Data Analytics (CS40003)

Lecture #7

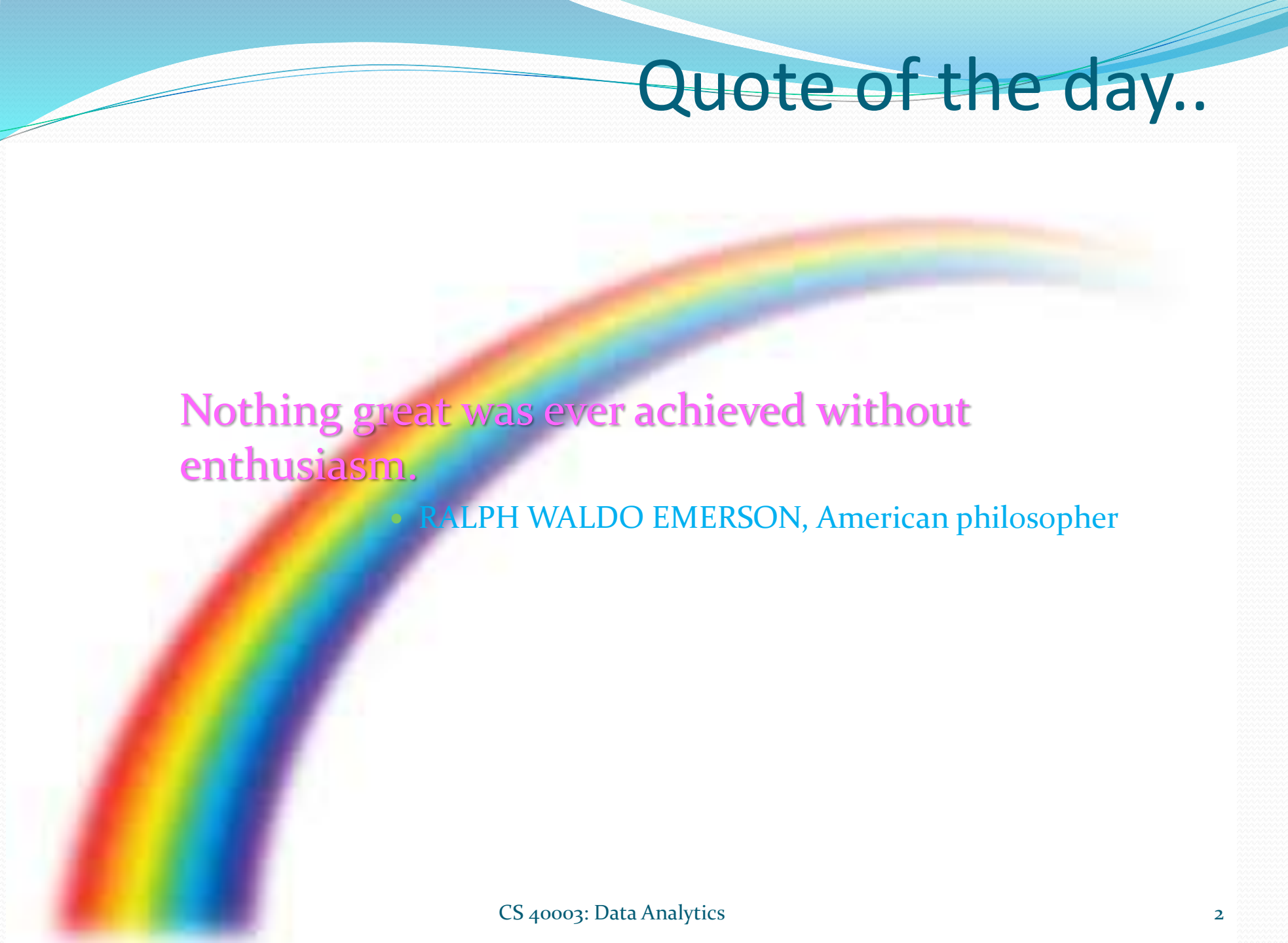
Relation Analysis

Dr. Debasis Samanta

Associate Professor

Department of Computer Science & Engineering

Quote of the day..



Nothing great was ever achieved without enthusiasm.

- RALPH WALDO EMERSON, American philosopher

This presentation includes…

- Introduction
- Measures of Relationship
- Correlation Analysis
  - χ^2 - Test
  - Spearman's Correlation Analysis
  - Pearson's Correlation Analysis
- Regression Analysis
  - Simple Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression Analysis
- Auto-Regression Analysis

Hypothesis Testing Strategies

- There are two types of tests of hypotheses
  - ✓Parametric tests (also called standard test of hypotheses).
  - ❑ Non-parametric tests (also called distribution-free test of hypotheses)

Parametric Tests : Applications

- Usually assume 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.

Hypothesis Testing : Non-Parametric Test

- ***Non-Parametric tests***
  - Does not under any assumption
  - Assumes only nominal or ordinal data

Note: Non-parametric tests need entire population (or very large sample size)

Relationship Analysis

- **Example: Wage Data**

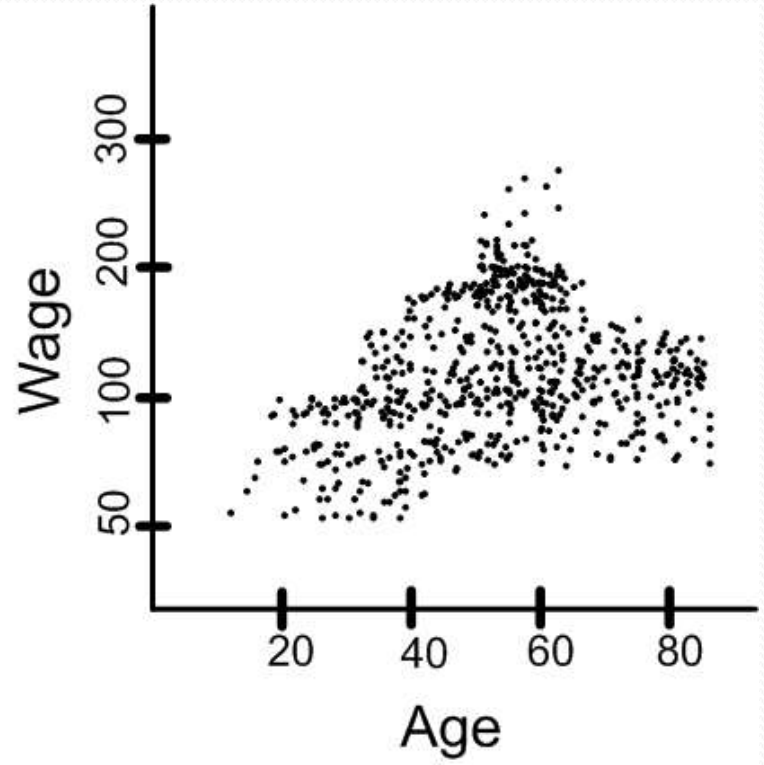
  A large data regarding the wages for a group of employees from the eastern region of India is given.

  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:

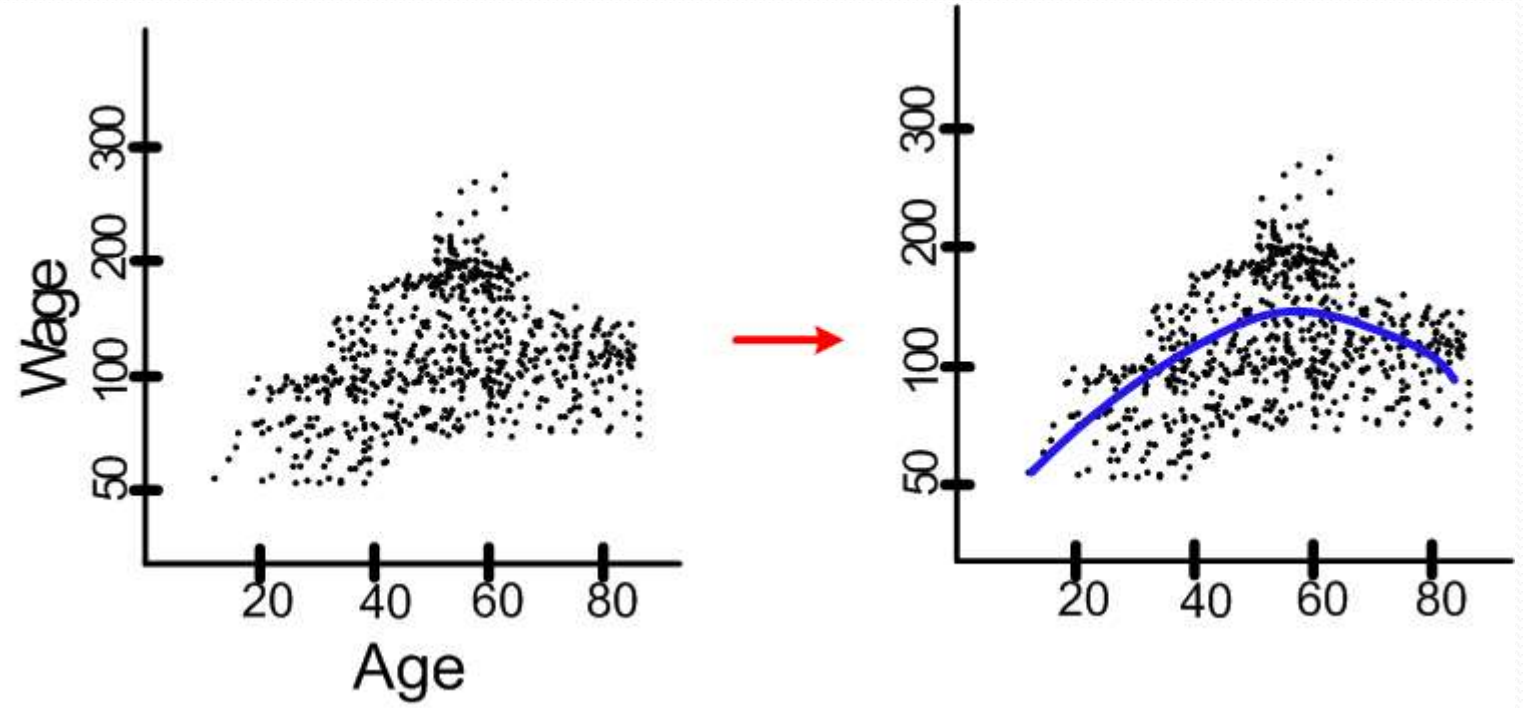


?

How wages vary with ages?

Relationship Analysis

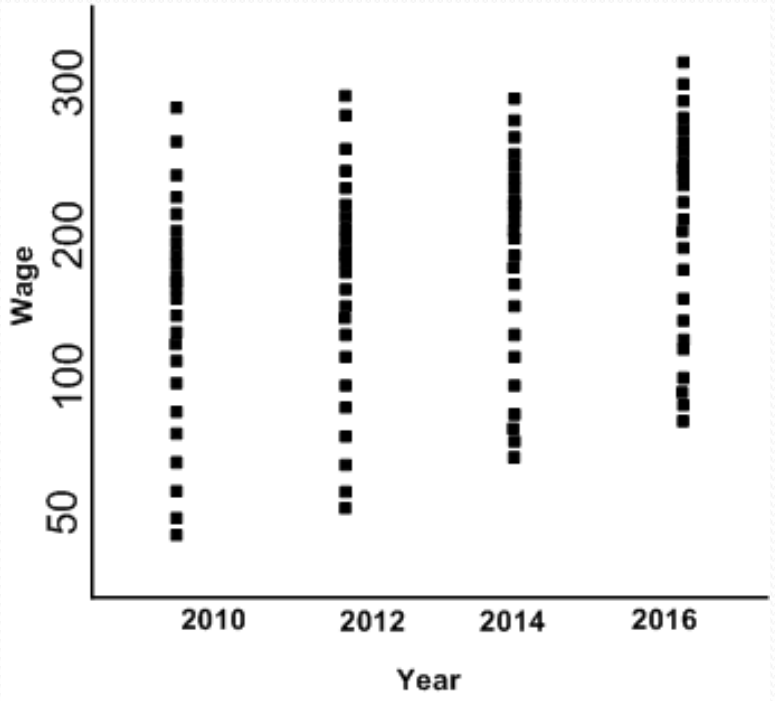
- Example: Wage Data
  - *Employee's age and wage:* How wages vary with ages?



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:

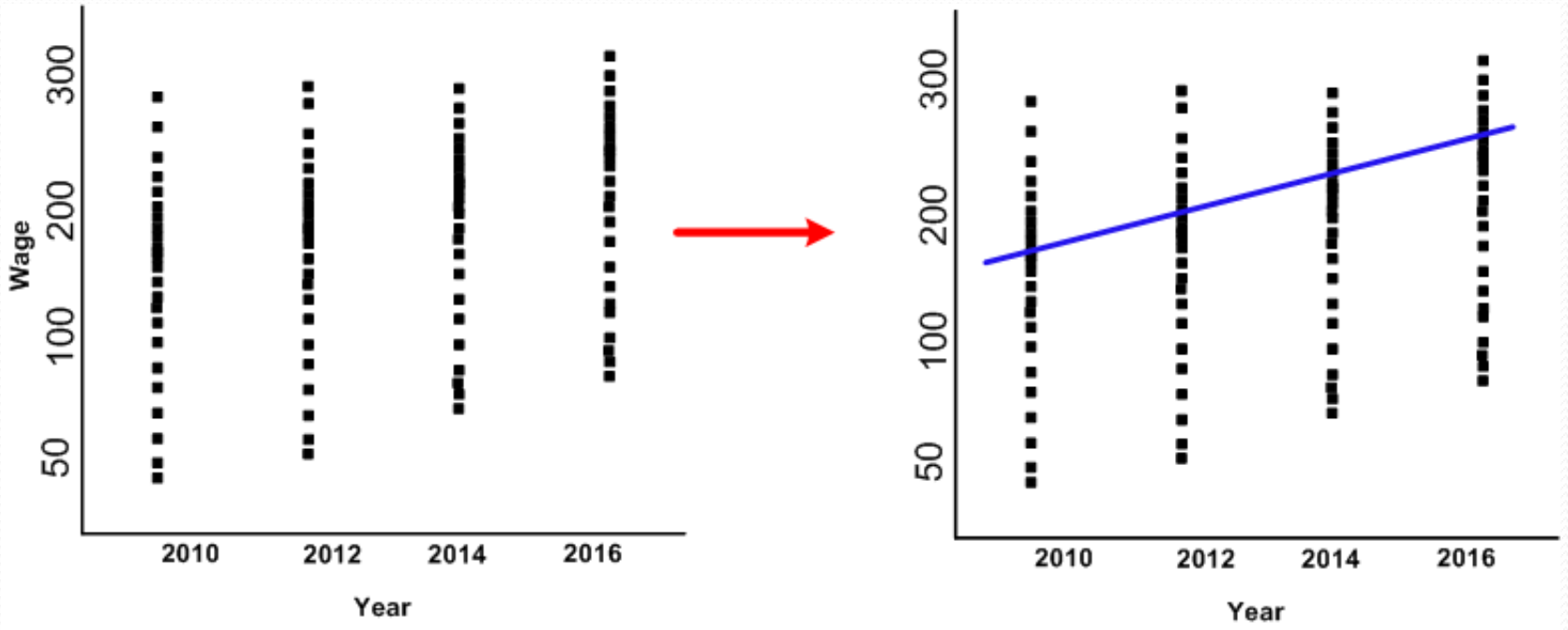


?

How wages vary with time?

Relationship Analysis

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

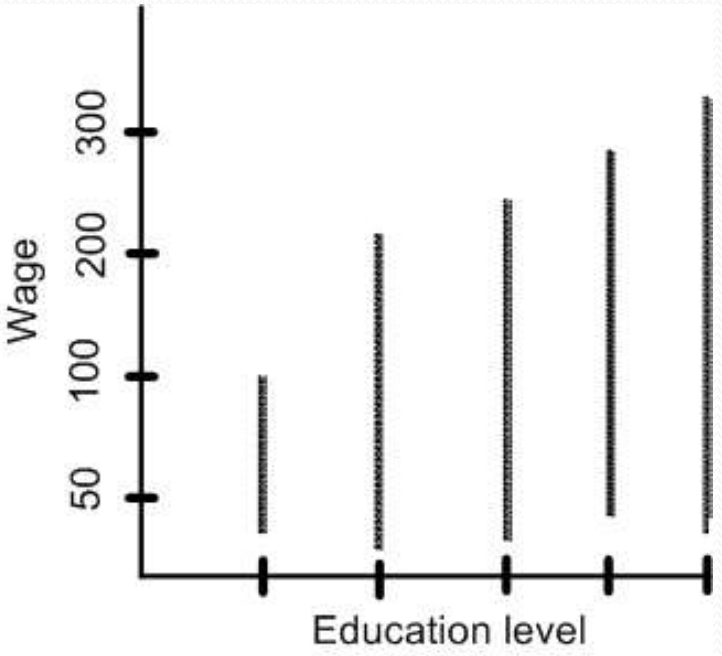


Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

.

Relationship Analysis

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

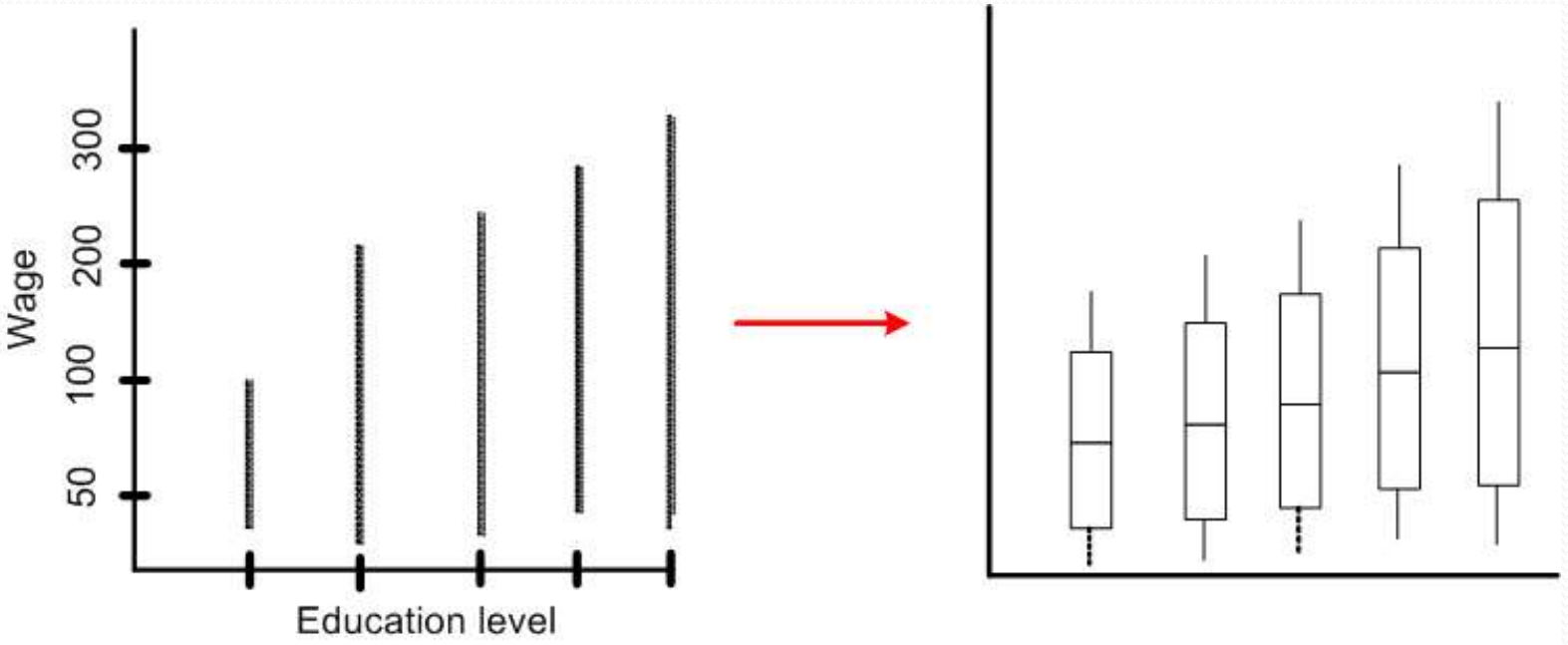


?

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.

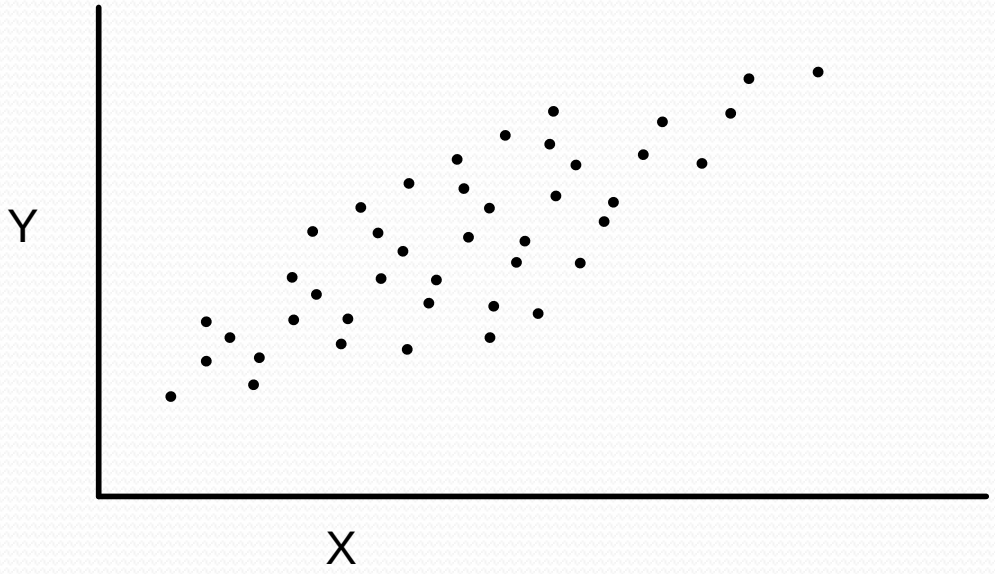
Relationship Analysis

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

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

etc….

An Open Challenge!

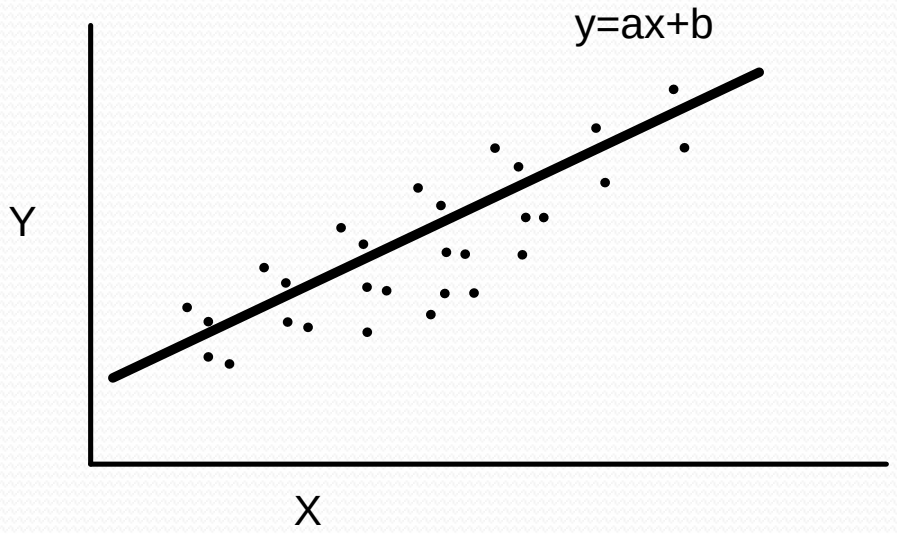


Suppose there are countably infinite points in the *XY plane*. We need a huge memory to store all such points.

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:* The population consisting of only one variable.

Temperature	20	30	21	18	23	45	52

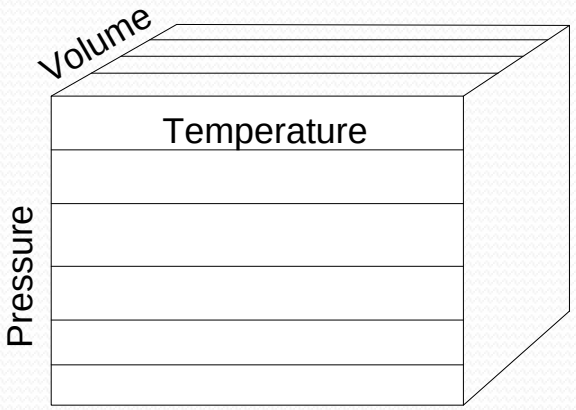
Here, statistical measures are suffice to find a relationship.

- *Bivariate population:* Here, the data happen to be on two variables.

Pressure	1	1.1	… …	0.8
Temperature	35	41		29

Measures of Relationship

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



? 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 correlation (i.e., association) between two (or more) variables?

If yes, of what degree?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or 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?

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**

Example:

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	… …

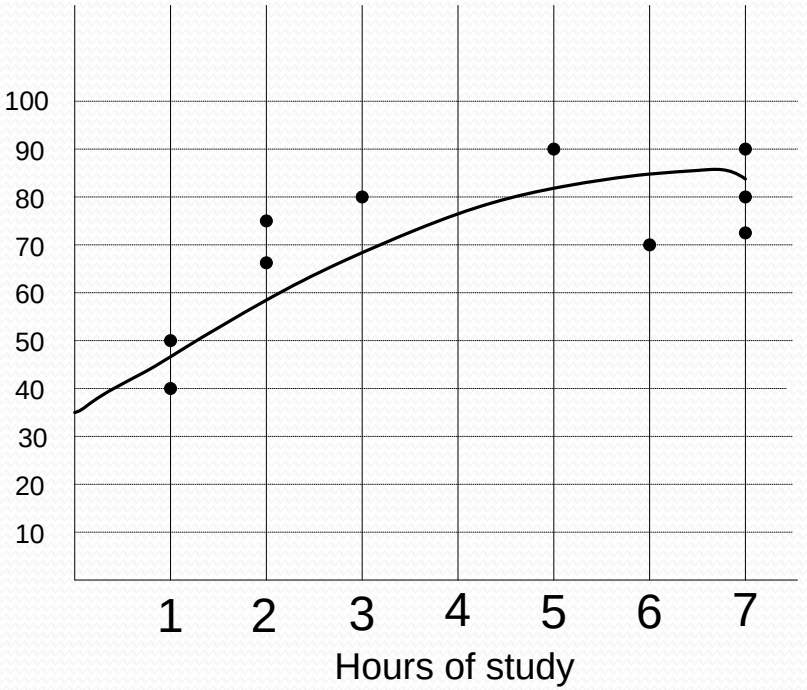
The correlation may be positive, negative or zero.

- **Positive correlation:** If the value of the attribute *A* increases with the increase in the value of the attribute *B* and vice-versa.
- **Negative correlation:** If the value of the attribute *A* decreases with the increase in the value of the attribute *B* and vice-versa.
- **Zero correlation:** When the values of attribute *A* varies at random with *B* and vice-versa.

Correlation Analysis

- In order to measure the degree of correlation between two attributes.

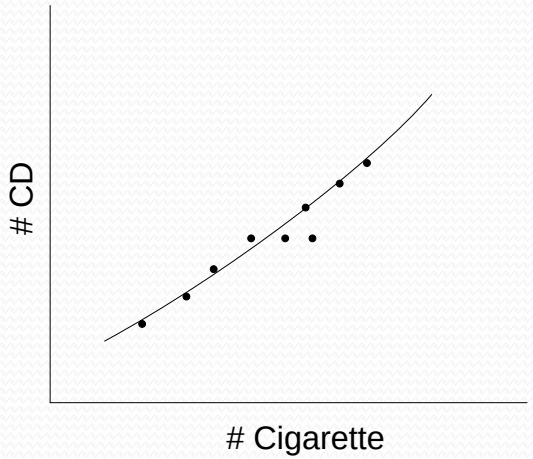
Hours Study	*Exam Score*
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100



Correlation Analysis

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

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

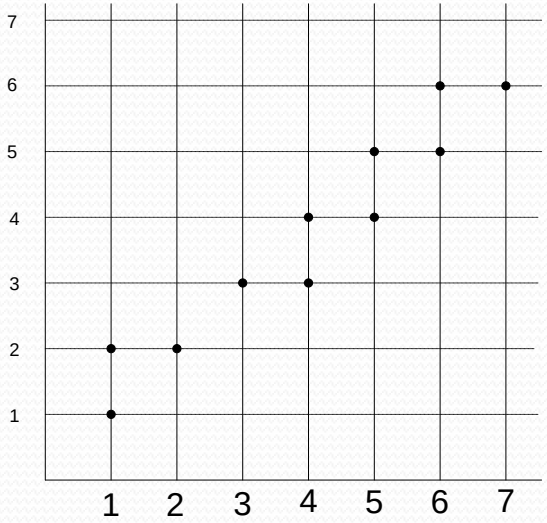


Note:

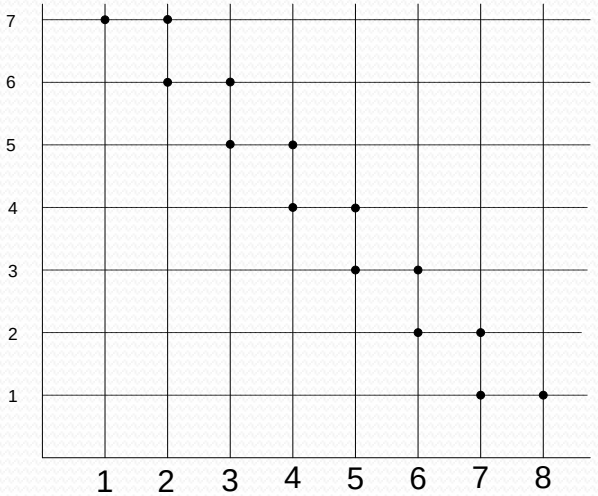
In data analytics, correlation analysis make sense only when relationship make sense. There should be a cause-effect relationship.

Correlation Analysis

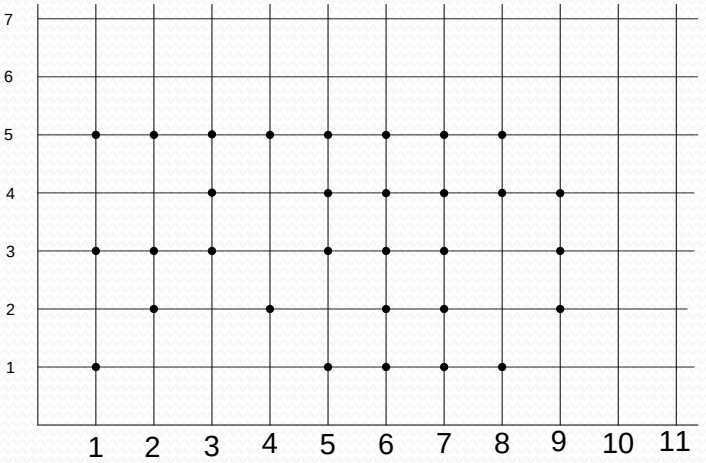
Positive correlation



Negative correlation



Zero correlation

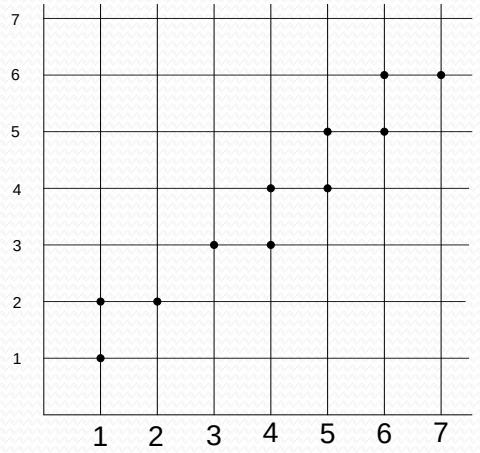


Correlation Coefficient

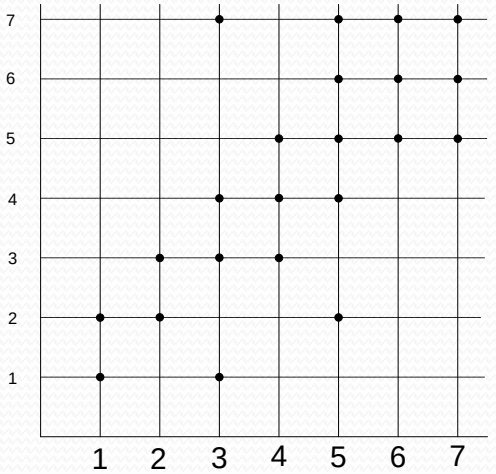
- **Correlation coefficient is used to measure the degree of association.**
- It is usually denoted by *r*.
- The value of *r* lies between +1 and -1.
- Positive values of *r* indicates positive correlation between two variables, whereas, negative values of *r* indicate negative correlation.
- $r = +1$ 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

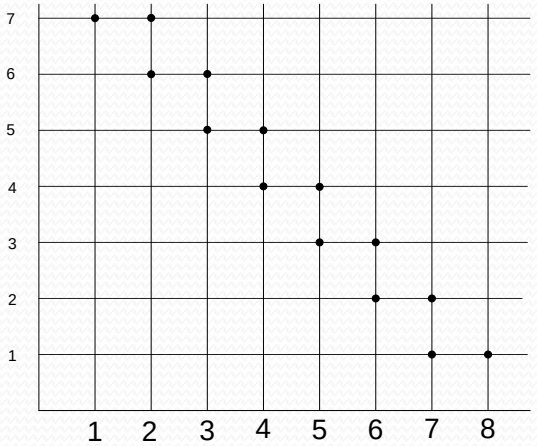
High Positive Correlation



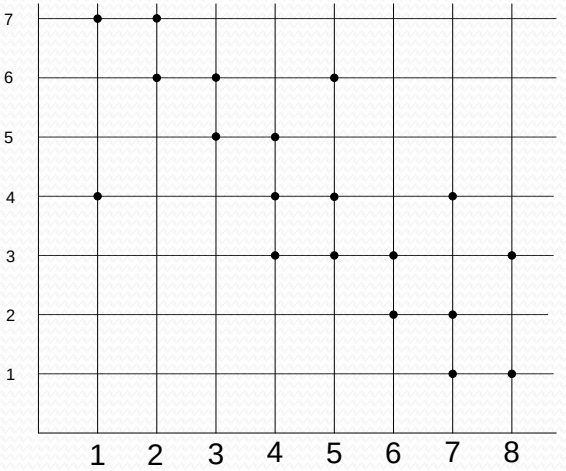
Low Positive Correlation



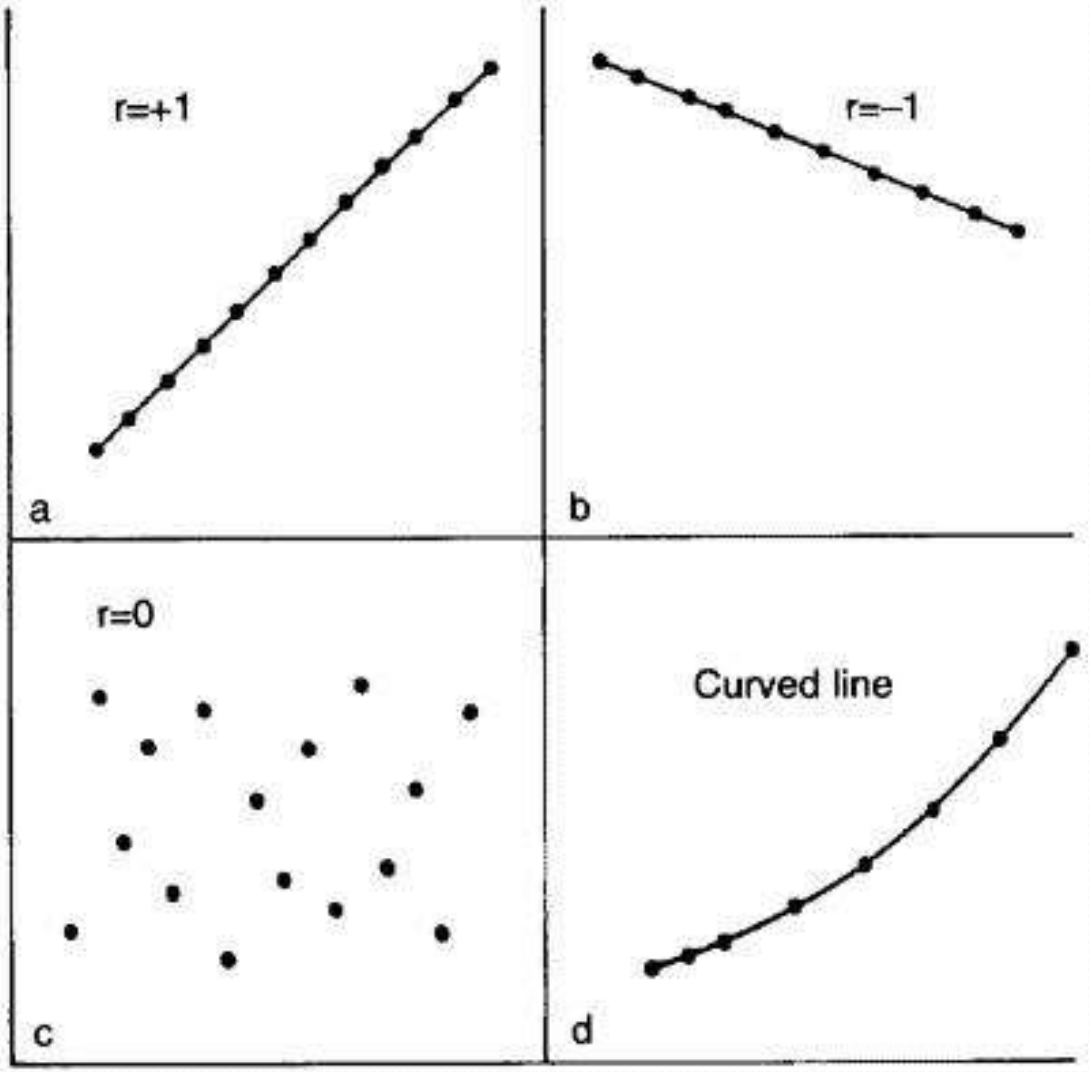
High Negative Correlation



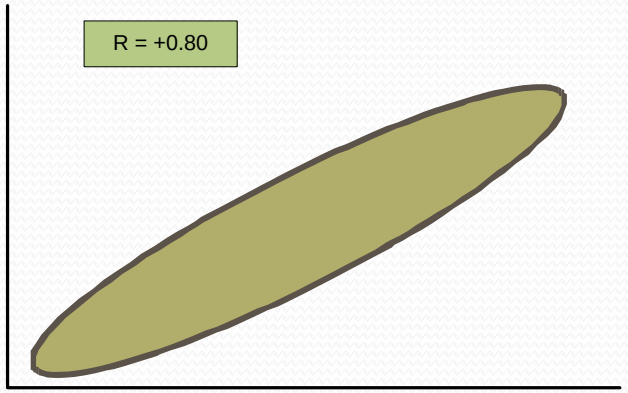
Low Negative Correlation

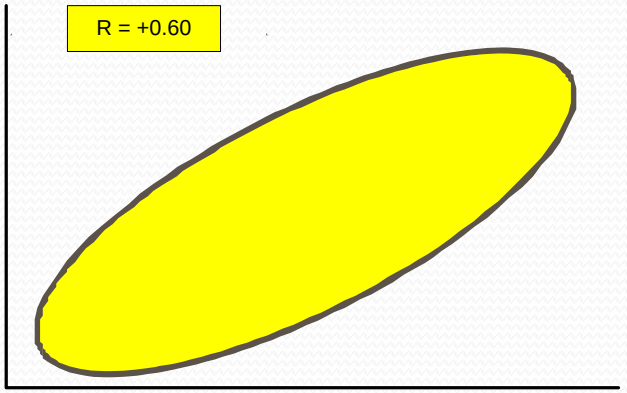


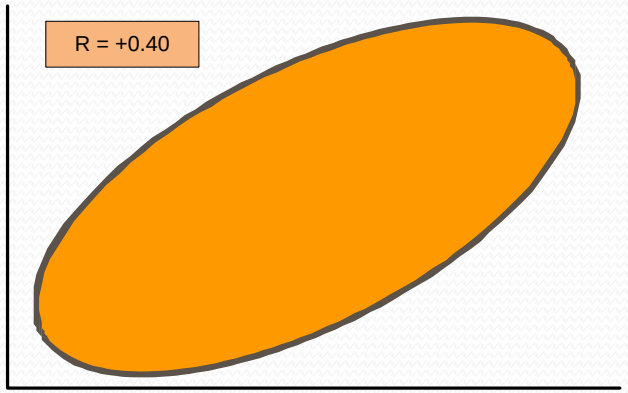
Correlation Coefficient

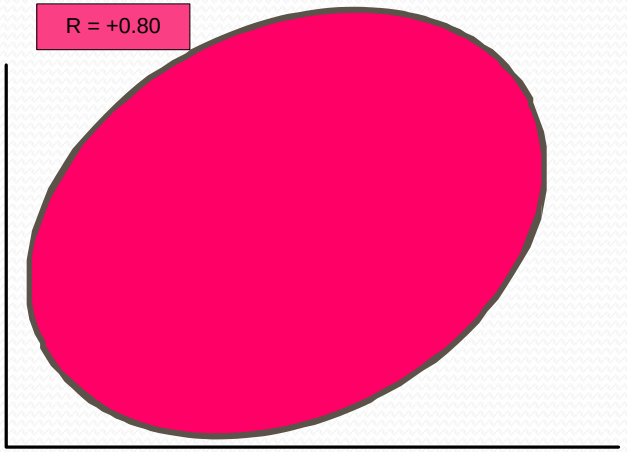


Correlation Coefficient









Measuring Correlation Coefficients

- There are three methods known to measure the correlation coefficients
  - Karl Pearson's coefficient of correlation
    - This method is applicable to find correlation coefficient between two **numerical** attributes
  - Charles Spearman's coefficient of correlation
    - This method is applicable to find correlation coefficient between two **ordinal** attributes
  - Chi-square coefficient of correlation
    - This method is applicable to find correlation coefficient between two **categorical** attributes

Pearson's Correlation Coefficient

Karl Pearson's Correlation Coefficient

- This is also called **Pearson's Product Moment Correlation**

Definition 7.1: 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

$$r^* = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X . \sigma_Y}$$

where $X_i = \text{i} - \text{th value of } X - \text{variable}$

$\bar{X} = \text{mean of } X$

$Y_i = \text{i} - \text{th value of } Y - \text{variable}$

$\bar{Y} = \text{mean of } Y$

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

$\sigma_X = \text{standard deviations of } X$

$\sigma_Y = \text{standard deviation of } Y$

Karl Pearson's coefficient of Correlation

Example 7.1: Correlation of Gestational Age and Birth Weight

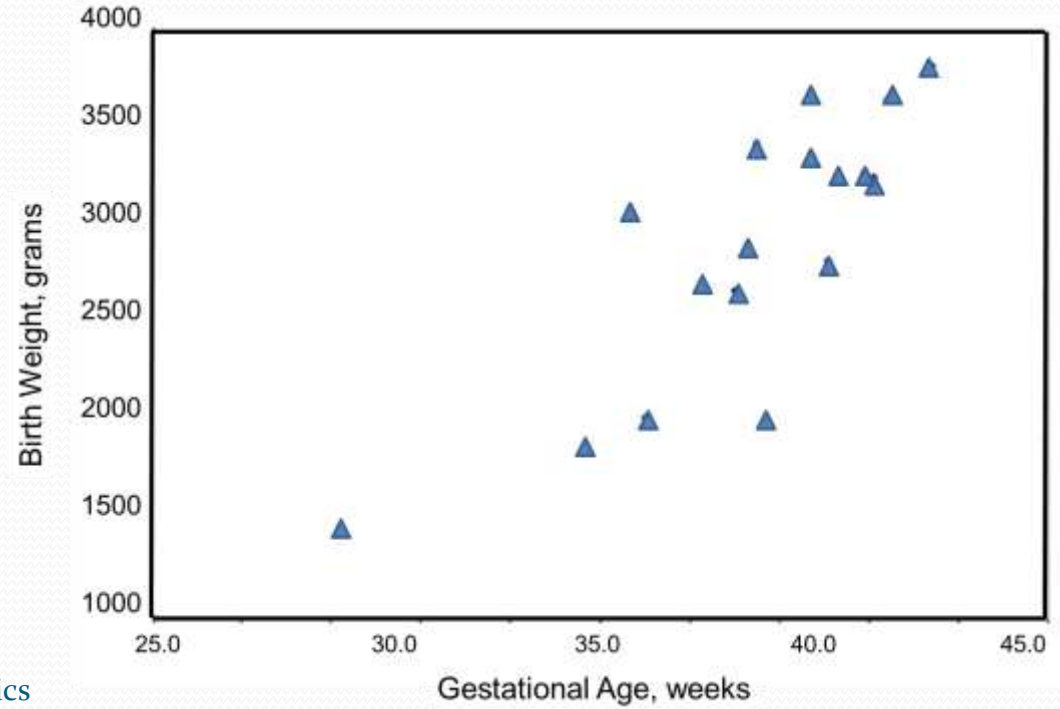
- A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

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 7.1: Correlation of Gestational Age and Birth Weight

- We wish to estimate the association between gestational age and infant birth weight.
- In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus Y = birth weight and X = gestational age.
- The data are displayed in a scatter diagram in the figure below.



Karl Pearson's coefficient of Correlation

Example 7.1: Correlation of Gestational Age and Birth Weight

- For the given data, it can be shown the following

$$\overline{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{49{,}334}{17} = 2902.$$

$$s_x^2 = \frac{\Sigma(X - \overline{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

$$s_y^2 = \frac{\Sigma(Y - \overline{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

$$r^* = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{s_x . s_y} = 0.82$$

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

Karl Pearson's coefficient of Correlation

Example 7.1: Correlation of Gestational Age and Birth Weight

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

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

  - 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 may be regarded as highly significant.

Rank Correlation Coefficient

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:

Height: [VS S L T VT]
1 2 3 4 5

T – shirt: [XS S L XL XXL]
11 12 13 14 15

Rank assigned

Charles Spearman's Correlation Coefficient

Definition 7.2: Charles Spearman's correlation coefficient

The rank correlation can be defined as

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

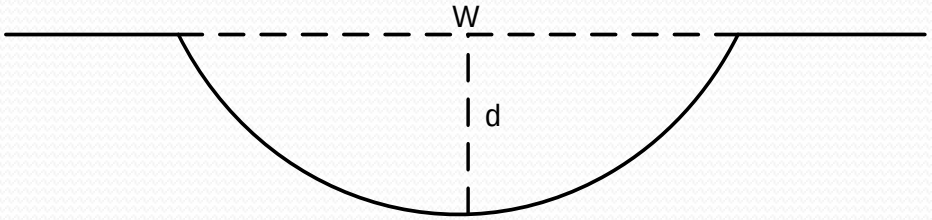
where d_i = Difference between ranks of i^{th} pair of the two variables

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 7.2: The hypothesis that the depth of a river **does not progressively increase** with the width of the river.



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

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

$$\sum d^2 = 4$$

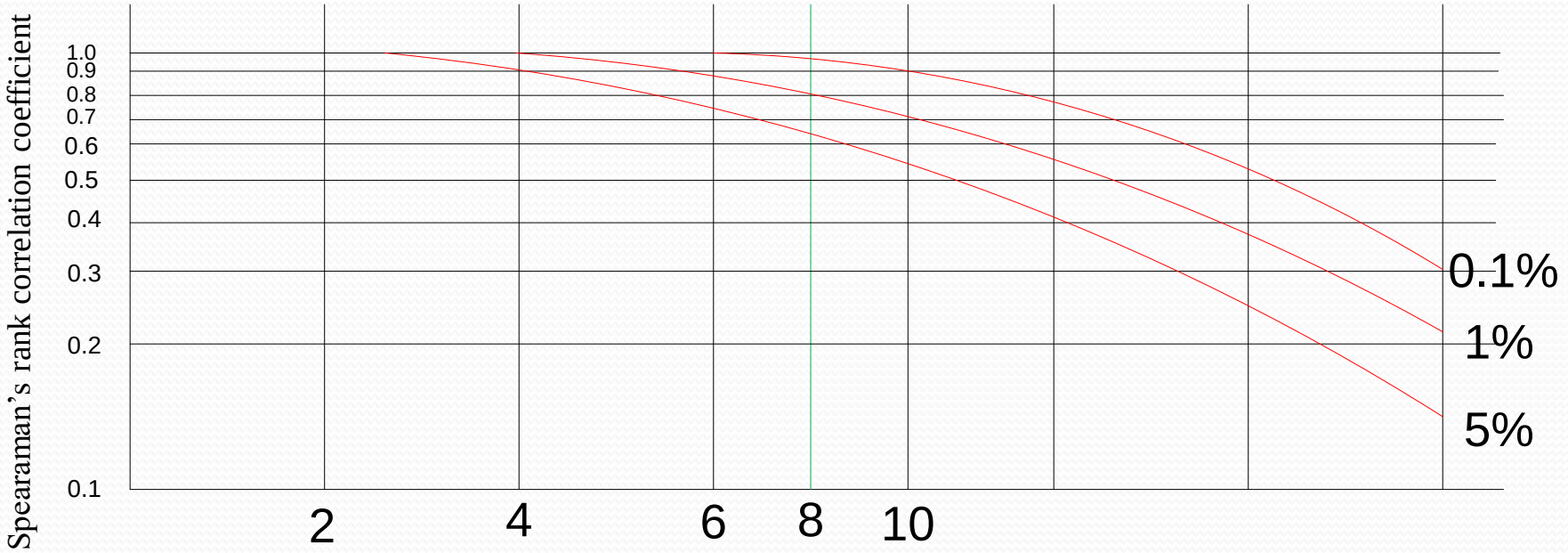
$$r_s = 0.9757$$

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 0.01% 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 with the width of the river.

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

χ^2 – Test

Definition 7.3: 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 7.3: 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 7.3: 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 7.3: 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		
		Male	Female	Total
HOBBY	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.

χ^2 – Test

Example 7.4: Hypothesis on "accident proneness" versus "driver's handedness".

- Consider the following contingency table on car accidents among left and right-handed drivers' of sample size 175.
- Hypothesis is that *"fatality of accidents is independent of driver's handedness"*

FATALITY	HANDEDNESS		
	Left-Handed	Right-Handed	Total
Non-Fatal	8	141	149
Fatal	3	23	26
Total	11	164	175

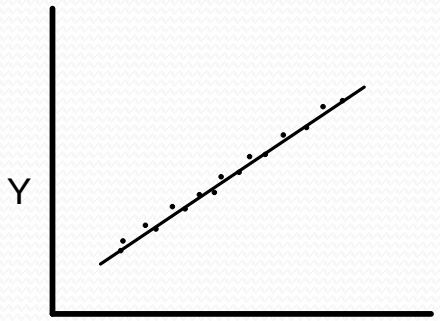
- Find the correlation between Fatality and Handedness and test the significance of the correlation with significance level 0.1%.

Regression Analysis

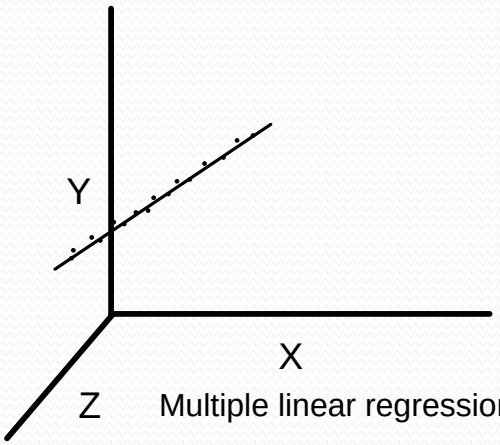
Regression Analysis

- 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 variables.

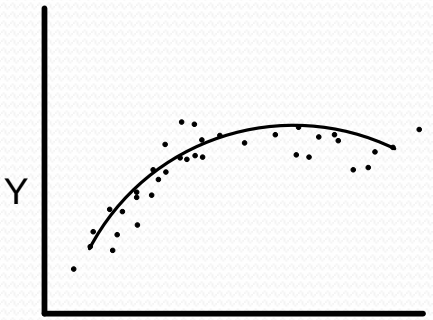
- **Classification of Regression Analysis Models**
  - Linear regression models
    1. Simple linear regression
    2. Multiple linear regression
  - Non-linear regression models



Simple linear regression



Multiple linear regression

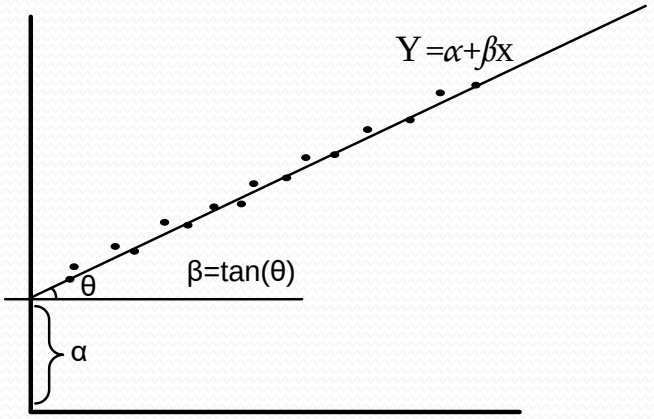


Non-linear regression

Simple Linear Regression Model

In simple linear regression, we have only two variables:

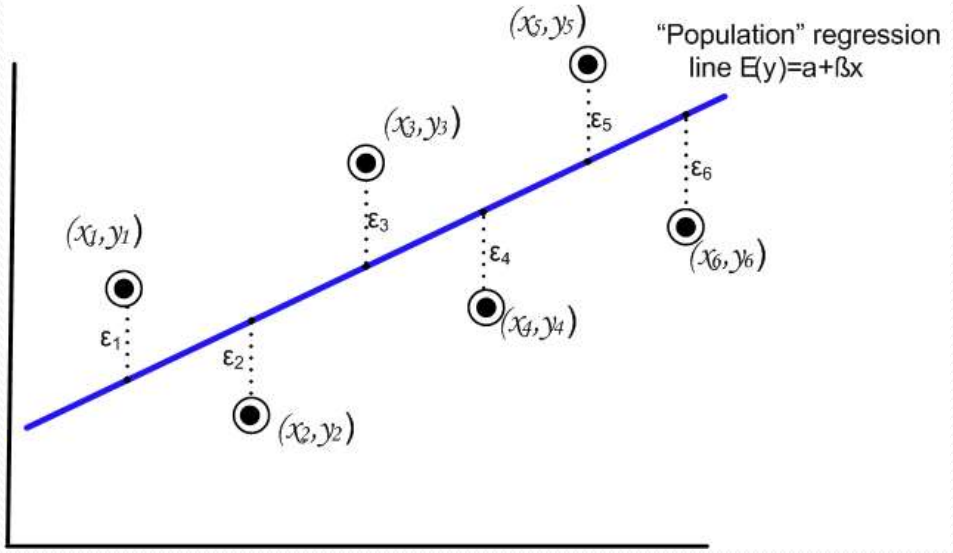
- Dependent variable (also called Response), usually denoted as Y.
- Independent variable (alternatively called Regressor), usually denoted as x.
- A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$



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 $[(x_i, y_i), i = 1,2, \ldots \ldots, n]$ of data involving n pairs of (x, y) values, our objective is 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**.

Note:

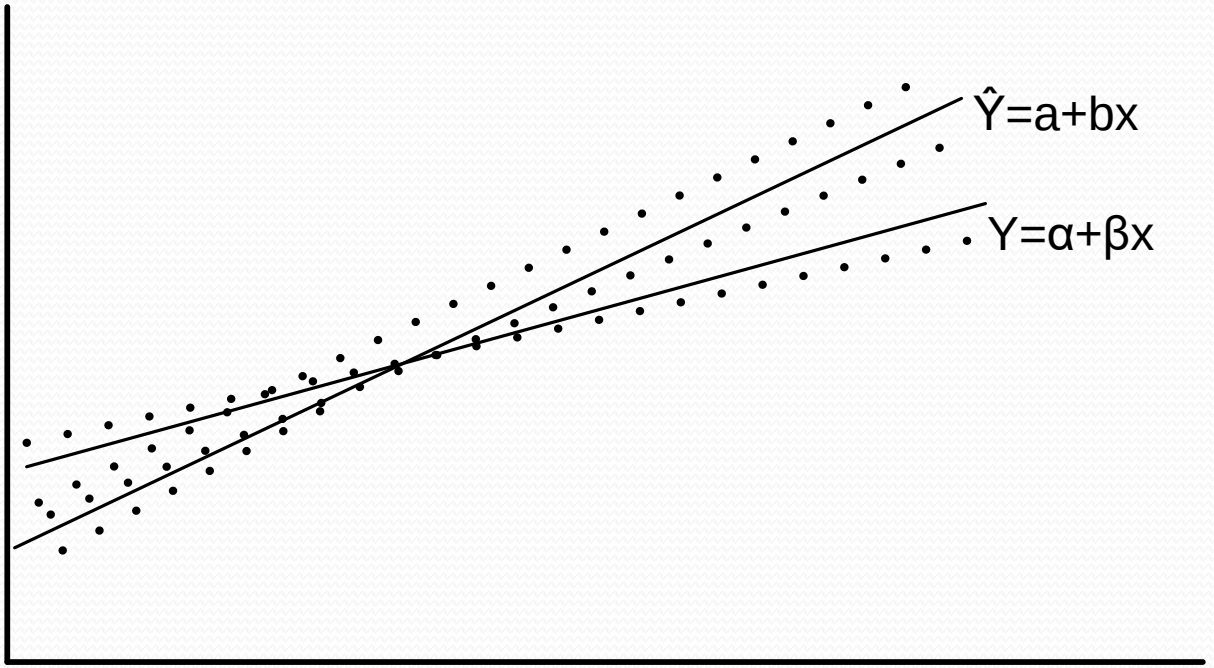
- $E(\in) = 0$ implies that at a specific x, the y values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- α and β are called **regression coefficients**.
- α and β values are to be estimated from the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

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

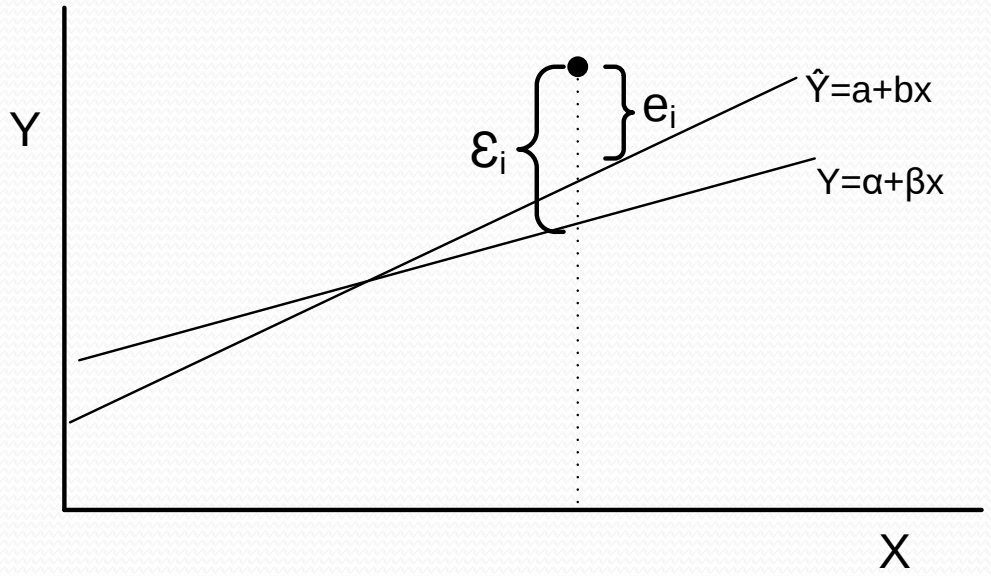
where $\hat{Y}$ is the predicted or fitted value.



Least Square Method to estimate α and β

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

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



Least Square method

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

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

- We are to minimize the value of SSE and hence to determine the parameters of *a* and *b*.
- Differentiating SSE with respect to *a* and *b*, we have

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

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

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

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

Least Square method to estimate α and β

Thus we set

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

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

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

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

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

R^2 : Measure of Quality of Fit

- A quantity R^2, is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.
- We have $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$
- It signifies the **variability due to error**.
- Now, let us define the total corrected sum of squares, defined as

$$\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

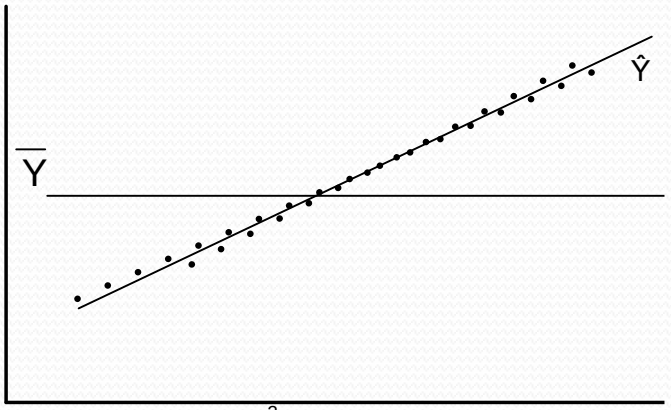
- SST represents the variation in the response values. The R^2 is

$$R^2 = 1 - \frac{SSE}{SST}$$

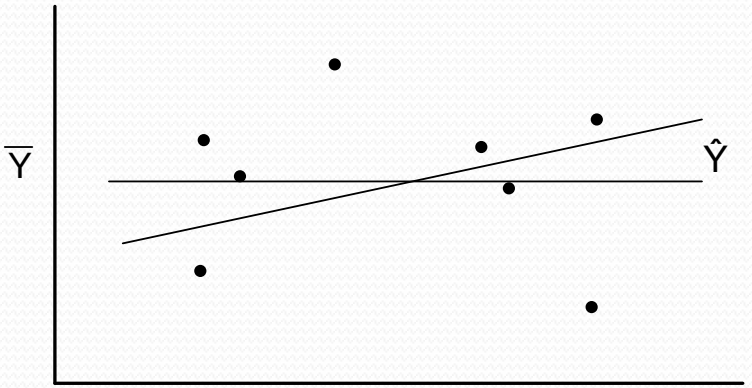
Note:

- 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)

R^2 : Measure of Quality of Fit



$R^2 \approx 1.0$ (Very good fit)



$R^2 \approx 0$ (Very poor fit)

Multiple Linear Regression

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

Let the data points given to us is

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

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 $x_1,\ x_2,\ x_3, \dots\dots\dots\dots\dots\dots, x_k$.

Thus,

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

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 $\hat{y}_i$.

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

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

Multiple Linear Regression

- 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.

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

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

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

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

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

- 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.
- Hence, the multiple linear regression model can be built.

Non Linear Regression Model

- When the regression equation is in terms of *r*-degree, *r*>1, then it is called nonlinear regression model. When more than one independent variables are there, then it is called Multiple Non linear Regression model. Also, 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

Given that (x_i, y_i); i = 1,2,…,n are n pairs of observations. Each observations would satisfy the equations:

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

and

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

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.

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.

Auto-Regression Analysis

Auto Regression Analysis

- Regression analysis for time-ordered data is known as Auto-Regression Analysis
- Time series data are data collected on the same observational unit at multiple time periods

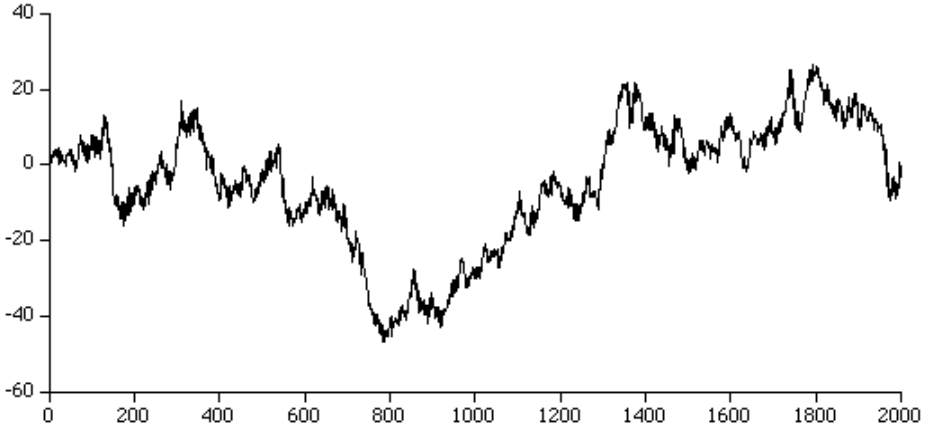
Example: Indian rate of price inflation

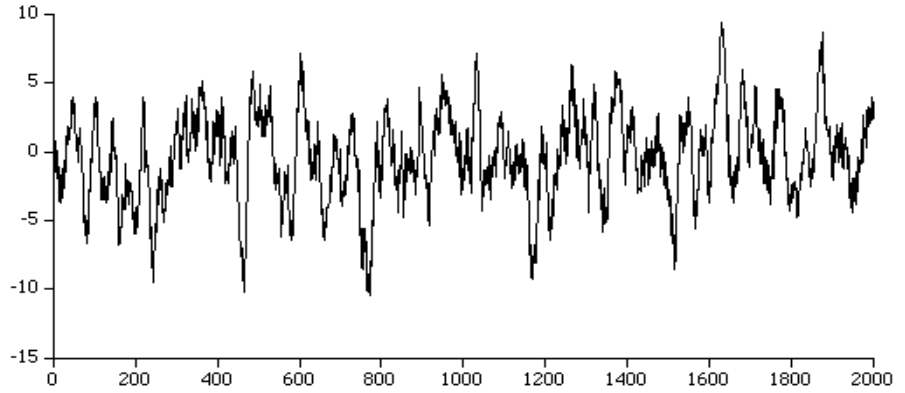
Auto Regression Analysis

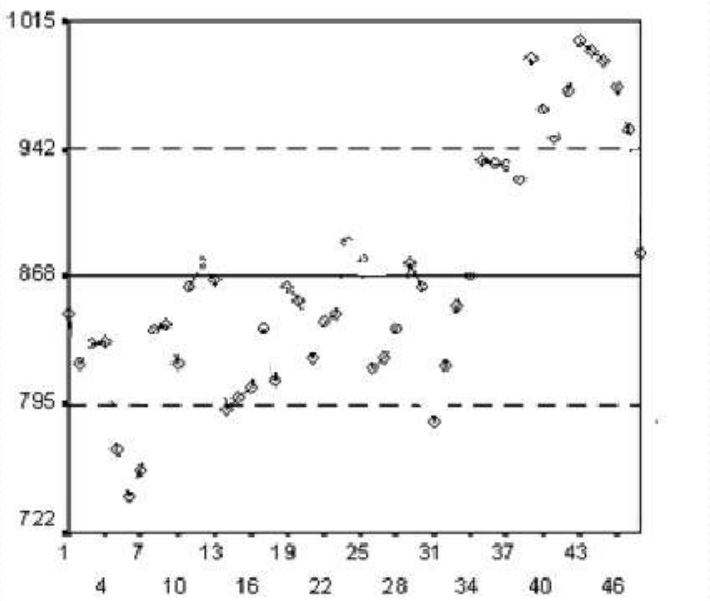
- **Examples:** Which of the following is a time-series data?
  - Aggregate consumption and GDP for a country (for example, 20 years of quarterly observations = 80 observations)
  - Yen/$, pound/$ and Euro/$ exchange rates (daily data for 1 year = 365 observations)
  - Cigarette consumption per capita in a state, by years
  - Rainfall data over a year
  - Sales of tea from a tea shop in a season

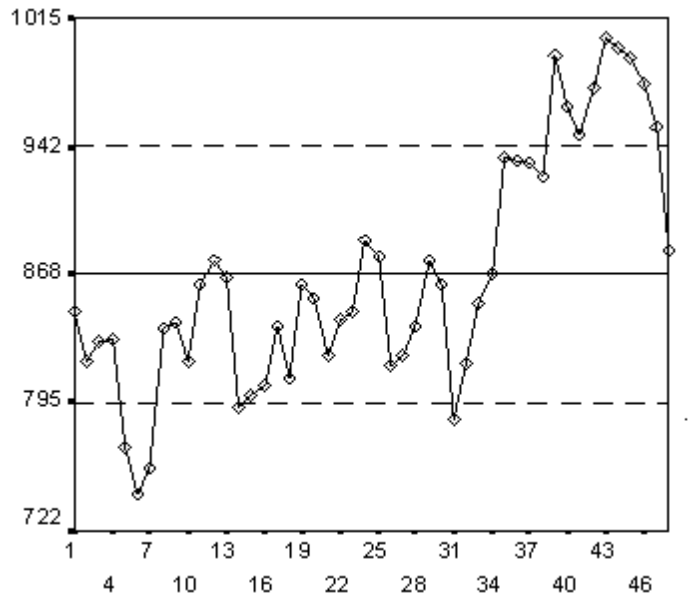
Auto Regression Analysis

- **Examples:** Which of the following graph is due to time-series data?









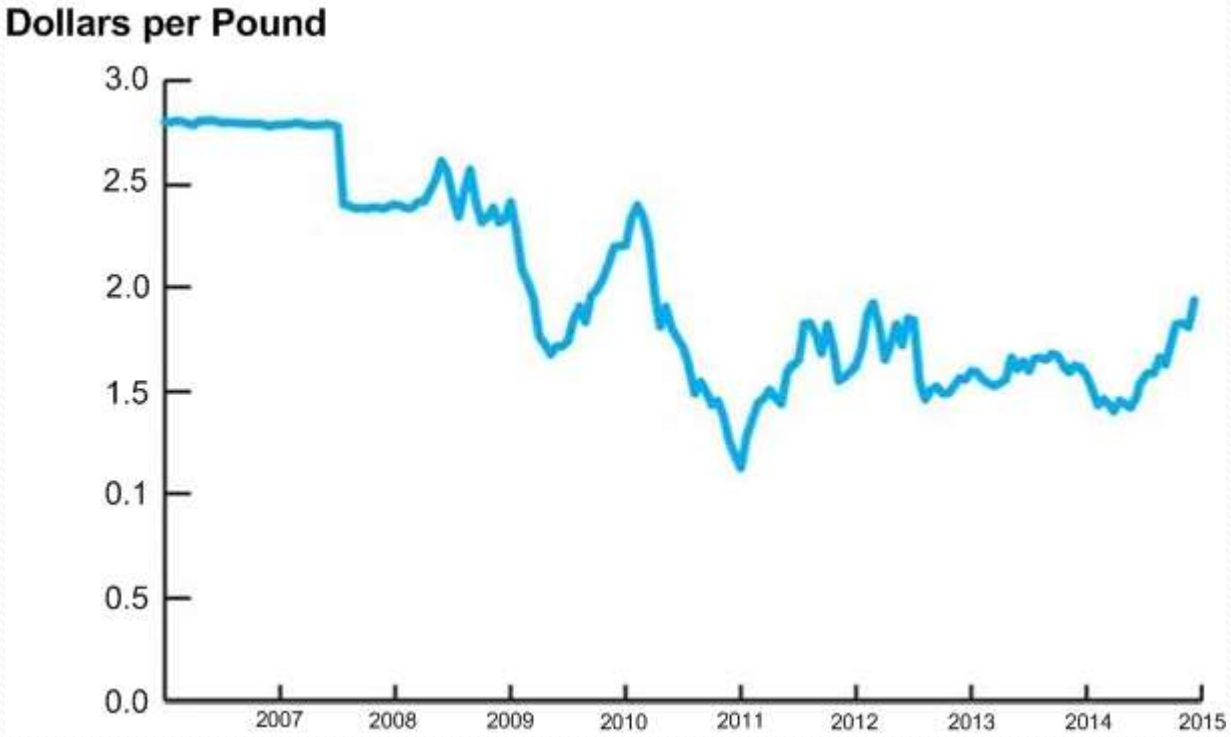
Use of Time Series Data

- To develop forecast model
  - What will the rate of inflation be next year?
- To estimate dynamic causal effects
  - If the rate of interest increases the interest rate now, 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 of a hike in the excise duty?
- Time dependent analysis
  - Rates of inflation and unemployment in the country can be observed only over time!

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

Auto-Regression Model for Forecasting



- Can we predict the tend at a time say 2017?

Some Notations and Concepts

- Y_t = Value of Y in a period t
- Data set [Y_1, Y_2, … Y_{T-1}, Y_T]: T observations on the time series random variable Y
- **Assumptions**
  - We consider only consecutive, evenly spaced observations
    - For example, monthly, 2000-2015, no missing months
  - 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+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 stationary.

Some Notations and Concepts

- There are four ways to have the time series data for AutoRegression 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:** $y_t = \frac{Y_{t-1}}{Y_t} \times 100$

Some Notations and Concepts

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

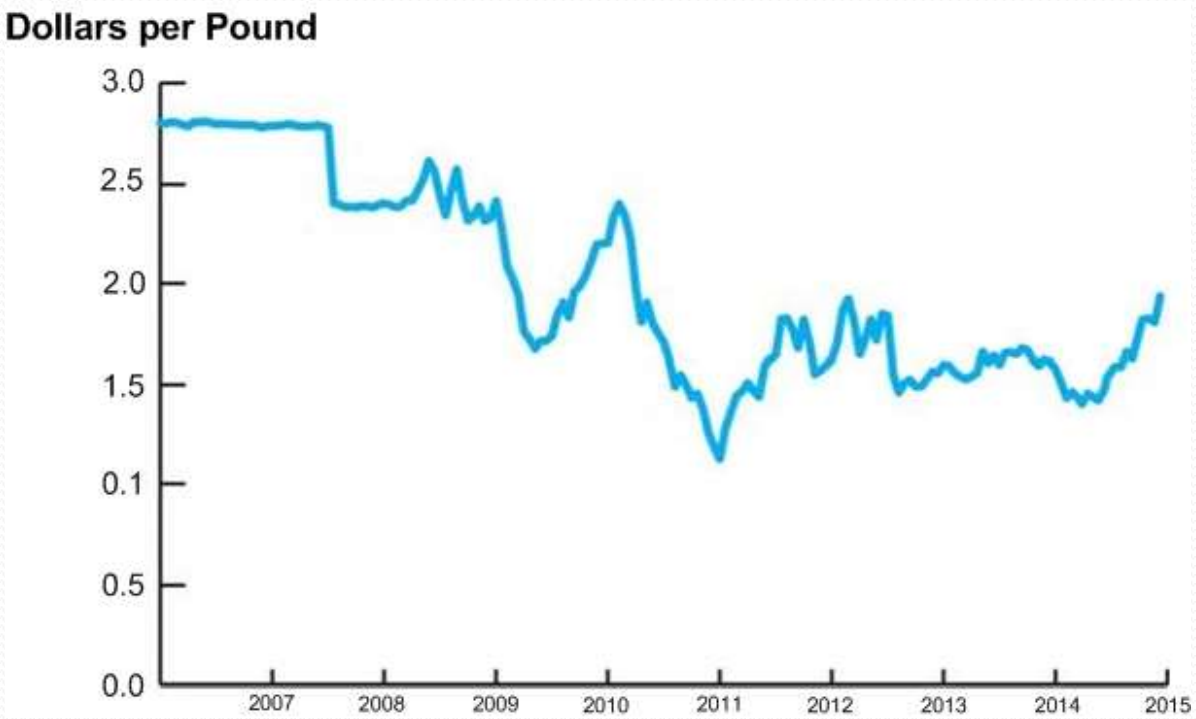
Definition 7.4: *j*-th Autocorrelation

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 autocovariance**

Some Notations and Concepts



- For the given data, say $\rho_1 = 0.84$
  - This implies that the Dollars per Pound is highly serially correlated
- Similarly, we can determine ρ_2 , ρ_3 .... etc., and hence different regression analyses

Auto-Regression Model for Forecatsing

- A natural starting point for forecasting model is to use past values of Y, that is, Y_{t-1}, Y_{t-2}, … to predict Y_t
- An autoregression 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 autoregression
  - In first order autoregression (denoted as AR(1)), Y_t is regressed against Y_{t-1}
  - In p-th order autoregression (denoted as AR(p)), Y_t is regressed against, Y_{t-1}, Y_{t-2}, …,Y_{t-p}

p-th Order AutoRegression Model

Definition 7.5: *p*-th AutoRegression Model

In general, the p-th order autregression 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 autoregression coefficients and ε_t is the noise term or residue and in practice it is assumed to Gausian 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 β_i $i = 0, 1, \ldots, p$ given a time series Y_t.

Computing AR Coefficients

- 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

$$\begin{bmatrix} 1 & r_1 & r_2 & r_3 & r_4 & \cdots & r_{p-2} & r_{p-1} \\ r_1 & 1 & r_1 & r_2 & r_3 & \cdots & r_{p-3} & r_{p-2} \\ r_2 & r_1 & 1 & r_1 & r_2 & \cdots & r_{p-4} & r_{p-3} \\ r_3 & r_2 & r_1 & 1 & r_2 & \cdots & r_{p-5} & r_{p-4} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ r_{p-1} & r_{p-2} & r_{p-3} & r_{p-4} & r_{p-5} & \cdots & r_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ r_{p-1} \\ r_p \end{bmatrix}$$

- Here, r_i (i = 1, 2 , 3, …, p-1) denotes the i-th auto correlation coefficient.
- β_0 can be chosen empirically, usually taken as zero.

Reference

- The detail material related to this lecture can be found in

  The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.

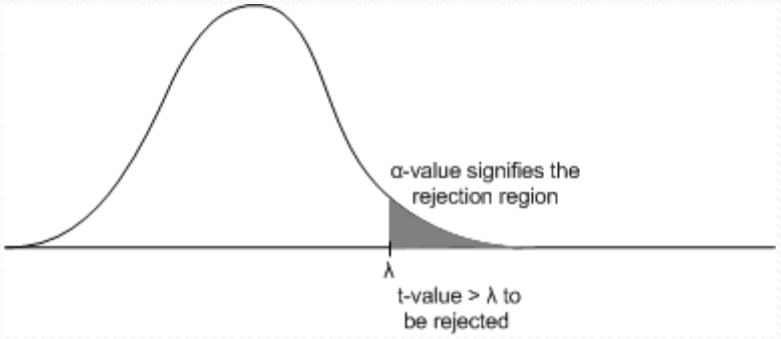
Any question?

You may post your question(s) at the "Discussion Forum" maintained in the course Web page!

Questions of the day…

1. For a given sample data the correlation coefficient according to the Karl Pearson's correlation analysis is found to be r = 0.79 with degree of freedom 69. Further, with significant test , the t-value is calculated as t = 2.36. From the t-test table, it is found that with degree of freedom 69, the t-value at 5% confidence level is 3.61. What is the inference that you can have in this case?

2. For a given degree of freedom, if α, the value of confidence level increases, then t-value increases. Is the statement correct? If not, what is the correct statement? Justify your answer. You can refer the following figure in your explanation.



Questions of the day…

3. Whether the Spearman's correlation coefficient analysis is applicable to the numeric data? If so, how?

4. Can χ^2–analysis be applied to ordinal data or numeric data? Justify your answer.

5. Briefly explain the following with reference to the χ^2 correlation analysis.
   a) Contingency table
   b) Observed frequency
   c) Expected frequency
   d) Expression for –vale calculation
   e) Hypothesis to be tested
   f) Degree of freedom of sample data