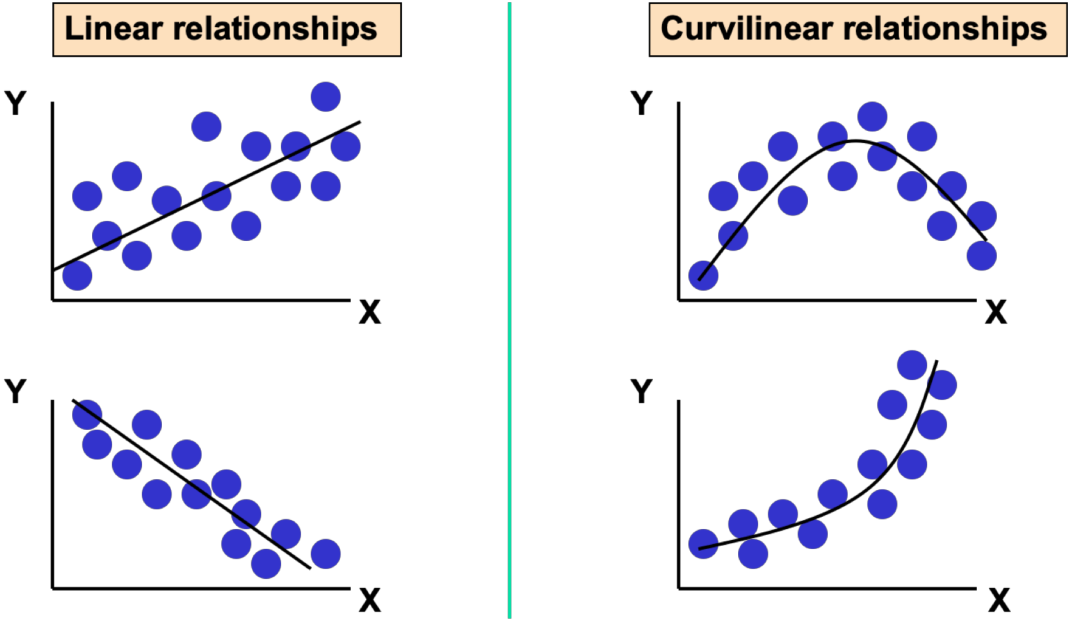
Linear Regression

Regression

Often in businesses and our lives we want to estimate the value of certain variables for e.g.

- A real estate agent might want to estimate the price of a house correctly based on other features of the house like no. of rooms, area of house etc
- A banker might want to estimate the credit worthiness of an individual based on his salary, cash flow, credit card, possessions etc.
- A retailer might want to estimate his sales based on promotional offers, footfall etc

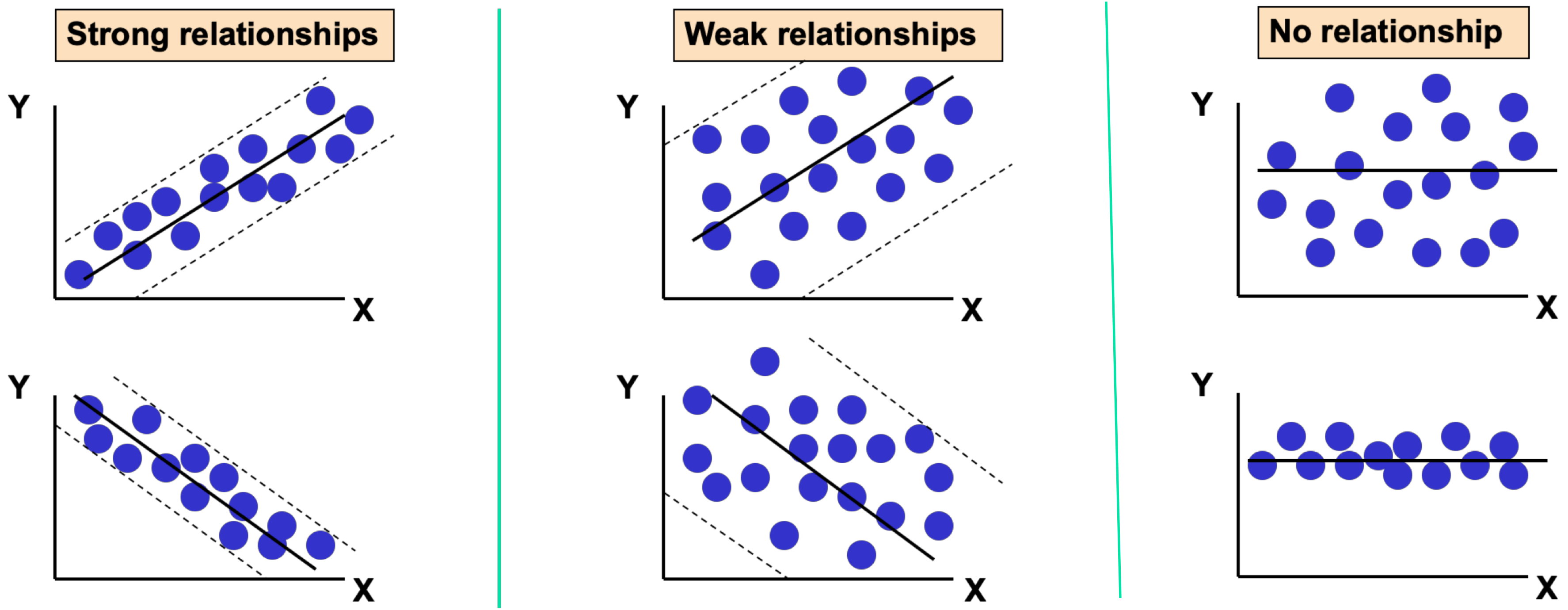
Types of Relationships



Linear relationships

Curvilinear relationships

Types of Relationships



Strong relationships

Weak relationships

No relationship

REGRESSION

WHY SHOULD WE USE THIS?

QUANTIFYING THE RELATIONSHIP BETWEEN TWO CONTINUOUS VARIABLES

PREDICT (OR FORECAST) THE VALUE OF ONE VARIABLE FROM KNOWLEDGE OF THE VALUE OF ANOTHER VARIABLE

REGRESSION

DETERMINISTIC/STOCHASTIC

COST OF DRIVING A SUV

MONTHTLY COST = EMI + FUEL_COST X DISTANCE
Y = 10000 + 7*D

NO WORK FOR A DATA SCIENTIST TO DO HERE—

THERE’S NOTHING RANDOM ABOUT THIS.

REGRESSION

SIMPLE LINEAR REGRESSION

In simple linear regression we generate an equation to calculate the value of a ***dependent variable*** (Y) from an ***independent variable*** (X)

What is a dependent variable?
What is an independent variable?

For e.g. Time taken to get to work (Y) is a function of the distance travelled (X)

Say you drive to work at an average of 60 km's/hour. It takes about 1 minute for every kilometre travelled…

Travel time = 1 minute×kilometres travelled

This is a ***mathematical model*** that represents the relationship between the two variables

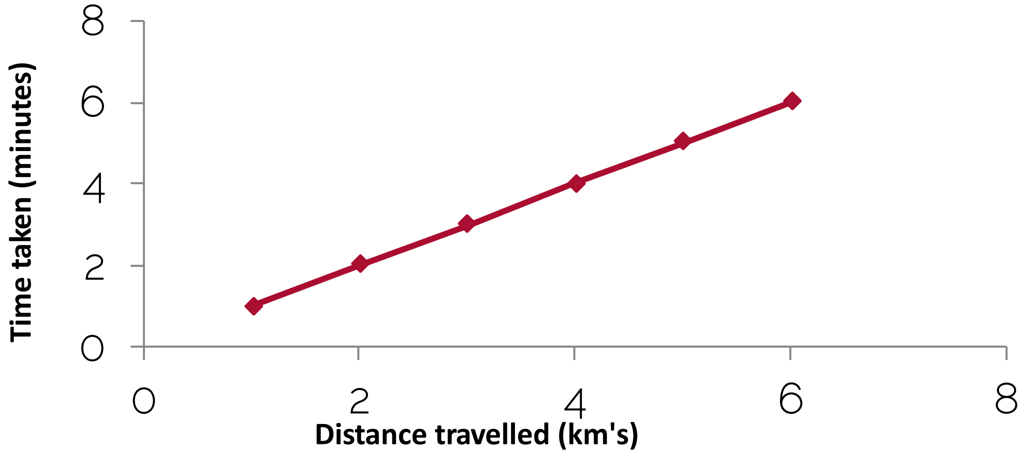
REGRESSION

THE FUNDAMENTAL EQUATION

Say you drive to work at an average of 60 km's/hour. It takes about 1 minute for every kilometre travelled…

Travel time = 1 minute×kilometres travelled

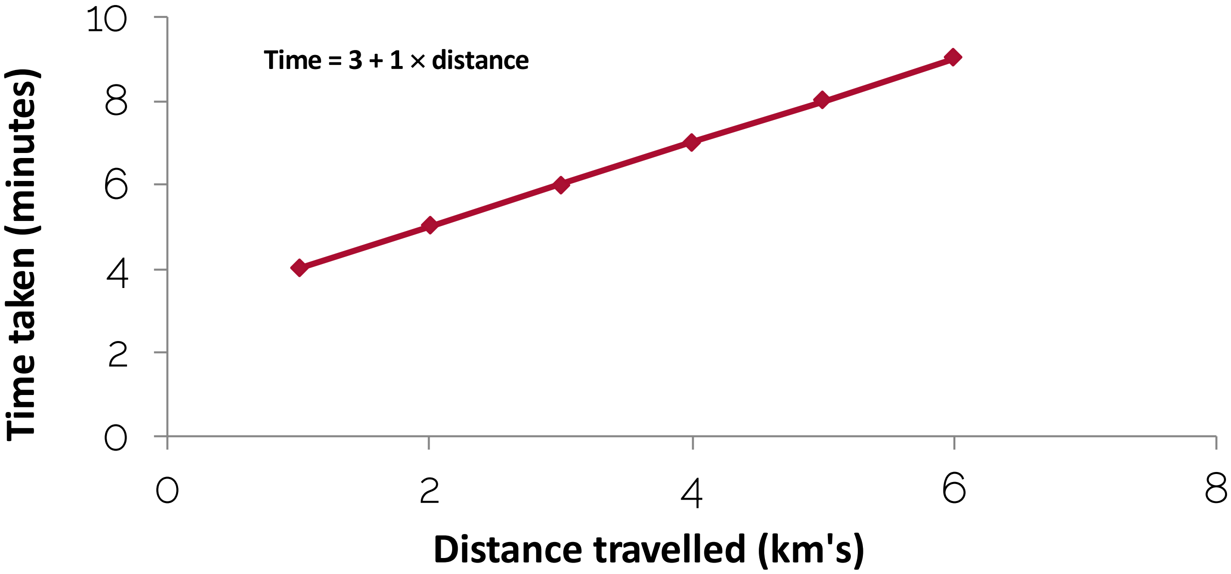
This is a ***mathematical model*** that represents the relationship between the two variables



REGRESSION

INTERCEPT

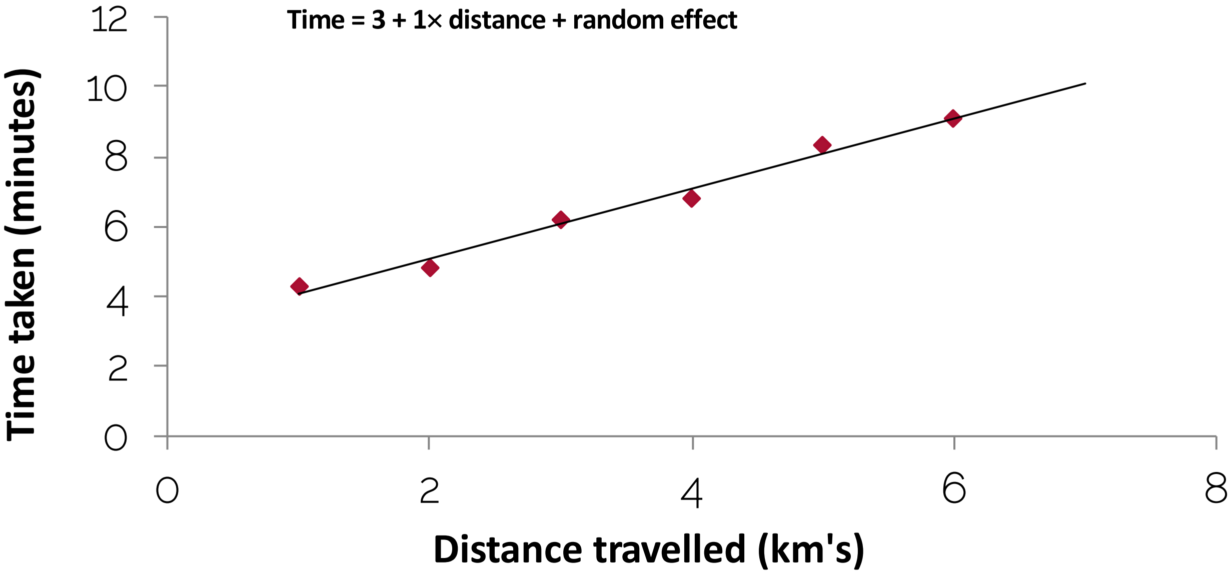
Actually, it won't be that simple, because there will some time taken to walk to your car and then walk from the car to work. Say this takes an extra 3 minutes per day



REGRESSION

ERROR

It also won't be that precise because there will be slight variations in time taken because of traffic, roads, etc.



REGRESSION

OVERALL EQUATION

In general, the regression equation takes the form;

$$y = \beta_o + \beta_1 x + \varepsilon$$

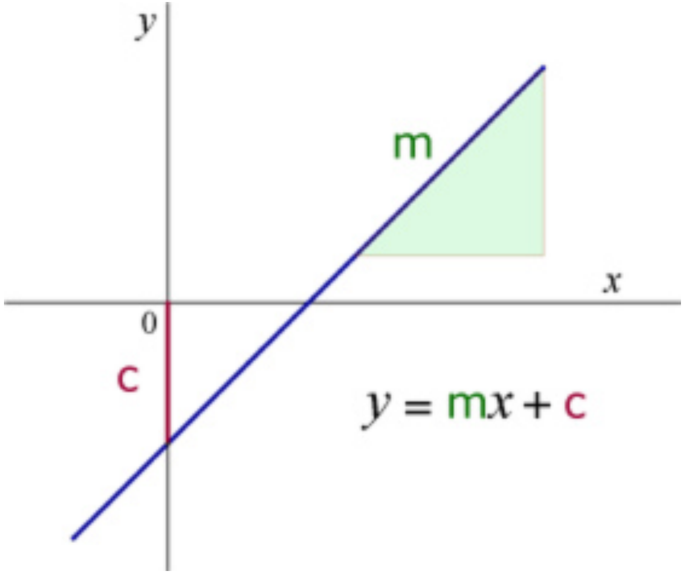
y = the dependent variable

x = the independent variable

β_o = The y-intercept

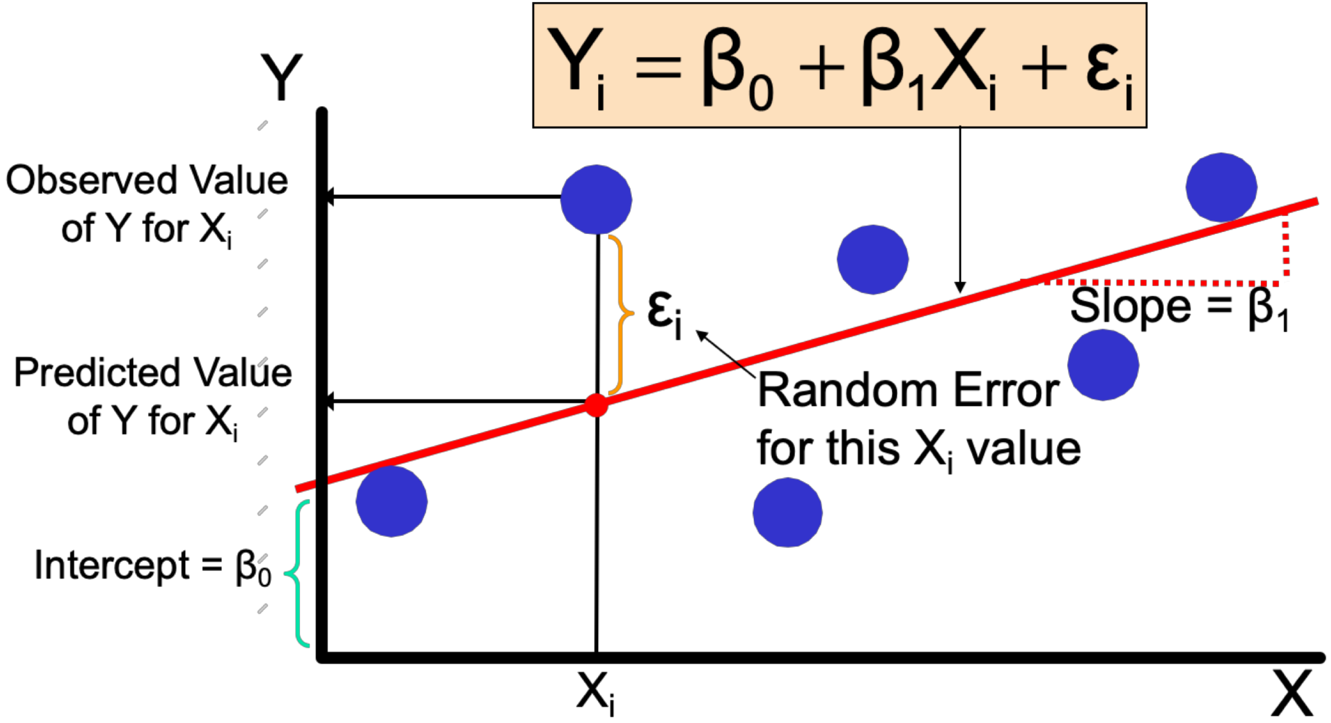
β_1 = The slope of the line

ε = random error term ~N(0, σ^2)



REGRESSION

OVERALL EQUATION



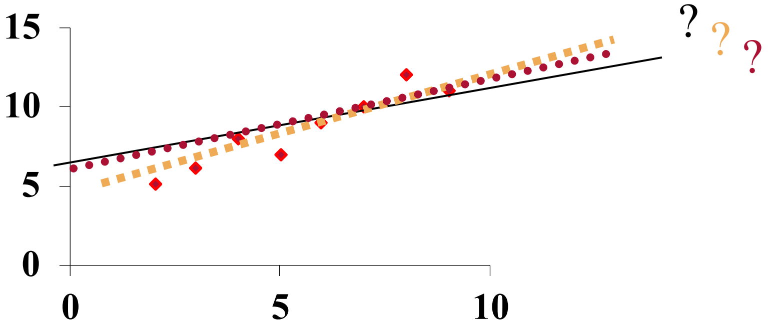
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

REGRESSION

LINE OF BEST FIT

Given a data set, we need to find a way of calculating the parameters of the equation. We have a set of data points and we need to estimate the best possible relation.

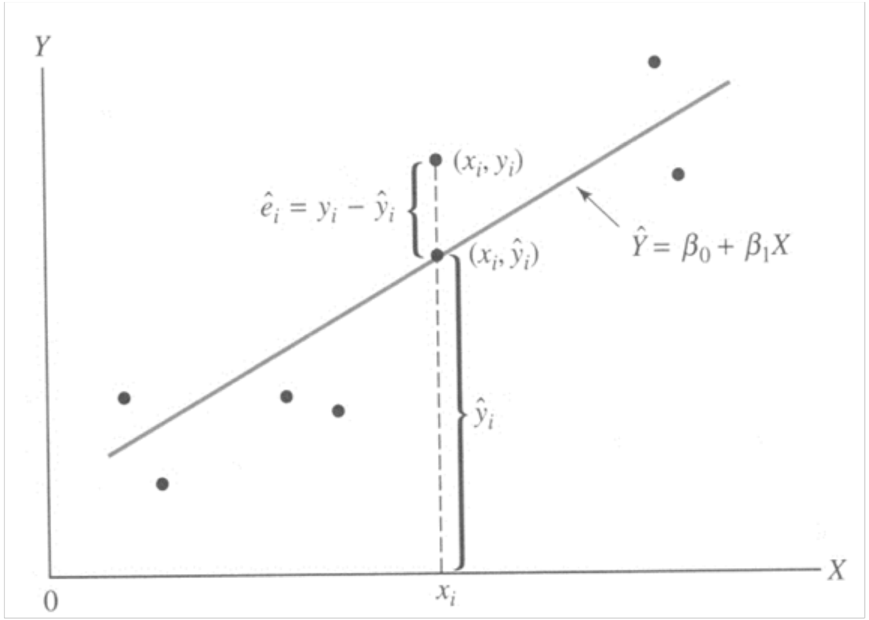


We need to fit a *line of best fit*

REGRESSION

LINE OF BEST FIT

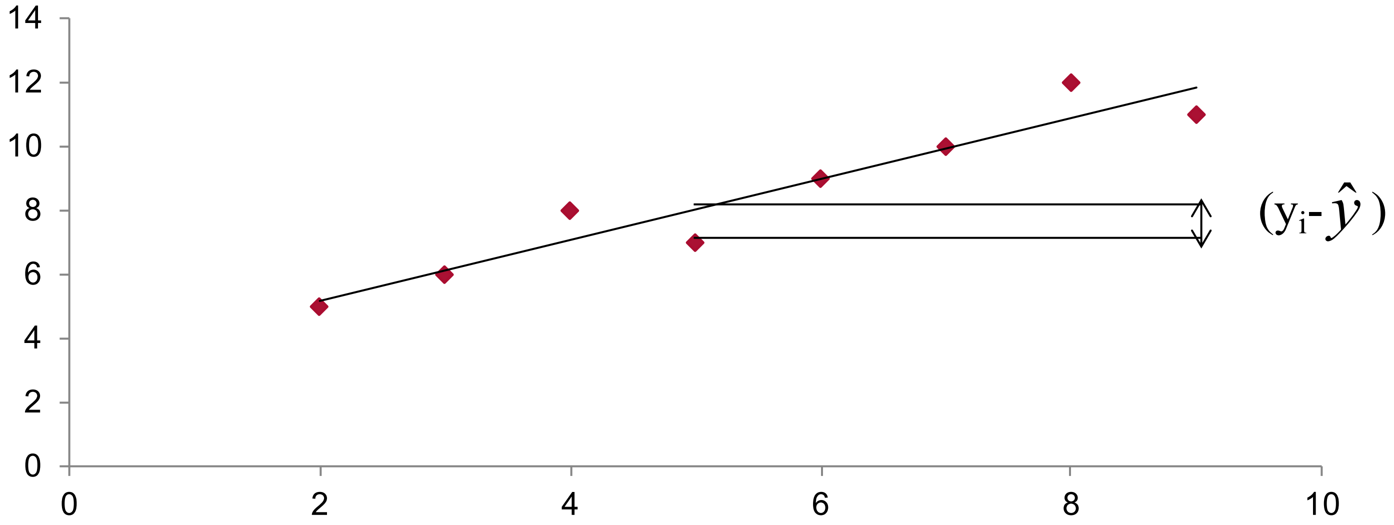
Error is the difference between predicted and actual values



REGRESSION

MINIMIZING THE ERROR

Because the line will seldom fit the data precisely, there is always some error associated with our line
The line of best fit is the line that minimises the spread of these errors



MEASURES OF VARIATION

- Total variation is made up of two parts:

$$SST \quad = \quad SSR \quad + \quad SSE$$

Total Sum of Squares

Regression Sum of Squares

Error Sum of Squares

$$SST = \sum (Y_i - \overline{Y})^2$$

$$SSR = \sum (\hat{Y}_i - \overline{Y})^2$$

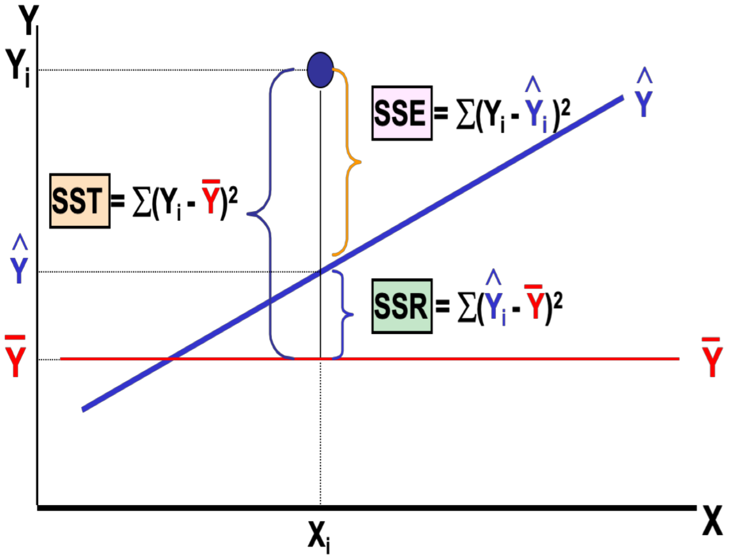
$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

where:

$\overline{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value



Y
Y_i
$\hat{Y}$
$\overline{Y}$
SST = $\Sigma(Y_i - \overline{Y})^2$
SSE = $\Sigma(Y_i - \hat{Y}_i)^2$
SSR = $\Sigma(\hat{Y}_i - \overline{Y})^2$
$\hat{Y}$
$\overline{Y}$
X_i
X

- SST (Total Variation) : Measures the variation of the Y_i values around their mean Y

- SSR (Explained Variation): Variation attributable to the relationship between X and Y

- SSE (Unexplained Variation): Variation in Y attributable to factors other than X

LET’S UNDERSTAND THIS BETTER THROUGH A CASE STUDY

Thanks!

Any *questions* ?

Next steps

- Attempt quiz –
- Share feedback –