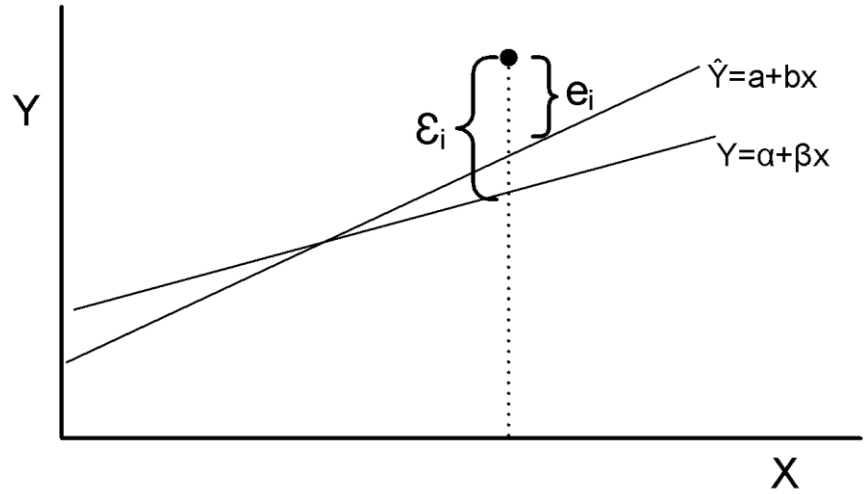


Least Square Method to estimate α and β

Concept of Residuals

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\hat{Y} = a + bx$. Thus, i^{th} residual is

$$e_i = Y_i - \hat{Y}_i, i = 1,2,3, \dots \dots, n$$



Least Square Method to estimate α and β

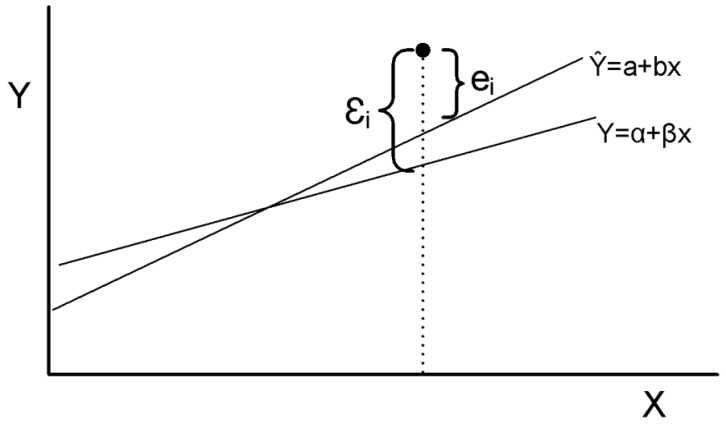
Sum of Squares Error (SSE)

The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as

SSE

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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



We need to **minimize the value of SSE** and hence to determine the parameters of *a* and *b*.

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 1: Differentiation

Differentiating SSE with respect to a and b, we have

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

$$\frac{\partial(SSE)}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i).x_i$$

Step 2: Equating the partial derivatives to zero

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$, and $\frac{\partial(SSE)}{\partial b} = 0$

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 2: Equating the partial derivatives to zero

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$, and $\frac{\partial(SSE)}{\partial b} = 0$

Thus we get,

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

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

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 2: Equating the partial derivatives to zero

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$,

and $\frac{\partial(SSE)}{\partial b} = 0$

Thus we get,

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

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

Step 3: Solving for a and b

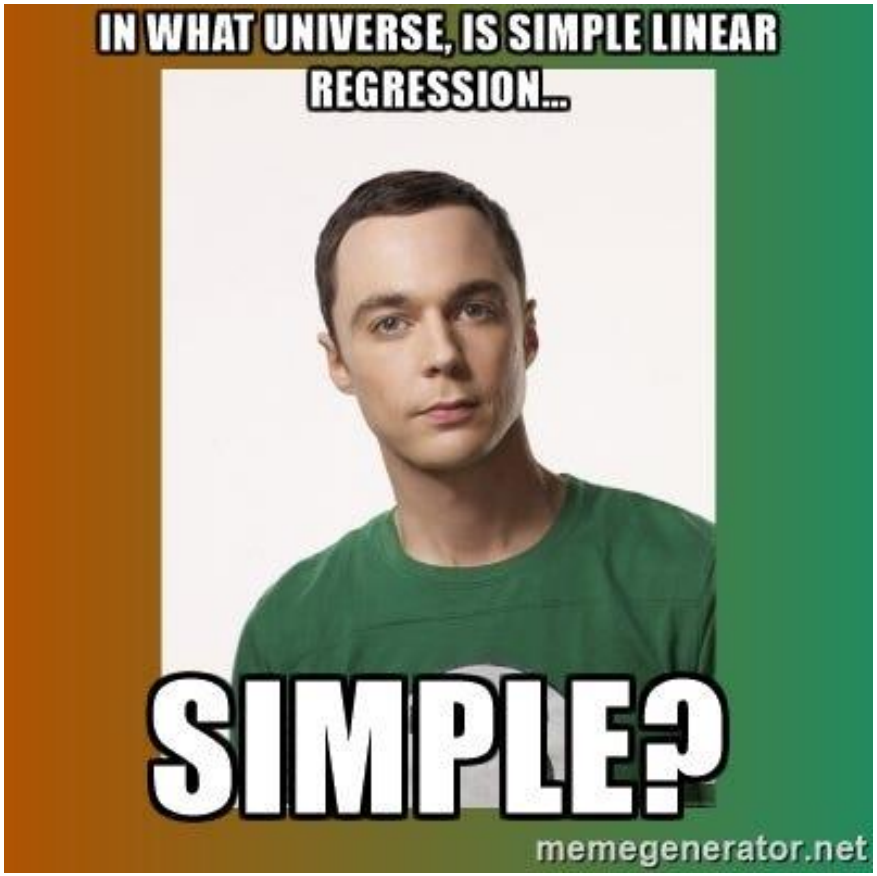
These two equations on the left can be solved to determine the values of a and b, and it can be calculated that

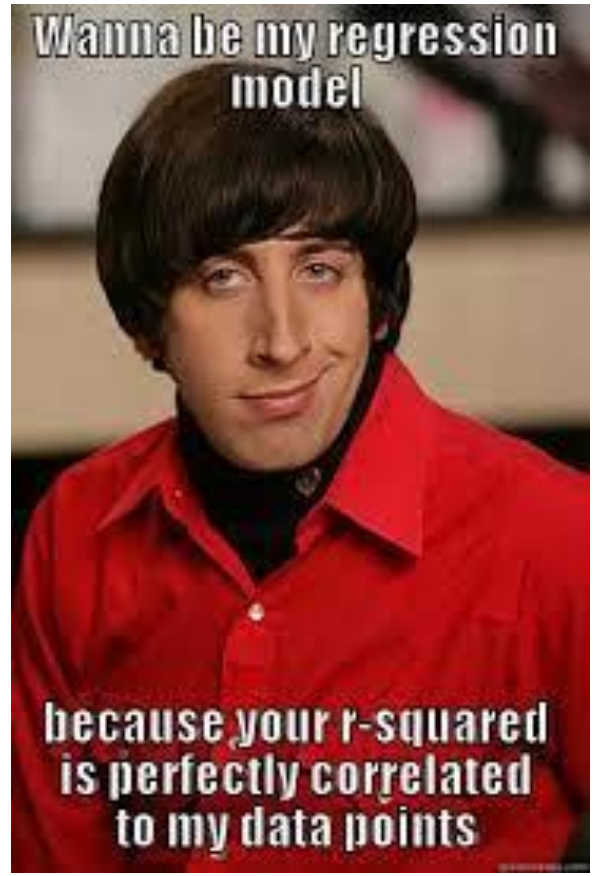
$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

R^2 : Measure of Quality of the Fitting

R2: Measure of Fit Quality





R2: Measure of Fit Quality

Coefficient of Determination

A quantity R^2, is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.

Total corrected sum of squares:

We have $\boldsymbol{SSE} = \sum_{i=1}^{n}(y_i - \hat{y})^2$. It signifies the **variability due to error**.

Now, the **total corrected sum of squares** is defined as $\mathbf{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$

R^2:

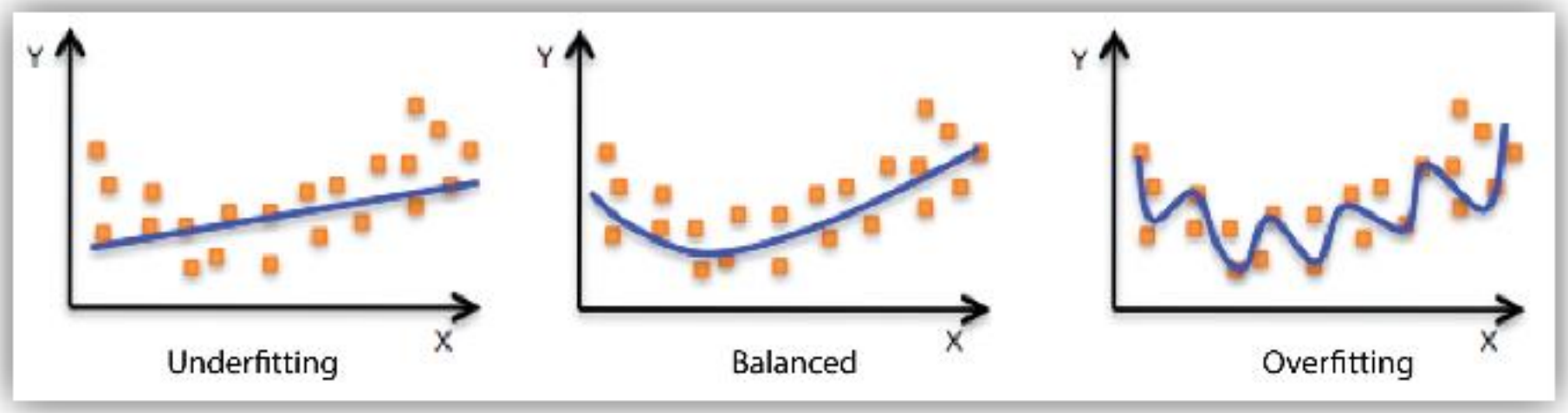
SST represents the variation in the response values. The R^2 is: $R^2 = 1 - \frac{SSE}{SST}$

R2: Measure of Quality Fit

Coefficient of Determination

Note

- If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)

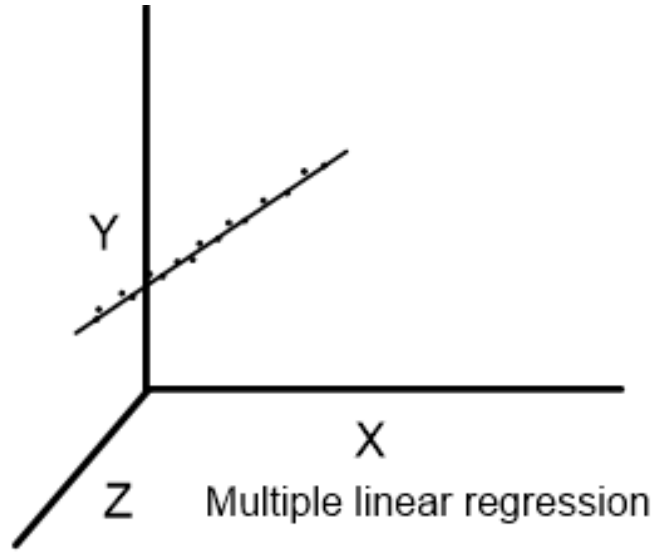


Multiple Linear Regression

Multiple Linear Regression

Definition:

- **Multiple** Regression Model**:** When more than one variable are independent variable, then the regression can be estimated as a multiple regression model
- Multiple **Linear** Regression**:** When this model is linear in coefficients, it is called multiple linear regression model



Multiple linear regression

Multiple Linear Regression

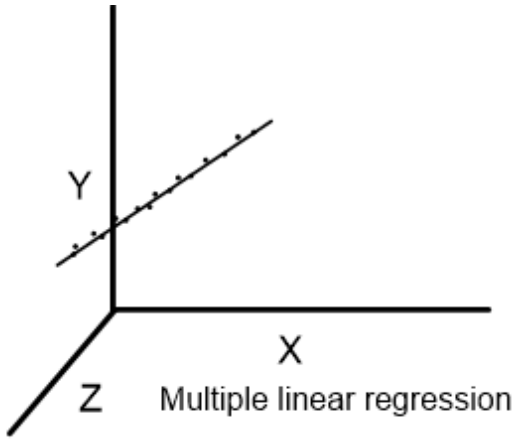
Formulation:

If *k*-independent variables $x_1, x_2, x_3 \ldots\ldots\ldots, x_k$ are associated, the multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \in$$

And the estimated response is obtained as

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + b_k x_k$$



Multiple Linear Regression

Estimating the coefficients: The data points

Let the data points given to us are

$$(x_{1i},\ x_{2i},\ x_{3i}, \dots\dots\dots\dots\dots\dots, x_{ki}, y_i) \quad i = 1,2, \dots\dots\dots, n, \qquad n > k$$

where y_i is the observed response to the values $x_{1i},\ x_{2i},\ x_{3i}, \dots\dots\dots\dots\dots\dots, x_{ki}$ of k independent variables $x_1,\ x_2,\ x_3, \dots\dots\dots\dots\dots\dots, x_k$.

Multiple Linear Regression

Estimating the coefficients: The model formulation

So, the regression model in this case is given by,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + \in_i$$

and

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_k x_{ki} + e_i$$

where $\in_i$ and e_i are the random error and residual error, respectively associated with true response y_i and fitted response $\hat{y}_i$.

Multiple Linear Regression

Estimating the coefficients: Minimization of the SSE

Using the concept of **Least Square Method** to estimate $b_0, b_1, b_2, \dots, b_k$, we minimize the expression

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

To minimize the SSE, we need to differentiate SSE in turn with respect to $b_0, b_1, b_2, \dots, b_k$ and equate to zero.

Multiple Linear Regression

Estimating the coefficients: Minimization of the SSE

Differentiating SSE in turn with respect to $b_0, b_1, b_2, \ldots, b_k$ and equating to zero, we generate the set of $(k+1)$ normal estimation equations for multiple linear regression.

$$nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} + \ldots + b_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i} \cdot x_{2i} + \ldots + b_k \sum_{i=1}^{n} x_{1i} \cdot x_{ki} = \sum_{i=1}^{n} x_i \cdot y_i$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$b_0 \sum_{i=1}^{n} x_{ki} + b_1 \sum_{i=1}^{n} x_{ki} \cdot x_{1i} + b_2 \sum_{i=1}^{n} x_{ki} \cdot x_{2i} + \ldots + b_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_i \cdot y_i$$

The system of linear equations can be solved for $b_0, b_1, \ldots, b_k$ by any appropriate method for solving system of linear equations.

Non-Linear Regression Model

Non-linear Regression Model

Definition and Formulation:

- **Non-linear Regression Model:** When the regression equation is in terms of r −degree, $r > 1$, then it is called non-linear regression model
- **Multiple Non-linear Regression Model:** When more than one independent variables are there, then it is called multiple non-linear regression model
  - It is alternatively termed as polynomial regression model.
- In general, it takes the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r + \epsilon$$

- The estimated response is obtained as

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_r x^r$$

Solving for Polynomial Regression Model

Model formulation:

Given that (x_i, y_i); $i = 1,2,\ldots,n$ are n pairs of observations.

Each observations would satisfy the equations:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \ldots + \beta_r {x_i}^r + \epsilon_i$$

and $$\hat{y}_i = b_0 + b_1 x_i + b_2 {x_i}^2 + \ldots + b_r {x_i}^r + e_i$$

where, r is the degree of polynomial
ϵ_i is the i^{th} random error
e_i is the i^{th} residual error

Note: The number of observations, n, must be at least as large as $r + 1$, the number of parameters to be estimated.

Solving for Polynomial Regression Model

Transformation to Linear Regression:

The polynomial model can be transformed into a general linear regression model setting $x_1 = x, x_2 = x^2, \ldots, x_n = x^r$.

Thus, the equation assumes the form:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r + \epsilon_i$$

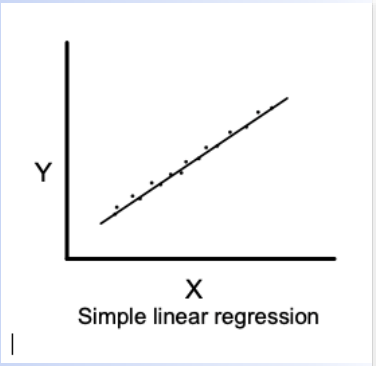
$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r + e_i$$

This model then can be solved using the procedure followed for multiple linear regression model.

Linear versus Non-Linear Regression

Simple linear regression model:

$$Y = \beta_0 + \beta_1 x$$



Simple non-linear regression model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r$$

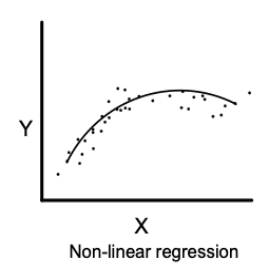
Linear versus Non-Linear Regression

Simple linear regression model:

$$Y = \beta_0 + \beta_1 x$$

Simple non-linear regression model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r$$



Issues:

a) Whether linear or non-linear model?
b) If non-linear, then what is its degree $r \geq 2$?

Solution:

Take the R^2 measures for all models (with r =1, 2, ...)
and then select that model with the higher value of R^2

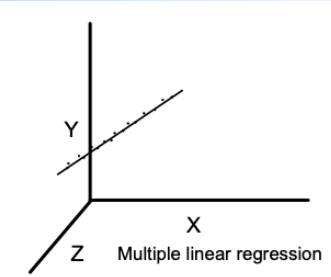
X	Y
x_i	y_i

Multiple Non-Linear Regression

Issues with Multiple Non-Linear Regression

Multiple non-linear regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \ldots + \beta_r x_k^r$$



Issues:

- Too complex to solve. Many parameters, many variations!
- Usually, used advanced machine learning models, such as SVM, kNN, ANN, etc.

X	Y	Z

Auto-regression Analysis

Time-series Data

Time-series Data

Time-series data: The data collected on the same observational unit at multiple time periods

Example: Rate of price inflation

Time-series Data

Examples of time-series data:

- Aggregate consumption and GDP for a country (for example, 20 years of quarterly observations = 80 observations)
- Yen/$, pound/$ and Euro/$ exchange rates (daily data for 1 year = 365 observations)
- Cigarette consumption per capita in a state, by years
- Rainfall data over a year or a period of years
- Sales of tea from a tea shop in a season

Use of Time-series Data

Use of time-series data

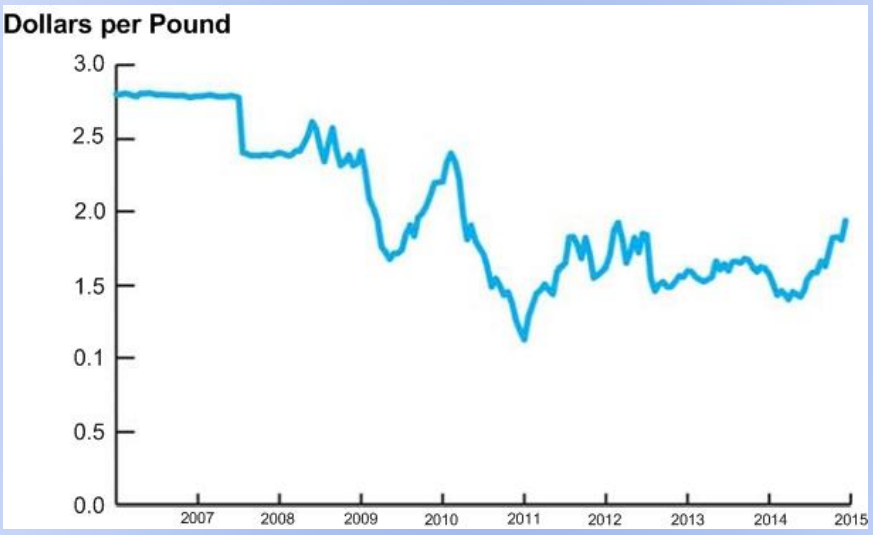
- To develop forecast model
  - What will be the rate of inflation in next year?
- To estimate dynamic causal effects
  - If the rate of interest increases, what will be the effect on the rates of inflation and unemployment in 3 months? in 12 months?
  - What is the effect over time on electronics good consumption due to a hike in the excise duty?
- Time dependent analysis
  - Rates of inflation and unemployment in the country can be observed over a time period.

Modeling with Time-series Data

- Correlation over time
  - Serial correlation, also called autocorrelation
  - Calculating standard error
- To estimate dynamic causal effects
  - Under which dynamic effects can be estimated?
  - How to estimate?
- Forecasting model
  - Forecasting model build on regression model

Modeling with Time-series Data: Example

Modeling with time-series data



- **Can we predict the trend at a time say 2017?**

Concept and Notations

Related Concepts and Notations

- Y_t = Value of Y in a period t
- Data set $[Y_1, Y_2, \ldots, Y_{T\_1}, Y_T]$: This is the T observations on the time series random variable Y

Example

Rainfall data in a region.

- Here, Y_t denotes the daily/ weekly/ monthly rainfall in a year. More precisely, for example, for the year 2021, 365 days data
- Data set: The rainfall data from 2012-2021, that is, the last 10 years data.

Related Concepts and Notations

- There are four ways to have the time series data for Auto-regression analysis
  - **Lag:** The first lag of Y_t is Y_{t-1}, its j^{th} lag is Y_{t-j}
  - **Difference:** The fist difference of a series, Y_t is its change between period t and $t-1$, that is, $y_t = Y_t - Y_{t-1}$
  - **Log difference:** $y_t = \log(Y_t) - \log(Y_{t-1})$
  - **Percentage:** $y_t = \frac{Y_{t-1}}{Y_t} \times 100$

Related Concepts and Notations

Assumptions

1. **Uniform:** We consider only consecutive, evenly spaced observations
   - For example, say monthly data in 2010-2021 for each year, and without any missing month(s); no other data, for example, on daily basis for a year is admissible.

2. **Stationarity:** A time series Y_t is stationary if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{i+1}, Y_{i+2}, Y_{i+3}, \ldots\ldots Y_{i+T})$ does not depend on i.

   - Stationary property implies that history is relevant. In other words, stationary requires the future to be like the past (in a probabilistic sense).

- Auto-regression analysis assumes that Y_t is **both** uniform and stationary.

Autocorrelation coefficient

Autocorrelation

The correlation of a series with its own lagged values is called autocorrelation (also called serial correlation)

Formula: j^{th} Autocorrelation

The j^{th} autocorrelation, denoted by ρ_j is defined as

$$\rho_j = \frac{COV(Y_t, Y_{t-j})}{\sqrt{\sigma_{Y_t}\sigma_{Y_{t-j}}}}$$

where, $COV(Y_t, Y_{t-j})$ is the j^{th} **auto-covariance**

Covariance

Formula: $COV(Yt, Yt\_j)$

The covariance between the variables Y_t and Y_{t-j}

$$COV(Y_t, Y_{t-j}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

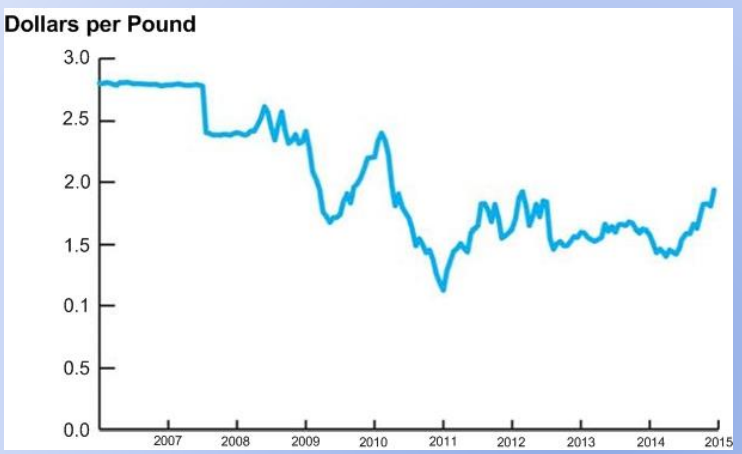
Y_{t-j}	...	Y_t
x_1		y_1
x_2		y_2
...		...
x_j		y_j
.		.
.		.
.		.
x_n		y_n

σ_X is the variance for the variable X

n is the number of observations

Example: Autocorrelation

Example



- For the given data, say $\rho_1 = 0.84$ between two given consecutive years
  - This implies that the Dollars per Pound is highly serially correlated
- Similarly, we can determine $\rho_2, \rho_3 \ldots$ etc.

Auto-Regression Model

Auto-Regression Model

Definition

An autoregressive model (also called AR model) is used to model a future behavior for a time-ordered data, using data from past behaviors.

- Essentially, it is a linear regression analysis of a dependent variable using one or more variables(s) in a given time-series data.

$$Y_t = f(Y_{t-1}, Y_{t-2}, \ldots\ldots, Y_{t-p})$$

Auto-Regression Model for Forecasting

Definition

- A natural starting point for forecasting model is to use past values of Y, that is, $Y_{t-1}, Y_{t-2}, \ldots\ldots$ to predict Y_t
- An auto-regression is a regression model in which Y_t is regressed against its own lagged values.
- The number of lags used as regressors is called the **order** of auto-regression
  - In first order auto-regression (denoted as AR(1)), Y_t is regressed against Y_{t-1}
  - In p^{th} order auto-regression (denoted as AR(p)), Y_t is regressed against, $Y_{t-1}, Y_{t-2}, \ldots\ldots, Y_{t-p}$

p^{th} order Auto-regression Model

Formula: p^{th} Order Auto-regression Model

In general, the p^{th} order auto-regression model is defined as

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t$$

where, $\beta_0, \beta_1, \ldots, \beta_p$ is called auto-regression coefficients and ε_t is the noise term or residue and in practice it is assumed to Gaussian white noise

For example, *AR(1)* is $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$

The task in AR analysis is to derive the "best" values for $\beta_i, i = 0, 1, \ldots, p$ given a time series$[Y_1, Y_2, \ldots YT_1, Y_T]$

Computing AR Coefficients

Computing AR(p) model

- A number of techniques known for computing the AR coefficients
- The most common method is called Least Squares Method (LSM)
- The LSM is based upon the **Yule-Walker equations**

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \cdots & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \cdots & \rho_{p-3} & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \cdots & \rho_{p-4} & \rho_{p-3} \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \cdots & \rho_{p-5} & \rho_{p-4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \rho_{p-4} & \rho_{p-5} & \cdots & \cdots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \vdots \\ \rho_{p-1} \\ \rho_p \end{bmatrix}$$

Computing AR Coefficients

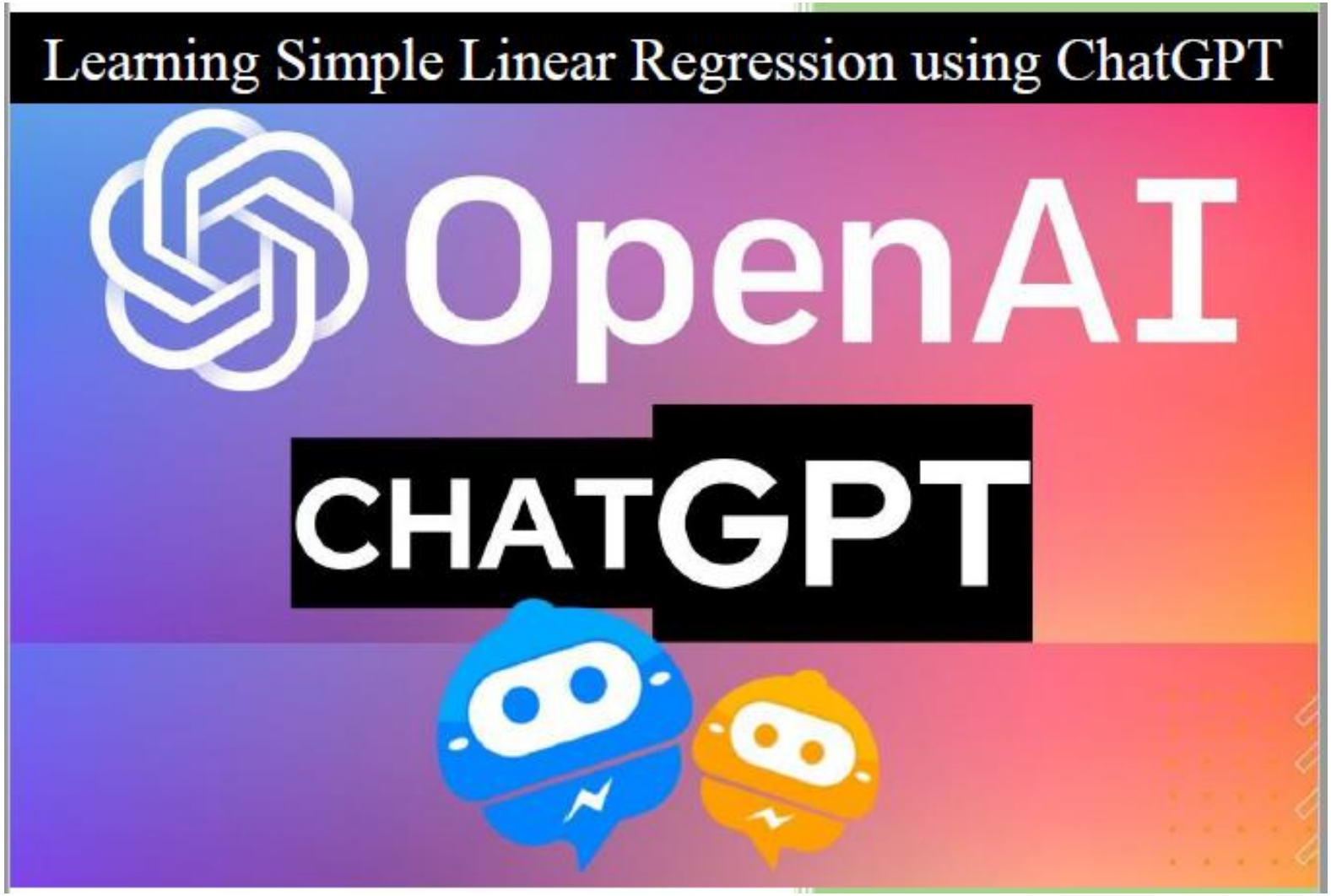
Computing AR (p): Yule-Walker Equations

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \cdots & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \cdots & \rho_{p-3} & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \cdots & \rho_{p-4} & \rho_{p-3} \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \cdots & \rho_{p-5} & \rho_{p-4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \rho_{p-4} & \rho_{p-5} & \cdots & \cdots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \vdots \\ \rho_{p-1} \\ \rho_p \end{bmatrix}$$

- Here, $\rho_i, i = 1, 2, \ldots, p$ denotes the i^{th} **auto correlation coefficient**.
- β_0 can be chosen empirically, usually taken as zero.

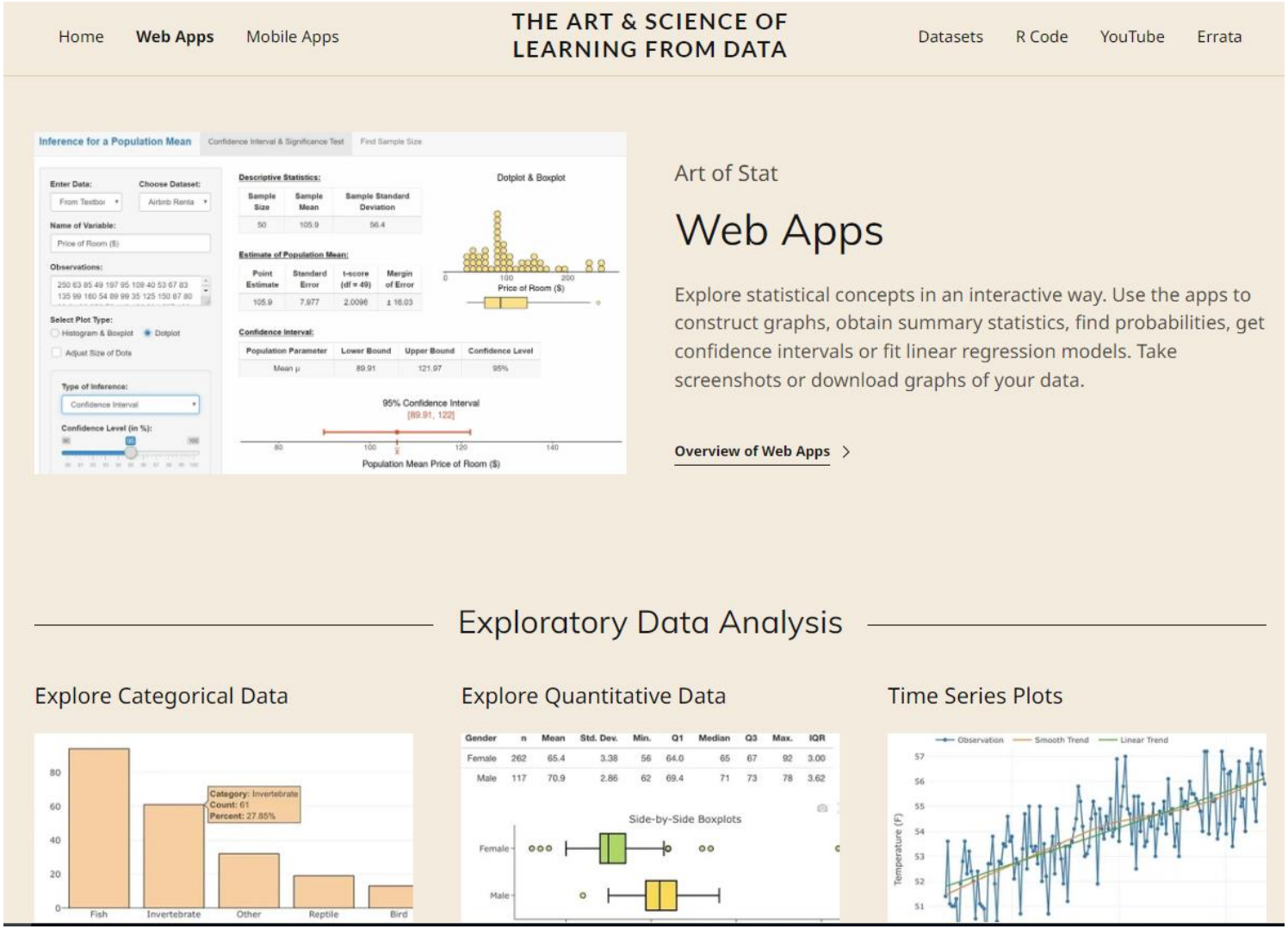
Web Applications

Demo : ChatGPT



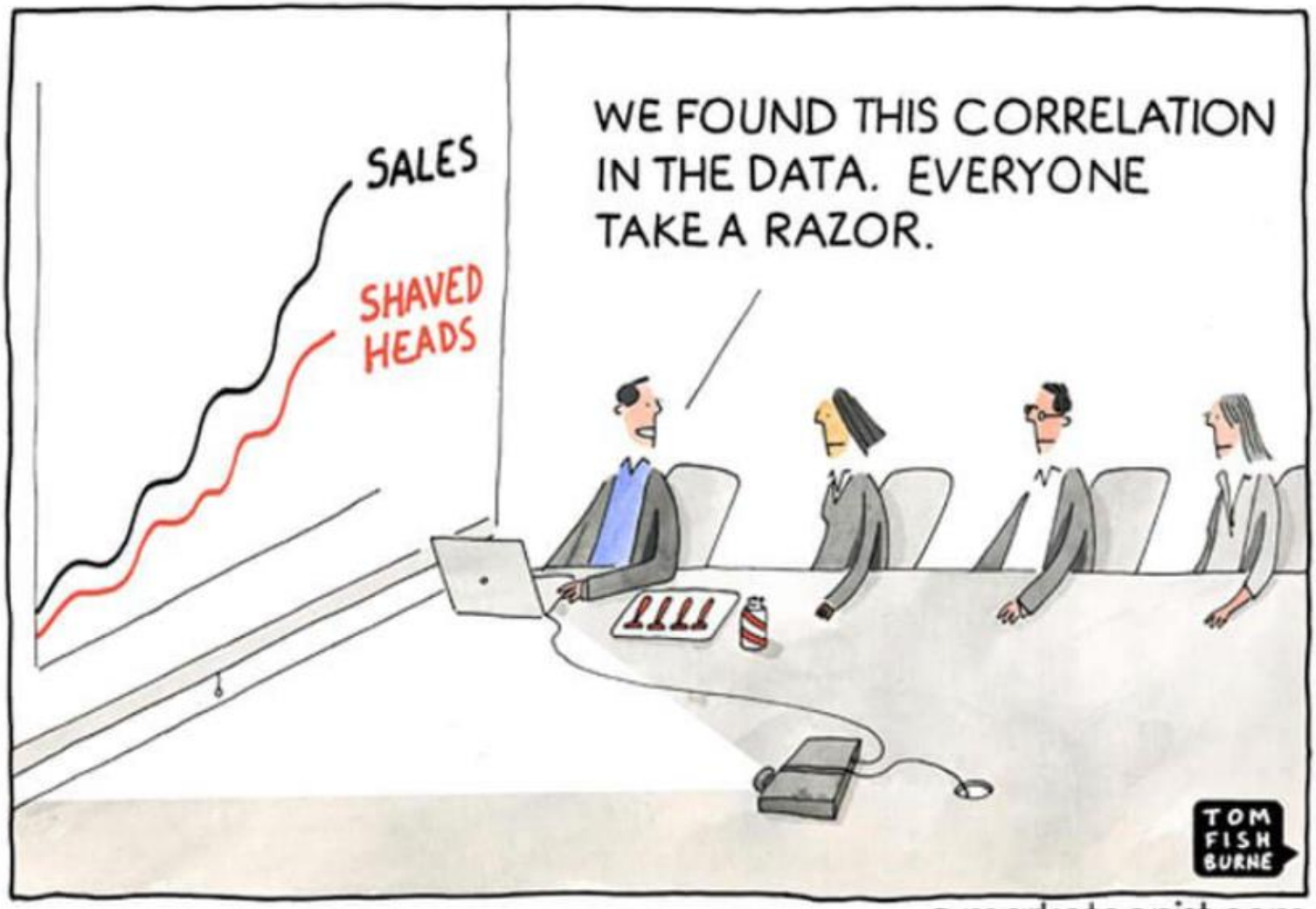
Link: https://chat.openai.com/chat

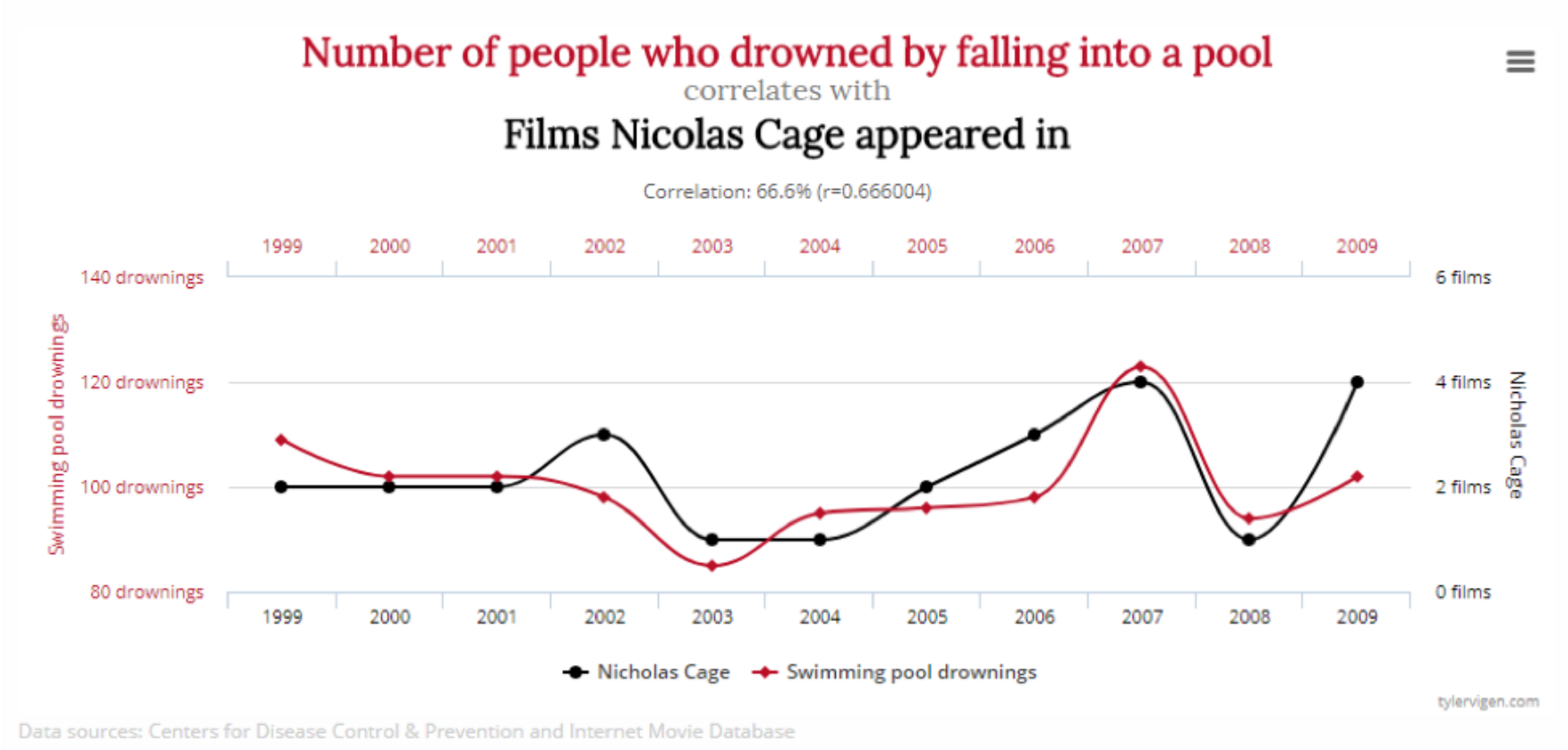
Demo : Art of Statistics



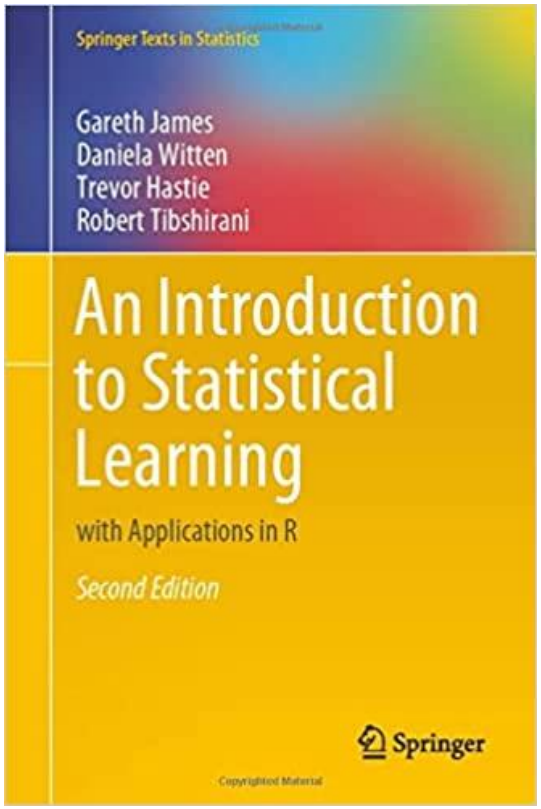
Link: https://artofstat.com/web-apps

Correlation does NOT imply Causation


SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE
©marketoonist.com


Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
Correlation: 66.6% (r=0.666004)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com
Data sources: Centers for Disease Control & Prevention and Internet Movie Database

References


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer

End of Course