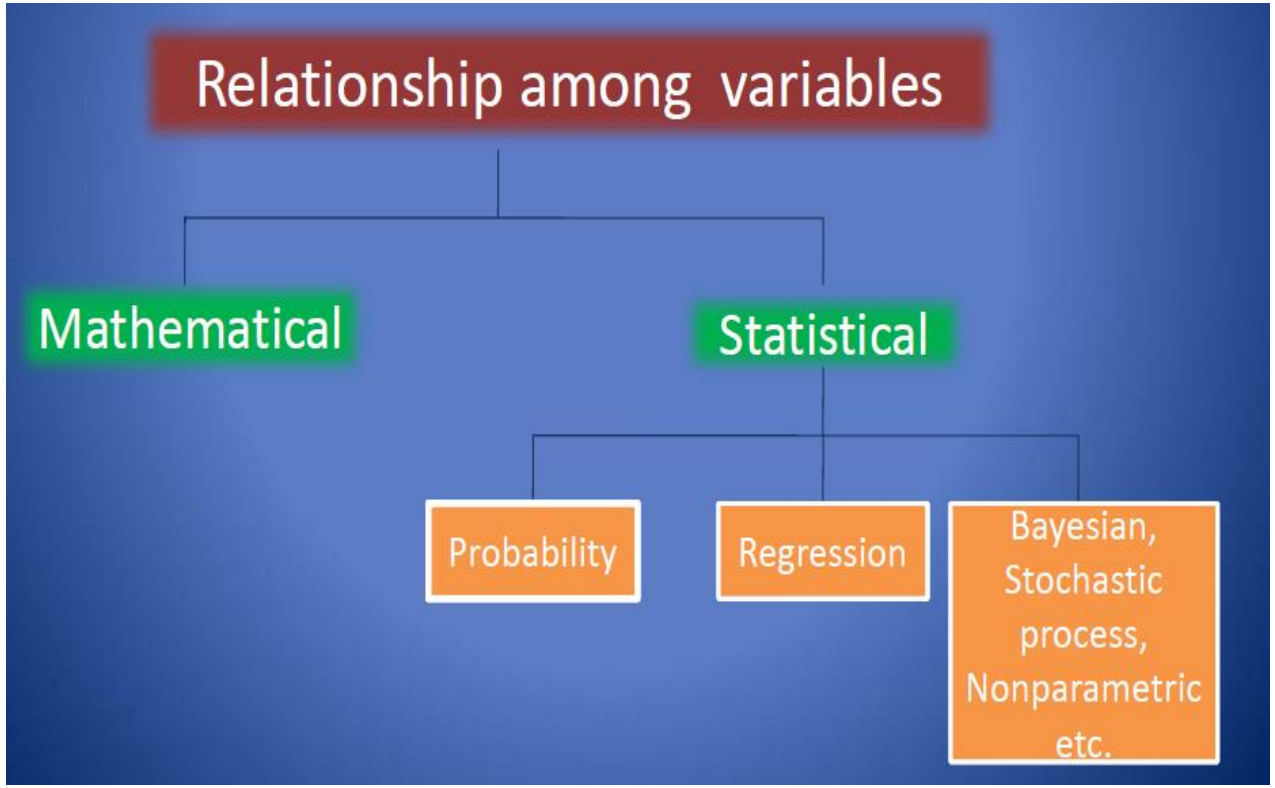
Regression Analysis

Introduction

- Statistics is a science of data.
- Data is numerical value.
- Data contains many information inside it.
- Using the data, a popular objective in applied sciences is to find out the relationship among different variables.
- Important objective: Modelling

What is a Model?

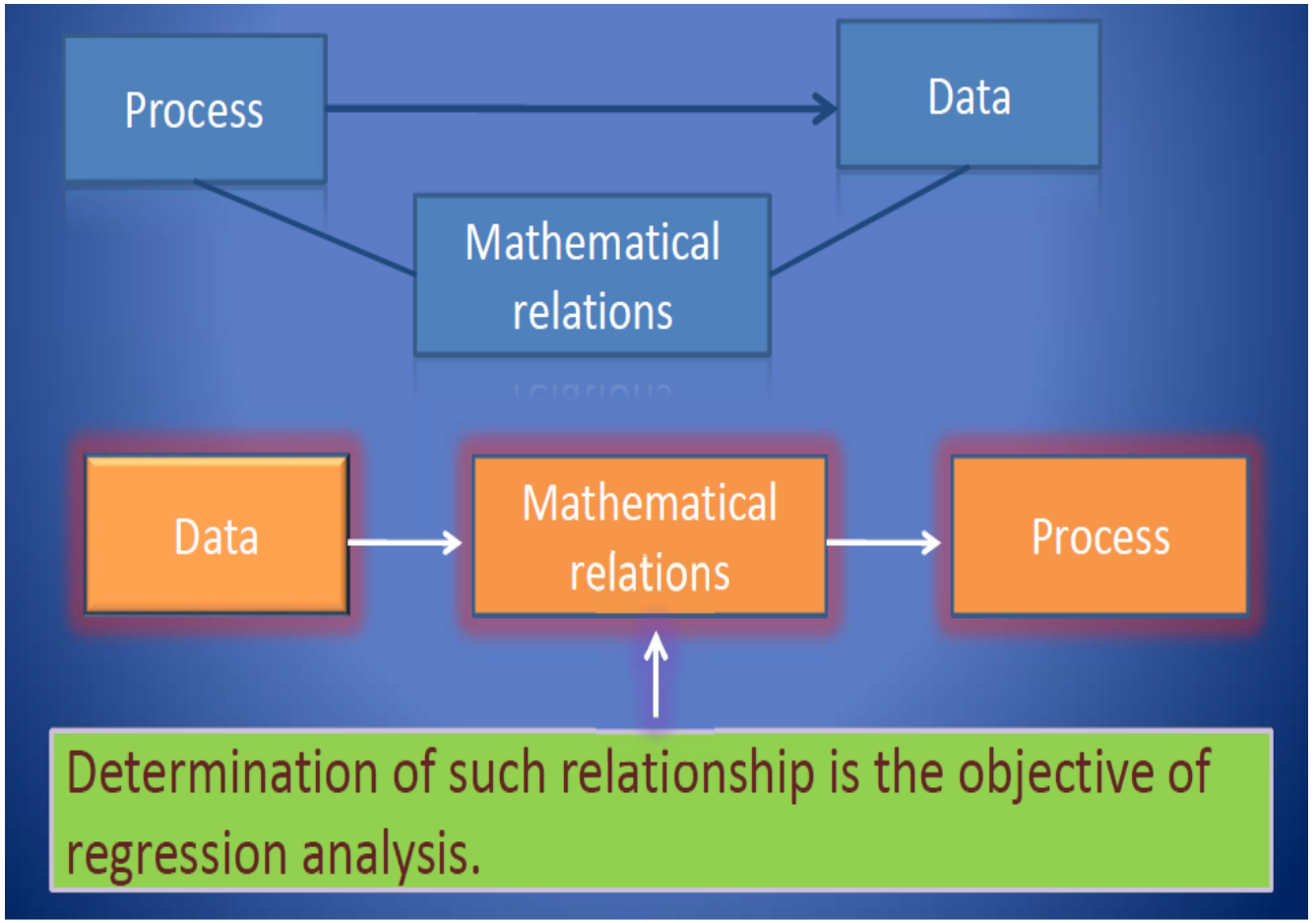
- Representation of a relationship
- All salient features of the process are incorporated



Process generates data → Correct

or

Data generates process → Wrong

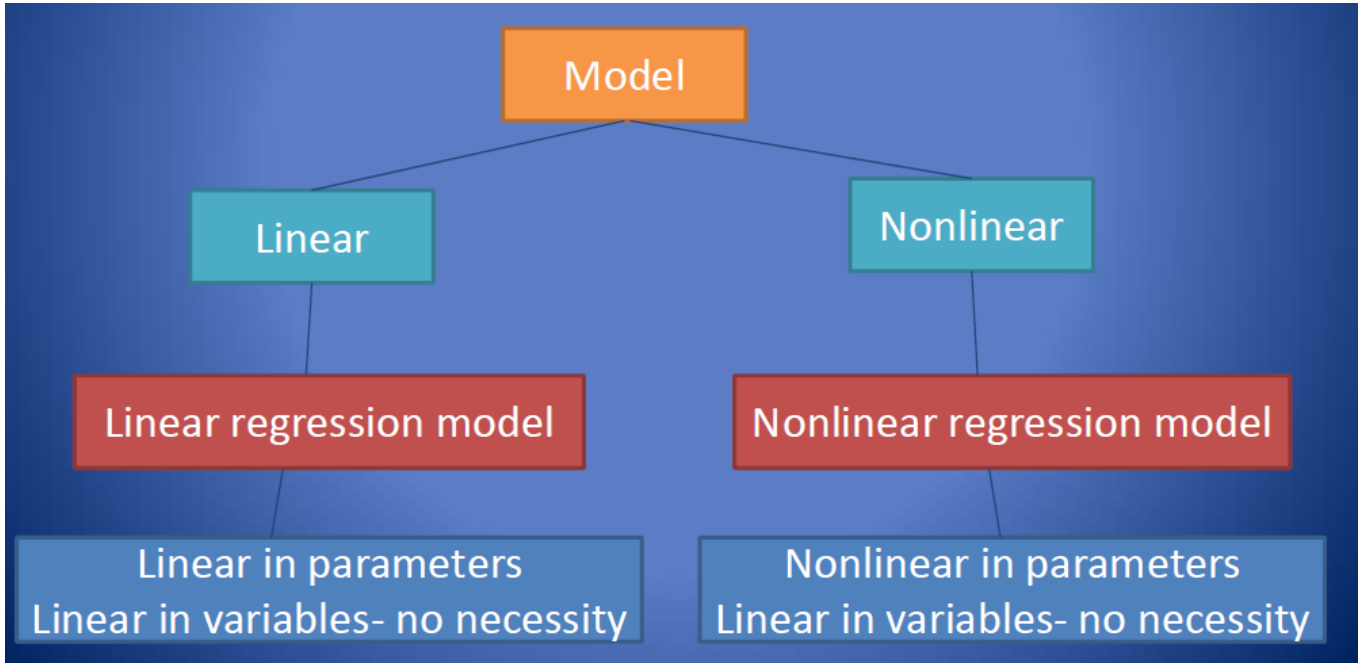


Definition

Regress: To move in backward direction

: Use data and get the statistical relations.

Model -> 2 components – variables & parameters



Difference between mathematical and statistical modelling?

Y= 3*x* + 2 Equation, regression model

y: Yield per hectare, x: Quantity of fertilizer (in Kg.)

x	1	2	5	10	50	100	1000
y	5	8	17	32	152	302	3002

y↑ x

Interpretation:

Increase quantity of fertilizer and obtain higher yield – Mathematical opinion

Increase quantity of fertilizer and obtain higher upto certain level, after that the crop will be destroyed – Statistical opinion

Regression Model

Regression Model ---- 2 types of variables

- Input variables or independent variables
- Output variables or dependent variables.

Objective: To determine a relationship between dependent and independent variables which describes the phenomenon/process in the best possible way.

Statistician's role:

- No right to change or alter the process.
- works only on the basis of a small sample.

Linearity in parameters

- Dependent variable – *y*, Independent variable – *x*

$$\frac{\partial(\text{dependent variable})}{\partial(\text{all parameters, seperately})} = \text{independent of parameters}$$

$$y = \exp(\beta x), \quad \frac{\partial y}{\partial \beta} = f(\beta) : \text{nonlinear}$$

$$y = \alpha + \beta x, \quad \frac{\partial y}{\partial \alpha} = 0, \quad \frac{\partial y}{\partial \beta} = x : \text{linear}$$

$$y = \alpha + \beta x^2, \quad \frac{\partial y}{\partial \beta} = x^2 : \text{linear}$$

$$y = \alpha^2 + \beta x, \quad \frac{\partial y}{\partial \alpha} = f(\alpha) : \text{nonlinear}$$

Regression Model

- It is easier to use mathematical tools on linear functions than on nonlinear functions.
- So linear regression model is more preferred over non-linear regression model.
- Linear function: dependent variable – *y*, independent variable – *x*

$$y=\alpha+\beta x$$

- *A- intercept term* β: slope or rate of change in *y* wrt *x*
- Relationship between *y* and *x* is linear

Model

- In practice, exact relationship may not hold due to several reasons.
- Introduce a random error component (u)

$$y=\alpha+\beta x+u$$

- u reflects the effect of
  - -Total effect of all variables/ relevant factors left
  - - randomness in human behaviour/ responses
  - - qualitative variables etc

Non-Linear Model

$$y = \alpha \exp(\beta x)$$

$$\log y = \log \alpha + \beta x$$

$$y^{*} = \alpha^{*} + \beta x$$

$$y = \alpha x^{\beta}$$

$$\log y = \log \alpha + \beta \log x$$

$$y^{**} = \alpha^{**} + \beta x^{**}$$

Assumptions about linear regression model

1. Relationship between y and X is linear.

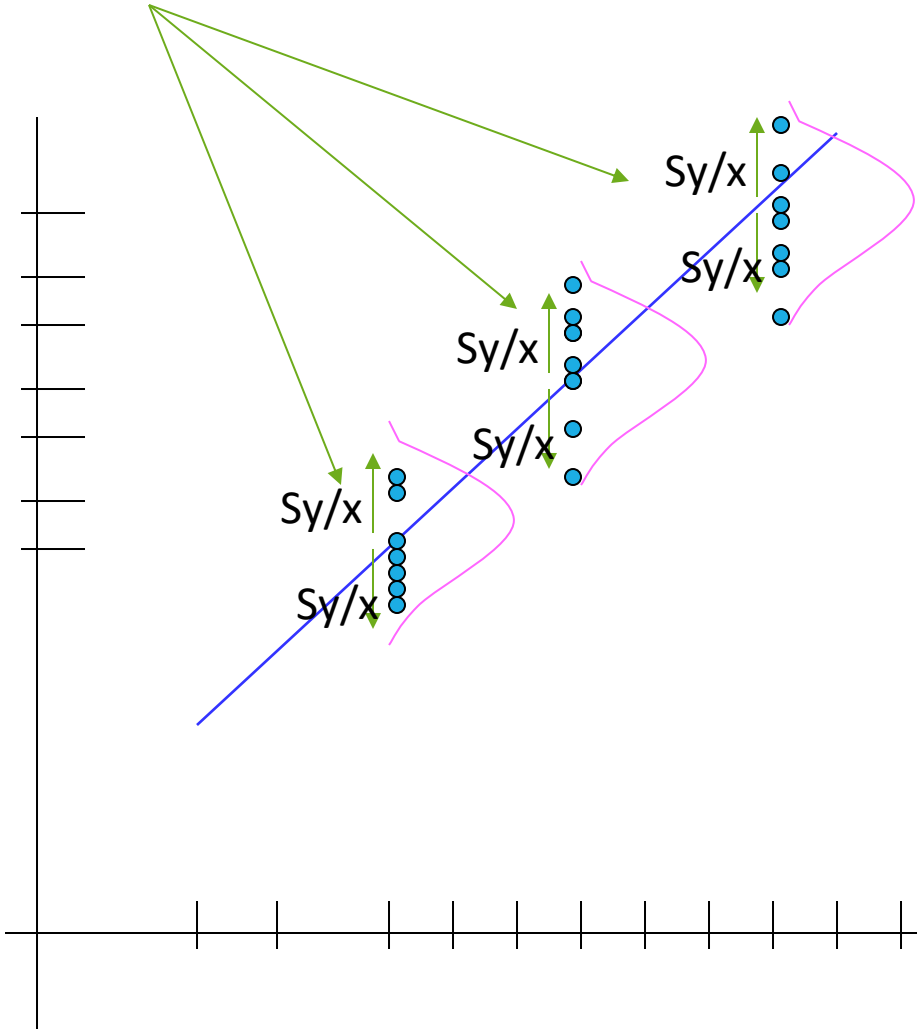
2. $E(u) = 0$ (mean of u is 0)

3. Variance of u is α^2 and $u1, u2, \ldots, un$ are independent of each other.

4. $u1, u2, \ldots, un$ are identically and independently distributed following $N(0, \alpha^2)$

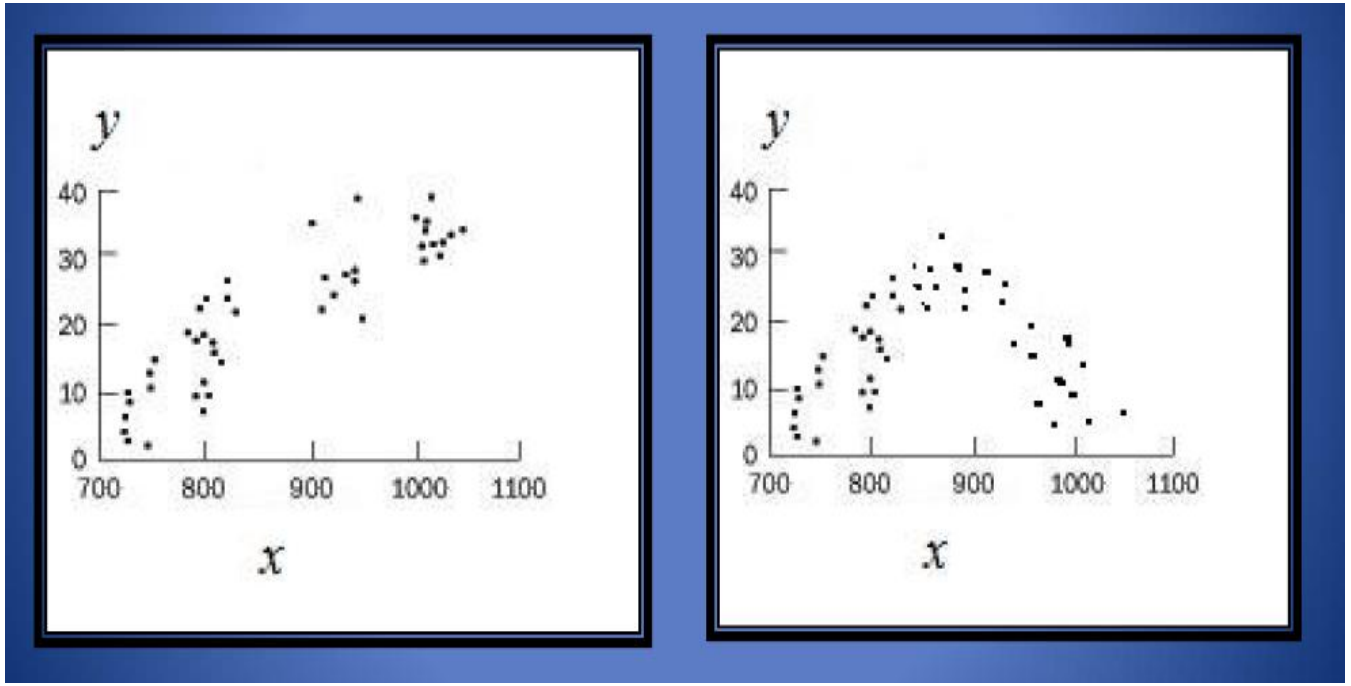
5. X is a nonstochastic variable

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.>



How to verify if there exist a linear relationship?

- Use scatter diagram



Linear trend *Nonlinear trend*

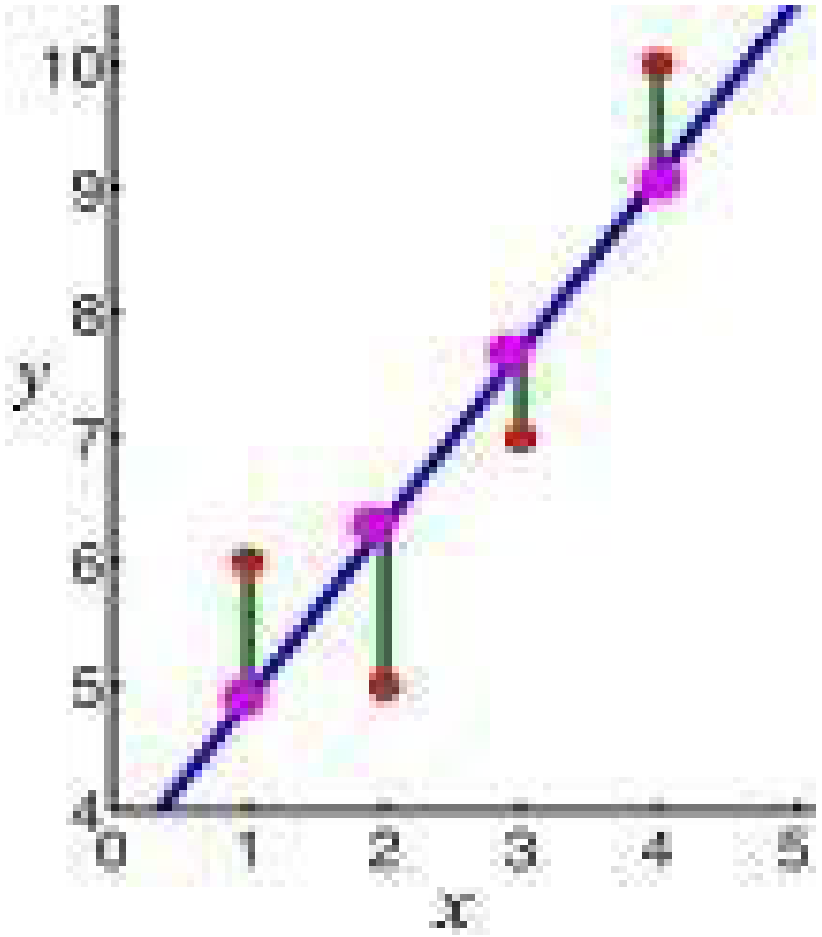
How to find parameters?

- Data on *n* values of study and explanatory variable is available
- (x_1,y_1), (x_2,y_2), (x_3,y_3)….. (x_n,y_n)

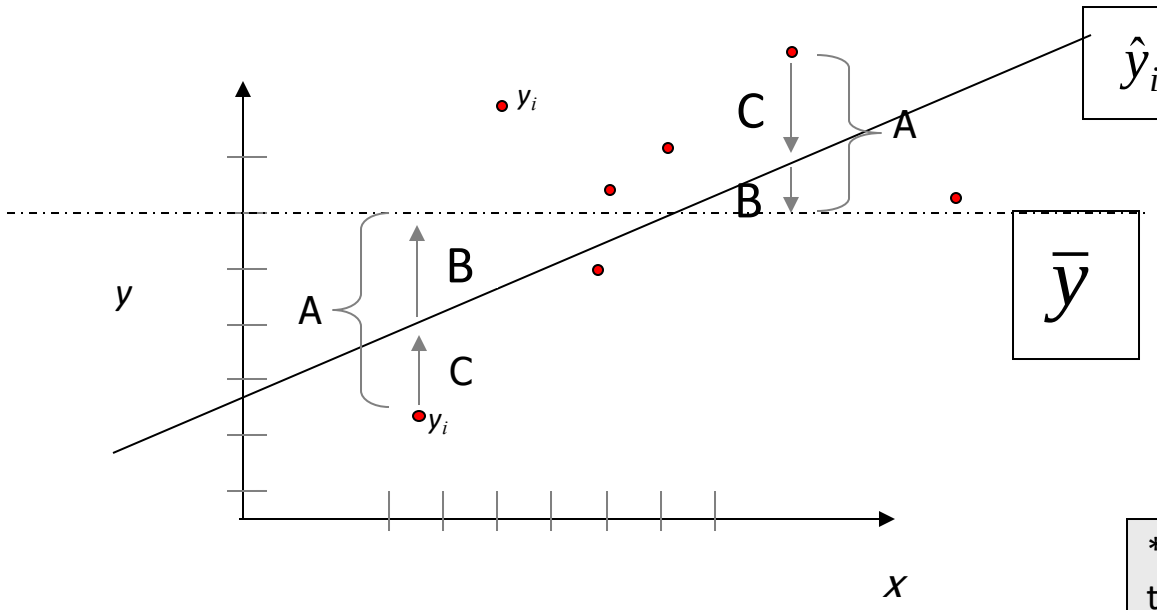
- Use principle of least squares (or maximum likelihood estimation)
- Other methods are also available.

Direct Regression

- Minimize the squared distance between the observed and fitted values



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

$R^2 = SS_{reg}/SS_{total}$

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

Lines of Regression

- Line of Regression of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

$b_{xy} = r\frac{\sigma_x}{\sigma_y}$ Regression Coefficient

- Line of Regression of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$ Regression Coefficient

Proof

$\sum y = na + b\sum x$ ----------1

$\sum xy = a\sum x + b\sum x^2 ---- -2$

Dividing equation 1 by n

$\frac{\sum y}{n} = \frac{n}{n}a + \frac{b}{n}\sum x$

$\frac{\sum y}{n} = a + \frac{b}{n}\sum x$

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

$y = a + bx$

y= $(\bar{y} - b\overline{x)} + b$x

$y - \bar{y} = b(x - \overline{x)}$ ----------3

$\sum y = na + b\sum x$ $-- -1$

Multiplying 1 by $\sum x$

$\sum x\sum y = na\sum x + b\sum x\sum x$

$\sum x\sum y = na\sum x + b\sum x^2$ $---- -4$

Multiplying equation 2 by n

$n\sum xy = na\sum x + nb\sum x^2$ ---------5

Subtracting 4 from 5

$n\sum xy - \sum x\sum y = nb\sum x^2 - b\sum x^2$

$n\sum xy - \sum x\sum y = b[n\sum x^2 - \sum x^2]$

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

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

$$b_{yx} = \frac{\text{Cov}\,(x, y)}{\sigma_x^2} = \frac{r\,\sigma_y}{\sigma_x}$$

$$b_{xy} = \frac{\text{Cov}\,(x, y)}{\sigma_y^2} = \frac{r\,\sigma_x}{\sigma_y}$$

$$y - \bar{y} = b_{yx}\,(x - \bar{x}),$$

$$x - \bar{x} = b_{xy}\,(y - \bar{y})$$

$$b_{yx} = \frac{\text{Cov}\,(x, y)}{\sigma_x^2} = \frac{\sum (x - \bar{x})\,(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b_{xy} = \frac{\text{Cov}\,(x, y)}{\sigma_y^2} = \frac{\sum (x - \bar{x})\,(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2}$$

Properties of Regression Coefficient

1. The coefficient of correlation is the geometric mean of the regression coefficients $r=\sqrt{b_{xy} * b_{yx}}$
2. If one regression coefficient is greater than 1, then the other will be less than 1.
3. AM of both regression coefficients is greater than or equal to the coefficient of correlation
4. They are not independent of the change of scale. There will be change in the regression coefficient if x and y are multiplied by any constant.
5. If b_{xy} is positive, then b_{yx} is also positive and vice versa.

Properties of Linear Regressions

- Two regression lines x on y and y o x always intersect at their means($\bar{x}, \bar{y}$)
- r,b_{xy} , b_{yx} all have same sign
- If r=0 then regression coefficient are zero
- The regression lines become identical if r=±1

Normal Equations

- The system of equations required to be solved for obtaining the values of constants known as Normal equations

y=a+bx

$$\sum y = Na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Example

x	y	x^2	xy
0	2	0	0
1	1	1	1
2	3	4	6
3	2	9	6
4	4	16	16
5	3	25	15
6	5	36	30
Sum 21	20	91	74

$$\sum y = Na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

20=7a+21b

74=21a+91b

a=1.357 b=0.5

y=1.357+0.5x

Exercise

x	y
0	1
1	1.8
2	3.3
3	4.5
4	6.3

Y=-4.29x+11.26

Normal Equations

- The system of equations required to be solved for obtaining the values of constants known as Normal equations

x=a+by

$$\sum x = Na + b\sum y$$

$$\sum xy = a\sum x + b\sum y^2$$

From the following data obtain two regression equations

x	6	2	10	4	8
y	9	11	5	8	7

	x	y	x^2	y^2	xy
	6	9	36	81	54
	2	11	4	121	22
	10	5	100	25	50
	4	8	16	64	32
	8	7	64	49	56
Sum	**30**	**40**	**220**	**340**	**214**
Mean	**6**	**8**	**44**	**68**	**42.8**

$$\sum y = Na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

40=5a+30b

214=30a+220b

a=11.9 b=-0.65

Y=11.9-0.65x

$$\sum x = Na + b\sum y$$

$$\sum xy = a\sum x + b\sum y^2$$

30=5a+40b

214=30a+340b

a=16.4 b=-1.3

y=16.4-1.3x

From the following data obtain regression equations taking deviations of items from the mean of x and y

x 6 2 10 4 8

y 9 11 5 8 7

	X	$X-\bar{X}$ x	x^2	Y	$Y-\bar{Y}$ y	y^2	xy
	6	0	0	9	1	1	0
	2	-4	16	11	3	9	-12
	10	4	16	5	-3	9	-12
	4	-2	4	8	0	0	0
	8	2	4	7	-1	1	-2
Sum	**30**	0	40	**40**	0	20	-26
Mean	**6**	0	8	**8**	0	4	-5.2

$$X-\bar{X}=r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

$$r\frac{\sigma_x}{\sigma_y}=\frac{\sum xy}{\sum y^2}=\frac{-26}{20}=-1.3$$

$X-6$ =-1.3 (Y −8)

X=-1.3Y+16.4

$$Y-\bar{Y}=r\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

$$r\frac{\sigma_y}{\sigma_x}=\frac{\sum xy}{\sum y^2}$$ =-5.2/40=-0.65

$Y-8$ =-0.65 (X−6)

Y=-0.65.3X+11.9

Example 9·1. *From the following data, obtain the two regression equations :*

Sales	:	91	97	108	121	67	124	51	73	111	57
Purchases	:	71	75	69	97	70	91	39	61	80	47

x	y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	$dxdy$
91	71	1	1	1	1	1
97	75	7	5	49	25	35
108	69	18	–1	324	1	–18
121	97	31	27	961	729	837
67	70	– 23	0	529	0	0
124	91	34	21	1156	441	714
51	39	–39	–31	1521	961	1209
73	61	–17	– 9	289	81	153
111	80	21	10	441	100	210
57	47	–33	–23	1089	529	759
$\sum x = 900$	$\sum y = 700$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 6360$	$\sum dy^2 = 2868$	$\sum dx\,dy = 3900$

$$\bar{x} = \frac{\sum x}{n} = \frac{900}{10} = 90 \text{ ;} \qquad \text{and} \qquad \bar{y} = \frac{\sum y}{n} = \frac{700}{10} = 70$$

$$b_{yx} = \frac{\sum (x - \bar{x})\,(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum dx\,dy}{\sum dx^2} = \frac{3900}{6360} = 0{\cdot}6132 \qquad b_{xy} = \frac{\sum (x - \bar{x})\,(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{\sum dx\,dy}{\sum dy^2} = \frac{3900}{2868} = 1{\cdot}361$$

$$y = 0{\cdot}6132x + 14{\cdot}812 \qquad x = 1{\cdot}361y - 5{\cdot}27$$

From the following data obtain regression equations taking deviations of X series from 5 and Y series from 7

x 6 2 10 4 8

y 9 11 5 8 7

	X	$X-5$ dx	dx^2
	6	1	1
	2	-3	9
	10	5	25
	4	-1	1
	8	3	9
Sum	**30**	5	45
Mean	**6**	1	9

	Y	$Y-7$ dy	dy^2	dxdy
	9	2	4	2
	11	4	16	-12
	5	-2	4	-10
	8	1	1	-1
	7	0	0	0
Sum	**40**	5	25	-21
Mean	**8**	1	5	-4.25

$$X-\bar{X}=r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})$$

$$r\frac{\sigma_x}{\sigma_y}=\frac{N\sum dxdy-(\sum dx)(\sum dy)}{N\sum dy^2-(\sum dy)^2}$$

$$=\frac{5(-21)-5*5}{5*25-25}=-1.3$$

$X-6$ =-1.3 (Y −8)

X=-1.3Y+16.4

$$Y-\bar{Y}=r\frac{\sigma_y}{\sigma_x}(X-\bar{X})$$

$$r\frac{\sigma_y}{\sigma_x}=\frac{N\sum dxdy-(\sum dx)(\sum dy)}{N\sum dx^2-(\sum dx)^2}$$

$$=\frac{5(-21)-5*5}{5*45-25}=-0.65$$

$Y-8$ =-0.65 (X−6)

Y=-0.65.3X+11.9

In a correlation study following values were obtained

	X	Y
Mean	65	67
Std. Dev	2.5	3.5
Coefficient of Correlation	0.8	

Find two regression equations that are associated with the values

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

X-65=0.8*(2.5/3.5)(Y-67)
X-65=0.57(Y-67)
X=0.57Y+26.81

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

Y-67=0.8*(3.5/2.5)(Y-65)
X-67=1.12(Y-65)
X=1.12Y-5.8

After 9/11 attack, a company could partially recover following record on correlation

Variance of X=9

Eqns. Of regression 8X-10Y+66=0

40X-18Y=214

Find on the basis of above information

1) Mean values of X and Y
2) Coefficient of Correlation
3) Standard Deviation of Y

Solving two equations we get mean of X and mean of Y

$\bar{X}$ =13 $\bar{Y}$ =17

8X-10Y=-66
X-(10/8)Y=-66/8
b_{xy}=10/8=1.25

8X-10Y=-66
10Y=8X+66
Y=(8/10)X+66/10
b_{yx}=8/10=0.8

40X-18Y=214
-Y=214/18-(40/18)X
b_{yx}=40/18=2.22

40X=18Y+214
Y=(18/40)X+214/18
b_{xy}=18/40=0.45

R=sqrt(b_{yx}*b_{xy})=sqrt(0.8*0.45)=0.6

$b_{xy} = r\frac{\sigma_x}{\sigma_y}$

0.45=0.6(3/σ_y)
σ_y=(0.6*3)/0.45
=4

For 50 students in a class, the regression equations for marks in statistics(X) and the marks in accountancy(Y) is 3Y-5X+180=0. The mean in accountancy is 44 and variance of marks in statistics(X) is (9/16)th of the marks in accountancy(Y) . Find the mean marks in statistics and coefficient of correlation between marks of two subjects.

3Y-5X=-180 X=(3Y+180)/5 X=(3*44+180)/5= 62.4

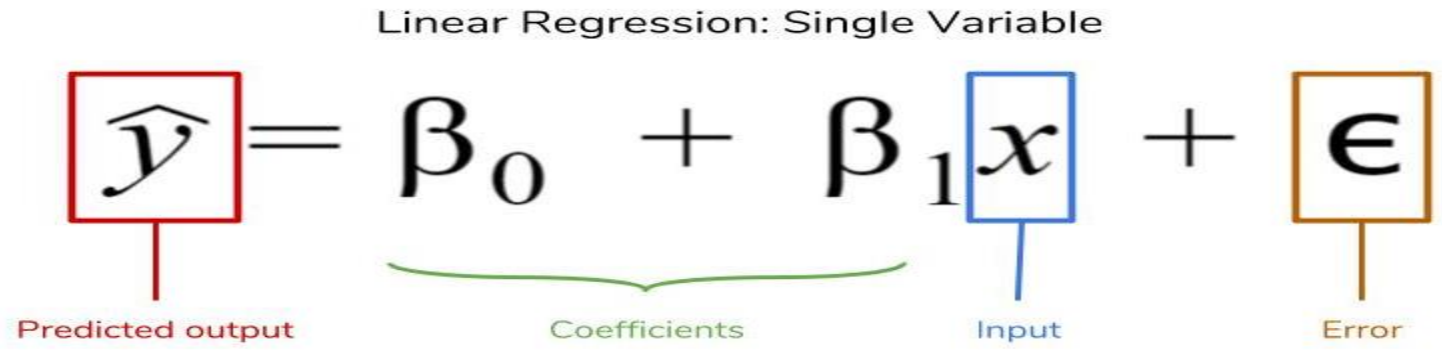
5X=3Y+180 X=(3/5)Y+(180)/5 bxy=(3/5)= 0.6

$b_{xy} = r\frac{\sigma_x}{\sigma_y}$ $0.6 = r\frac{\sqrt{9}}{\sqrt{16}}$ r =(0.6*4)/3
=0.8

Errors in regression

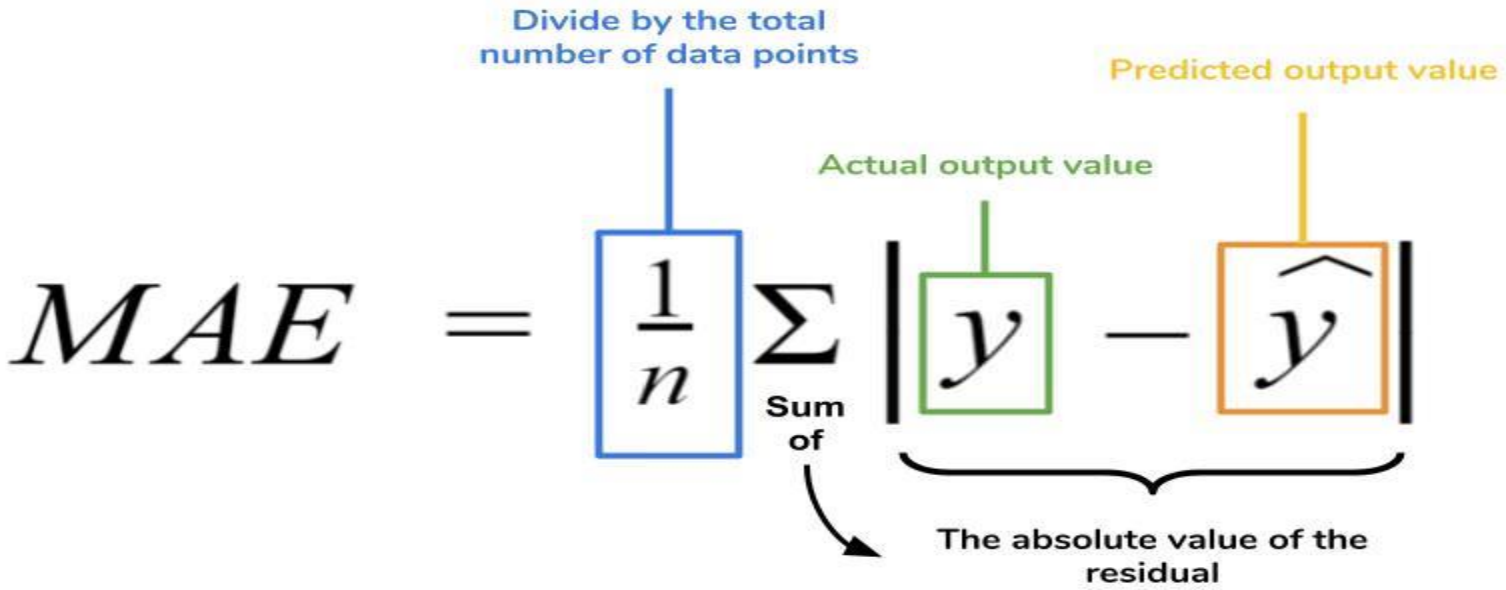
- **Residual-** the difference between the actual value and the model's estimate
- If our collection of residuals are small, it implies that the model does a good job at predicting our output of interest.
- Conversely, if these residuals are generally large, it implies that model is a poor estimator.
- Statisticians have developed summary measurements that take our collection of residuals and condense them into a *single* value that represents the predictive ability of our model.
  - Mean Absolute Error
  - Mean Square Error
  - Mean Absolute Percentage Error
  - Mean Percentage Error

epsilon ϵ

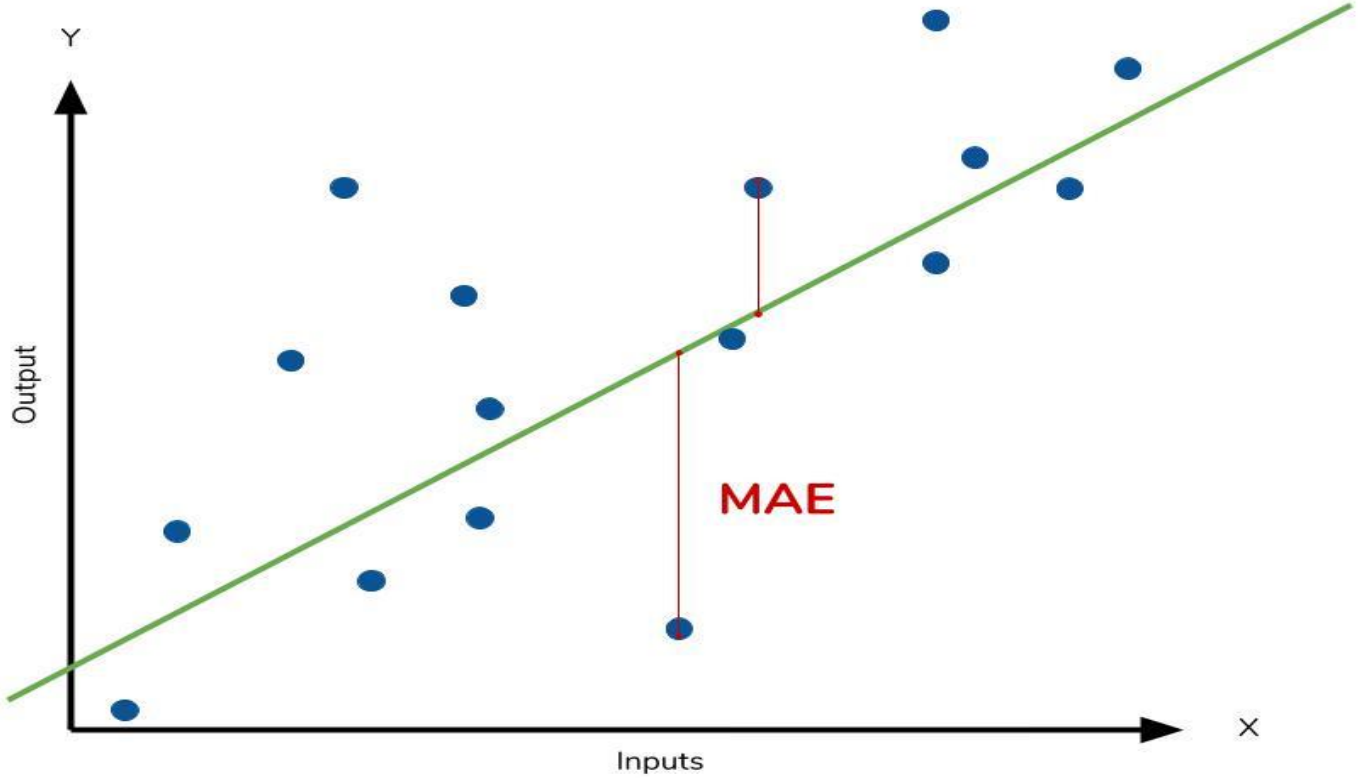


- ϵ represents error that comes from sources out of our control, causing the data to deviate slightly from their *true* position.
- Error metrics will be able to judge the differences between prediction and actual values, but we cannot know how much the error has contributed to the discrepancy.
- While we cannot ever completely eliminate epsilon, it is useful to retain a term for it in a linear model.
- Population parameter

Mean Absolute Error



Mean Absolute Error



Mean Absolute Error

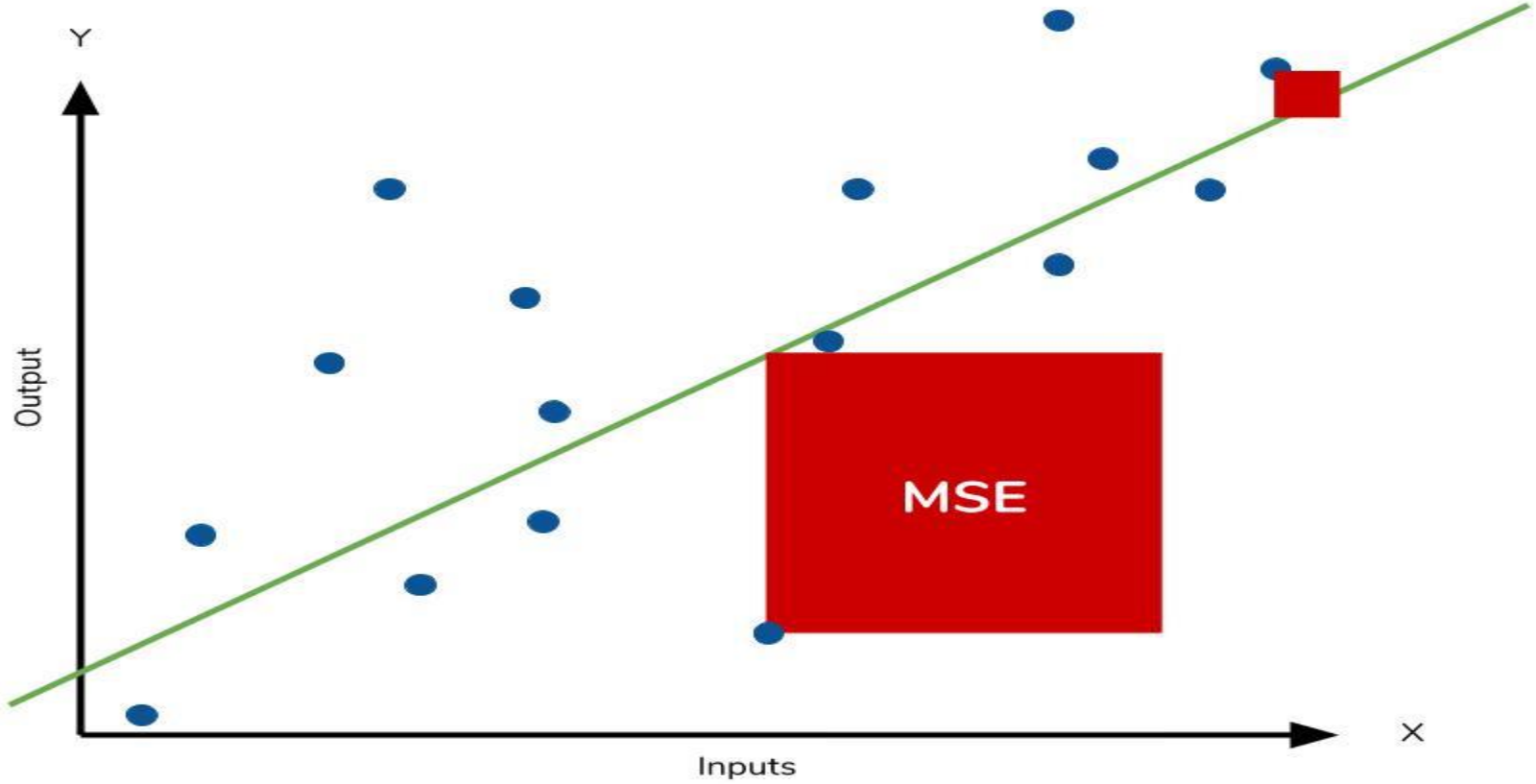
- MAE is easily interpretable-simplest
- Because we use the absolute value of the residual, the MAE does not indicate *underperformance* or *overperformance* of the model.
- Each residual contributes proportionally to the total amount of error, meaning that larger errors will contribute linearly to the overall error.
- A small MAE suggests the model is great at prediction, while a large MAE suggests that your model may have trouble in certain areas. A MAE of 0 means that your model is a **perfect** predictor of the outputs.

Mean Square Error

$$MSE = \frac{1}{n}\sum\left(\underbrace{y - \hat{y}}_{\text{The square of the difference between actual and predicted}}\right)^2$$

- Because we are squaring the difference, the MSE will almost always be bigger than the MAE.
- The effect of the square term in the MSE equation is most apparent with the presence of outliers in our data.
- While each residual in MAE contributes **proportionally** to the total error, the error grows **quadratically** in MSE.
- This means that outliers in our data will contribute to much higher total error in the MSE than they would the MAE.
- Similarly, our model will be penalized more for making predictions that differ greatly from the corresponding actual value.

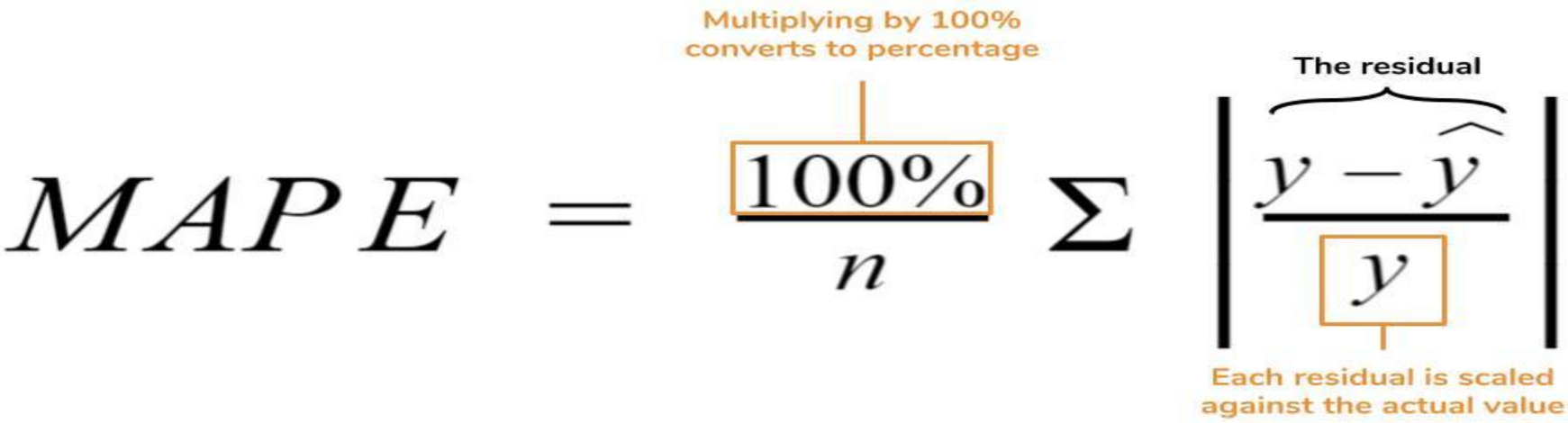
Mean Square Error



Root mean squared error (RMSE)

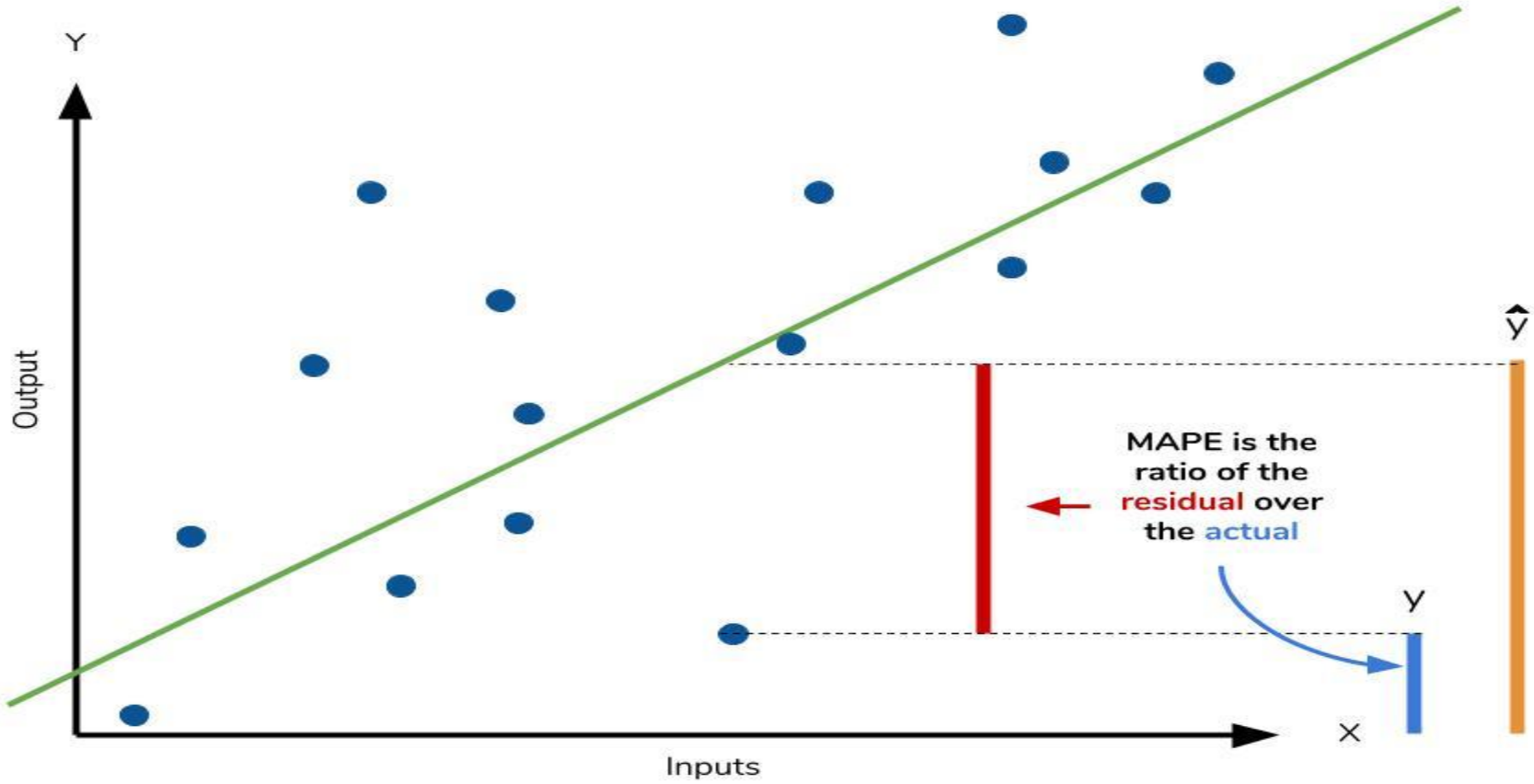
- Square root of the MSE.
- Because the MSE is squared, its units do not match that of the original output.
- Researchers will often use RMSE to convert the error metric back into similar units, making interpretation easier.
- Since the MSE and RMSE both square the residual, they are similarly affected by outliers.
- The RMSE is analogous to the standard deviation (MSE to variance) and is a measure of how large your residuals are spread out.
- Both MAE and MSE can range from 0 to positive infinity, so as both of these measures get higher, it becomes harder to interpret how well your model is performing.
- Another way we can summarize our collection of residuals is by using percentages so that each prediction is scaled against the value it's supposed to estimate.

Mean Absolute Percentage Error



- Easy to interpret in percentage
- Limited in use

Mean absolute percentage error



Mean Absolute Percentage Error

MAPE is undefined for data points where the value is 0.
MAPE can grow unexpectedly large if the actual values are exceptionally small themselves.
MAPE is biased towards predictions that are systematically less than the actual values themselves.
MAPE will be lower when the prediction is lower than the actual compared to a prediction that is higher by the same amount.

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

$\hat{y}$ is smaller than the actual value

n = 1 $\hat{y}$ = 10 y = 20

MAPE = 50%

$\hat{y}$ is greater than the actual value

n = 1 $\hat{y}$ = 20 y = 10

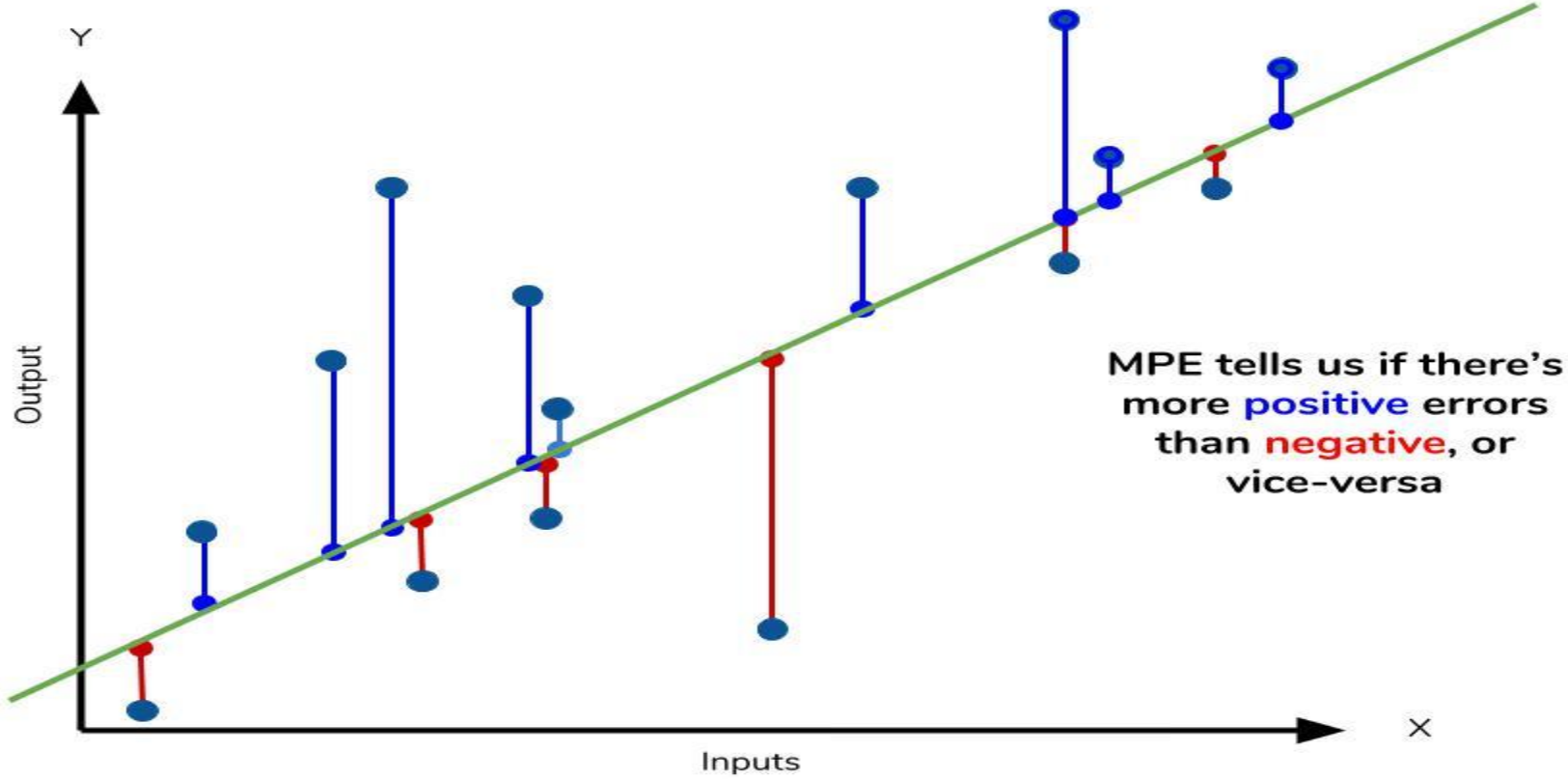
MAPE = 100%

Mean Percentage Error

$$MPE = \frac{100\%}{n} \sum \left(\frac{y - \hat{y}}{y} \right)$$

- Since positive and negative errors will cancel out, we cannot make any statements about how well the model predictions perform overall.
- However, if there are more negative or positive errors, this bias will show up in the MPE.
- Unlike MAE and MAPE, MPE is useful to us because it allows us to see if our model systematically **underestimates** (more negative error) or **overestimates** (positive error).

Mean Percentage Error



Summary

Acronym	Full Name	Residual Operation?	Robust To Outliers?
MAE	Mean Absolute Error	Absolute Value	Yes
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MAPE	Mean Absolute Percentage Error	Absolute Value	Yes
MPE	Mean Percentage Error	N/A	Yes