Introduction to Linear Regression

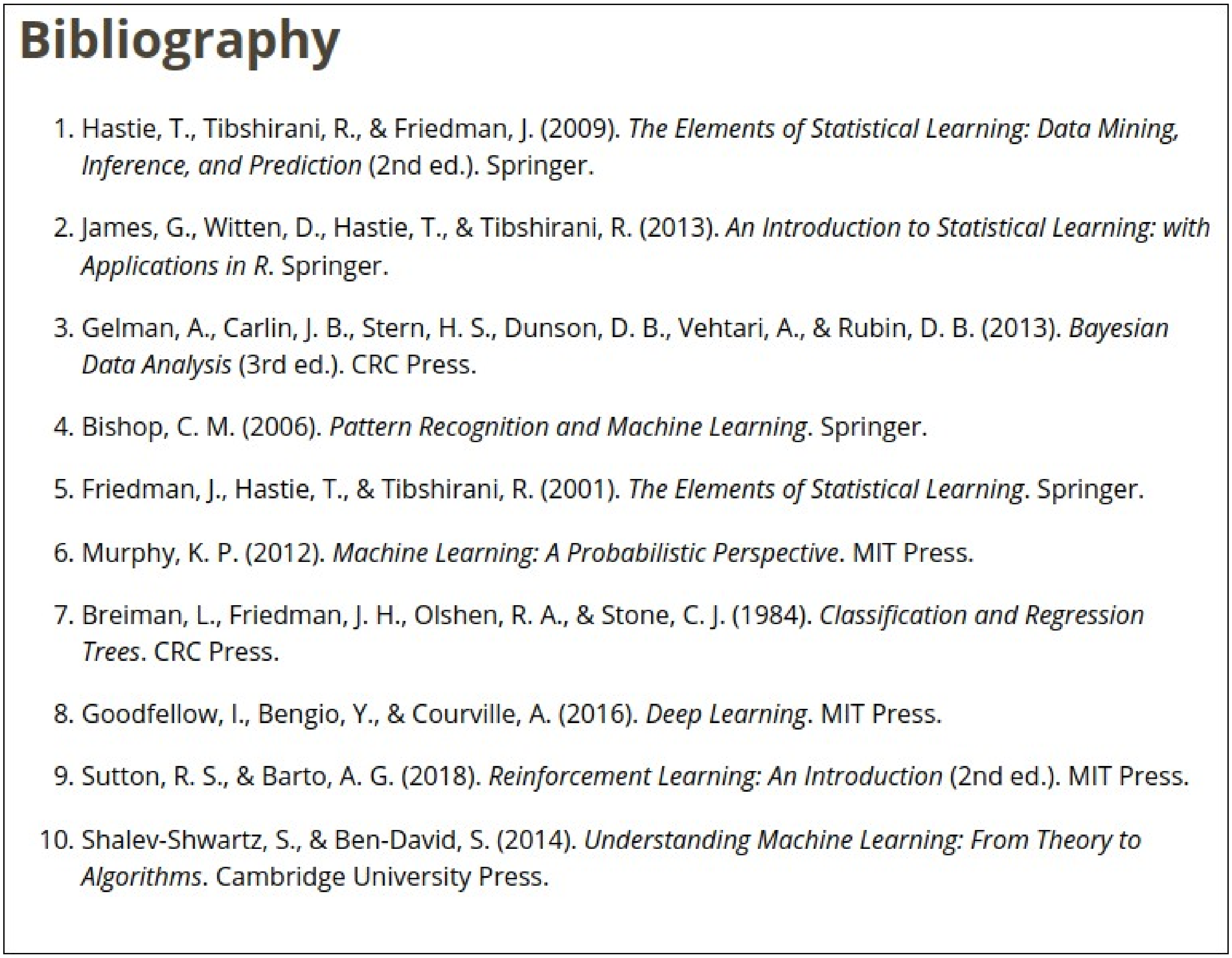
KATHMANDU MODEL
SECONDARY SCHOOL
(KMSS)

Submitted by:
Bhashkar Paudyal
Grade XII ‘D3’
Roll no 12 or 13

Submmission Signature

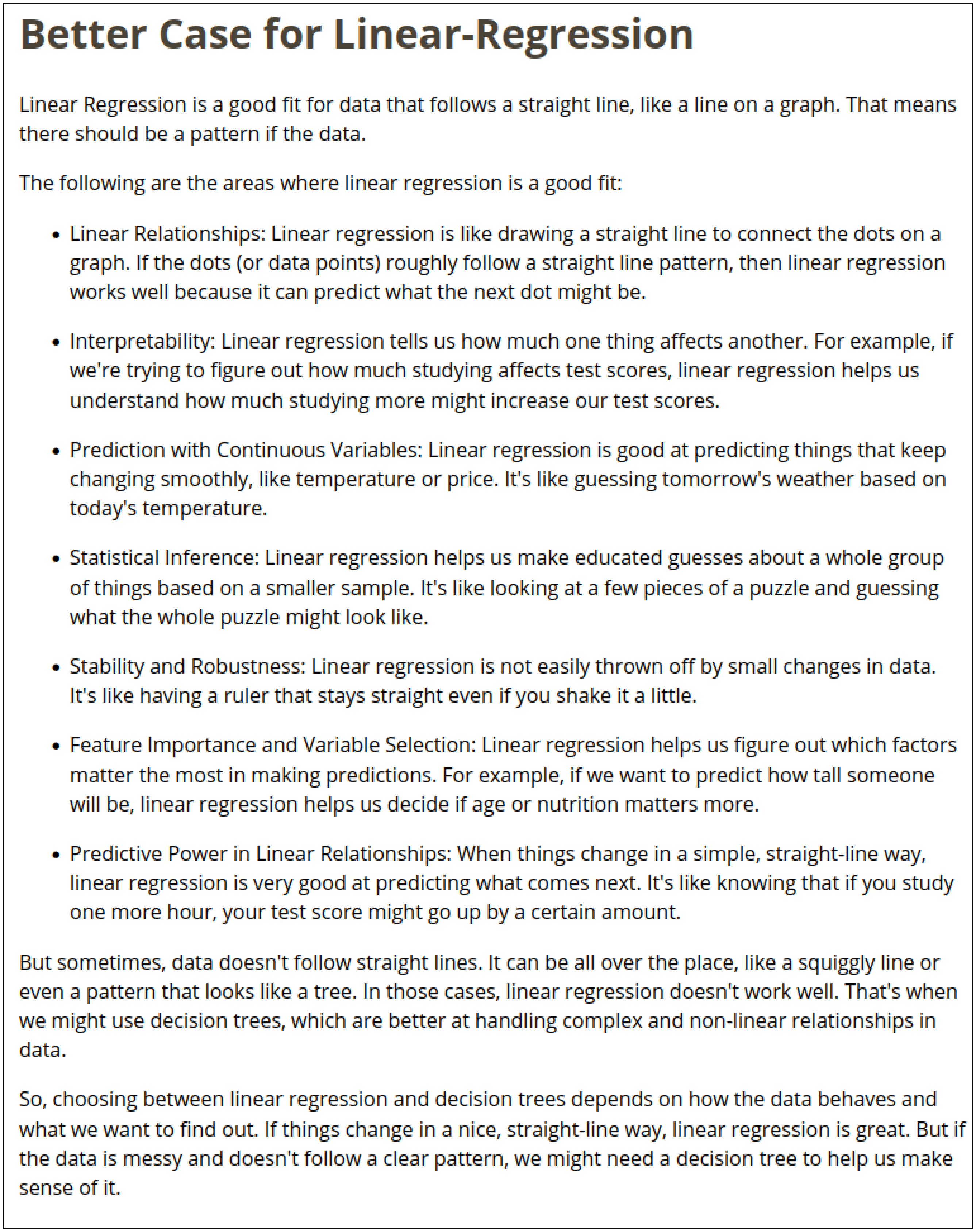
Submitted to:
Math Department

Aprroval Signature



Bibliography

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.
3. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
5. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer.
6. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
7. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. CRC Press.
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
9. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.
10. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.



Better Case for Linear-Regression

Linear Regression is a good fit for data that follows a straight line, like a line on a graph. That means there should be a pattern if the data.

The following are the areas where linear regression is a good fit:

- Linear Relationships: Linear regression is like drawing a straight line to connect the dots on a graph. If the dots (or data points) roughly follow a straight line pattern, then linear regression works well because it can predict what the next dot might be.
- Interpretability: Linear regression tells us how much one thing affects another. For example, if we're trying to figure out how much studying affects test scores, linear regression helps us understand how much studying more might increase our test scores.
- Prediction with Continuous Variables: Linear regression is good at predicting things that keep changing smoothly, like temperature or price. It's like guessing tomorrow's weather based on today's temperature.
- Statistical Inference: Linear regression helps us make educated guesses about a whole group of things based on a smaller sample. It's like looking at a few pieces of a puzzle and guessing what the whole puzzle might look like.
- Stability and Robustness: Linear regression is not easily thrown off by small changes in data. It's like having a ruler that stays straight even if you shake it a little.
- Feature Importance and Variable Selection: Linear regression helps us figure out which factors matter the most in making predictions. For example, if we want to predict how tall someone will be, linear regression helps us decide if age or nutrition matters more.
- Predictive Power in Linear Relationships: When things change in a simple, straight-line way, linear regression is very good at predicting what comes next. It's like knowing that if you study one more hour, your test score might go up by a certain amount.

But sometimes, data doesn't follow straight lines. It can be all over the place, like a squiggly line or even a pattern that looks like a tree. In those cases, linear regression doesn't work well. That's when we might use decision trees, which are better at handling complex and non-linear relationships in data.

So, choosing between linear regression and decision trees depends on how the data behaves and what we want to find out. If things change in a nice, straight-line way, linear regression is great. But if the data is messy and doesn't follow a clear pattern, we might need a decision tree to help us make sense of it.

Linear Regression

Linear Regression, refers to a statistical method used to model the relationship between one or more independent variables and a dependent variable having a linear relationship.

The core idea of Linear regression analysis is to estimate the parameters of a linear equation that best describes the relationship between the variables. The model can then be used for prediction.

The Linear regression equation expresses the relationship between the independent variables and the dependent variable. It is always linear as this regression assumes linearity of the data. i.e when x changes y changes with a constant ratio and a bias factor which is seen as the constant in the linear equation.

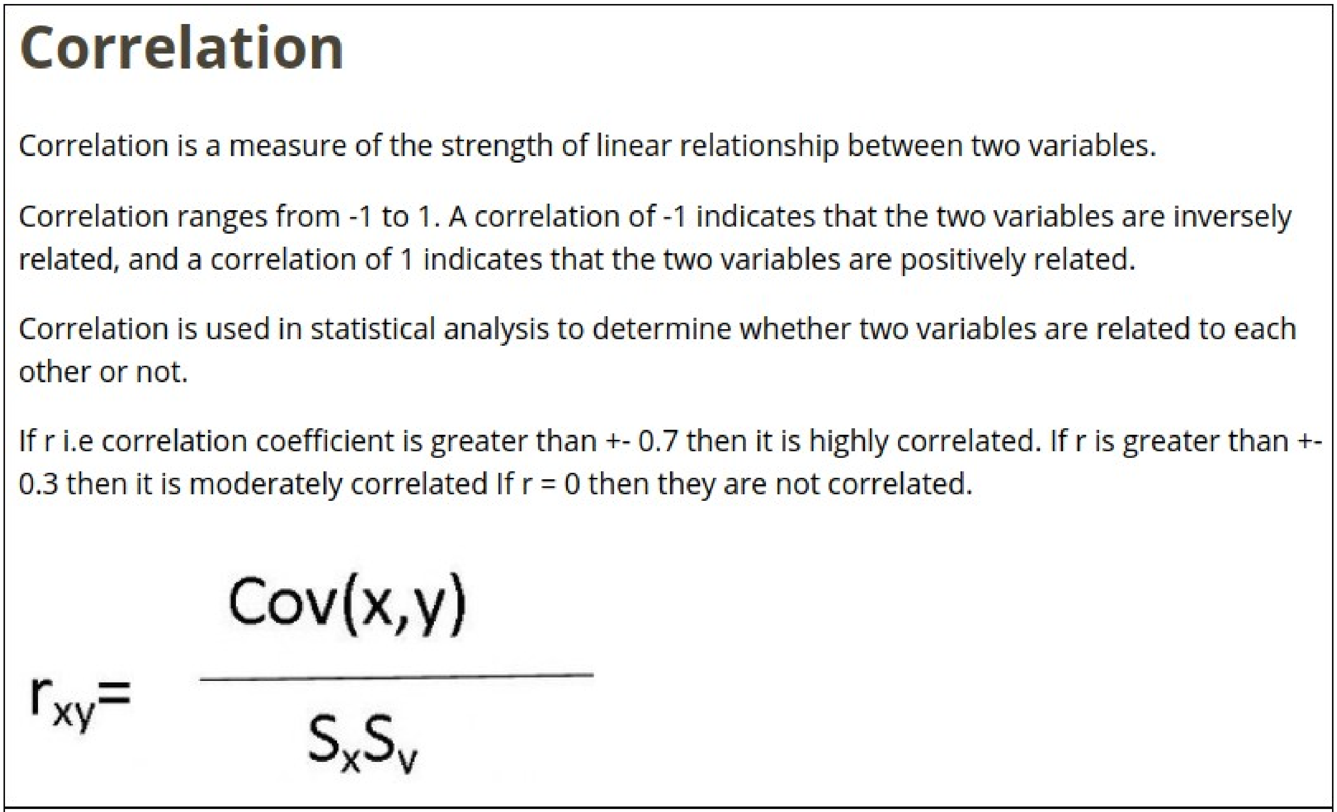
So if the variables have a exponential relationship, the equation will give a very inaccurate prediction.

So we need to determine if the relationship is linear or not before doing linear regression analysis.

So In this case the best and most used methods are :

1. Scatter plot
2. Correlation coefficient

If there are certain segments where the ratio and bias are different, one can use different correlation coefficients to fit into many linear regression equation.



Correlation

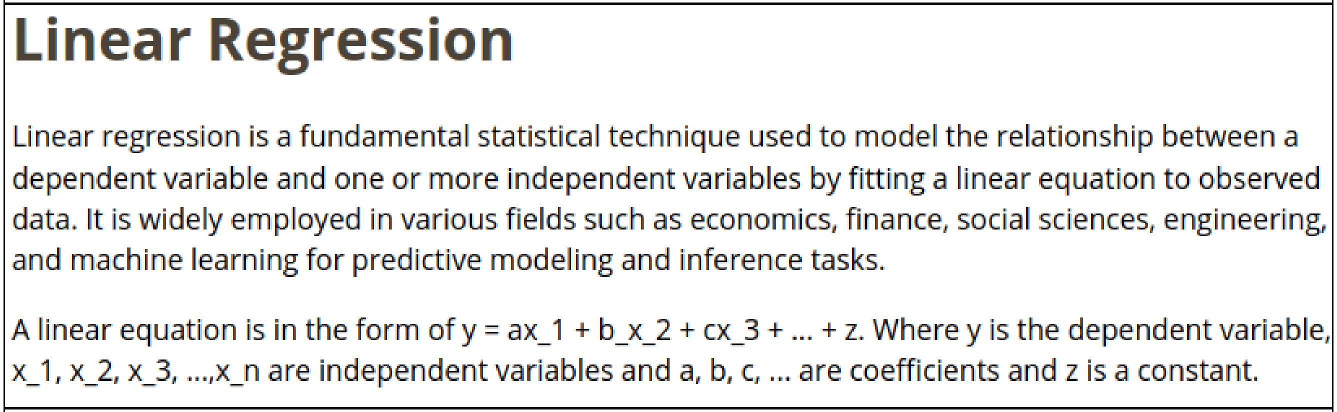
Correlation is a measure of the strength of linear relationship between two variables.

Correlation ranges from -1 to 1. A correlation of -1 indicates that the two variables are inversely related, and a correlation of 1 indicates that the two variables are positively related.

Correlation is used in statistical analysis to determine whether two variables are related to each other or not.

If r i.e correlation coefficient is greater than +- 0.7 then it is highly correlated. If r is greater than +- 0.3 then it is moderately correlated If r = 0 then they are not correlated.

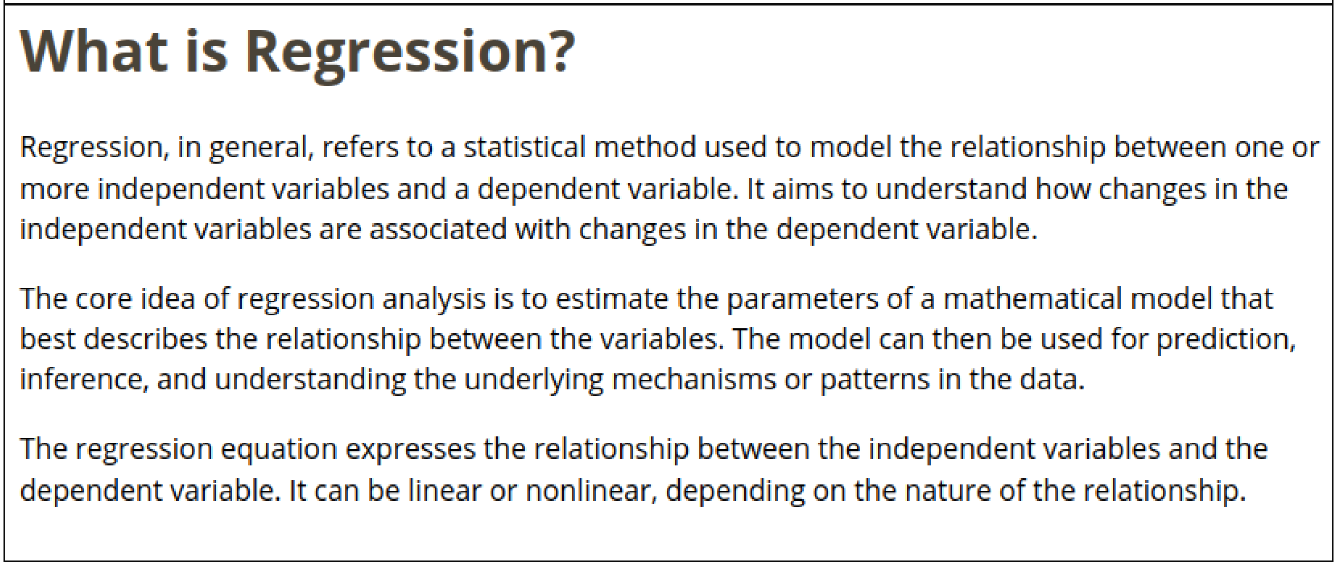
$$r_{xy} = \frac{Cov(x,y)}{S_x S_v}$$



Linear Regression

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It is widely employed in various fields such as economics, finance, social sciences, engineering, and machine learning for predictive modeling and inference tasks.

A linear equation is in the form of y = ax_1 + b_x_2 + cx_3 + ... + z. Where y is the dependent variable, x_1, x_2, x_3, ...,x_n are independent variables and a, b, c, ... are coefficients and z is a constant.



What is Regression?

Regression, in general, refers to a statistical method used to model the relationship between one or more independent variables and a dependent variable. It aims to understand how changes in the independent variables are associated with changes in the dependent variable.

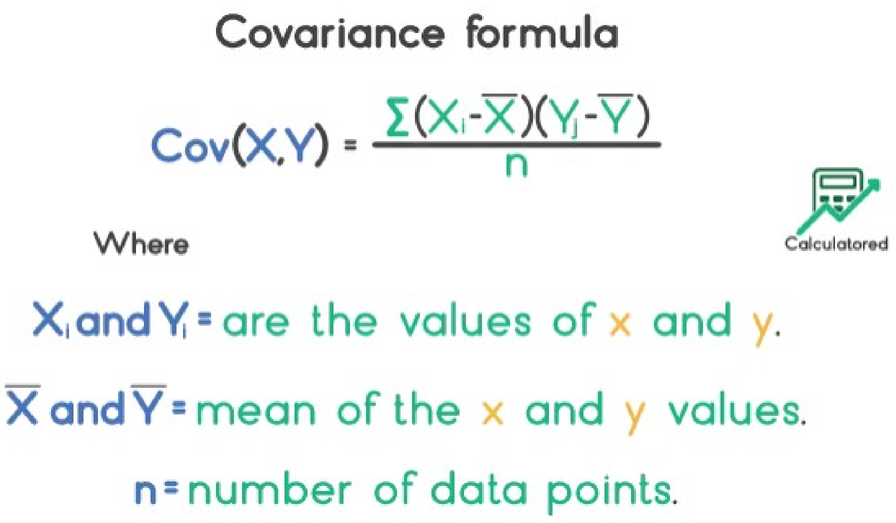
The core idea of regression analysis is to estimate the parameters of a mathematical model that best describes the relationship between the variables. The model can then be used for prediction, inference, and understanding the underlying mechanisms or patterns in the data.

The regression equation expresses the relationship between the independent variables and the dependent variable. It can be linear or nonlinear, depending on the nature of the relationship.

Multi Variable Statistics

Multi-variable statistics, also known as multivariate statistics, deals with the analysis of data sets that involve more than one variable. In contrast to univariate statistics, which focus on analyzing a single variable at a time, multivariate statistics examine the relationships and interactions between multiple variables simultaneously.

Covariance



Covariance formula

$$\mathrm{Cov}(X,Y) = \frac{\Sigma(X_i-\overline{X})(Y_j-\overline{Y})}{n}$$

Where

Calculatored

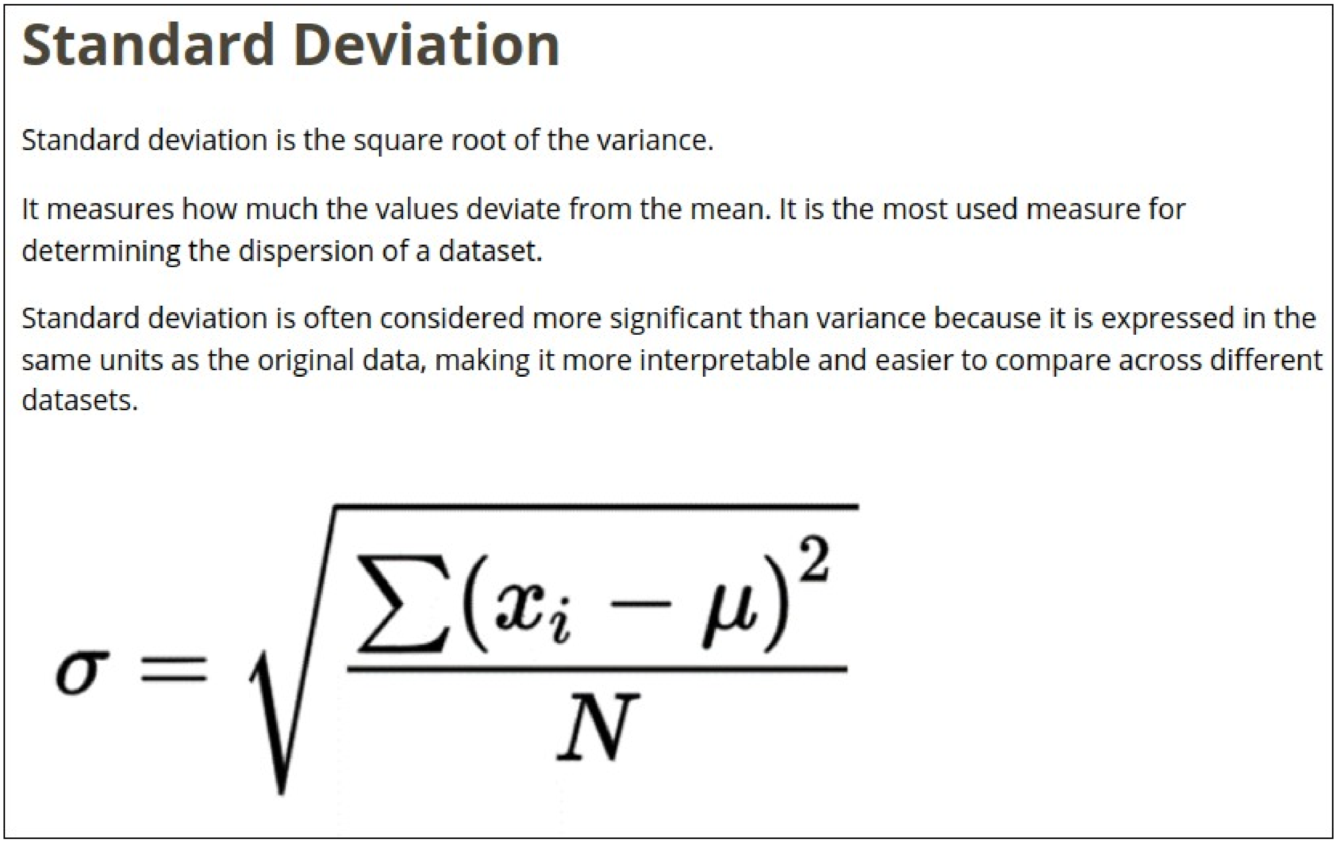
X_i and Y_i = are the values of x and y.

$\overline{X}$ and $\overline{Y}$ = mean of the x and y values.

n = number of data points.

Source : Calculatored

- Covariance is a statistical tool used to determine the relationship between the movements of two random variables
- When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.
- Covariance is different from the correlation coefficient, a measure of the strength of a correlative relationship.



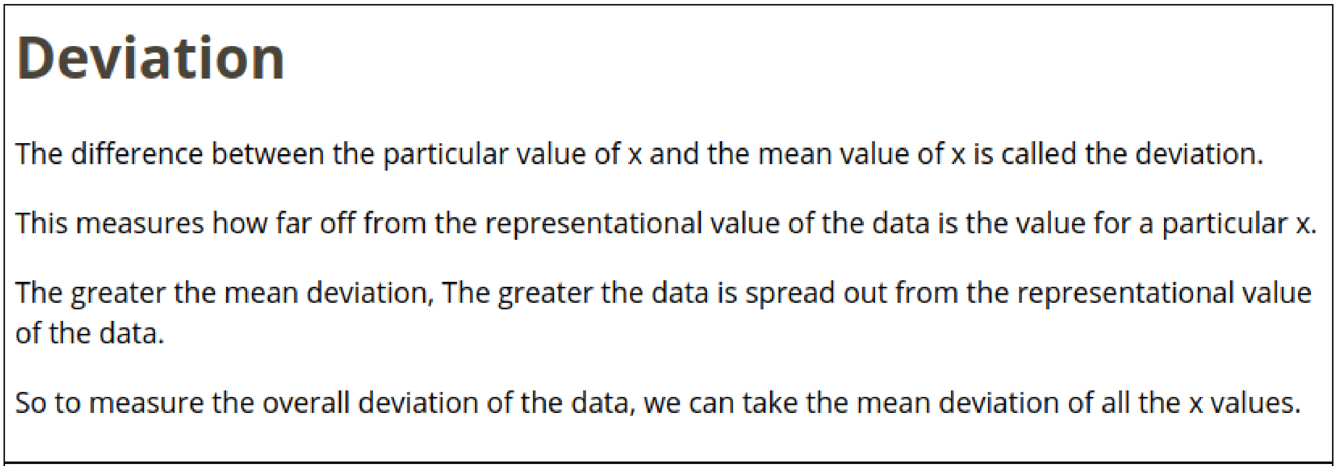
Standard Deviation

Standard deviation is the square root of the variance.

It measures how much the values deviate from the mean. It is the most used measure for determining the dispersion of a dataset.

Standard deviation is often considered more significant than variance because it is expressed in the same units as the original data, making it more interpretable and easier to compare across different datasets.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$



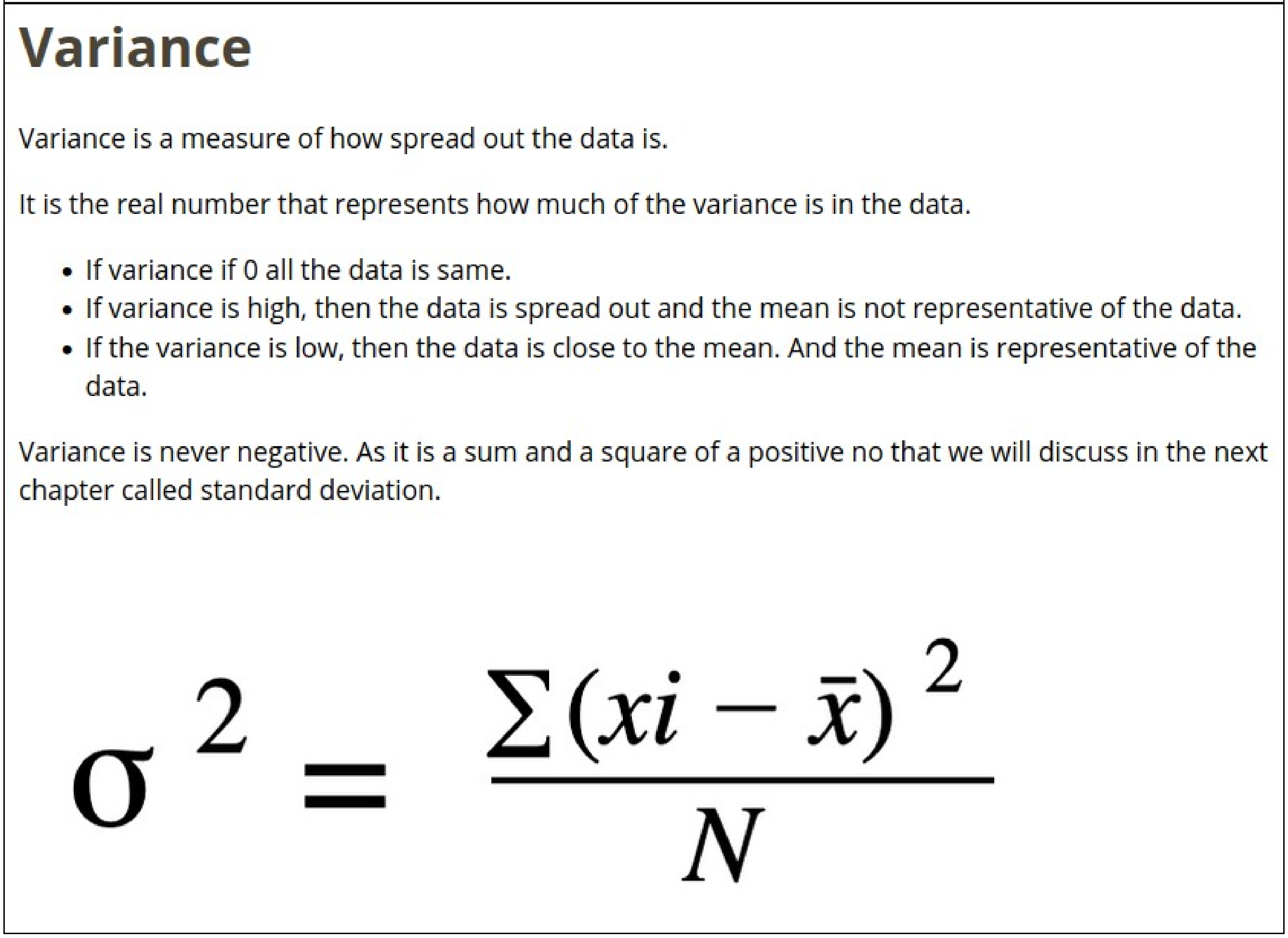
Deviation

The difference between the particular value of x and the mean value of x is called the deviation.

This measures how far off from the representational value of the data is the value for a particular x.

The greater the mean deviation, The greater the data is spread out from the representational value of the data.

So to measure the overall deviation of the data, we can take the mean deviation of all the x values.



Variance

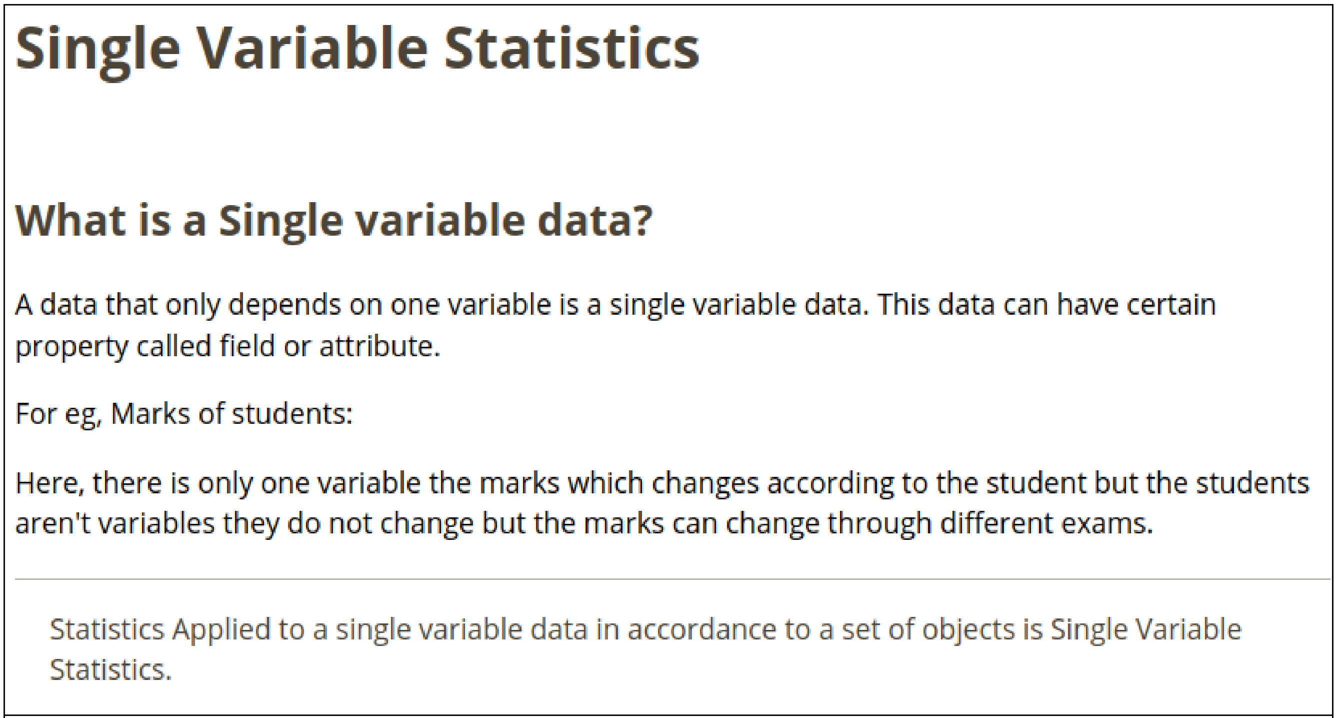
Variance is a measure of how spread out the data is.

It is the real number that represents how much of the variance is in the data.

- If variance if 0 all the data is same.
- If variance is high, then the data is spread out and the mean is not representative of the data.
- If the variance is low, then the data is close to the mean. And the mean is representative of the data.

Variance is never negative. As it is a sum and a square of a positive no that we will discuss in the next chapter called standard deviation.

$$\sigma^2 = \frac{\sum(xi - \bar{x})^2}{N}$$



Single Variable Statistics

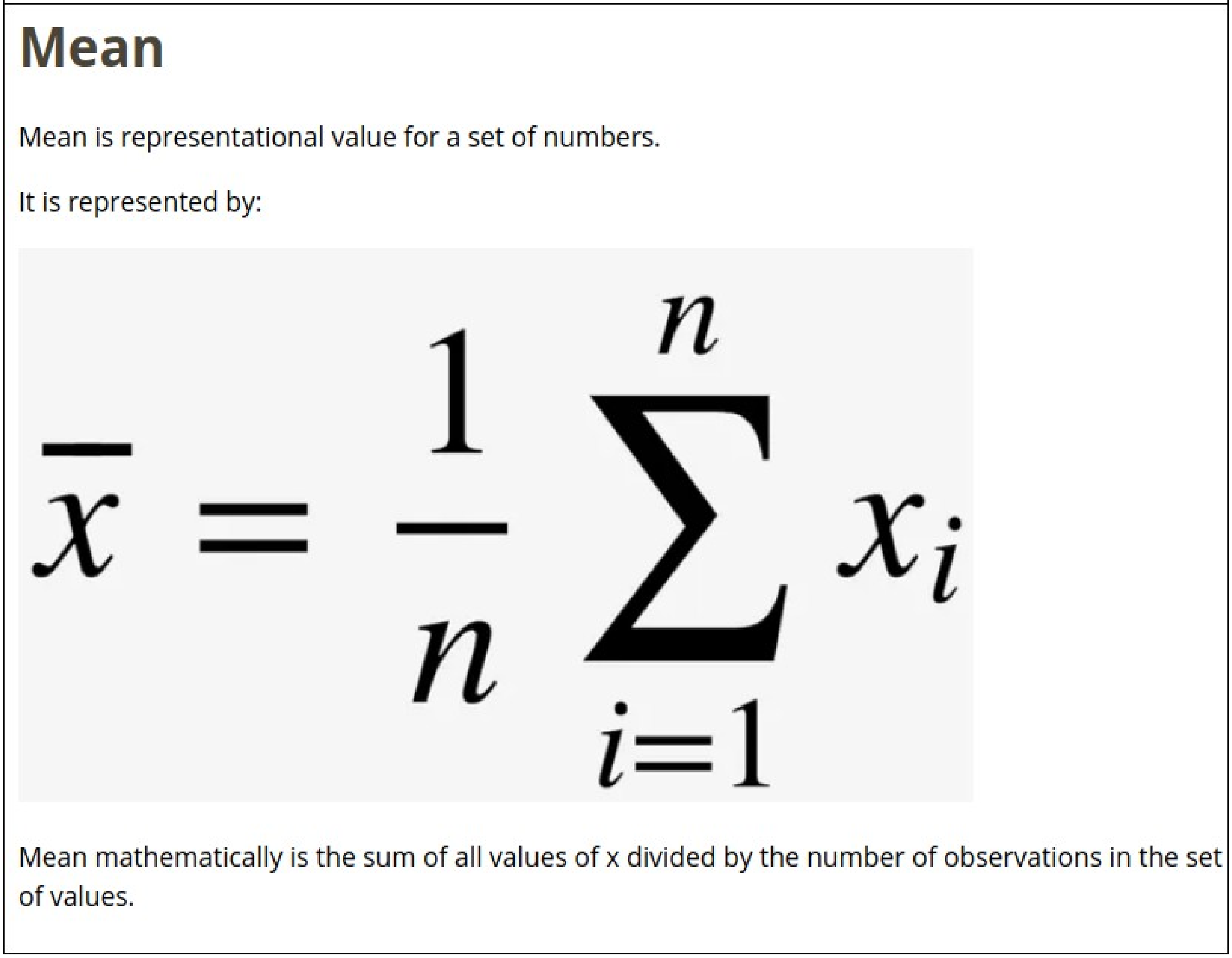
What is a Single variable data?

A data that only depends on one variable is a single variable data. This data can have certain property called field or attribute.

For eg, Marks of students:

Here, there is only one variable the marks which changes according to the student but the students aren't variables they do not change but the marks can change through different exams.

> Statistics Applied to a single variable data in accordance to a set of objects is Single Variable Statistics.



Mean

Mean is representational value for a set of numbers.

It is represented by:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

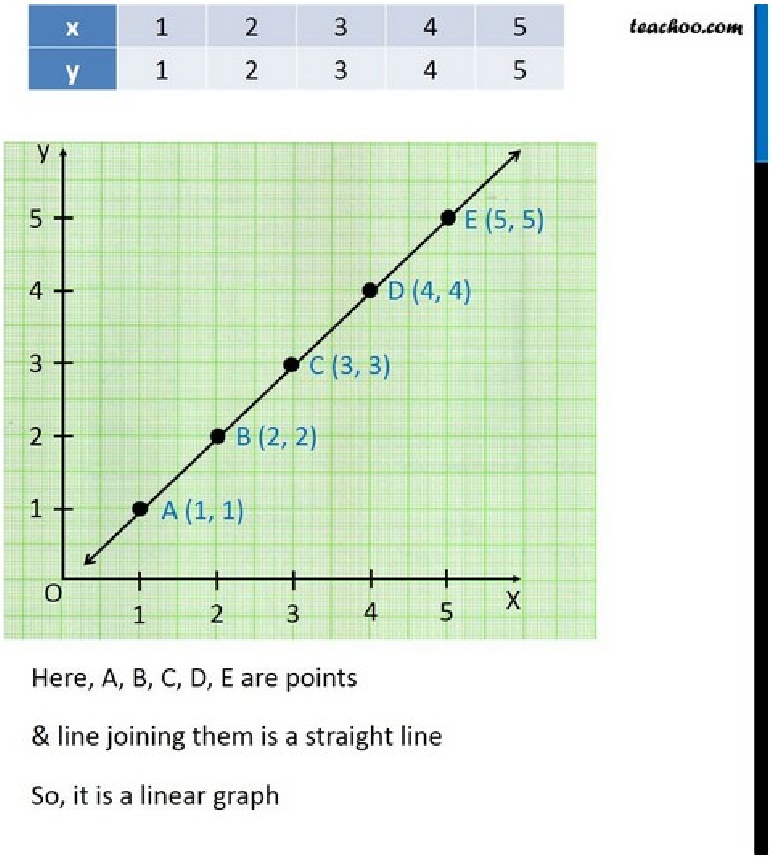
Mean mathematically is the sum of all values of x divided by the number of observations in the set of values.

Linear Graph

A line can be represented in the form of a graph. This is the most common graph representation. This type of graph is called linear graph.

The linear graph corresponds to two dimensions. comprising of two variables.

Here one is the function of other. Generally, the one that is the function is shown in vertical axis and the other in horizontal axis.



x	1	2	3	4	5
y	1	2	3	4	5

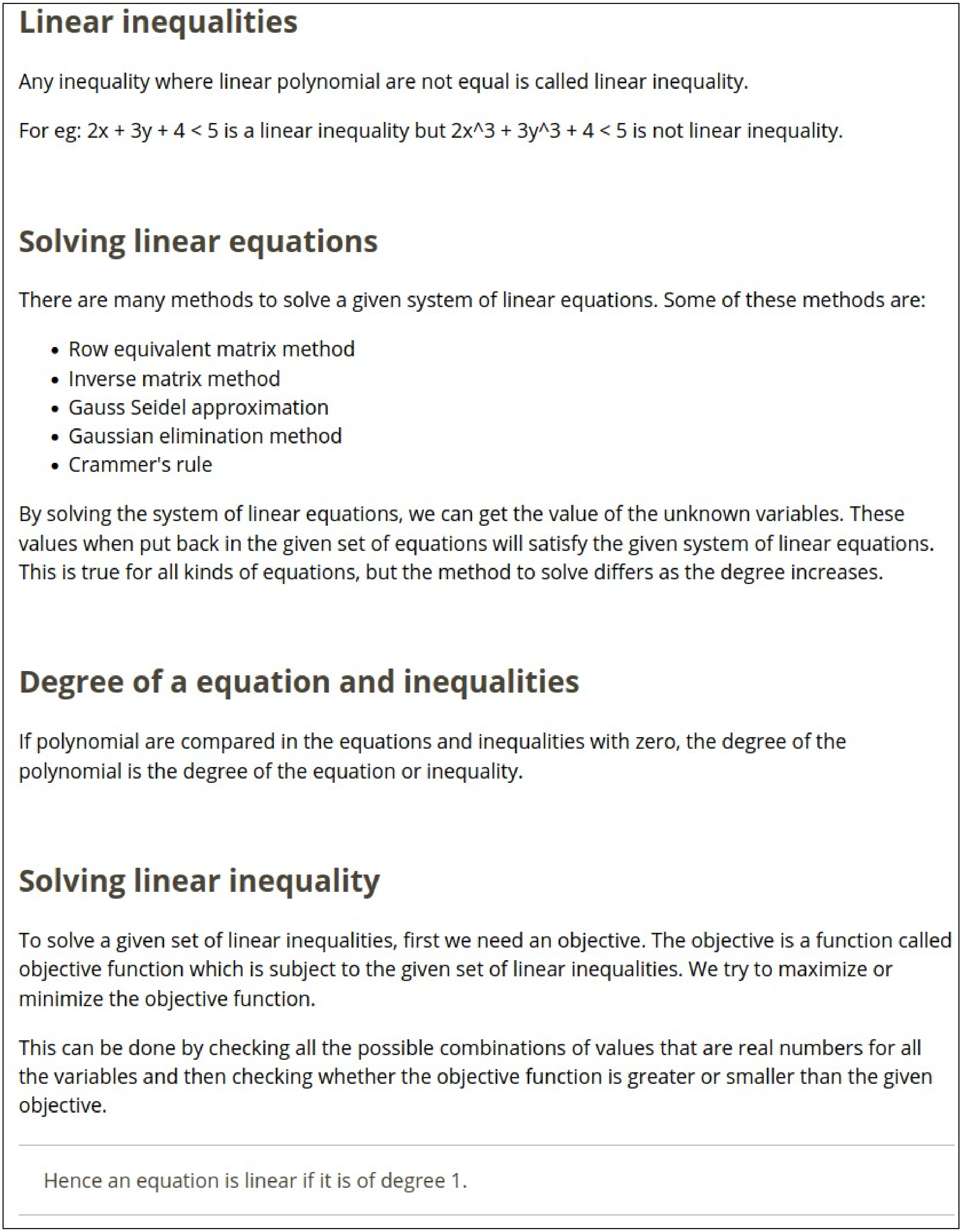
Here, A, B, C, D, E are points

& line joining them is a straight line

So, it is a linear graph

As show in the figure above, two point join to make a straight line.

Here, y is the function of x.



Linear inequalities

Any inequality where linear polynomial are not equal is called linear inequality.

For eg: $2x + 3y + 4 < 5$ is a linear inequality but $2x^3 + 3y^3 + 4 < 5$ is not linear inequality.

Solving linear equations

There are many methods to solve a given system of linear equations. Some of these methods are:

- Row equivalent matrix method
- Inverse matrix method
- Gauss Seidel approximation
- Gaussian elimination method
- Crammer's rule

By solving the system of linear equations, we can get the value of the unknown variables. These values when put back in the given set of equations will satisfy the given system of linear equations. This is true for all kinds of equations, but the method to solve differs as the degree increases.

Degree of a equation and inequalities

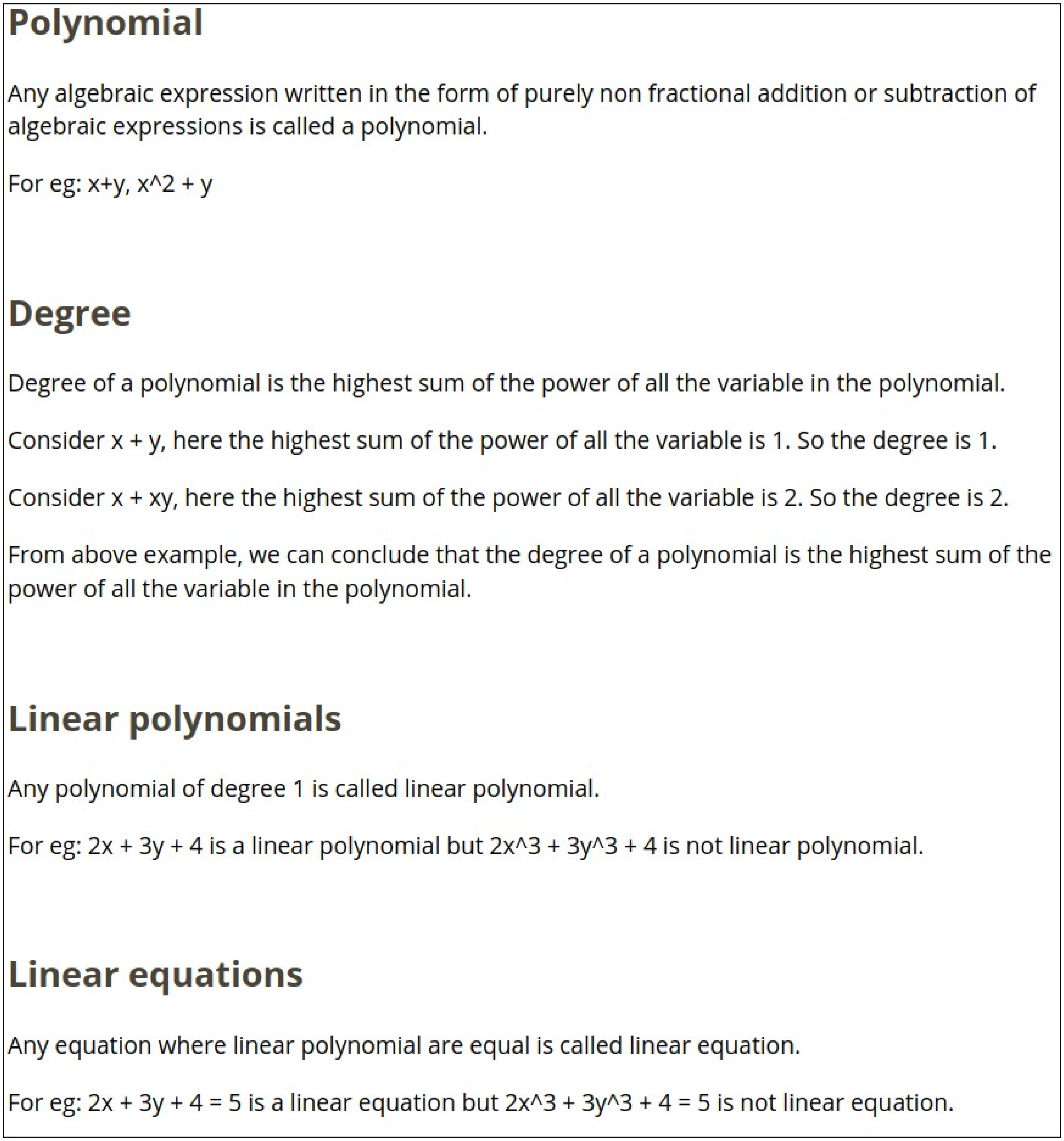
If polynomial are compared in the equations and inequalities with zero, the degree of the polynomial is the degree of the equation or inequality.

Solving linear inequality

To solve a given set of linear inequalities, first we need an objective. The objective is a function called objective function which is subject to the given set of linear inequalities. We try to maximize or minimize the objective function.

This can be done by checking all the possible combinations of values that are real numbers for all the variables and then checking whether the objective function is greater or smaller than the given objective.

> Hence an equation is linear if it is of degree 1.



Polynomial

Any algebraic expression written in the form of purely non fractional addition or subtraction of algebraic expressions is called a polynomial.

For eg: x+y, x^2 + y

Degree

Degree of a polynomial is the highest sum of the power of all the variable in the polynomial.

Consider x + y, here the highest sum of the power of all the variable is 1. So the degree is 1.

Consider x + xy, here the highest sum of the power of all the variable is 2. So the degree is 2.

From above example, we can conclude that the degree of a polynomial is the highest sum of the power of all the variable in the polynomial.

Linear polynomials

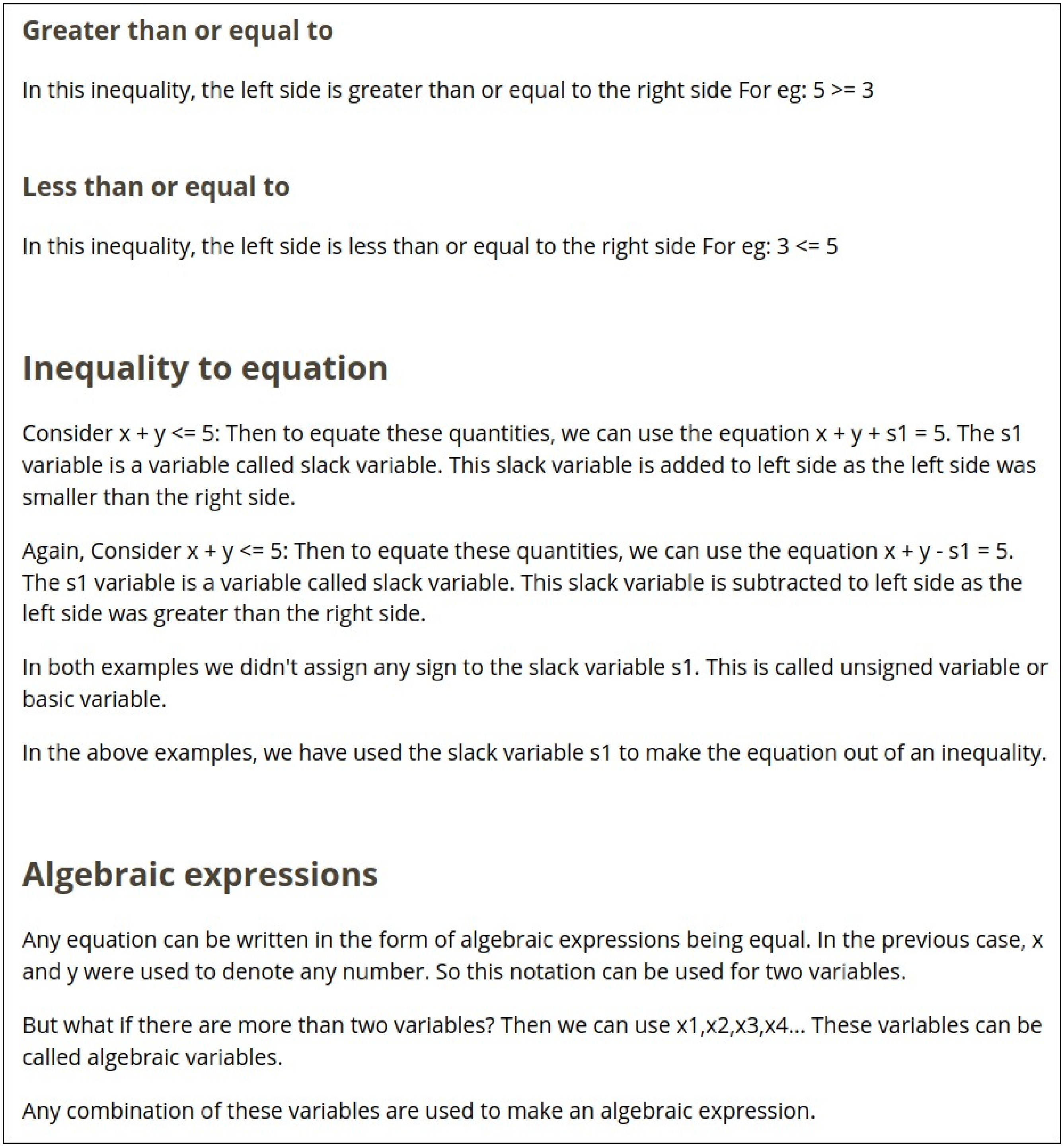
Any polynomial of degree 1 is called linear polynomial.

For eg: 2x + 3y + 4 is a linear polynomial but 2x^3 + 3y^3 + 4 is not linear polynomial.

Linear equations

Any equation where linear polynomial are equal is called linear equation.

For eg: 2x + 3y + 4 = 5 is a linear equation but 2x^3 + 3y^3 + 4 = 5 is not linear equation.



Greater than or equal to

In this inequality, the left side is greater than or equal to the right side For eg: 5 >= 3

Less than or equal to

In this inequality, the left side is less than or equal to the right side For eg: 3 <= 5

Inequality to equation

Consider x + y <= 5: Then to equate these quantities, we can use the equation x + y + s1 = 5. The s1 variable is a variable called slack variable. This slack variable is added to left side as the left side was smaller than the right side.

Again, Consider x + y <= 5: Then to equate these quantities, we can use the equation x + y - s1 = 5. The s1 variable is a variable called slack variable. This slack variable is subtracted to left side as the left side was greater than the right side.

In both examples we didn't assign any sign to the slack variable s1. This is called unsigned variable or basic variable.

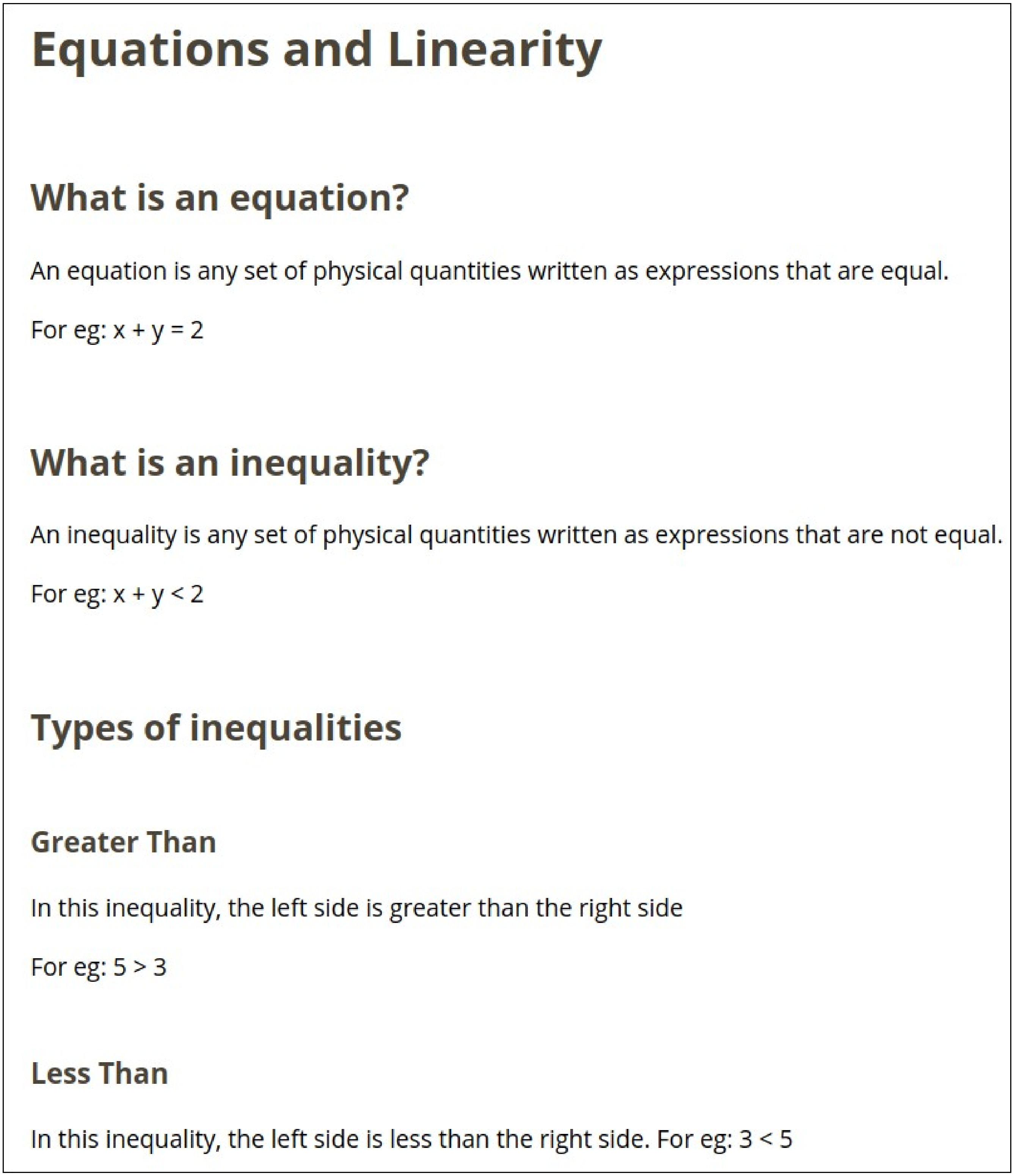
In the above examples, we have used the slack variable s1 to make the equation out of an inequality.

Algebraic expressions

Any equation can be written in the form of algebraic expressions being equal. In the previous case, x and y were used to denote any number. So this notation can be used for two variables.

But what if there are more than two variables? Then we can use x1,x2,x3,x4... These variables can be called algebraic variables.

Any combination of these variables are used to make an algebraic expression.



Equations and Linearity

What is an equation?

An equation is any set of physical quantities written as expressions that are equal.

For eg: $x + y = 2$

What is an inequality?

An inequality is any set of physical quantities written as expressions that are not equal.

For eg: $x + y < 2$

Types of inequalities

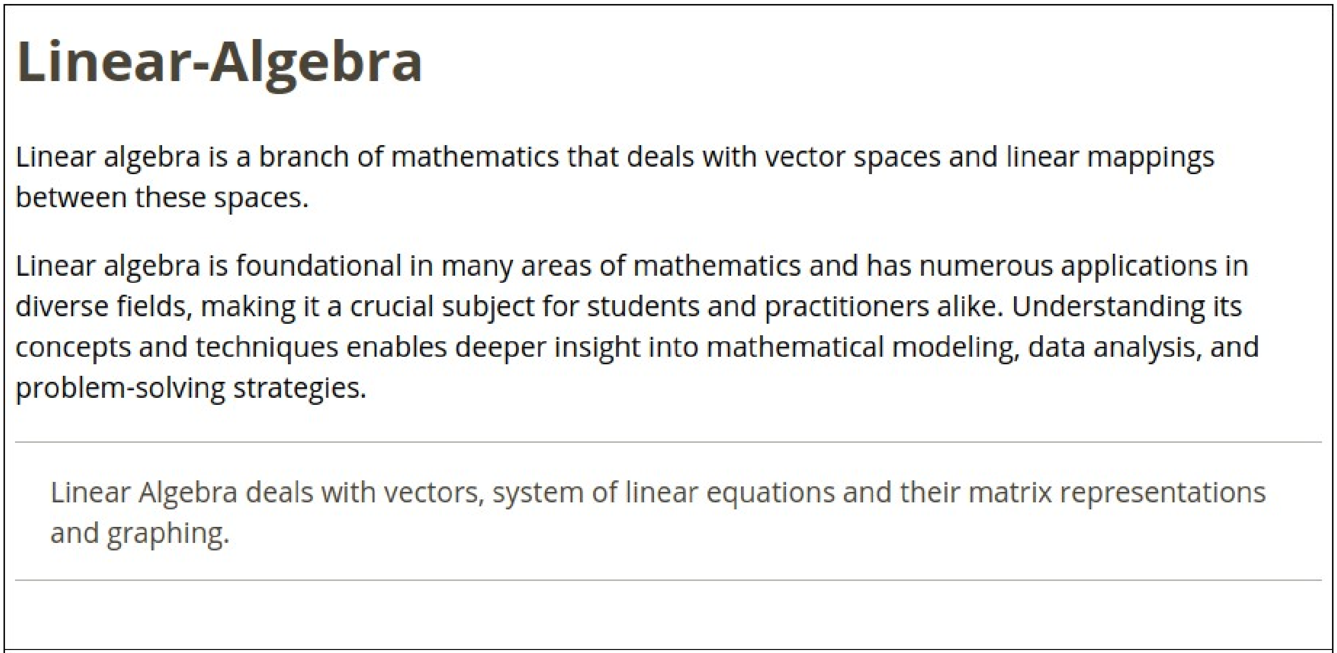
Greater Than

In this inequality, the left side is greater than the right side

For eg: $5 > 3$

Less Than

In this inequality, the left side is less than the right side. For eg: $3 < 5$

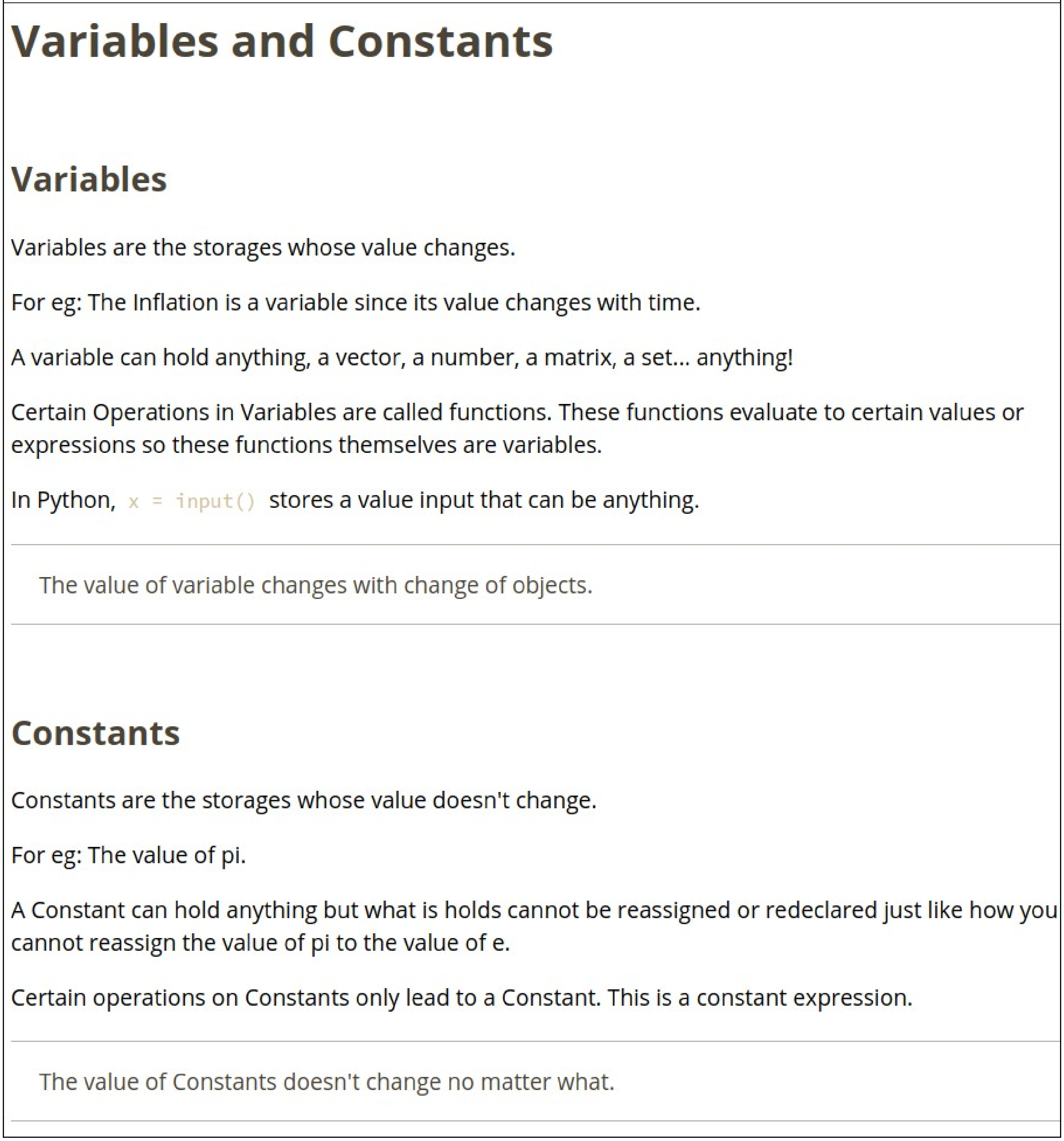


Linear-Algebra

Linear algebra is a branch of mathematics that deals with vector spaces and linear mappings between these spaces.

Linear algebra is foundational in many areas of mathematics and has numerous applications in diverse fields, making it a crucial subject for students and practitioners alike. Understanding its concepts and techniques enables deeper insight into mathematical modeling, data analysis, and problem-solving strategies.

> Linear Algebra deals with vectors, system of linear equations and their matrix representations and graphing.



Variables and Constants

Variables

Variables are the storages whose value changes.

For eg: The Inflation is a variable since its value changes with time.

A variable can hold anything, a vector, a number, a matrix, a set... anything!

Certain Operations in Variables are called functions. These functions evaluate to certain values or expressions so these functions themselves are variables.

In Python, `x = input()` stores a value input that can be anything.

> The value of variable changes with change of objects.

Constants

Constants are the storages whose value doesn't change.

For eg: The value of pi.

A Constant can hold anything but what is holds cannot be reassigned or redeclared just like how you cannot reassign the value of pi to the value of e.

Certain operations on Constants only lead to a Constant. This is a constant expression.

> The value of Constants doesn't change no matter what.

Acknowledgements

I extend my sincere gratitude and appreciation to all individuals and resources that contributed to the completion of this document.

I am grateful to the authors and researchers whose works have served as sources of inspiration and knowledge for this document. Their contributions to the field of Machine Learning have been instrumental in shaping my understanding and perspective on the subject matter.

I would also like to acknowledge the support and understanding of my family, friends, and colleagues, whose encouragement and patience have been a constant source of motivation and inspiration throughout this endeavor.

Finally, I extend my appreciation to all individuals, organizations, and institutions whose contributions, directly or indirectly, have played a significant role in the completion of this document.

Thank you.

Bhashkar Paudyal

Declaration

I, Bhashkar Paudyal hereby declare the following:

- The content material provided on this file, consisting of however now no longer confined to text, diagrams, and references, is unique and has been created totally for the motive of instructional or informational dissemination.

- Any outside reassets, consisting of books, articles, websites, and publications, applied withinside the advent of this file had been as it should be referred to and referenced withinside the "References and Citations" section.

- I take complete duty for the accuracy, completeness, and integrity of the statistics furnished herein. Any mistakes or omissions are accidental and do now no longer undermine the general validity of the file.

- This file does now no longer infringe upon any highbrow assets rights, consisting of copyright, trademarks, or patents. All images, illustrations, and text used from outside reassets had been acquired with right authorization or fall below truthful use guidelines.

- The evaluations, insights, and interpretations expressed on this file are primarily based totally on non-public knowledge, research, and information of the issue matter. They do now no longer always replicate the perspectives or evaluations of any organization, institution, or person related to the author.

- I verify my dedication to uphold instructional integrity and scholarly concepts in all factors of my work.

Bhashakar Paudyal

2024, 02 Feb, Friday.