



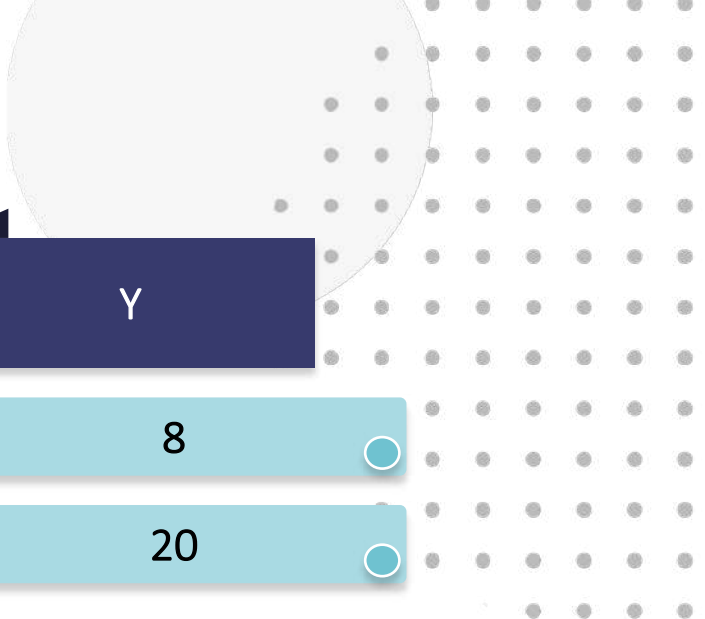
Mohamed Noordeen Alaudeen

Linear Regression



Relationship

$Y = \text{??????????}$



| X | Y |
|---|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |

Relationship

$$Y = 2 + 3(X)$$

| X | Y |
|---|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |

What is 2 here?

$$Y = 2 + 3(X)$$

| X | Y |
|---|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |

Find the Y in ?

$$Y = 2 + 3X$$

| X | Y |
|----|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |
| 10 | ? |
| 1 | ? |

Value for Y with given X

$Y = 2 + 3X$

| X | Y |
|----|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |
| 10 | 32 |
| 1 | 5 |

Terminology

$$Y = 2 + 3X$$

● Y = Model

| X | Y |
|----|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |
| 10 | 32 |
| 1 | 5 |

Terminology

$$Y = 2 + 3X$$

- Y = Model
- 2 = Intercept

| X | Y |
|----|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |
| 10 | 32 |
| 1 | 5 |

Terminology

$$Y = 2 + 3X$$

- Y = Model
- 2 = Intercept
- 3 = Slope

| X | Y |
|----|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |
| 10 | 32 |
| 1 | 5 |

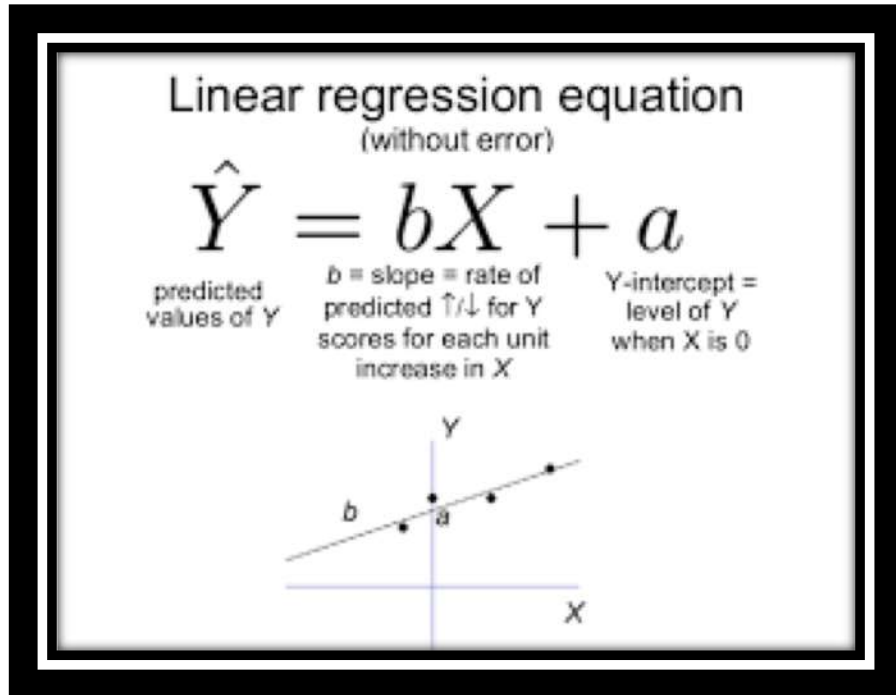
Terminology

$$Y = 2 + 3X$$

- Y = Model
- 2 = Intercept
- 3 = Slope
- X = Input

| X | Y |
|----|----|
| 2 | 8 |
| 6 | 20 |
| 4 | 14 |
| 3 | 11 |
| 7 | 23 |
| 4 | 14 |
| 2 | 8 |
| 5 | 17 |
| 10 | 32 |
| 1 | 5 |

Formula For a Line



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

Dependent Variable $\rightarrow Y_i$

Population Y intercept $\rightarrow \beta_0$

Population Slope Coefficient $\rightarrow \beta_1$

Independent Variable $\rightarrow X_i$

Random Error term $\rightarrow \epsilon_i$

Linear component: $\beta_0 + \beta_1 X_i$

Random Error component: ϵ_i

Linear Regression



Welcome to the world of data science

What is Linear?



What is Linear?



A Straight Line

What is Regression?



What is Regression?



Relationship between two points

What is Linear Regression?



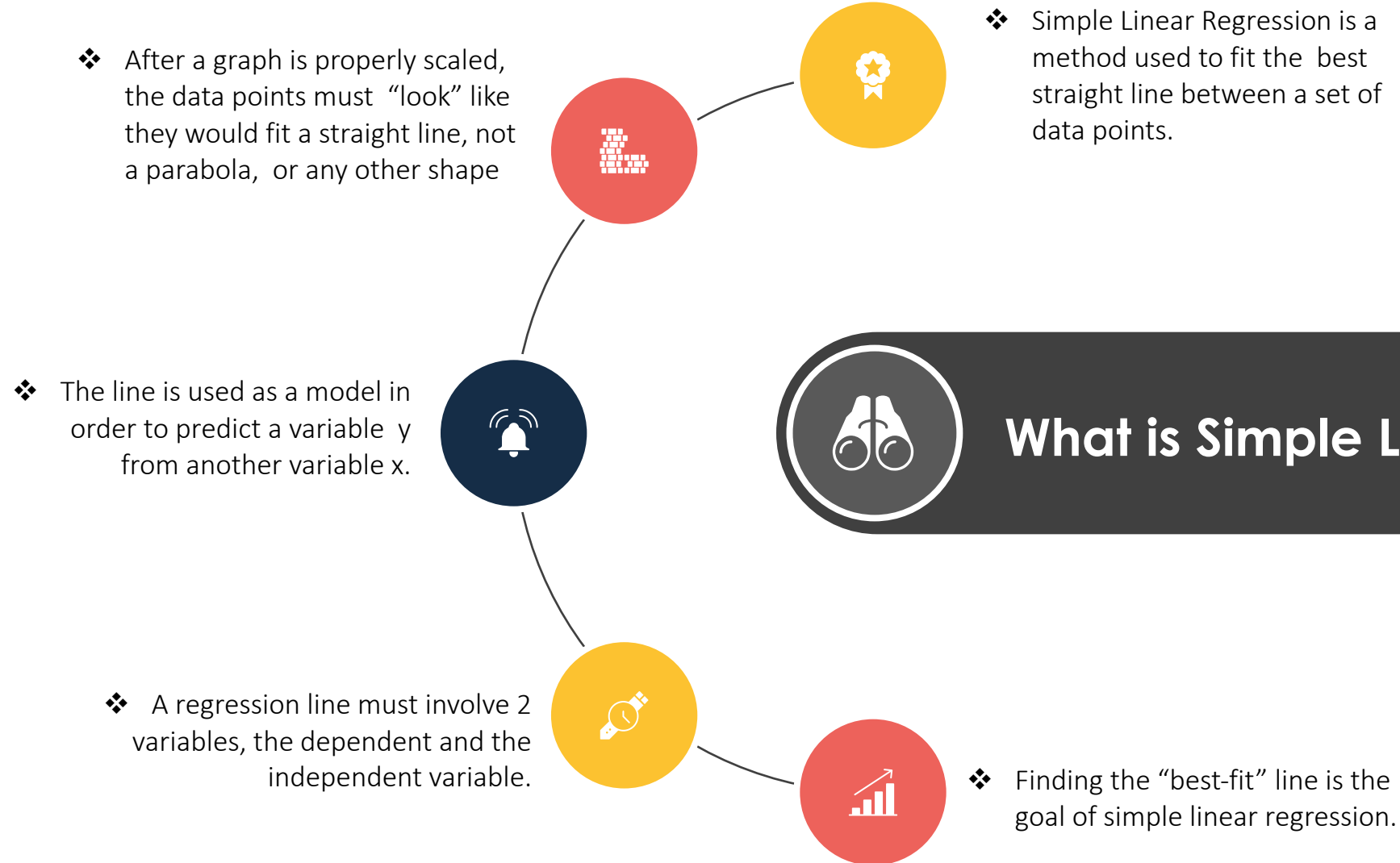
What is Linear Regression?



A Straight line that attempts to predict the relationship between two points



What is Simple Linear Regression?



Definitions



Input, Predictive, Or Independent Variable X

3 names mean the same thing. This is the variable whose value that is believed to influence the value of another variable. This variable should not be dependent on another variable (by definition)



Best-Fit Line

Represents our model. It is the line that “best fits” our data points. The line represents the best estimate of the y value for every given input of x..



Output, Response, Or Dependent Variable Y

3 names mean the same thing. This is the variable whose value that is believed to be influenced by the value of another variable. It is by definition, dependent on another variable



Sum Of Squares

An important calculation we will use to find the best-fit line.



One Variable

Problem:

A waiter wants to predict his next tip, but he forgot to record the bill amounts for previous tips

Here is a graph of his tips. The tips is the only variable.

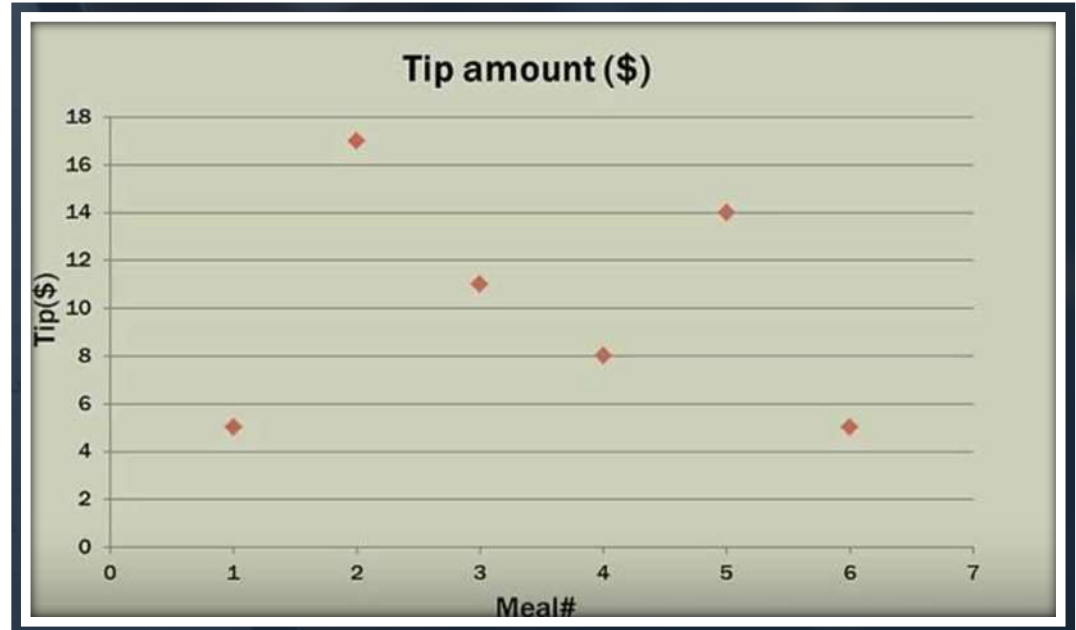
Let's call it the y variable

Meal# is not a variable. It is simply used to identify a tip.



y variable

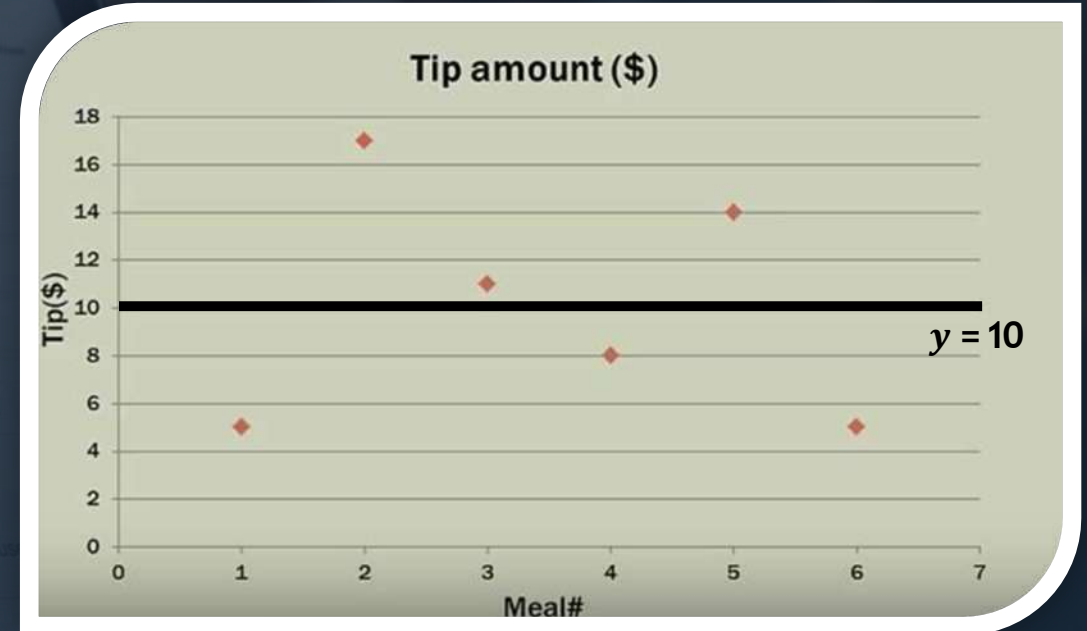
| Meal# | Tip amount (\$) |
|-------|-----------------|
| 1 | 5.00 |
| 2 | 17.00 |
| 3 | 11.00 |
| 4 | 8.00 |
| 5 | 14.00 |
| 6 | 5.00 |



Can we come up with a model for this problem with only 1 variable?

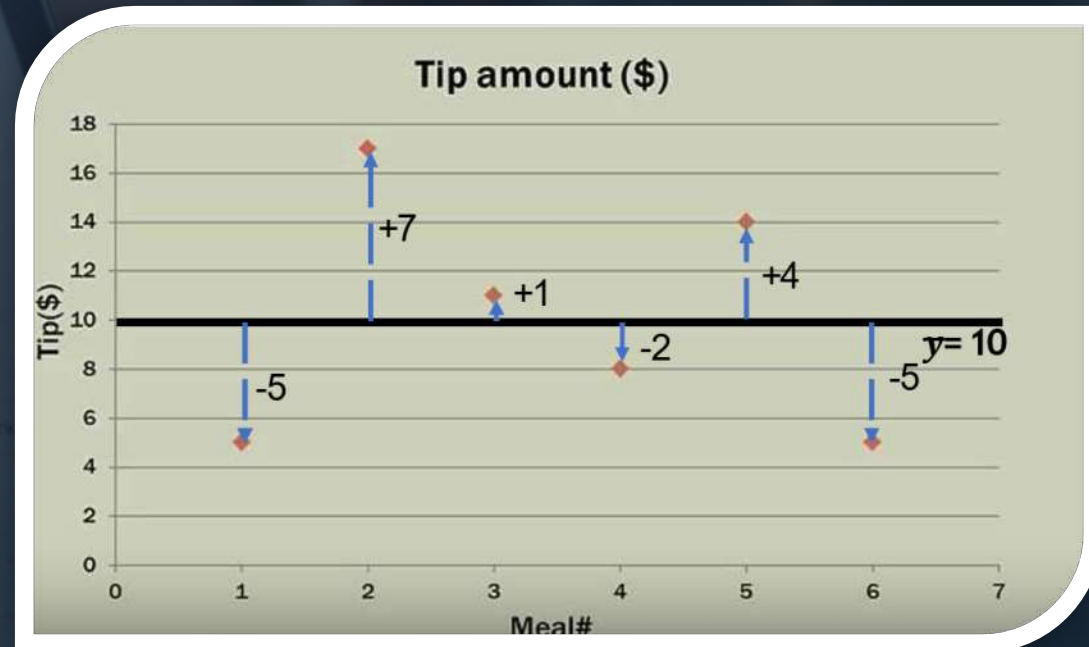
- ❑ The only option for our model is to use the mean of the Tips(\$)
- ❑ Tips are on the y axis. We would call the mean \bar{y} (y bar).
- ❑ The mean for the tip amounts is 10.
- ❑ The model for our problem is simply $\bar{y} = 10$.
- ❑ $\bar{y} = 10$ is our best fit line (represented by bold black line).

| Meal# | Tip amount (\$) |
|-------|-----------------|
| 1 | 5.00 |
| 2 | 17.00 |
| 3 | 11.00 |
| 4 | 8.00 |
| 5 | 14.00 |
| 6 | 5.00 |



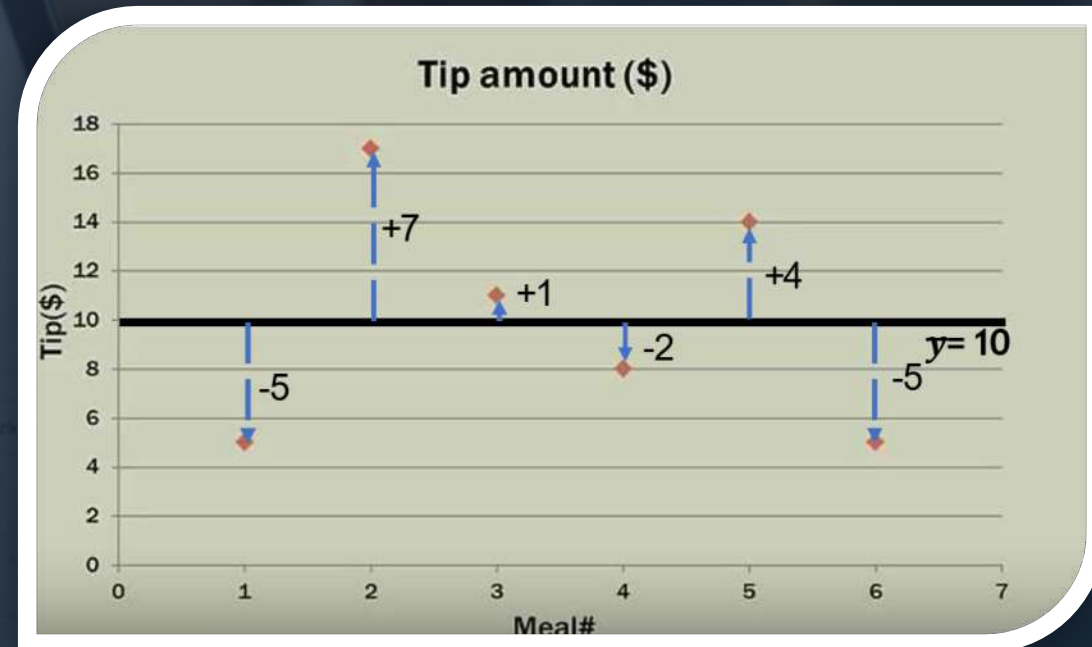
- ✓ Now, let's talk about goodness of fit. This will tell us how good our data points fit the line.
- ✓ We need to calculate the residuals (errors) For each point.

| Meal# | Tip amount (\$) |
|-------|-----------------|
| 1 | 5.00 |
| 2 | 17.00 |
| 3 | 11.00 |
| 4 | 8.00 |
| 5 | 14.00 |
| 6 | 5.00 |



- ✓ The best fit line is the one that minimizes the sum of the squares of the residuals (errors).
- ✓ The error is the difference between the actual data point and the point on the line.
- ✓ SSE (Sum Of Squared Errors) = $(-5)^2 + 7^2 + 1^2 + (-2)^2 + 4^2 + (-5)^2 = 120$

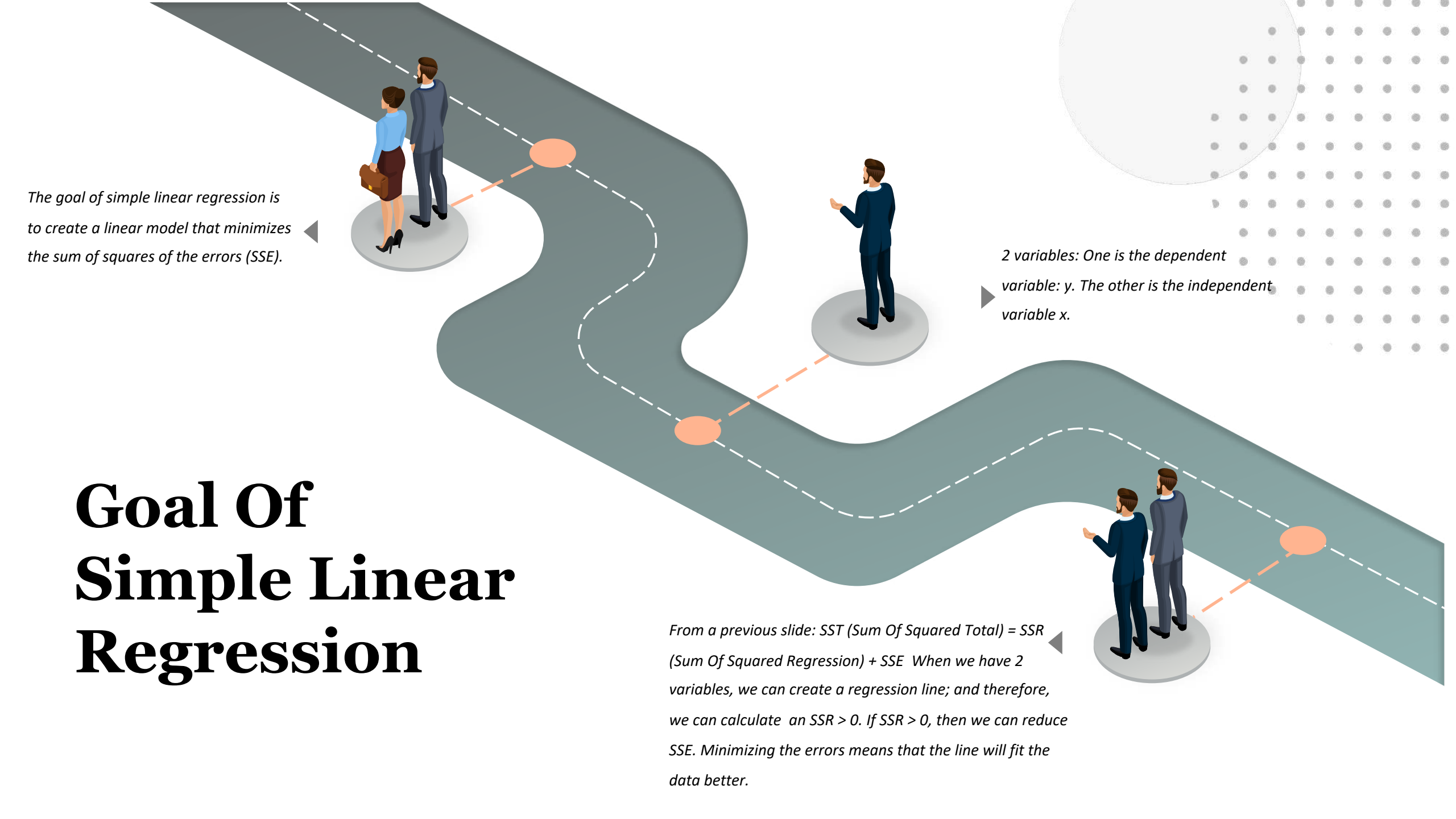
| Meal# | Tip amount (\$) |
|-------|-----------------|
| 1 | 5.00 |
| 2 | 17.00 |
| 3 | 11.00 |
| 4 | 8.00 |
| 5 | 14.00 |
| 6 | 5.00 |



- ✓ SST (Sum Of Squared Total) = SSR (Sum Of Squared Regression) + SSE is the Sum Of Squares Equation.
- ✓ Since there is no regression line (as we only have 1 variable), we can not make the SSE any smaller than 120, because $SSR = 0$.



Two Variables



The goal of simple linear regression is to create a linear model that minimizes the sum of squares of the errors (SSE).

2 variables: One is the dependent variable: y . The other is the independent variable x .

Goal Of Simple Linear Regression

From a previous slide: SST (Sum Of Squared Total) = SSR (Sum Of Squared Regression) + SSE . When we have 2 variables, we can create a regression line; and therefore, we can calculate an $SSR > 0$. If $SSR > 0$, then we can reduce SSE . Minimizing the errors means that the line will fit the data better.

So now, we need to introduce a little math.....

- Remember from the previous slide that we want to minimize the SSE.

We write this mathematically this way.:

$$\min \sum (y_i - \hat{y}_i)^2$$

y_i = observed value of dependent variable (tip amount)

\hat{y}_i = estimated(predicted) value of the dependent variable (predicted tip amount)

- y is often referred to as y -hat.

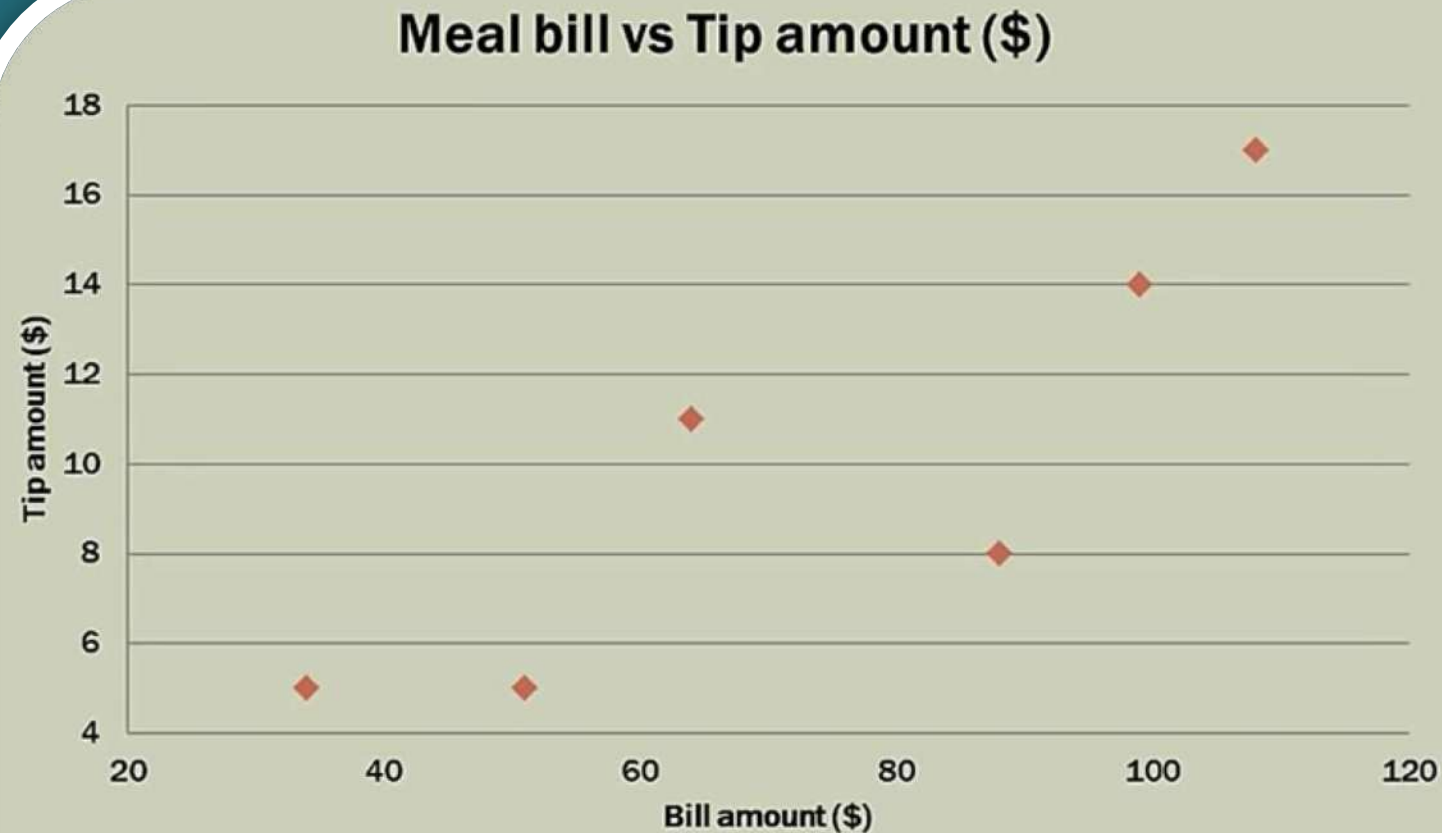
- Repeating the Problem: As a waiter, how do we predict the tips we will receive for service rendered?
- Let's say, we didn't forget to record the bill amount.

Independent Variable
(x)

| Total bill (\$) | Tip amount (\$) |
|-----------------|-----------------|
| 34.00 | 5.00 |
| 108.00 | 17.00 |
| 64.00 | 11.00 |
| 88.00 | 8.00 |
| 99.00 | 14.00 |
| 51.00 | 5.00 |

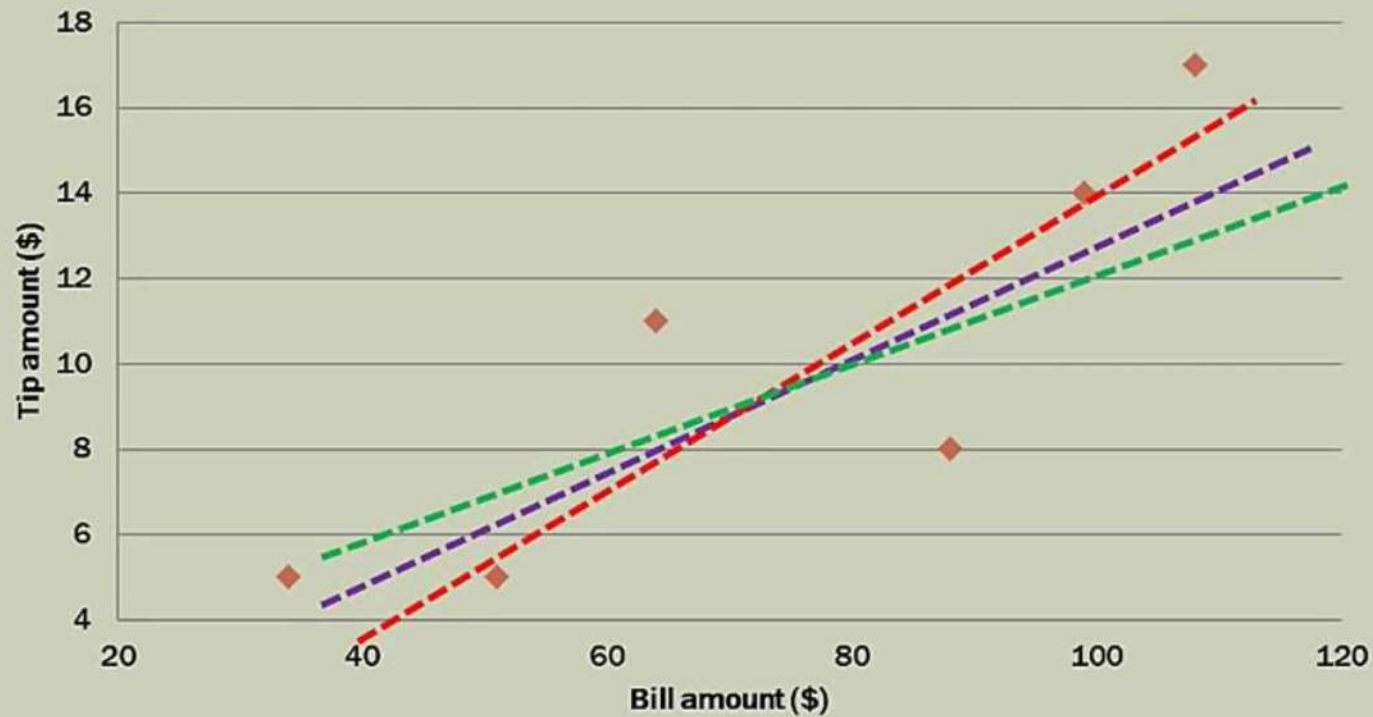
Dependent Variable (y)

If we scale the graph according to the data points available, we can then plot the points.



| Bill (\$) | Tip (\$) |
|-----------|----------|
| 34.00 | 5.00 |
| 108.00 | 17.00 |
| 64.00 | 11.00 |
| 88.00 | 8.00 |
| 99.00 | 14.00 |
| 51.00 | 5.00 |

Meal bill vs Tip amount (\$)

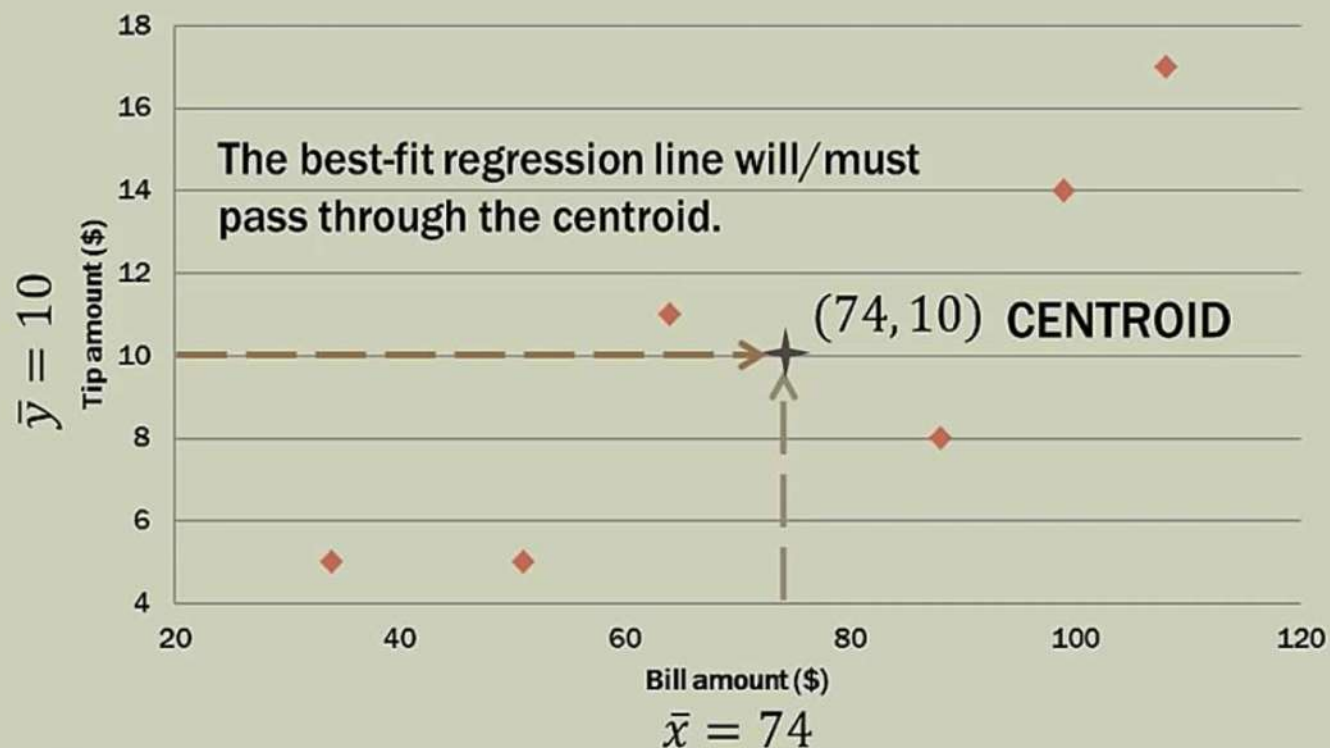


Does the data seem to fall along a line?

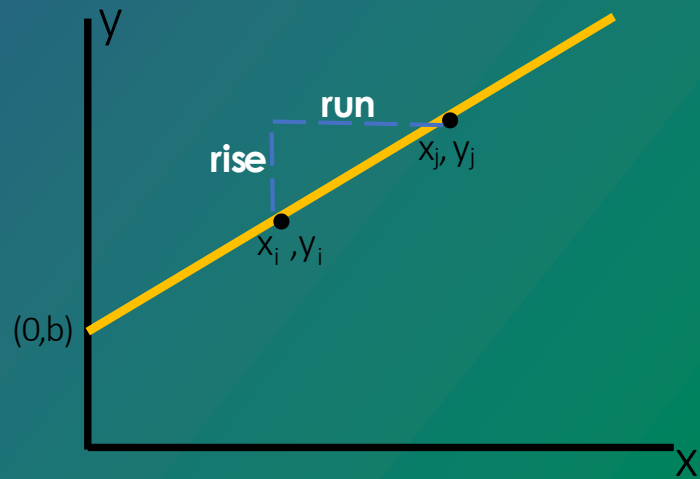
In this case, YES! Proceed.

If not...if it's a BLOB with no linear pattern, then stop.

Meal bill vs Tip amount (\$)



| Bill (\$) | Tip (\$) |
|----------------|----------------|
| 34.00 | 5.00 |
| 108.00 | 17.00 |
| 64.00 | 11.00 |
| 88.00 | 8.00 |
| 99.00 | 14.00 |
| 51.00 | 5.00 |
| $\bar{x} = 74$ | $\bar{y} = 10$ |



Equation Of A Straight Line

$$Y = mX + b$$

Y-intercept

Slope

$$m = (\text{rise/run}) = (y_j - y_i) / (x_j - x_i)$$

b is the point where $x=0$ and y intersects the y-axis

Regression Line - Slope

The formula for the slope, m , of the best-fitting line is

$$m = r \left(\frac{s_y}{s_x} \right)$$

where r is the correlation between X and Y , and s_x and s_y are the standard deviations of the x -values and the y -values, respectively. You simply divide s_y by s_x and multiply the result by r .

Think of s_y divided by s_x as the variation (resembling change) in Y over the variation in X , in units of X and Y . Rise Over Run!

\bar{x} = mean of the independent variable

\bar{y} = mean of the dependent variable

$$\hat{y}_i = b_0 + b_1 x_i$$

Slope

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

x_i = value of independent variable

y_i = value of dependent variable

Let's create a table of calculations that we can use to calculate the slope of the regression (best-fit) line.

| Meal | Total bill (\$) | Tip amount (\$) | Bill deviation | Tip Deviations | Deviation Products | Bill Deviations Squared |
|------|-----------------|-----------------|-----------------|-----------------|----------------------------------|-------------------------|
| | x | y | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
| 1 | 34 | 5 | -40 | -5 | 200 | 1600 |
| 2 | 108 | 17 | 34 | 7 | 238 | 1156 |
| 3 | 64 | 11 | -10 | 1 | -10 | 100 |
| 4 | 88 | 8 | 14 | -2 | -28 | 196 |
| 5 | 99 | 14 | 25 | 4 | 100 | 625 |
| 6 | 51 | 5 | -23 | -5 | 115 | 529 |
| | | | | | | |
| | $\bar{x} = 74$ | $\bar{y} = 10$ | | | $\sum = 615$ | $\sum = 4206$ |

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_1 = \frac{615}{4206}$$

$$b_1 = 0.1462$$

| Deviation Products | Bill Deviations Squared |
|----------------------------------|-------------------------|
| $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
| 200 | 1600 |
| 238 | 1156 |
| -10 | 100 |
| -28 | 196 |
| 100 | 625 |
| 115 | 529 |
| | |
| $\sum = 615$ | $\sum = 4206$ |

- ✓ Now that we have the slope, we can calculate the y-intercept because we know 1 point on the line already (74,10).
- ✓ What do we call 74,10?

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = 0.1462$$

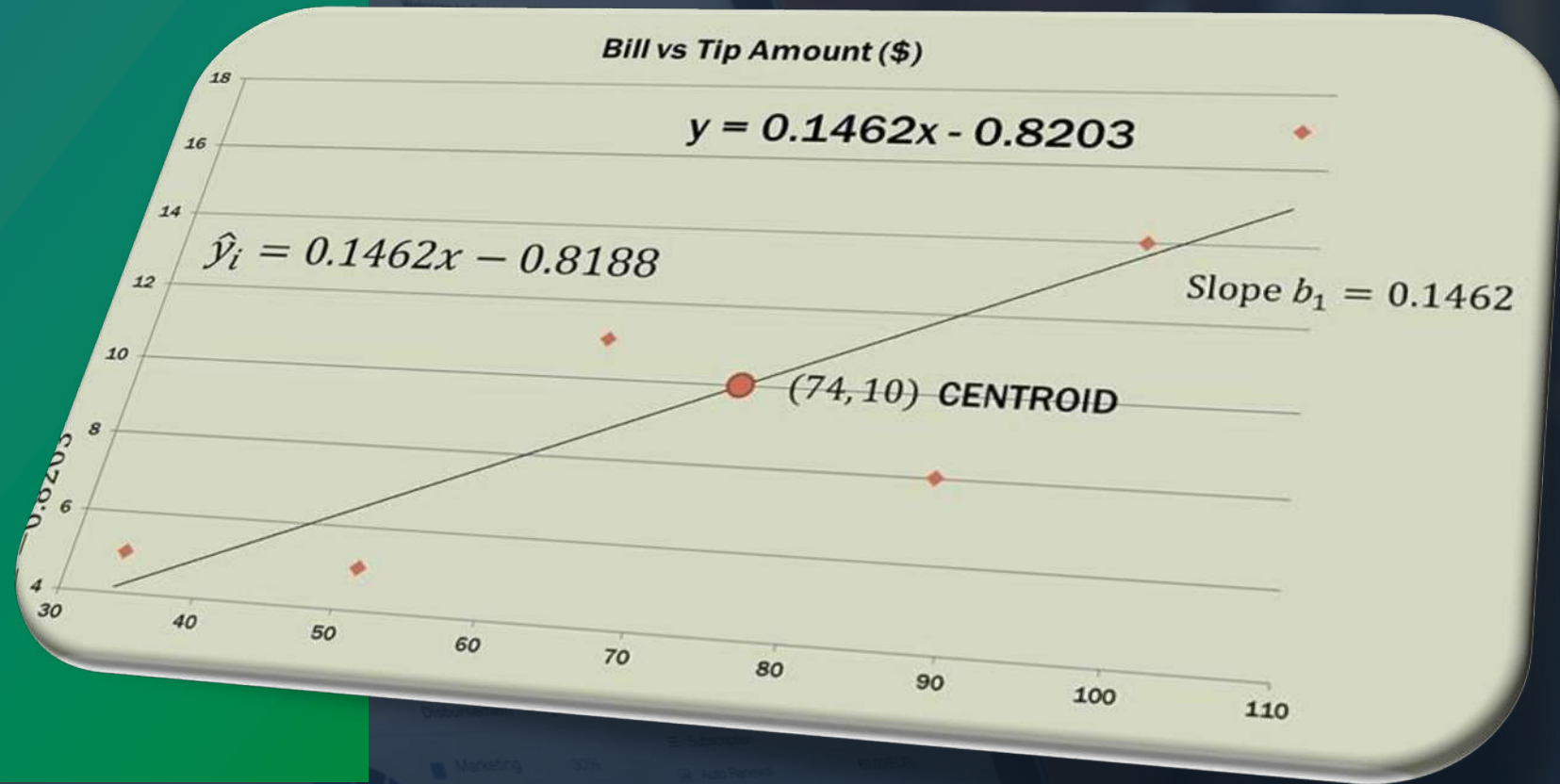
$$b_0 = 10 - 0.1462(74)$$

$$b_0 = 10 - 10.8188$$

$$b_0 = -0.8188$$

| Total bill (\$) | Tip amount (\$) |
|-----------------|-----------------|
| x | y |
| 34 | 5 |
| 108 | 17 |
| 64 | 11 |
| 88 | 8 |
| 99 | 14 |
| 51 | 5 |
| | |
| $\bar{x} = 74$ | $\bar{y} = 10$ |

- ✓ (74,10) is the Centroid.
- ✓ For comparison, Excel has calculated the regression equation very close to our manual calculation.



$$\hat{y}_i = 0.1462x - 0.8188$$

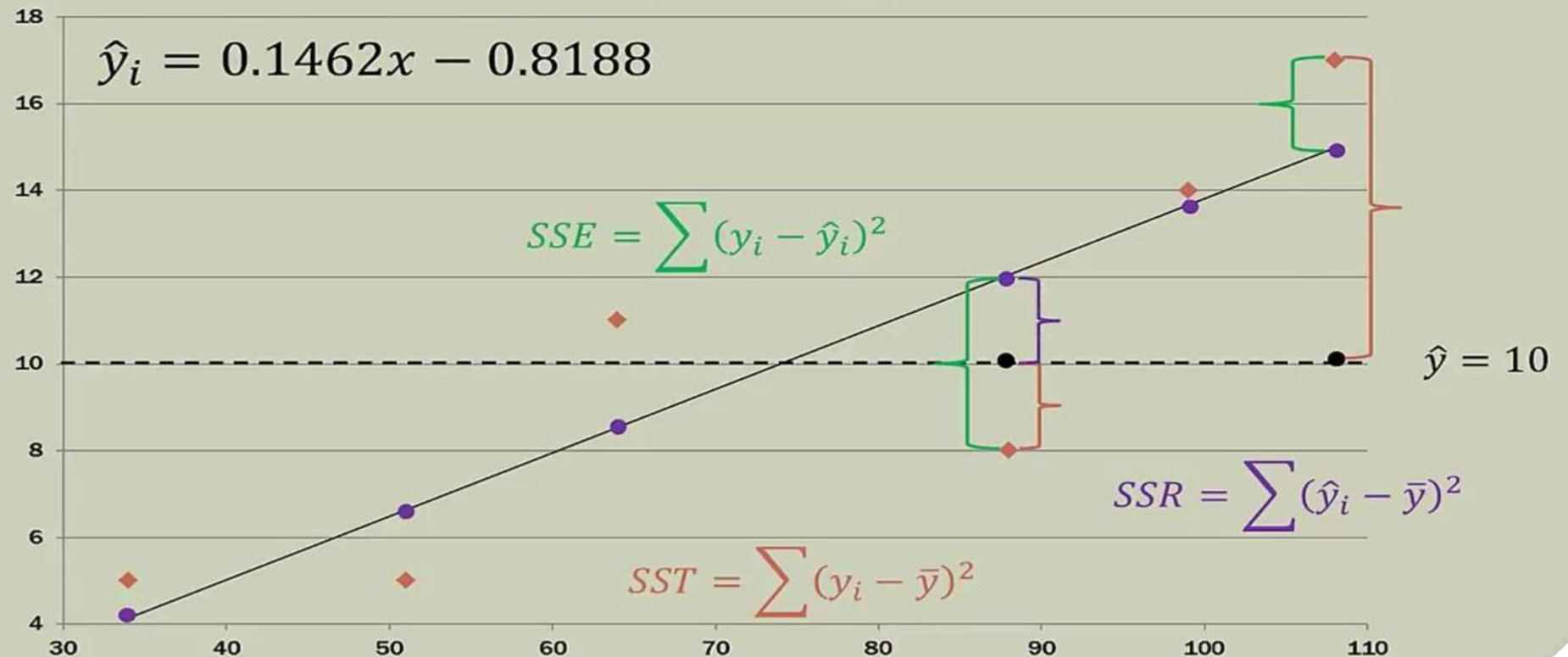
For every \$1 the bill amount (x) increases, we would expect the tip amount to increase by \$0.1462 or about 15-cents.

If the bill amount (x) is zero, then the expected/predicted tip amount is \$-0.8188 or negative 82-cents! Does this make sense? NO. The intercept may or may not make sense in the “real world.”

$$SST = SSR + SSE$$

Bill vs Tip Amount (\$)

3 Squared Differences



| Meal | Total bill (\$) | Observed tip amount (\$) | \hat{y}_i (predicted tip amount) | Error ($y - \hat{y}_i$) | Squared Error ($y - \hat{y}_i$) ² |
|------|-----------------|--------------------------|------------------------------------|---------------------------|--|
| | x | y | | | |
| 1 | 34 | 5 | 4.1505 | 0.8495 | 0.7217 |
| 2 | 108 | 17 | 14.9693 | 2.0307 | 4.1237 |
| 3 | 64 | 11 | 8.5365 | 2.4635 | 6.0688 |
| 4 | 88 | 8 | 12.0453 | -4.0453 | 16.3645 |
| 5 | 99 | 14 | 13.6535 | 0.3465 | 0.1201 |
| 6 | 51 | 5 | 6.6359 | -1.6359 | 2.6762 |
| | | | | | |
| | $\bar{x} = 74$ | $\bar{y} = 10$ | | SSE = $\sum = 30.075$ | |

General Regression using Linear Algebra

*Find the best-fitting line
for the data points*

| X | Y |
|---|----|
| 1 | 2 |
| 2 | 3 |
| 4 | 7 |
| 5 | 5 |
| 7 | 11 |

– Linear Algebra



Solve the following system of linear equations for x , y , and z :

$$\begin{array}{rclcrcl} x & + & 2y & + & z & = & 12 \\ 2x & - & y & + & z & = & 1 \\ x & + & y & - & 3z & = & -4 \end{array}.$$

While there are many ways to solve these types of systems, one of special interest is by treating the three lines as a linear transformation and looking at its corresponding matrix:

$$\begin{array}{rcrcrcrcrcl} x & + & 2y & + & z & = & 12 \\ 2x & - & y & + & z & = & 1 \\ x & + & y & - & 3z & = & -4 \end{array} \implies \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 1 \\ 1 & 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 12 \\ 1 \\ -4 \end{pmatrix}.$$

Then, $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ is an element of the vector space \mathbb{R}^3 , and the matrix describes a linear transformation from \mathbb{R}^3 to itself. Finding the matrix's inverse then yields the answer:

$$\begin{pmatrix} \frac{2}{19} & \frac{7}{19} & \frac{3}{19} \\ \frac{7}{19} & -\frac{4}{19} & \frac{1}{19} \\ \frac{3}{19} & \frac{1}{19} & -\frac{5}{19} \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 1 \\ 2 & -1 & 1 \\ 1 & 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{2}{19} & \frac{7}{19} & \frac{3}{19} \\ \frac{7}{19} & -\frac{4}{19} & \frac{1}{19} \\ \frac{3}{19} & \frac{1}{19} & -\frac{5}{19} \end{pmatrix} \cdot \begin{pmatrix} 12 \\ 1 \\ -4 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \\ 3 \end{pmatrix}. \quad \square$$

Solving with Linear algebra

The previous example can be rewritten in matrix language: we seek a least-squares approximation to the equation

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \\ 5 & 1 \\ 7 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 7 \\ 5 \\ 11 \end{pmatrix}.$$

This equation has no solutions (since no line goes through all five points), but the least squares solution is given by multiplying both sides by A^T and solving

$$\begin{pmatrix} 1 & 2 & 4 & 5 & 7 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 4 & 1 \\ 5 & 1 \\ 7 & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 1 & 2 & 4 & 5 & 7 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 7 \\ 5 \\ 11 \end{pmatrix}$$
$$\begin{pmatrix} 95 & 19 \\ 19 & 5 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} 138 \\ 28 \end{pmatrix},$$

which is the same system of equations we got by taking partial derivatives, and leads again to the unique solution $m = \frac{79}{57}$ and $b = \frac{1}{3}$. \square