CHENNAI INSTITUTE OF TECHNOLOGY

(Affiliated to Anna University, Approved by AICTE, Accredited by NAAC & NBA)
Sarathy Nagar, Kundrathur, Chennai – 600069, India.

Lecture Notes UNIT III

DEPARTMENT OF INFORMATION TECHNOLOGY

Subject: CS 3353-Foundations of Data Science

Dr.A.R.Kavitha

II Year IT/III SEMESTER

UNIT III DESCRIBING RELATIONSHIPS

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations – regression towards the mean

3.1 Scatterplot

- ❖ Define scatter plot(2M)
- ❖ Explain scatter plot with example(16M)
- ❖ How to interpret scatter plots(16M)

- The most useful graph for displaying the relationship between two quantitative variables is a scatterplot.

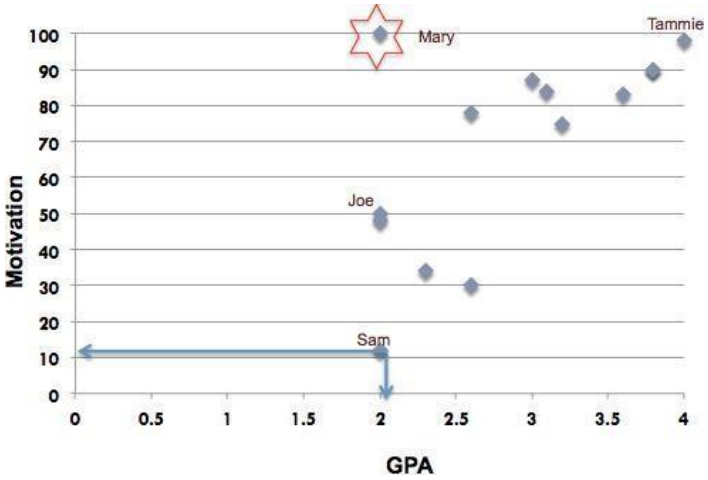
> A scatterplot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the othervariable appear on the vertical axis. Each individual in the data appears as a point on the graph.

- Many research projects are correlational studies because they investigate the relationships that may exist between variables. Prior to investigating the relationship between two quantitative variables, it is always helpful to create a graphical representation that includes both of these variables. Such a graphical representation is called a scatterplot.

3.1.1 Scatterplot

What is the relationship between students' achievement motivation and GPA?

Student	Student GPA	Motivation
Joe	2.0	50
Lisa	2.0	48
Mary	2.0	100
Sam	2.0	12
Deana	2.3	34
Sarah	2.6	30
Jennifer	2.6	78
Gregory	3.0	87
Thomas	3.1	84
Cindy	3.2	75
Martha	3.6	83
Steve	3.8	90
Jamell	3.8	90
Tammie	4.0	98



- In this example, the relationship between students' achievement motivation and their GPA is being investigated.
- The table on the left includes a small group of individuals for whom GPA and scores on amotivation scale have been recorded. GPAs can range from 0 to 4 and motivation scores in this example range from 0 to 100. Individuals in this table were ordered based on their GPA.
- Simply looking at the table shows that, in general, as GPA increases, motivation scores also increase.
- However, with a real set of data, which may have hundreds or even thousands of individuals, a pattern cannot be detected by simply looking atthe numbers. Therefore, a very useful strategy is to represent the two variables graphically to illustrate the relationship between them.
- A graphical representation of individual scores on two variables is called a scatterplot.
- The image on the right is an example of a scatterplot and displays the datafrom the table on the left. GPA scores are displayed on the horizontal axisand motivation scores are displayed onthe vertical axis.
- Each dot on the scatterplot represents one individual from the data set. The location of each point on the graph depends on both the GPA and motivation scores. Individuals with higher GPAs are located further to the right and individuals with higher motivation scores are located higherup on the graph.
- Sam, for example, has a GPA of 2 so his point is located at 2 on the right. He also has a motivation score of 12, so his point is located at 12 going up.
- Scatterplots are not meant to be used in great detail because there areusually hundreds of individuals in a data set.
- The purpose of a scatterplot is to provide a general illustration of the relationship between the two variables.
- In this example, in general, as GPA increases so does an individual'smotivation score.
- One of the students in this example does not seem to follow the general pattern: Mary. She is one of the students with the lowest GPA, but she has the maximum score on the motivation scale. This makes her an exception oran outlier.

Solved Examples on Scatter Diagram

Question: Draw the scatter diagram for the given pair of variables and understand the type of correlation between them.

No. of Students	Marks obtained (out of 100)
12	40-50
10	50-60
8	60-70
7	70-80
5	80-90
2	90-100

Solution:

Here, we take the two variables for consideration as:

M: The marks obtained out of 100
S: Number of students

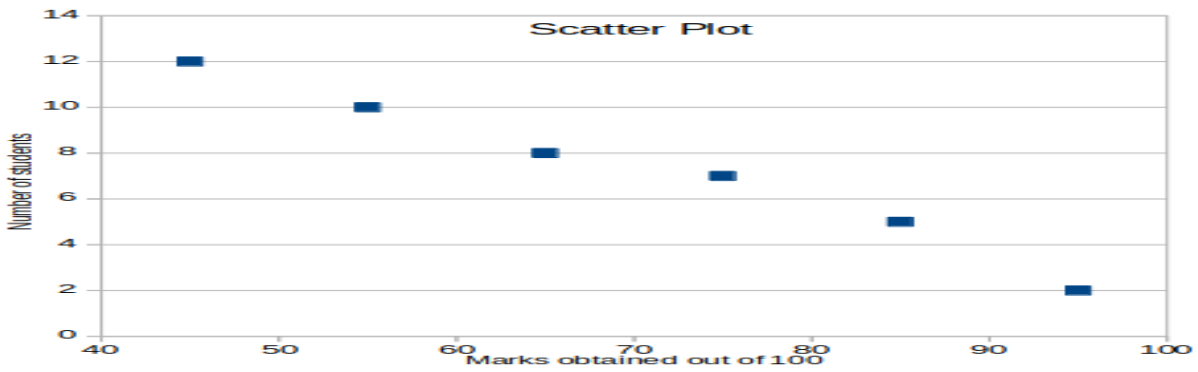
Since the values of M is in the form of bins, we can use the centre point of each class in the scatter diagram instead. So let us first choose the axes of our diagram.

X-axis – Marks obtained out of 100
Y-axis – Number of Students

The data points that we need to plot according to the given dataset are –

(45,12), (55,10), (65,8), (75,7), (85,5), (95,2)

Here's how the plot will look like –



3.1.2 Interpreting Scatterplots

How to Examine a Scatterplot

As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

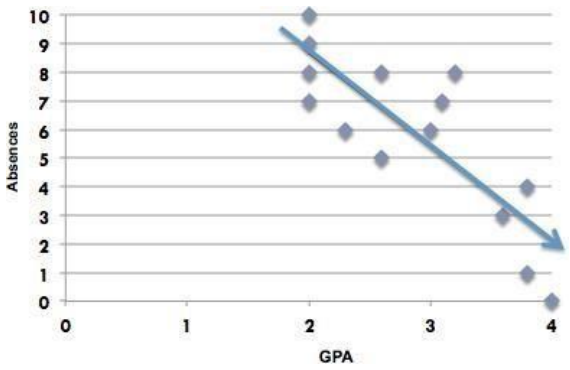
- The overall pattern of a scatterplot can be described by the **direction**, **form**, and **strength** of the relationship.
- An important kind of departure is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Interpreting Scatterplots: Direction

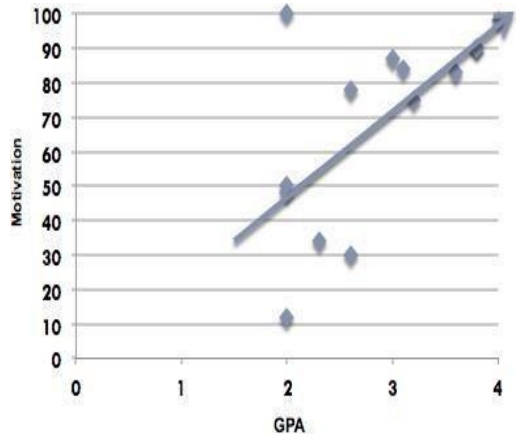
- One important component to a scatterplot is the direction of the relationshipbetween the two variables.

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.



This example comparesstudents' achievement
motivation and their GPA. These two variables have a positive association becauseas GPA increases, so does motivation.

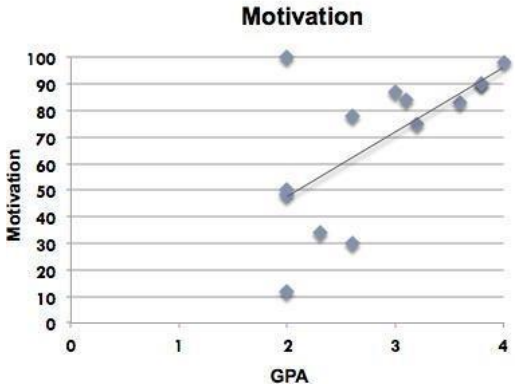


This example compares students' GPA and their numberof absences. These two variables have a negative association because, in general,as a student's number of absences decreases, their GPA increases

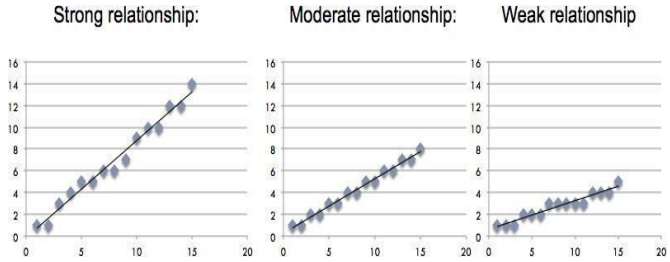
Interpreting Scatterplots: Form

- Another important component to a scatterplot is the form of the relationshipbetween the two variables.

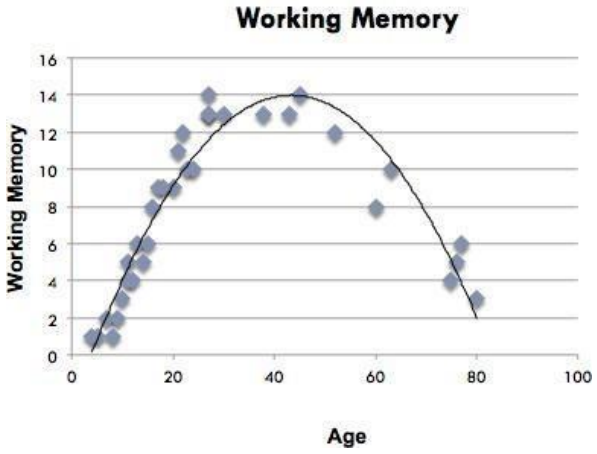
Linear relationship:



This example illustrates a linear relationship. This means that the points on the scatterplot closely resemble a straight line. A relationship is linear if one variable increases by approximately the same rate as theother variables changes by one unit.



Curvilinear relationship:



This example illustrates a relationship that has the form of acurve, rather than a straight line. This is due to the fact that one variable does not increase at a constant rate and may even start decreasing after a certain point.

This example describes a curvilinear relationship between the variable "age" and the variable "working memory." In this example, working memory increases throughout childhood, remains steady in adulthood, and begins decreasing around age 50.

Interpreting Scatterplots: Strength

- Another important component to a scatterplot is the strength of therelationship between the two variables.
- The slope provides information on the strength of the relationship.
- The strongest linear relationship occurs when the slope is 1. This means thatwhen one variable increases by one, the other variable also increases by the same amount. This line is at a 45 degree angle.
- The strength of the relationship between two variables is a crucial piece of information. Relying on the interpretation of a scatterplot is too subjective. More precise evidence is needed, and this evidence is obtained by computinga coefficient that measures the strength of the relationship under investigation.

Measuring Linear Association

- A scatterplot *displays* the strength, direction, and form of the relationshipbetween two quantitative variables.
- A correlation coefficient *measures* the strength of that relationship.

The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

Pearson r:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- *r* is always a number between -1 and 1.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship.

- Calculating a Pearson correlation coefficient requires the assumption that therelationship between the two variables is linear.
- There is a rule of thumb for interpreting the strength of a relationship basedon its r value (use the absolute value of the r value to make all values positive):

Absolute Value of r	Strength of Relationship
r < 0.3	None or very weak
0.3 < r <0.5	Weak
0.5 < r < 0.7	Moderate
r > 0.7	Strong

- The relationship between two variables is generally considered strong whentheir r value is larger than 0.7.

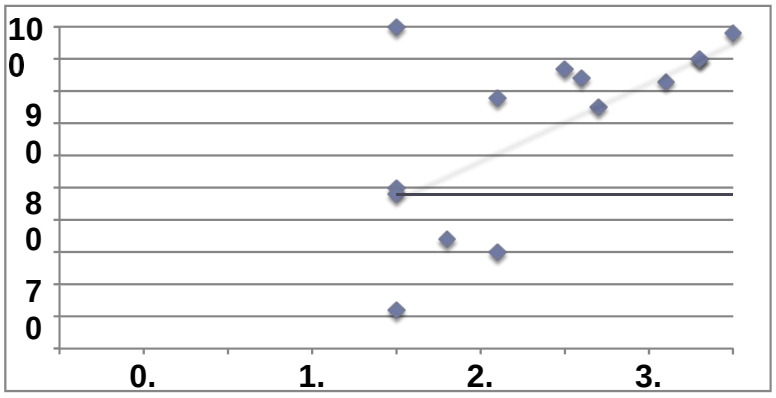
3.2 Correlations

- What is correlations?(2M)
- Facts about correlations(2M)

Example: There is a moderate, positive, linear relationship between GPA andachievement motivation.

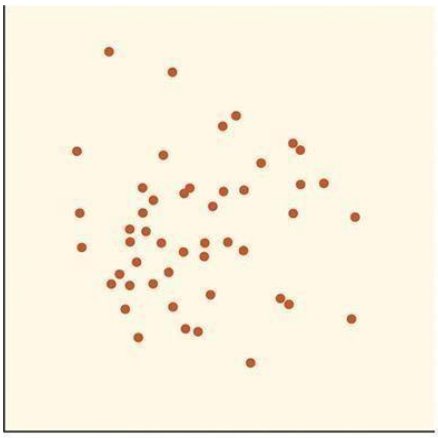
- Based on the criteria listed on the previous page, the value of r in this case (r = 0.62) indicates that there is a positive, linear relationship of moderate strength between achievement motivation and GPA.

r = 0.62

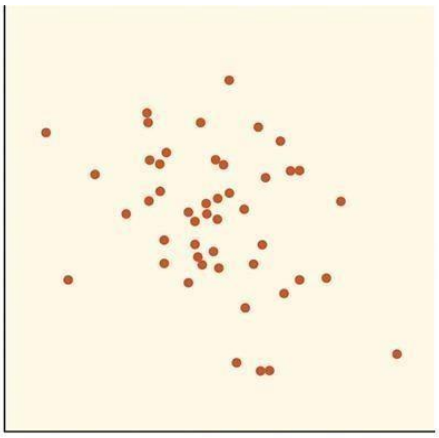


Correlation

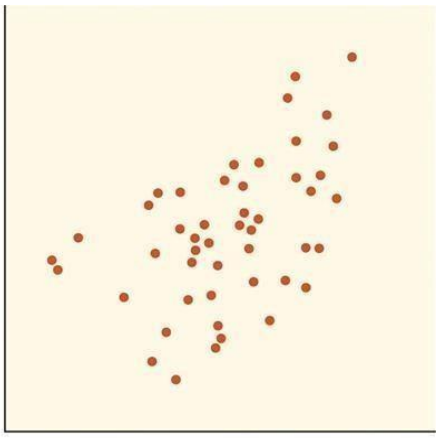
- The images below illustrate what the relationships might look like atdifferent degrees of strength (for different values of r).



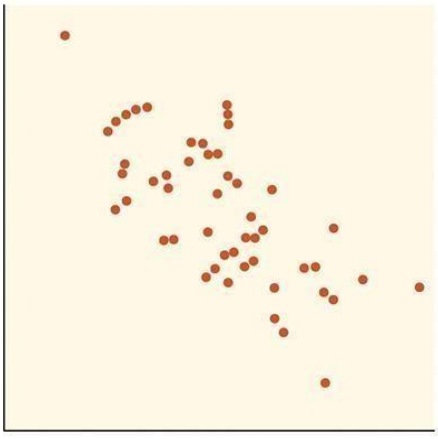
Correlation $r = 0$



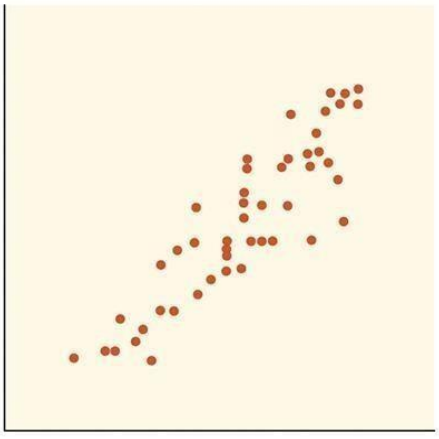
Correlation $r = -0.3$



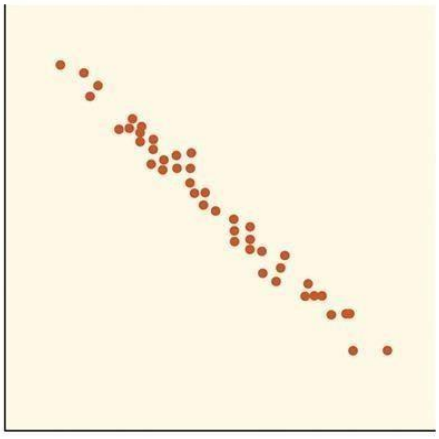
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

- For a correlation coefficient of zero, the points have no direction, the shape isalmost round, and a line does not fit to the points on the graph.
- As the correlation coefficient increases, the observations group closertogether in a linear shape.
- The line is difficult to detect when the relationship is weak (e.g., r = -0.3), butbecomes more clear as relationships become stronger (e.g., r = -0.99)

Facts About Correlation

1) The order of variables in a correlation is not important.
2) Correlations provide evidence of association, not causation.
3) *r* has no units and does not change when the units of measure of *x*, *y*, or bothare changed.
4) Positive *r* values indicate positive association between the variables, andnegative *r* values indicate negative associations.
5) The correlation *r* is always a number between -1 and 1.

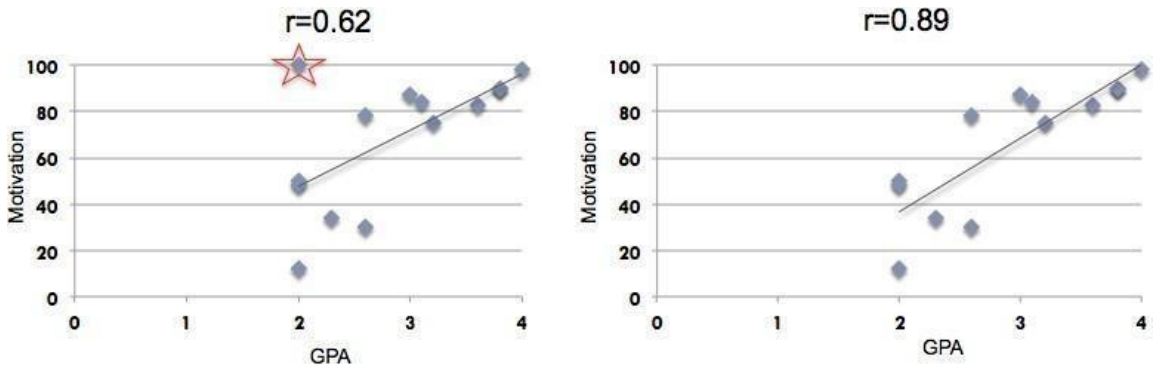
Pearson *r*: Assumptions

Assumptions:

- Correlation requires that both variables be quantitative.
- Correlation describes *linear* relationships. Correlation does not describe curve relationships between variables, no matter how strong the relationshipis.

Cautions:

- Correlation is not resistant. *r* is strongly affected by outliers.
- Correlation is not a complete summary of two-variable data.
- For example:



- The correlation coefficient is based on means and standard deviations, so it isnot robust to outliers; it is strongly affected by extreme observations. These individuals are sometimes referred to as *influential observations* because they have a strong impact on the correlation coefficient.
- For instance, in the above example the correlation coefficient is 0.62 on the left when the outlier is included in the analysis. However, when this outlier is removed, the correlationcoefficient increases significantly to 0.89.
- This one case, when included in the analysis, reduces a strong relationship toa moderate relationship.
- This case makes such a big difference in this example because the data set contains a very small number of individuals. As a general rule, as the size ofthe sample increases, the influence of extreme observations decreases.
- When describing the relationship between two variables, correlations are just one piece of the puzzle. This information is necessary, but not sufficient.Other analyses should also be conducted to provide more information.

CORRELATION COEFFICIENT:

- What does a correlation coefficient tell you?(2M)
- Significance of correlation coefficient(2M)
- How to interpret correlation coefficient?(8M)
- Types of Correlation coefficient(16M)
- What are the assumptions our data has to meet for pearson'sr?(2M)
- Give Pearson's r formula with explaination(2M)
- Give spearmans's rho formula(2M)

A correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables. In other words, it reflects how similar the measurements of two or more variables are across a dataset.

Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.

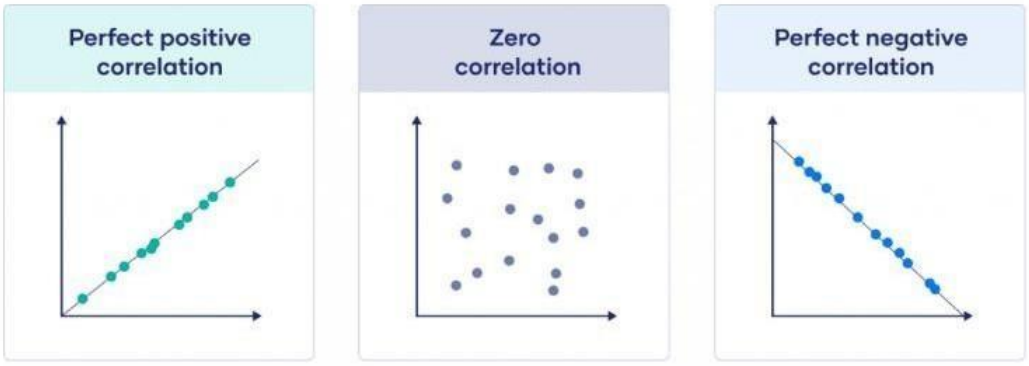


Figure : Co-relation

3.3 Correlation Coefficients

The Statistical Significance of Correlation Coefficients:

- Correlation coefficients have a probability (p-value), which shows the probability that the relationship between the two variables is equal tozero (null hypotheses; no relationship).
- Strong correlations have low p-values because the probability that they haveno relationship is very low.
- Correlations are typically considered statistically significant if the p-value islower than 0.05 in the social sciences, but the researcher has the liberty to decide the p-value for which he or she will consider the relationship to be significant.
- The value of p for which a correlation will be considered statisticallysignificant is called the alpha level and must be reported.
- SPSS notation for p values: Sig. (2 tailed)

In the previous example, r = 0.62 and p-value = 0.03. The p-value of 0.03 is less thanthe acceptable alpha level of 0.05, meaning the correlation is statistically significant.

Four things must be reported to describe a relationship:

1) The strength of the relationship given by the correlation coefficient.
2) The direction of the relationship, which can be positive or negative based onthe sign of the

correlation coefficient.

3) The shape of the relationship, which must always be linear to computer aPearson correlation coefficient.
4) Whether or not the relationship is statistically significant, which is basedon the p-value

What does a correlation coefficient tell you?
Correlation coefficients summarize data and help you compare results between studies.

Summarizing data

A correlation coefficient is a descriptive statistic. That means that it summarizes sample data without letting you infer anything about the population. A correlation coefficient is a bivariate statistic when it summarizes the relationship between two variables, and it's a multivariate statistic when you have more than two variables.

If your correlation coefficient is based on sample data, you'll need an inferential statistic if you want to generalize your results to the population. You can use an F test or a t test to calculate a test statistic that tells you the statistical significance of your finding.

Comparing studies

A correlation coefficient is also an effect size measure, which tells you the practical significance of a result. Correlation coefficients are unit-free, which makes it possible to directly compare coefficients between studies.

Using a correlation coefficient

In correlational research, you investigate whether changes in one variable are associated with changes in other variables.

Correlational research example

You investigate whether standardized scores from high school are related to academic grades in college. You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

After data collection, you can visualize your data with a scatterplot by plotting one variable on the x-axis and the other on the y-axis. It doesn't matter which variable you place on either axis.

Visually inspect your plot for a pattern and decide whether there is a linear or non-linear pattern between variables. A linear pattern means you can fit a straight line of best fit between the data points, while a non- linear or curvilinear pattern can take all sorts of different shapes, such as a U-shape or a line with a curve.

Visual inspection example

You gather a sample of 5,000 college graduates and survey them on their high school SAT scores and college GPAs. You visualize the data in a scatterplot to check for a linear

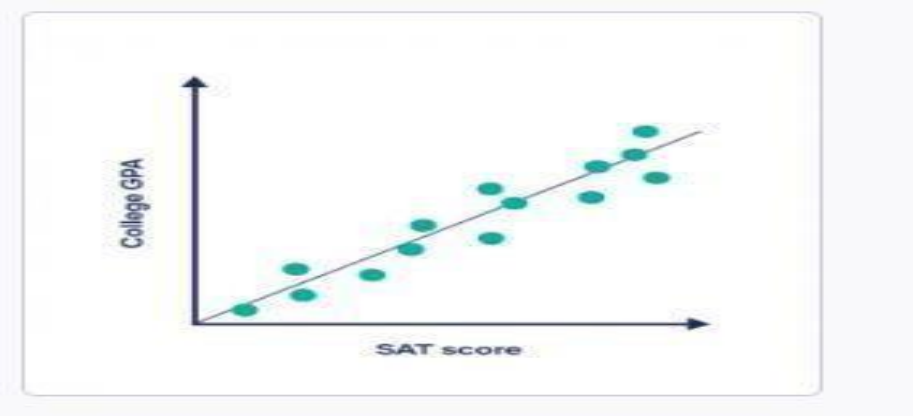


Figure :

There are many different correlation coefficients that you can calculate. After removing any outliers, select a correlation coefficient that's appropriate based on the general shape of the scatter plot pattern. Then you can perform a correlation analysis to find the correlation coefficient for your data.

You calculate a correlation coefficient to summarize the relationship between variables without drawing any conclusions about causation.

Correlation analysis example

You check whether the data meet all of the assumptions for the Pearson's r correlation test.
Both variables are quantitative and normally distributed with no outliers, so you calculate a Pearson's r correlation coefficient.

The correlation coefficient is strong at .58

Interpreting a correlation coefficient

The value of the correlation coefficient always ranges between 1 and -1, and you treat it as a general indicator of the strength of the relationship between variables.

The sign of the coefficient reflects whether the variables change in the same or opposite directions: a positive value means the variables change together in the same direction, while a negative value means theychange together in opposite directions.

The absolute value of a number is equal to the number without its sign. The absolute value of a correlation coefficient tells you the magnitude of the correlation: the greater the absolute value, the stronger the correlation.

There are many different guidelines for interpreting the correlation coefficient because findings can vary a lot between study fields. You can use the table below as a general guideline for interpreting correlation strength from the value of the correlation coefficient.

While this guideline is helpful in a pinch, it's much more important to take your research context and purpose into account when forming conclusions. For example, if most studies in your field have correlation coefficients nearing .9, a correlation coefficient of .58 may be low in that context.

Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

Table :

Visualizing linear correlations

The correlation coefficient tells you how closely your data fit on a line. If you have a linear relationship, you'll draw a straight line of best fit that takes all of your data points into account on a scatter plot.

The closer your points are to this line, the higher the absolute value of the correlation coefficient and the stronger your linear correlation.

If all points are perfectly on this line, you have a perfect correlation.

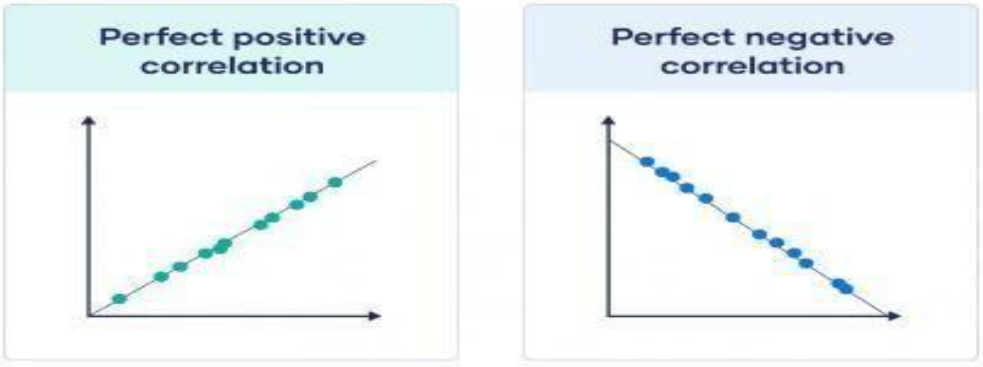


Figure :

If all points are close to this line, the absolute value of your correlation coefficient is high.

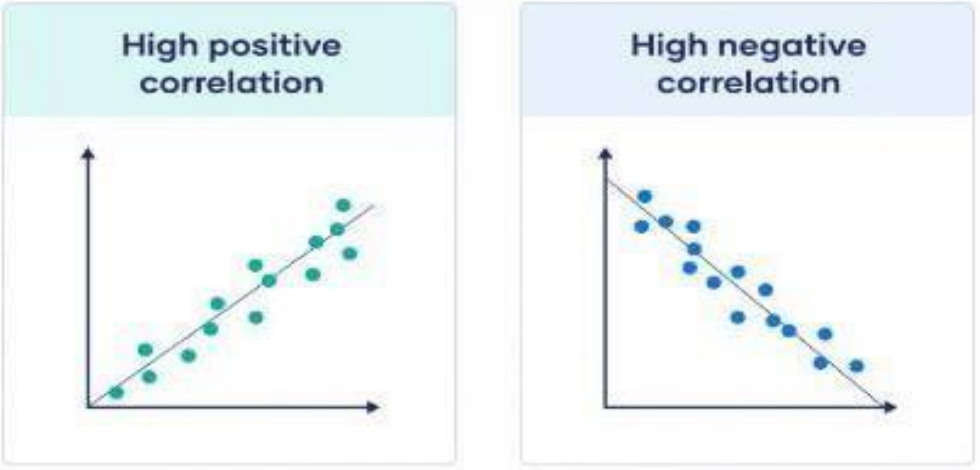


Figure:

If these points are spread far from this line, the absolute value of your correlation coefficient is low.

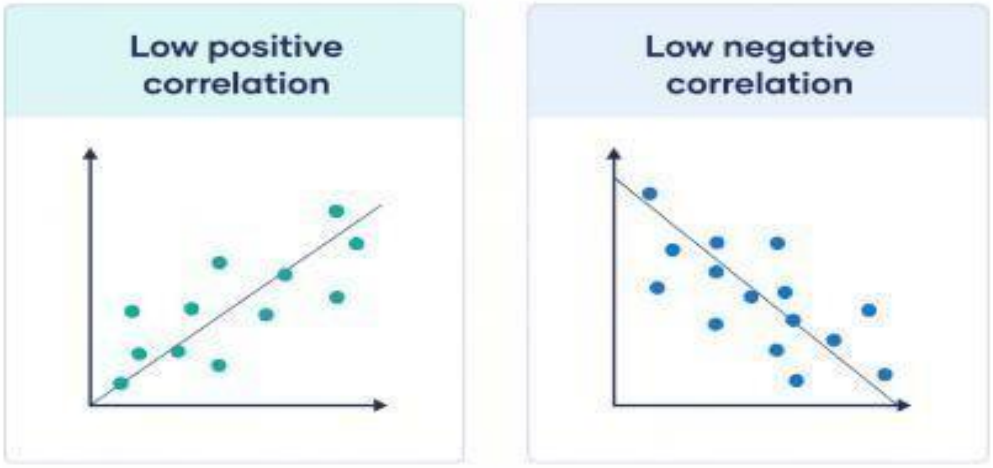


Figure:

Note that the steepness or slope of the line isn't related to the correlation coefficient value. The correlation coefficient doesn't help you predict how much one variable will change based on a given change in the other, because two datasets with the same correlation coefficient value can have lines
with very different slopes.

3.3.1. Types of correlation coefficients

You can choose from many different correlation coefficients based on the linearity of the relationship, the level of measurement of your variables, and the distribution of your data.

For high statistical power and accuracy, it's best to use the correlation coefficient that's most appropriate for your data.

The most commonly used correlation coefficient is Pearson's r because it allows for strong inferences. It's parametric and measures linear relationships. But if your data do not meet all assumptions for this test, you'll need to use a non-parametric test instead.

Non-parametric tests of rank correlation coefficients summarize non-linear relationships between variables. The Spearman's rho and Kendall's tau have the same conditions for use, but Kendall's tau is generally preferred for smaller samples whereas Spearman's rho is more widely used.

The table below is a selection of commonly used correlation coefficients, and we'll cover the two most widely used coefficients in detail in this article.

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Table

3.3.1.1 Pearson's r

The Pearson's product-moment correlation coefficient, also known as Pearson's r, describes the linear relationship between two quantitative variables.

These are the assumptions your data must meet if you want to use Pearson's r:

- Both variables are on an interval or ratio level of measurement
- Data from both variables follow normal distributions
- Your data have no outliers
- Your data is from a random or representative sample
- You expect a linear relationship between the two variables

The Pearson's r is a parametric test, so it has high power. But it's not a good measure of correlation if your variables have a nonlinear relationship, or if your data have outliers, skewed distributions, or come from categorical variables. If any of these assumptions are violated, you should consider a rank correlation measure.

The formula for the Pearson's r is complicated, but most computer programs can quickly churn out the correlation coefficient from your data. In a simpler form, the formula divides the covariance between the variables by the product of their standard deviations.

Formula	Explanation
$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$	• r_{xy} = strength of the correlation between variables x and y • n = sample size • $\sum$ = sum of what follows... • X = every x-variable value • Y = every y-variable value • XY = the product of each x-variable score and the corresponding y-variable score

3.3.1.1.1 Pearson sample vs population correlation coefficient formula

When using the Pearson correlation coefficient formula, you'll need to consider whether you're dealing with data from a sample or the whole population.

The sample and population formulas differ in their symbols and inputs. A sample correlation coefficient is called r, while a population correlation coefficient is called rho, the Greek letter ρ.

The sample correlation coefficient uses the sample covariance between variables and their sample standard deviations.

Sample correlation coefficient formula	Explanation
$r_{xy} = \frac{cov(x,y)}{s_x s_y}$	• r_{xy}= strength of the correlation between variables x and y • $cov(x,y)$ = covariance of x and y • s_x = sample standard deviation of x • s_y = sample standard deviation of y

The population correlation coefficient uses the population covariance between variables and their population standard deviations.

Population correlation coefficient formula	Explanation
$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$	• ρ_{XY}= strength of the correlation between variables X and Y • $cov(X,Y)$ = covariance of X and Y • σ_X = population standard deviation of X • σ_Y = population standard deviation of Y

There are 2 stocks – A and B. Their share prices on particular days are as follows:

There are 2 stocks – A and B. Their share prices on particular days are as follows:

Find out the Pearson correlation coefficient from the above data.

Stock A (x)	Stock B (y)
45	9
50	8
53	8
58	7
60	5

Solution:

First, we will calculate the following values.

	A	B	C	D	E
1	Stock A (x)	Stock B (y)	x*y	x^2	y^2
2	45	9	405	2025	81
3	50	8	400	2500	64
4	53	8	424	2809	64
5	58	7	406	3364	49
6	60	5	300	3600	25
7	**266**	**37**	**1935**	**14298**	**283**
8					

The calculation of the Pearson coefficient is as follows,

B10 : =(B9*C7-A7*B7)/((B9*D7-A7^2)*(B9*E7-B7^2))^0.5

	A	B	C	D	E	F
1	Stock A (x)	Stock B (y)	x*y	x^2	y^2	
2	45	9	405	2025	81	
3	50	8	400	2500	64	
4	53	8	424	2809	64	
5	58	7	406	3364	49	
6	60	5	300	3600	25	
7	**266**	**37**	**1935**	**14298**	**283**	
8						
9	n	5				
10	r	-0.9088				
11						

- r = (5*1935-266*37)/((5*14298-(266)^2)*(5*283-(37)^2))^0.5
- = -0.9088

Therefore the Pearson correlation coefficient between the two stocks is -0.9088

3.3.1.2 Spearman's rho

Spearman's rho, or Spearman's rank correlation coefficient, is the most common alternative to Pearson's r. It's a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

You should use Spearman's rho when your data fail to meet the assumptions of Pearson's r. This happens when at least one of your variables is on an ordinal level of measurement or when the data from one or both variables do not follow normal distributions.

While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships.

In a linear relationship, each variable changes in one direction at the same rate throughout the data range. In a monotonic relationship, each variable also always changes in only one direction but not necessarily at the same rate.

Positive monotonic: when one variable increases, the other also increases.
Negative monotonic: when one variable increases, the other decreases.
Monotonic relationships are less restrictive than linear relationships.

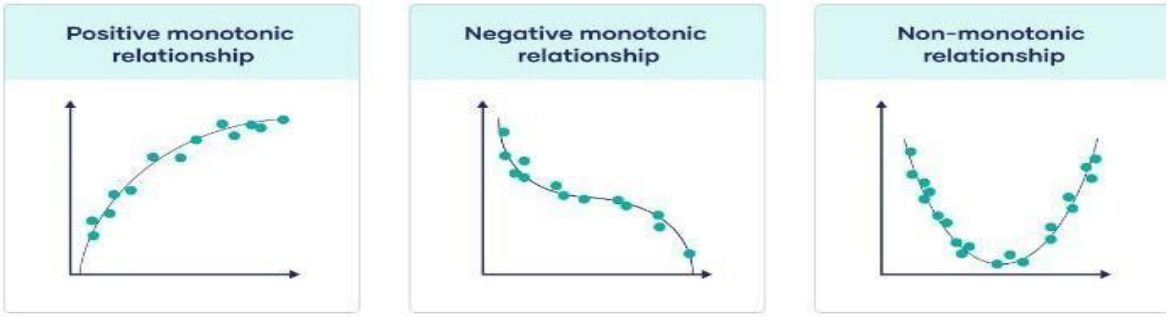


Figure :

3.3.1.2.1 Spearman's rank correlation coefficient formula

The symbols for Spearman's rho are ρ for the population coefficient and rs for the sample coefficient. The formula calculates the Pearson's r correlation coefficient between the rankings of the variable data.

To use this formula, you'll first rank the data from each variable separately from low to high: every datapoint gets a rank from first, second, or third, etc.

Then, you'll find the differences (di) between the ranks of your variables for each data pair and take that as the main input for the formula.

Spearman's rank correlation coefficient formula	Explanation
$r_s = 1 - \frac{6\sum d_i^2}{(n^3 - n)}$	• r_s= strength of the rank correlation between variables • d_i = the difference between the x-variable rank and the y-variable rank for each pair of data • $\sum d^2_i$ = sum of the squared differences between x- and y-variable ranks • n = sample size

If you have a correlation coefficient of 1, all of the rankings for each variable match up for every data pair. If you have a correlation coefficient of -1, the rankings for one variable are the exact opposite of the ranking of the other variable. A correlation coefficient near zero means that there's no monotonic relationship between the variable rankings.

Spearman rank correlation

The scores for nine students in physics and math are as follows:

- Physics: 35, 23, 47, 17, 10, 43, 9, 6, 28
- Mathematics: 30, 33, 45, 23, 8, 49, 12, 4, 31

Compute the student's ranks in the two subjects and compute the Spearman rank correlation.

Step 1: Find the ranks for each individual subject. I used the Excel rank function to find the ranks. If you want to rank by hand, order the scores from greatest to smallest; assign the rank 1 to the highest score, 2 to the next highest and so on:

Physics	Rank	Math	Rank
35	3	30	5
23	5	33	3
47	1	45	2
17	6	23	6
10	7	8	8
43	2	49	1
9	8	12	7
6	9	4	9
28	4	31	4

Step 2: Add a third column, d, to your data. The d is the difference between ranks. For example, the first student's physics rank is 3 and math rank is 5, so the difference is 2 points. In a fourth column, square your d values.

Physics	Rank	Math	Rank	d	d squared
35	3	30	5	2	4
23	5	33	3	2	4
47	1	45	2	1	1
17	6	23	6	0	0
10	7	8	8	1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0

Step 3: Sum (add up) all of your d-squared values.
4 + 4 + 1 + 0 + 1 + 1 + 1 + 0 + 0 = 12. You'll need this for the formula (the Σd^2 is just "the sum of d-squared values").

Step 4: Insert the values into the formula. These ranks are not tied, so use the first formula:

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

= 1 – (6*12)/(9(81-1))
= 1 – 72/720
= 1-0.1
= 0.9
The Spearman Rank Correlation for this set of data is 0.9.

POINT BISERIAL CORRELATION

The formula for the point biserial correlation coefficient is:

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{pq}$$

M_1 = mean (for the entire test) of the group that received the positive binary variable (i.e. the "1").

- M_0 = mean (for the entire test) of the group that received the negative binary variable (i.e. the "0").
- S_n = standard deviation for the entire test.
- p = Proportion of cases in the "0" group.
- q = Proportion of cases in the "1" group.

Cramer's V is a measure of the strength of association between two nominal variables.

It ranges from 0 to 1 where:

- **0** indicates no association between the two variables.
- **1** indicates a strong association between the two variables.

It is calculated as:

Cramer's V = $\sqrt{(X^2/n) / \min(c-1, r-1)}$

where:

- **X^2:** The Chi-square statistic
- **n:** Total sample size
- **r:** Number of rows
- **c:** Number of columns

Kendall's Tau is a non-parametric measure of relationships between columns of ranked data. The Tau correlation coefficient returns a value of 0 to 1, where:

- 0 is no relationship,
- 1 is a perfect relationship.

A quirk of this test is that it can also produce negative values (i.e. from -1 to 0). Unlike a linear graph, a negative relationship doesn't mean much with ranked columns (other than you perhaps switched the columns around), so just remove the negative sign when you're interpreting Tau.

Kendall's Tau = (C – D / C + D)

Where C is the number of concordant

pairs and D is the number of discordant

airs.

x	2	2	6	8	10
y	0	1	2	3	3

The Least Squares Regression LineGoodness of Fit of a Straight Line to Data

Once the scatter diagram of the data has been drawn and the model assumptions described in the previous sections at least visually verified (and perhaps the correlation coefficient r computed to quantitatively verify the linear trend), the next step in the analysis is to find the straight line that best fits the data. We will explain how to measure how well a straight line fits a collection of points by examining how well the line y=12x−1 fits the data set

(which will be used as a running example for the next three sections). We will write the equation of this line as yˆ=12x−1 with an accent on the y to indicate that the y-values computed using this equation are not from the data. We will do this with all lines approximating data sets. The line yˆ=12x−1 was selected as one that seems to fit the data reasonably well.

The idea for measuring the goodness of fit of a straight line to data is illustrated in Figure 10.6 "Plot of the Five-Point Data and the Line ", in which the graph of the line yˆ=12x−1 has been superimposed on the scatter plot for the sample data set.

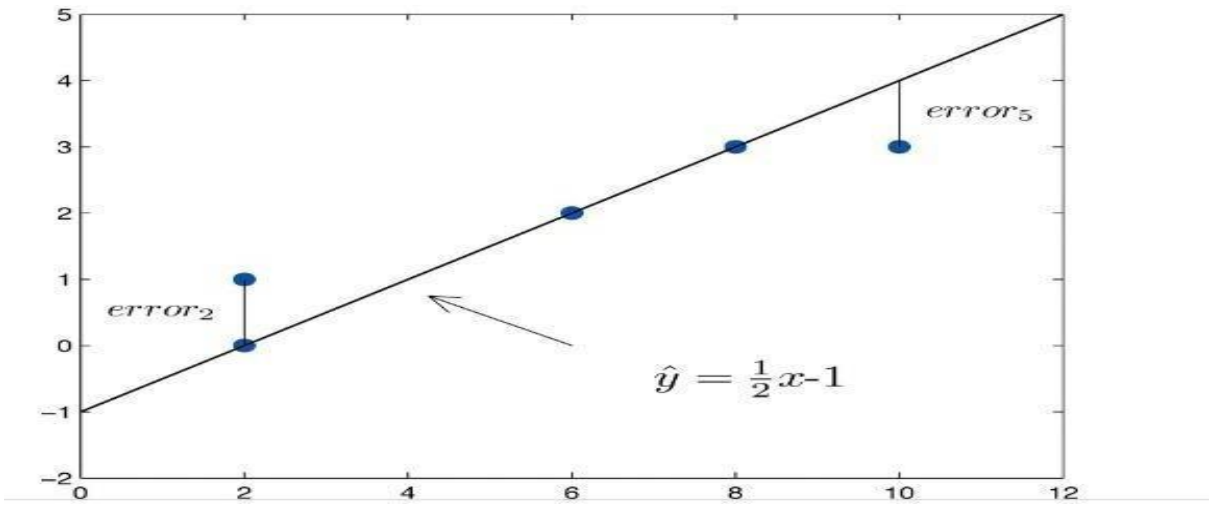


Figure Plot of the Five-Point Data and the Line yˆ=12x−1

To each point in the data set there is associated an “error,” the positive or negative vertical distance from the point to the line: positive if the point is above the line and negative if it is below the line. The error canbe computed as the actual y-value of the point minus the y-value yˆ that is “predicted” by inserting the x- value of the data point into the formula for the line:

erroratdatapoint(x,y)=(truey)−(predictedy)=y−yˆ

	x	y	$\hat{y} = \frac{1}{2}x - 1$	$y - \hat{y}$	$(y - \hat{y})^2$
	2	0	0	0	0
	2	1	0	1	1
	6	2	2	0	0
	8	3	3	0	0
	10	3	4	−1	1
Σ	-	-	-	0	2

Table The Errors in Fitting Data with a Straight Line

A first thought for a measure of the goodness of fit of the line to the data would be simply to add the errors at every point, but the example shows that this cannot work well in general. The line does not fit the data perfectly (no line can), yet because of cancellation of positive and negative errors the sum of the errors (the fourth column of numbers) is zero. Instead goodness of fit is measured by the sum of the squares of the errors. Squaring eliminates the minus signs, so no cancellation can occur. For the data and line in Figure 10.6"Plot of the Five-Point Data and the Line " the sum of the squared errors (the last column of numbers) is 2. This number measures the goodness of fit of the line to the data.

Definition

The **goodness of fit** *of a line* y^=mx+by^=mx+b *to a set of n pairs* (x,y)(x,y) *of numbers in a sample is the sum of the squared errors*

$$\Sigma(y-\hat{y})^2 \Sigma(y-y\hat{})2$$

(n terms in the sum, one for each data pair).

The Least Squares Regression Line

Given any collection of pairs of numbers (except when all the x-values are the same) and the corresponding scatter diagram, there always exists exactly one straight line that fits the data better than any other, in the sense of minimizing the sum of the squared errors. It is called the least squares regression line. Moreover there are formulas for its slope and y-intercept.

Given a collection of pairs (x,y)(x,y) *of numbers (in which not all the x-values are the same), there is a line* y^=β^1x+β^0y^=β^1x+β^0 *that best fits the data in the sense of minimizing the sum of the squared errors. It is called the* least squares regression line. *Its slope* β^1β^1 *and y-intercept* β^0β^0 *are computed using the formulas*

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad and \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2, \quad SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)$$

x— —x- *is the mean of all the x-values,* y— —y- *is the mean of all the y-values, and n is the number of pairs in the data set.*

The equation y^=β^1x+β^0y^=β^1x+β^0 *specifying the least squares regression line is called the* least squares regression equation.

Remember from Section 10.3 "Modelling Linear Relationships with Randomness Present" that the line with the equation y=β1x+β0 is called the population regression line. The numbers β^1 and β^0 are statistics that estimate the population parameters β1 and β

EXAMPLE 1

Find the least squares regression line for the five-point data set and verify that it fits the data better than the line $\hat{y}=12x-1$ considered in Section 10.4.1 "Goodness of Fit of a Straight Line to Data".

x	2	2	6	8	10
y	0	1	2	3	3

Solution:

In actual practice computation of the regression line is done using a statistical computation package. In order to clarify the meaning of the formulas we display the computations in tabular form.

	x	y	x^2	xy
	2	0	4	0
	2	1	4	2
	6	2	36	12
	8	3	64	24
	10	3	100	30
Σ	28	9	208	68

In the last line of the table we have the sum of the numbers in each column. Using them we compute:

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2 = 208 - \frac{1}{5}(28)^2 = 51.2$$

$$SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y) = 68 - \frac{1}{5}(28)(9) = 17.6$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{5} = 5.6$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{9}{5} = 1.8$$

so that

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375 \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.8 - (0.34375)(5.6) = -0$$

The least squares regression line for these data is

$$\hat{y} = 0.34375x - 0.125$$

The computations for measuring how well it fits the sample data are given in Table 10.2 "The Errors in Fitting Data with the Least Squares Regression Line". The sum of the squared errors is the sum of the numbers in the last column, which is 0.75. It is less than 2, the sum of the squared errors for the fit of theline yˆ=12x−1 to this data set.

THE ERRORS IN FITTING DATA WITH THE LEAST SQUARES REGRESSION LINE

x	y	$\hat{y} = 0.34375x - 0.125$	$y - \hat{y}$	$(y - \hat{y})^2$
2	0	0.5625	−0.5625	0.31640625
2	1	0.5625	0.4375	0.19140625
6	2	1.9375	0.0625	0.00390625
8	3	2.6250	0.3750	0.14062500
10	3	3.3125	−0.3125	0.09765625

3.4 What is Regression?

- ❖ Define regression with example(2M,8M)
- ❖ Application of regression in real life(2M)

Regression allows researchers to predict or explain the variation in one variable based on another variable.

The variable that researchers are trying to explain or predict is called the response variable. It is also sometimes called the dependent variable because it depends on another variable.

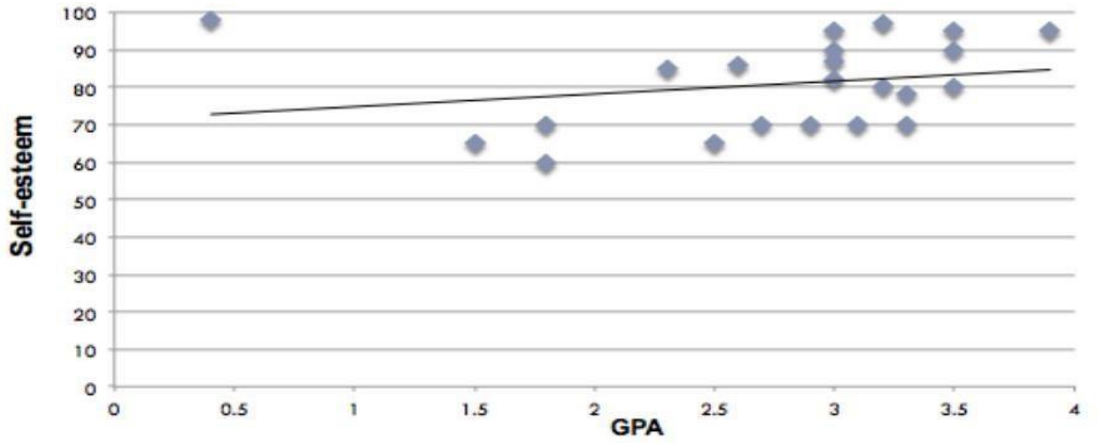
The variable that is used to explain or predict the response variable is called the explanatory variable. It is also sometimes called the independent variable because it is independent of the other variable.

In regression, the order of the variables is very important. The explanatory variable (or the independent variable) always belongs on the x-axis. The response variable (or the dependent variable) always belongs on the y-axis.

Example:

If it is already known that there is a significant correlation between students' GPA and their self-esteem, the next question researchers might ask is: Can students' scores on a self-esteem scale be predicted based on GPA? In other words, does GPA explain self-esteem? These are the types of questions that regression responds to.

**Note that these questions do not imply a causal relationship. In this example, GPA is the explanatory variable (or the independent variable) and self-esteem is the response variable (or the dependent variable). GPA belongs on the x-axis and self-esteem belongs on the y-axis.



Regression is essential for any machine learning problem that involves continuous numbers, which includes a vast array of real-life applications:

1. Financial forecasting, such as estimating housing or stock prices
2. Automobile testing
3. Weather analysis
4. Time series forecasting

3.4.2 Types of Regression

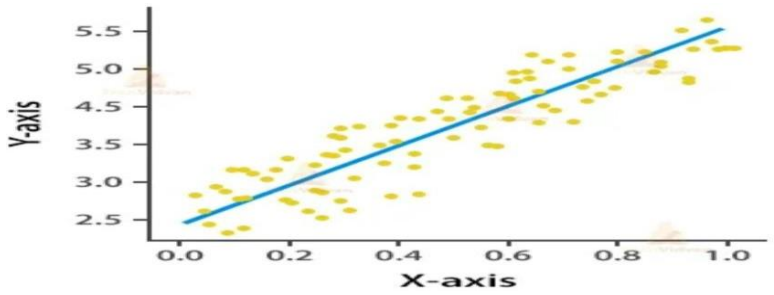
- ❖ Types of regression(2M,16M)
- ❖ What are the three approaches in stepwise regression?(2M)

- ➢ Linear Regression
- ➢ Logistic Regression
- ➢ Polynomial Regression
- ➢ Stepwise Regression
- ➢ Ridge Regression
- ➢ Lasso Regression
- ➢ Elastic Net Regression

3.4.2 .1 LINEAR REGRESSION:

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.



Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\sum X = 28$	$\sum Y = 77$	$\sum X^2 = 140$	$\sum Y^2 = 875$	$\sum XY = 334$

Table 9.7

$\overline{X} = \frac{\Sigma X}{N} = \frac{28}{7} = 4$,

$\overline{Y} = \frac{\Sigma Y}{N} = \frac{77}{7} = 11$

Regression coefficient of X on Y

$$b_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{7(334) - (28)(77)}{7(875) - (77)^2}$$

$$= \frac{2338 - 2156}{6125 - 5929}$$

$$= \frac{182}{196}$$

$$b_{xy} = 0.929$$

(i) Regression equation of X on Y

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

$$X - 4 = 0.929(Y - 11)$$

$$X - 4 = 0.929Y - 10.219$$

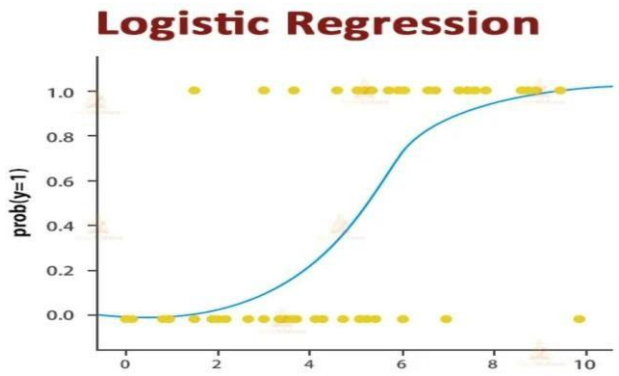
$\therefore$ The regression equation X on Y is $X = 0.929Y - 6.219$

3.4.2.2 Logistic regression

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

A logistic regression model can take into consideration multiple input criteria. In the case of college acceptance, the logistic function could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into one of two outcome categories.

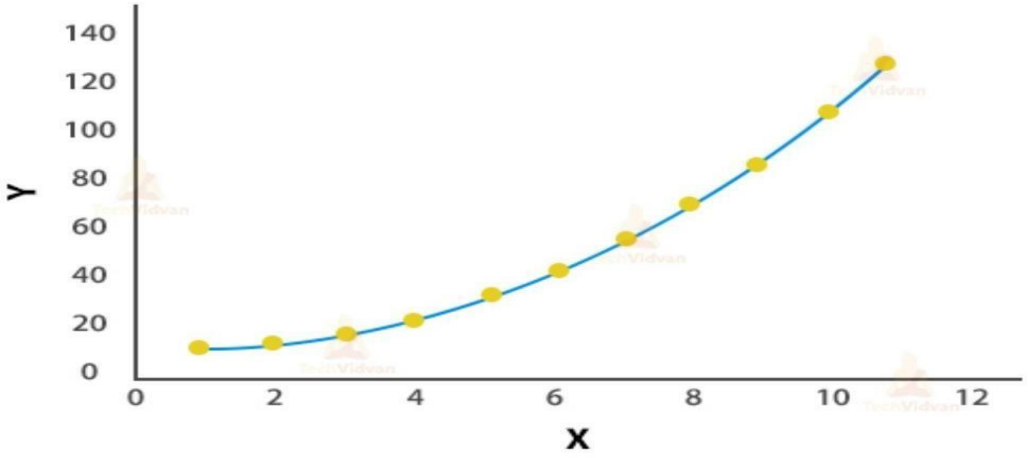


3.4.2.2 POLYNOMIAL REGRESSION:

In a polynomial regression, the power of the independent variable is more than 1. The equation below represents a polynomial equation:

$$y = a + bx2$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



3.4.2.3 Stepwise regression

Stepwise is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potentialexplanatory variables in succession and testing for statistical significance after each iteration.

The availability of statistical software packages makes stepwise regression possible, even in models with hundreds of variables.

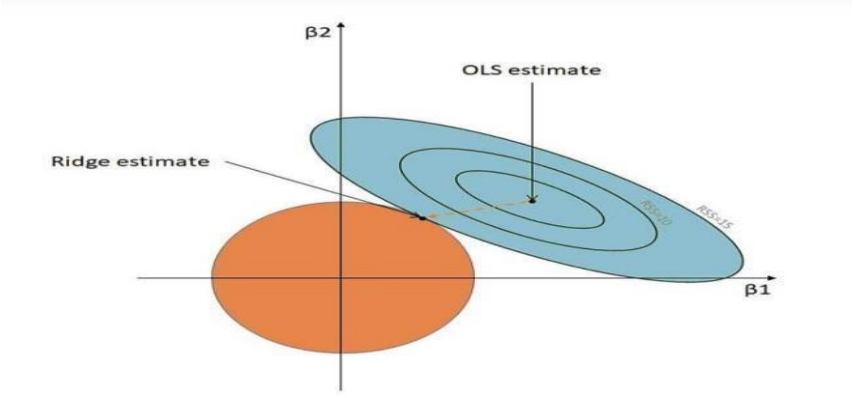
The underlying goal of stepwise regression is, through a series of tests (e.g. F-tests, t-tests) to find a set of independent variables that significantly influence the dependent variable.

there are three approaches to stepwise regression:

- Forward selection begins with no variables in the model, tests each variable as it is added to the model, then keeps those that are deemed most statistically significant—repeating the process until the results are optimal.
- Backward elimination starts with a set of independent variables, deleting one at a time, then testing to see if the removed variable is statistically significant.
- Bidirectional elimination is a combination of the first two methods that test which variables should be included or excluded.

3.4.2.4 RIDGE REGRESSION:

Ridge regression is a type of linear regression technique that is used in machine learning to reduce the overfitting of linear models. Recall that Linear regression is a method of modeling data that represents relationships between a response variable and one or more predictor variables. Ridge regression is used when there are multiple variables that are highly correlated. It helps to prevent overfitting by penalizing the coefficients of the variables. Ridge regression reduces the overfitting by adding a penalty term to the error function that shrinks the size of the coefficients. The penalty term is called the L2 norm.Ridge regression is similar to ordinary least squares regression, but the penalty term ensures that the coefficients do not become too large. This can be beneficial when there is a lot of noise in the data, as it prevents the model from being too sensitive to individual data points.



Below is the equation used to denote the Ridge Regression, λ (lambda) resolves the multicollinearity issue:

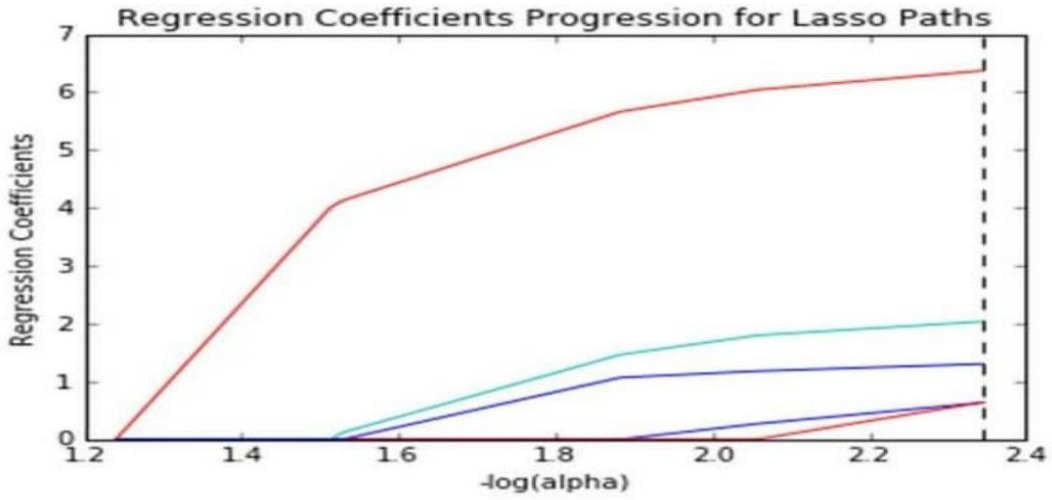
$$\beta = (X^{T}X + \lambda*I)^{-1}X^{T}y$$

3.4.2.5 LASSO REGRESSION:

3.4.2.6 LASSO REGRESSION:

The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator.

In short, Lasso Regression is like Ridge Regression regarding its use. However, the only difference is that the data is being fed is not normal. In the case of Lasso Regression, only the required parameters are used, and the rest is made zero. This helps avoid the overfitting in the model. But if independent variables are highly collinear, then Lasso regression chooses only one variable and makes other variables reduce to zero.



3.4.2.7 Elastic Net Regression

Elastic Net regression is being utilized in the case of dominant independent variables being more than one amongst many correlated independent variables.

Also, seasonality & time value factors are made to work together to identify the type of regression.

Elastic Net Regression is a combination of Lasso Regression and Ridge Regression methods. It is prepared with L1 and L2 earlier as regularizer.

The equation represents as:

ElasticNet Regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

A clear advantage of trade-off among Lasso and Ridge is that it permits Elastic-Net to acquire a portion of Ridge's dependability under rotation.

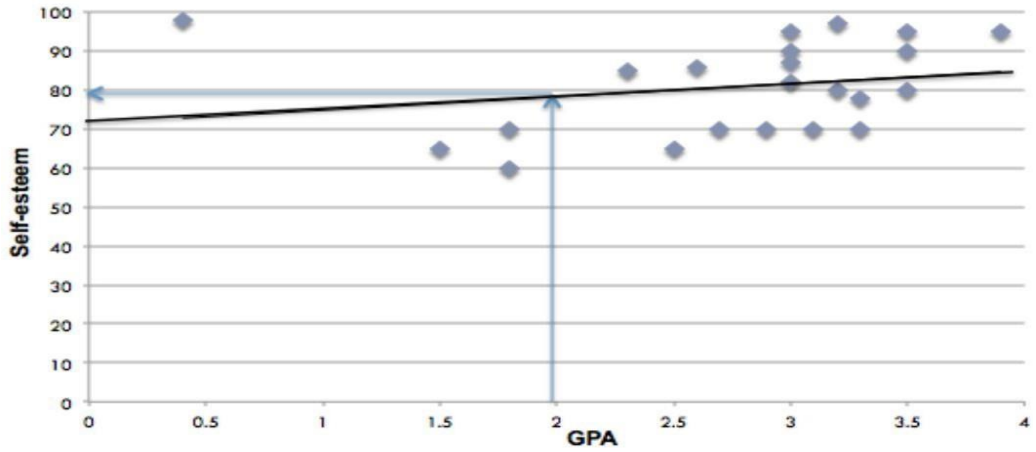
3.4.3 REGRESSION LINE:

- Define regression lines or why regression lines are important?(2M)
- Explain regression line and give an example(13M)

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. A regression line can be used to predict the value of y for a given value of x.

Regression analysis identifies a regression line. The regression line shows how much and in what direction the response variable changes when the explanatory variable changes. Most individuals in the sample are not located exactly on the line; the line closely approximates all the points. The way this line is computed will be described in more detail later

Example: Predict a student's self-esteem score based on her GPA



- The purpose of a regression line is to make predictions.

- In the above example, it is known how GPA is related to self-esteem. Therefore, if a student's self-esteem has not been measured, but her GPA is known, her self-esteem score can be predicted based on her GPA.
- As an example, if a student has a GPA of 2.0, this score matches up with a score of approximately 78 or 79 on the self-esteem scale. This score has been estimated by looking at the graph.

  ✓ Draw a straight line up from the point that represents a 2.0 GPA and find where this line intersects with the regression line.

  ✓ Then, draw a line straight from this point to the self-esteem axis to find the corresponding self-esteem score.

3.4.4 LEAST SQUARE METHOD:

- ❖ Define least square method(2M)
- ❖ What is the formula to calculate Least Square Regression?(6M)
- ❖ Define Least Square Regression Line(2M)
- ❖ Explain Least Square Regression line With example(13M)

The least squares method is a form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points. Each point of data represents the relationship between a known independent variable and an unknown dependent variable.

This method of regression analysis begins with a set of data points to be plotted on an x- and y-axis graph. An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

The least squares method is used in a wide variety of fields, including finance and investing. For financial analysts, the method can help to quantify the relationship between two or more variables—such as a stock's share price and its earnings per share (EPS). By performing this type of analysis investors often try to predict the future behavior of stock prices or other factors.

3.4.4.1 FORMULA TO CALCULATE LEAST SQUARE REGRESSION:

The regression line under the Least Squares method is calculated using the following formula –

$y = a + bx$

Where,

y = dependent variable

x = independent variable

a = y-intercept

b = slope of the line

The slope of line b is calculated using the following formula –

$$b = \frac{\Sigma(x-\bar{x})\,(y-\overline{y})}{\Sigma(x-\overline{x})^2}$$

Or

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Y-intercept, 'a' is calculated using the following formula –

$$a = \frac{\sum y - (b \sum x)}{n}$$

Where,

y = dependent variable

x = independent variable

a = y-intercept

b = slope of the line

The slope of line b is calculated using the following formula –

$$b = \frac{\sum(x-\bar{x})\,(y-\overline{y})}{\sum(x-\overline{x})^2}$$

Or

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Y-intercept, 'a' is calculated using the following formula –

$$a = \frac{\sum y - (b \sum x)}{n}$$

3.4.4.2 LEAST SQUARE REGRESSION LINE:

If the data shows a leaner relationship between two variables, the line that best fits this linear relationship is known as a least-squares regression line, which minimizes the vertical distance from the data points to the regression line. The term "least squares" is used because it is the smallest sum of squares of errors, which is also called the "variance."

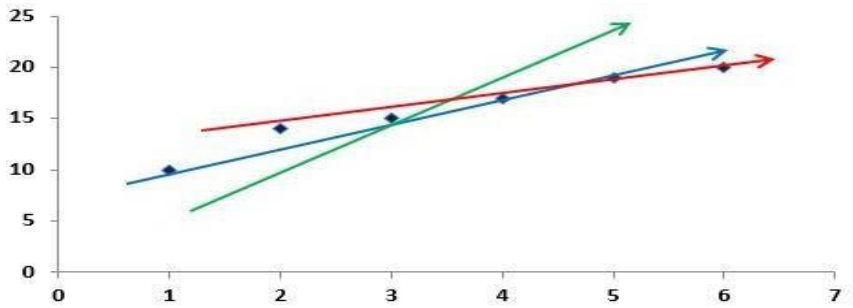
In regression analysis, dependent variables are illustrated on the vertical y-axis, while independent variables are illustrated on the horizontal x-axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.

In contrast to a linear problem, a non-linear least-squares problem has no closed solution and is generally solved by iteration.

EXAMPLE:

The line of best fit is a straight line drawn through a scatter of data points that best represents the relationship between them.

Let us consider the following graph wherein a set of data is plotted along the x and y-axis. These data points are represented using the blue dots. Three lines are drawn through these points – a green, a red, and a blue line. The green line passes through a single point, and the red line passes through three data points. However, the blue line passes through four data points, and the distance between the residual points to the blue line is minimal as compared to the other two lines.



In the above graph, the blue line represents the line of best fit as it lies closest to all the values and the distance between the points outside the line to the line is minimal (i.e., the distance between the residuals to the line of best fit – also referred to as the sums of squares of residuals). In the other two lines, the orange and the green, the distance between the residuals to the lines is greater as compared to the blue line.

3.4.5 STANDARD ERROR OF ESTIMATE:

- ❖ Define Standard Error of Estimate(2M)
- ❖ How Standard Error of Estimate is calculated?(6M)

The standard error of the estimate is a way to measure the accuracy of the predictions made by a regression model.

Likewise, a standard deviation which measures the variation in the set of data from its mean, the standard error of estimate also measures the variation in the actual values of Y from the computed values of Y (predicted) on the regression line. It is computed as a standard deviation, and here the deviations are the vertical distance of every dot from the line of average relationship.

Often denoted σest, it is calculated as:

$\sigma est = \sqrt{\Sigma(y - y)2/n}$

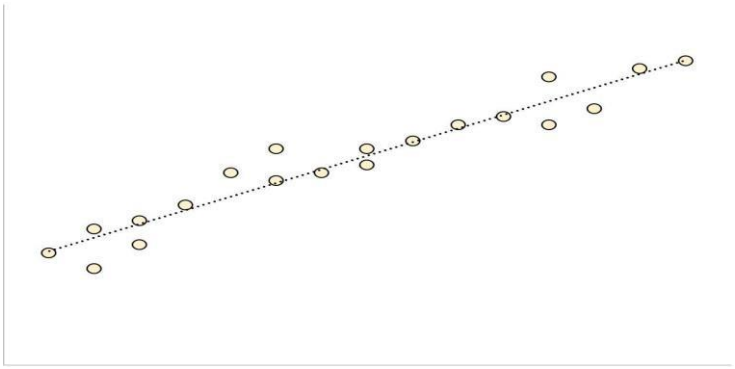
where:

- y: The observed value
- y: The predicted value
- n: The total number of observations

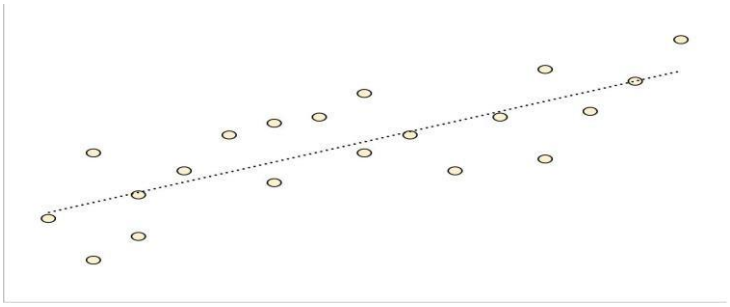
The standard error of the estimate gives us an idea of how well a regression model fits a data set. In particular:

- The smaller the value, the better the fit.
- The larger the value, the worse the fit.

For a regression model that has a small standard error of the estimate, the data points will be closely packed around the estimated regression line:



Conversely, for a regression model that has a large standard error of the estimate, the data points will be more loosely scattered around the regression line:



3.4.6 R-SQUARED:

- Explain R-Squared(2M)
- What is the formula to calculate R-Squared?
- How to interpret R-Squared?(16M)

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

After fitting a linear regression model, you need to determine how well the model fits the data.For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good.

R-squared is always between 0 and 100%:

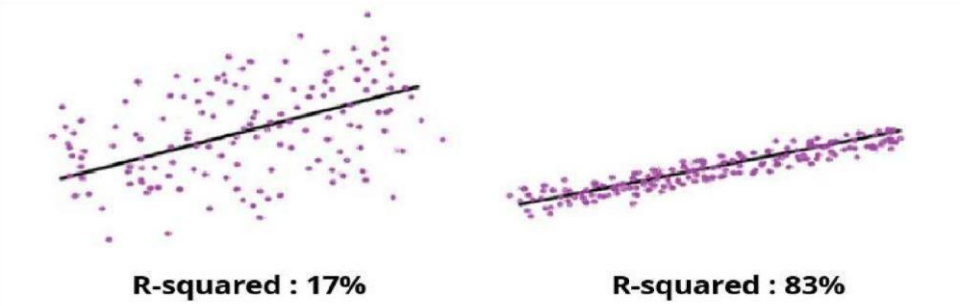
- 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.
- 100% represents a model that explains all the variation in the response variable around its mean.

Usually, the larger the R2, the better the regression model fits your observations.

3.4.6.1 INTERPRETATION OF R2:

Visual Representation of R-squared

You can have a visual demonstration of the plots of fitted values by observed values in a graphical manner. It illustrates how R-squared values represent the scatter around the regression line.



As observed in the pictures above, the value of R-squared for the regression model on the left side is 17%, and for the model on the right is 83%. In a regression model, when the variance accounts to be high, the data points tend to fall closer to the fitted regression line.

However, a regression model with an R2 of 100% is an ideal scenario which is actually not possible. In such a case, the predicted values equal the observed values and it causes all the data points to fall exactly on the regression line.

How to Interpret R squared

The simplest r squared interpretation is how well the regression model fits the observed data values. Let us take an example to understand this.

Consider a model where the R2 value is 70%. Here r squared meaning would be that the model explains 70% of the fitted data in the regression model. Usually, when the R2 value is high, it suggests a better fit for the model.

The correctness of the statistical measure does not only depend on R2 but can depend on other several factors like the nature of the variables, the units on which the variables are measured, etc. So, a high R-squared value is not always likely for the regression model and can indicate problems too.

A low R-squared value is a negative indicator for a model in general. However, if we consider the other factors, a low R2 value can also end up in a good predictive model.

Calculation of R-squared

R- squared can be evaluated using the following formula:

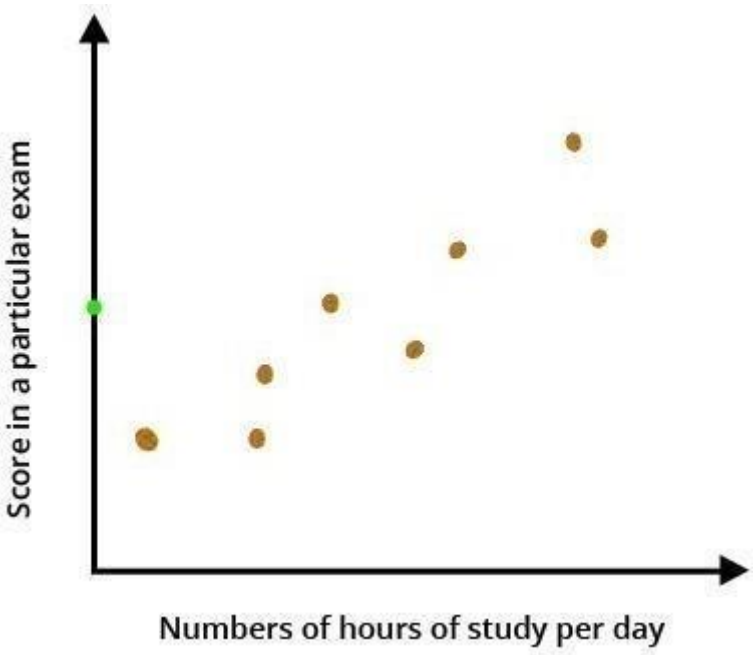
$$\text{R-squared} = \frac{SS_{regresion}}{SS_{total}}$$

Where:

- SSregression – Explained sum of squares due to the regression model.
- SStotal – The total sum of squares.

The sum of squares due to regression assesses how well the model represents the fitted data and the total sum of squares measures the variability in the data used in the regression model.

Now let us come back to the earlier situation where we have two factors: number of hours of study per day and the score in a particular exam to understand the calculation of R-squared more effectively. Here, the target variable is represented by the score and the independent variable by the number of hours of study per day.

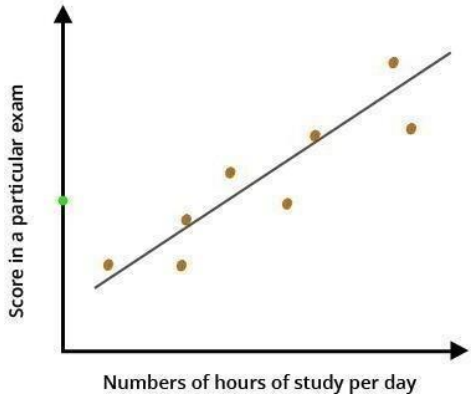


In this case, we will need a simple linear regression model and the equation of the model will be as follows:

$$y = w1x1 + b$$

The parameters w1 and b can be calculated by reducing the squared error over all the data points. The following equation is called the least square function:

$$\text{minimize } \sum(yi - w1x1i - b)2$$



Now, to calculate the goodness-of-fit, we need to calculate the variance:

$$var(u) = 1/n\sum(ui - \bar{u})2$$

where, n represents the number of data points.

Now, R-squared calculates the amount of variance of the target variable explained by the model, i.e. function of the independent variable.

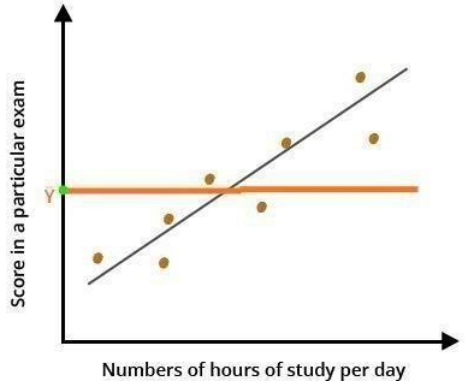
However, in order to achieve that, we need to calculate two things:

- Variance of the target variable:

$var(avg) = \sum(yi - \bar{y})2$

- Variance of the target variable around the best-fit line:

$var(model) = \sum(yi - y)2$



Finally, we can calculate the equation of R-squared as follows:

$R2 = 1 - [var(model)/var(avg)] = 1 - [\sum(yi - y)2/\sum(yi - \bar{y})2]$

3.4.7 MULTIPLE REGRESSION:

- ❖ Explain Multiple regression?(2M)
- ❖ Explain linear regression and multiple regression equation with example(13M)
- ❖ Assumptions of Multiple Regression Equations(2M)
- ❖ Benefits of Multiple Regression Equations(2M)

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

$Y=a+b1X1+b2X3+...+bnXn$

Here Y is the dependent variable, and X1,…,Xn are the *n* independent variables. In calculating the weights, a, b1,…,bn, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables. This is usually done by least squares estimation.

In the case of linear regression, although it is used commonly, it is limited to just one independent and one dependent variable. Apart from that, linear regression restricts the training data set and does not predict a non-linear regression.

For the same limitations and to cover them, we use multiple regression. It focuses on overcoming one particular limitation and that is allowing to analyze more than one independent variable.

3.4.7.1 Multiple regression equation

We will start the discussion by first taking a look at the linear regression equation:

$y = bx + a$

Where,

y is a dependent variable we need to find, x is an independent variable. The constants a and b drive the equation. But according to our definition, as the multiple regression takes several independent variables (x), so for the equation we will have multiple x values too:

$y = b1x1 + b2x2 + \dots bnxn + a$

Here, to calculate the value of the dependent variable y, we have multiple independent variables x1, x2, and so on. The number of independent variables can grow till n and the constant b with every variable denotes its numeric value. The purpose of the constant a is to denote the dependent variable's value in case when all the independent variable values turn to zero.

Example: A researcher decides to study students' performance at a school over a period of time. He observed that as the lectures proceed to operate online, the performance of students started to decline as well. The parameters for the dependent variable "decrease in performance" are various independent variables like "lack of attention, more internet addiction, neglecting studies" and much more.

So for the above example, the multiple regression equation would be:

$y = b1 * attention + b2 * internet\ addiction + b3 * technology\ support + \dots bnxn + a$

3.4.7.2 ASSUMPTIONS OF MULTIPLE REGRESSION ANALYSIS:

- ✧ The variables considered for the model should be relevant and the model should be reliable.
- ✧ The model should be linear and not non-linear.
- ✧ Variables must have a normal distribution
- ✧ The variance should be constant for all levels of the predicted variable.

3.4.7.3 BENEFITS OF MULTIPLE REGRESSION ANALYSIS:

- ✧ Multiple regression analysis helps us to better study the various predictor variables at hand.

- ✧ It increases reliability by avoiding dependency on just one variable and having more than one independent variable to support the event.
- ✧ Multiple regression analysis permits you to study more formulated hypotheses that are possible.

3.4.8 REGRESSION TOWARDS THE MEAN:

- ❖ Define regression towards mean(2M)
- ❖ Explain regression towards mean with example(13M)

In statistics, regression toward the mean (also called reversion to the mean, and reversion to mediocrity) is a concept that refers to the fact that if one sample of a random variable is extreme, the next sampling of the same random variable is likely to be closer to its mean. Furthermore, when many random variables are sampled and the most extreme results are intentionally picked out, it refers to the fact that (in many cases) a second sampling of these picked-out variables will result in "less extreme" results, closer to the initial mean of all of the variables.

Regression to the mean usually happens because of sampling error. A good sampling technique is to randomly sample from the population. If you don't (i.e. if you asymmetrically sample), then your results may be abnormally high or low for the average and therefore would regress back to the mean. Regression to the mean can also happen because you take a very small, unrepresentative sample (say, the highest 1 percent of the population or the lowest ten percent).

Formula for the Percent of Regression to the Mean:

You can use the following formula to find the percent for any set of data:

Percent of Regression to the Mean = 100(1-r)

where r is the correlation coefficient.

Why 1-r?
Note: In order to understand this discussion you should be very familiar with r, the correlation coefficient.

The percent of regression to the mean takes into account the correlation between the variables. Take two extremes:
If r=1 (i.e. perfect correlation), then 1-1 = 0 and the regression to the mean is zero. In other words, if your data has perfect correlation, it will never regress to the mean.
With an r of zero, there is 100 percent regression to the mean. In other words, data with an r of zero will *always* regress to the mean.

EXAMPLE:

If your favorite team won the championship last year, what does that mean for their chances for winning next season? This is an important question, often with money or pride on the line (The League, anyone?). To the extent this is due to skill (the team is in good condition, top coach etc.), their win signals that it's more likely they'll win next year. But the greater the extent this is due to luck (other teams embroiled in a drug scandal, favourable draw, draft picks turned out well etc.), the less likely it is they'll win next year. This is because of the statistical concept of regression to the mean.

Another example,

because of regression toward the mean, we would expect that students who made the top five scores on the first statistics exam would not make the top five scores on the second statistics exam. Although all five students might score above the mean on the second exam, some of their scores would regress back toward the mean. Most likely, the top five scores on the first exam reflect two components. One relatively permanent component reflects the fact that these students are superior because of good study habits, a strong aptitude for quantitative reasoning, and so forth. The other relatively transitory component reflects the fact that, on the day of the exam, at least some of these students were very lucky because all sorts of little chance factors, such as restful sleep, a pleas ant commute to campus, etc., worked in their favor. On the second test, even though the scores of these five students continue to reflect an above-average permanent component, some of their scores will suffer because of less good luck or even bad luck. The net effect is that the scores of at least some of the original five top students will drop below the top five scores—that is, regress *back* toward the mean—on the second exam. (When signifi cant regression toward the mean occurs after a spectacular performance by, for example, a rookie athlete or a first-time author, the term *sophomore jinx* often is invoked.)
There is good news for those students who made the five lowest scores on the first exam. Although all five students might score below the mean on the second exam, some of their scores probably will regress *up* toward the mean. On the second exam, some of them will not be as unlucky. The net effect is that the scores of at least some ofthe original five lowest scoring students will move above the bottom five scores—that is, regress up toward the mean—on the second