Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Machine Learning

Linear Regression

Edwin Puertas, Ph.D(c).
epuerta@utb.edu.co

Introduction

- The Pearson correlation measures the degree to which a set of data points form a straight-line relationship.
- **Regression** is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.
- Any straight line can be represented by an equation of the form $Y = bX + a$, where ***b*** and ***a*** are constants.
- The value of ***b*** is called the slope constant and determines the direction and degree to which the line is tilted.
- The value of a is called the Y-intercept and determines the point where the line crosses the Y-axis.

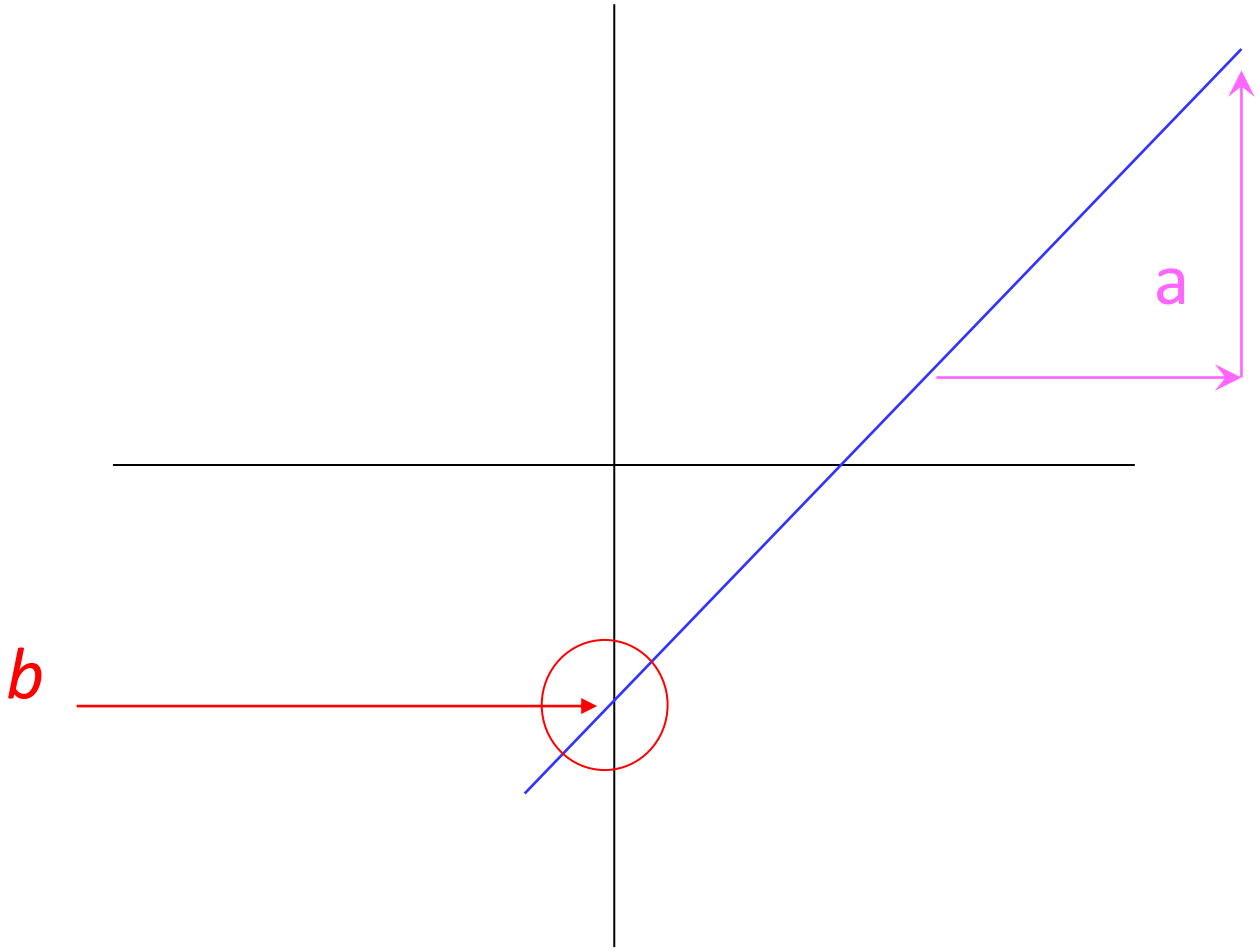
Introduction

- *X* = independent (explanatory) variable
- *Y* = dependent (response) variable
- Use instead of correlation
  - when distribution of *X* is fixed by researcher (i.e., set number at each level of *X*)
  - studying functional dependency between *X* and *Y*

What is “Linear”?

Remember this:

Y=aX+b?



What's Slope?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

Prediction

If you know something about X, this knowledge helps you predict something about Y. (Sound familiar?…sound like conditional probabilities?)

Regression equation…

Expected value of y at a given level of *x*=

$$E(y_i / x_i) = \alpha + \beta x_i$$

Predicted value for an individual…

$$y_i = \underbrace{\alpha + \beta * x_i}_{\text{Fixed – exactly on the line}} + \boxed{\text{random error}_i}$$

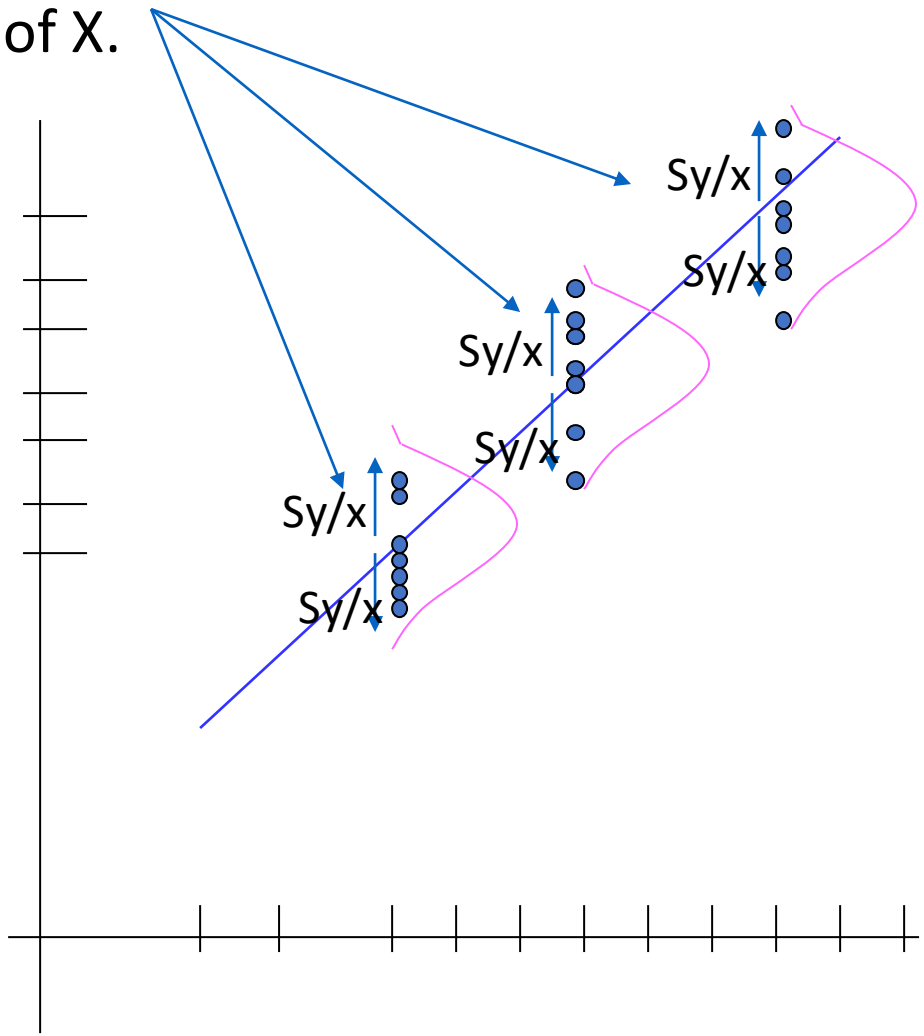
Follows a normal distribution

Assumptions

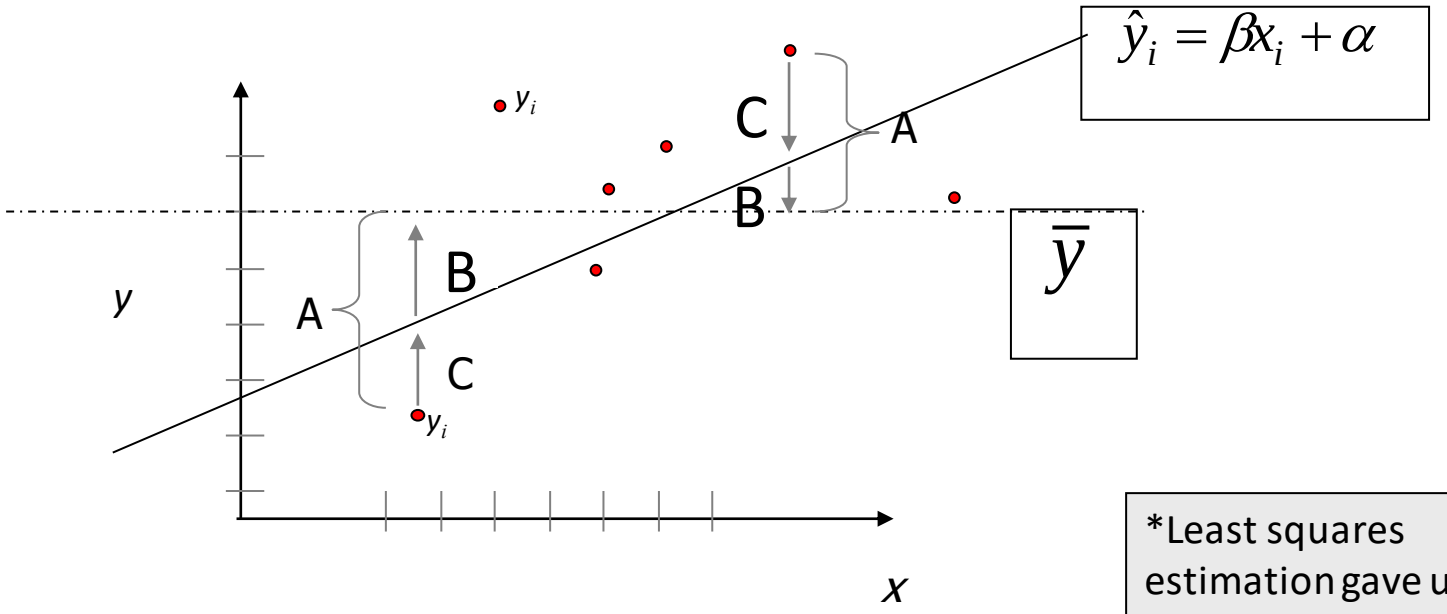
Linear regression assumes that…

1. The relationship between X and Y is linear
2. Y is distributed normally at each value of X
3. The variance of Y at every value of X is the same (homogeneity of variances)
4. The observations are independent

The standard error of Y given X is the average variability around the regression line at any given value of X. It is assumed to be equal at all values of X.



Regression Picture



*Least squares estimation gave us the line (β) that minimized C^2

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

A^2 B^2 C^2

R^2=SSreg/SStotal

SS_{total}
Total squared distance of observations from naïve mean of y
Total variation

SS_{reg}
Distance from regression line to naïve mean of y
Variability due to x (regression)

$SS_{residual}$
Variance around the regression line
Additional variability not explained by x—what least squares method aims to minimize