Prediction:
Correlation and Regression

Prediction

- Predicting one characteristic based on another:
  - Given my height, how tall will I be next year?
  - Given my height, how tall will my kid be as an adult?
  - Given my height, how much will I spend on a boat?

- There's something I know, and something I want to determine
  - Characteristics of an example: known and unknown

- Assumption of prediction: for some sample, we know all the characteristics

Relation Between Two Variables

- Association
- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape”
  - Linear
  - Non-linear
- Good protocol: visualize first, then quantify

(Demo)

The Correlation Coefficient, *r*

- Measures linear association
- Based on standard units
- -1 <= *r* <= 1
  - *r* = 1: scatter plot is perfect straight line sloping up
  - *r* = -1: scatter plot is perfect straight line sloping down
  - *r* = 0: no linear association; *uncorrelated*

(Demo)

The Correlation Coefficient, *r*

1. Convert both variables to standard units
   - Subtract off the mean, divide by the standard deviation
2. Multiply them together
3. Average the products
   - That's *r*

The Correlation Coefficient, *r*

- *r* is a pure number, with no units
- *r* is not affected by changing units of measurement
- *r* is not affected by switching the the horizontal and vertical axes

X	Y
1.00	2.00
2.00	3.00
3.00	1.00
4.00	5.00
5.00	2.00
6.00	7.00

Computing Correlation

Step 1: ?

X	Y
1.00	2.00
2.00	3.00
3.00	1.00
4.00	5.00
5.00	2.00
6.00	7.00

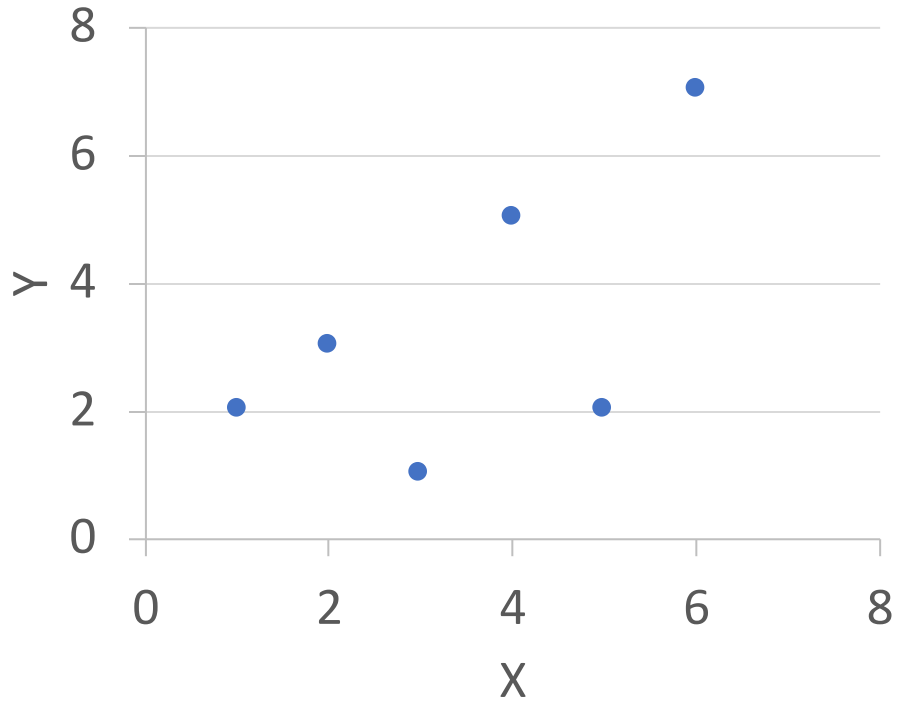
Computing Correlation

Step 1: Visualize!

X	Y
1.00	2.00
2.00	3.00
3.00	1.00
4.00	5.00
5.00	2.00
6.00	7.00

Computing Correlation

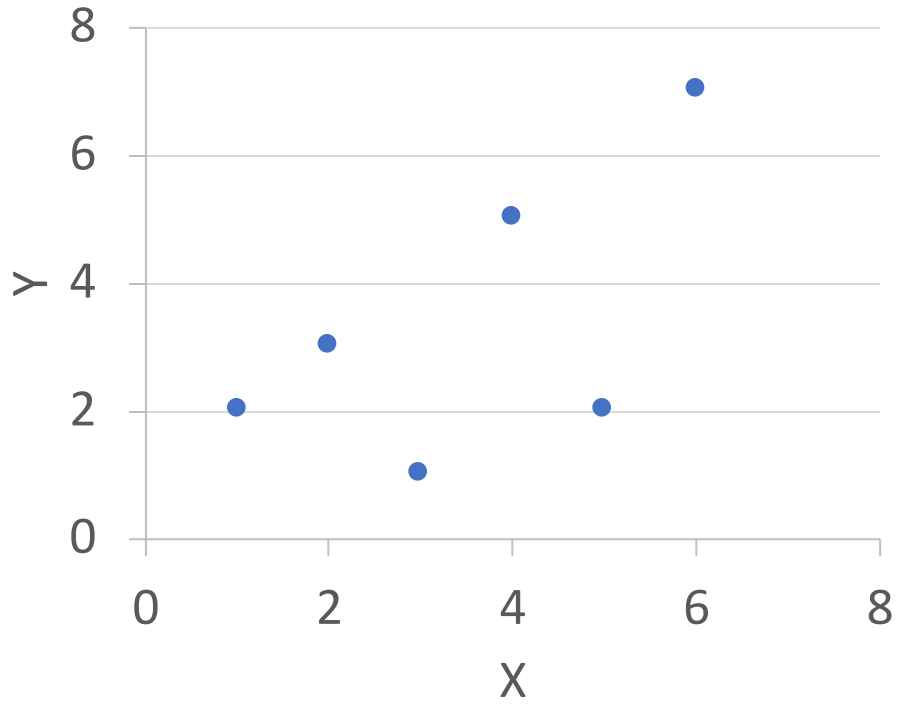
Step 1: Visualize!



X	Y
1.00	2.00
2.00	3.00
3.00	1.00
4.00	5.00
5.00	2.00
6.00	7.00

Computing Correlation

Step 1: Visualize!
Step 2: ?



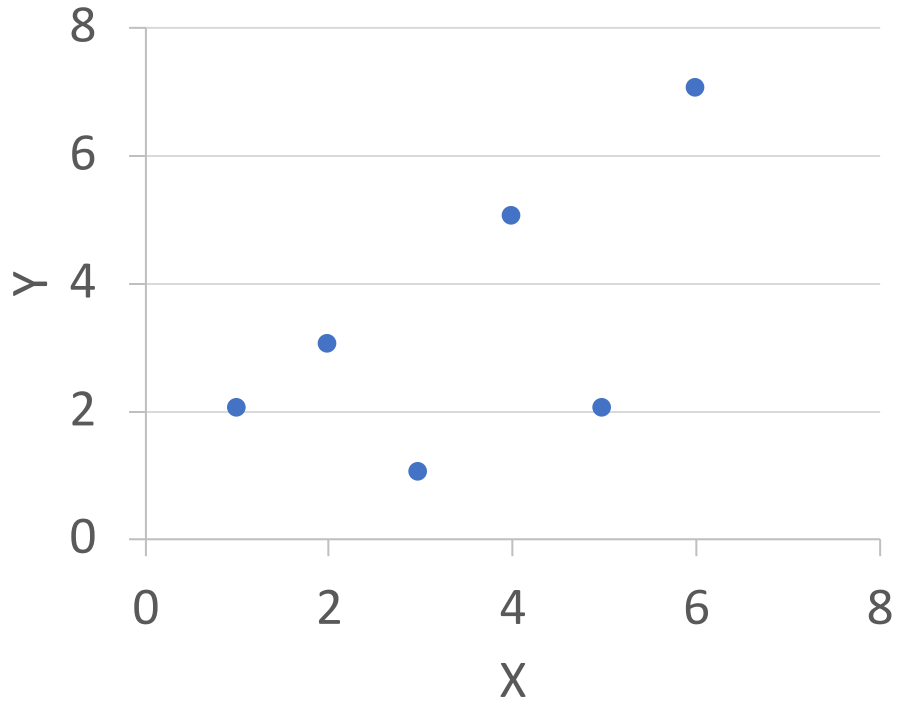
X	Y
1.00	2.00
2.00	3.00
3.00	1.00
4.00	5.00
5.00	2.00
6.00	7.00

Computing Correlation

Step 1: Visualize!

Step 2: Convert to standard units

Subtract off the mean and divide by the standard deviation



	X	Y
	1.00	2.00
	2.00	3.00
	3.00	1.00
	4.00	5.00
	5.00	2.00
	6.00	7.00
Mean	3.50	3.33
St.D.	1.87	2.25

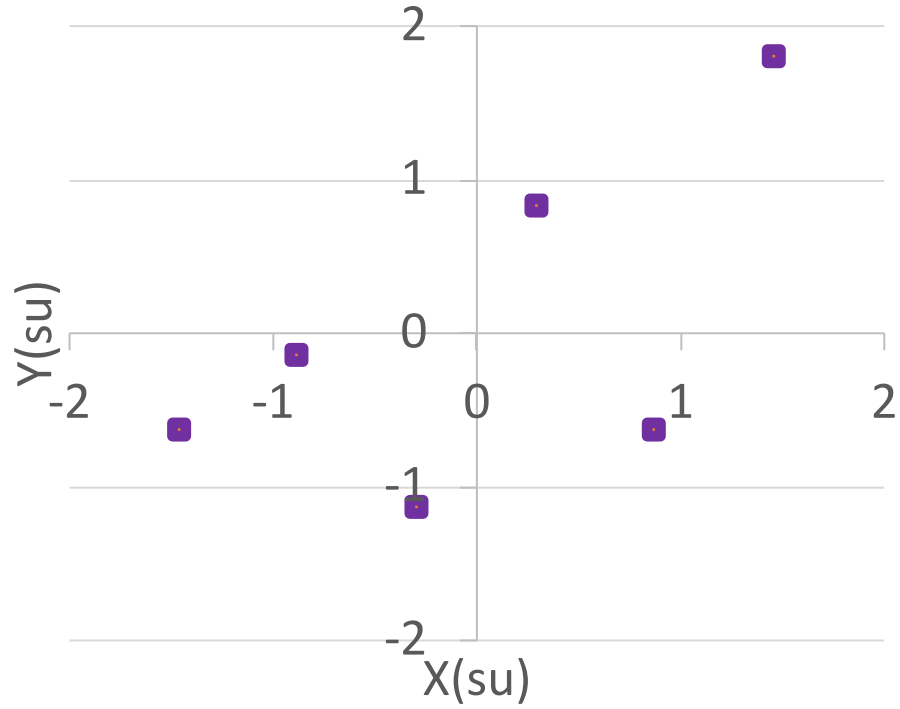
Computing Correlation

Step 1: Visualize!

Step 2: Convert to standard units

Subtract off the mean and divide by the standard deviation

X(su)	Y(su)
-1.34	-0.59
-0.80	-0.15
-0.27	-1.04
0.27	0.74
0.80	-0.59
1.34	1.63



	X	Y
	1.00	2.00
	2.00	3.00
	3.00	1.00
	4.00	5.00
	5.00	2.00
	6.00	7.00
Mean	3.50	3.33
St.D.	1.87	2.25

Computing Correlation

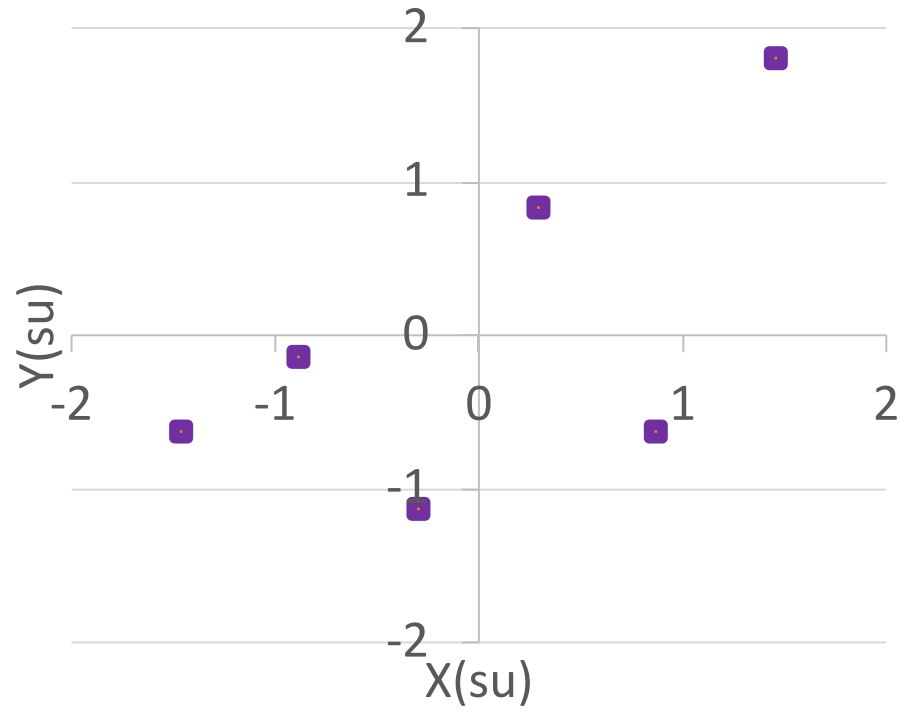
Step 1: Visualize!

Step 2: Convert to standard units

Subtract off the mean and divide by the standard deviation

Step 3: ?

X(su)	Y(su)
-1.34	-0.59
-0.80	-0.15
-0.27	-1.04
0.27	0.74
0.80	-0.59
1.34	1.63



	X	Y
	1.00	2.00
	2.00	3.00
	3.00	1.00
	4.00	5.00
	5.00	2.00
	6.00	7.00
Mean	3.50	3.33
St.D.	1.87	2.25

Computing Correlation

Step 1: Visualize!
Step 2: Convert to standard units
Subtract off the mean and divide by the standard deviation
Step 3: Multiply X(su) * Y(su)
Step 4: ?

X(su)	Y(su)	Product
-1.34	-0.59	0.79
-0.80	-0.15	0.12
-0.27	-1.04	0.28
0.27	0.74	0.20
0.80	-0.59	-0.47
1.34	1.63	2.18

	X	Y
	1.00	2.00
	2.00	3.00
	3.00	1.00
	4.00	5.00
	5.00	2.00
	6.00	7.00
Mean	3.50	3.33
St.D.	1.87	2.25

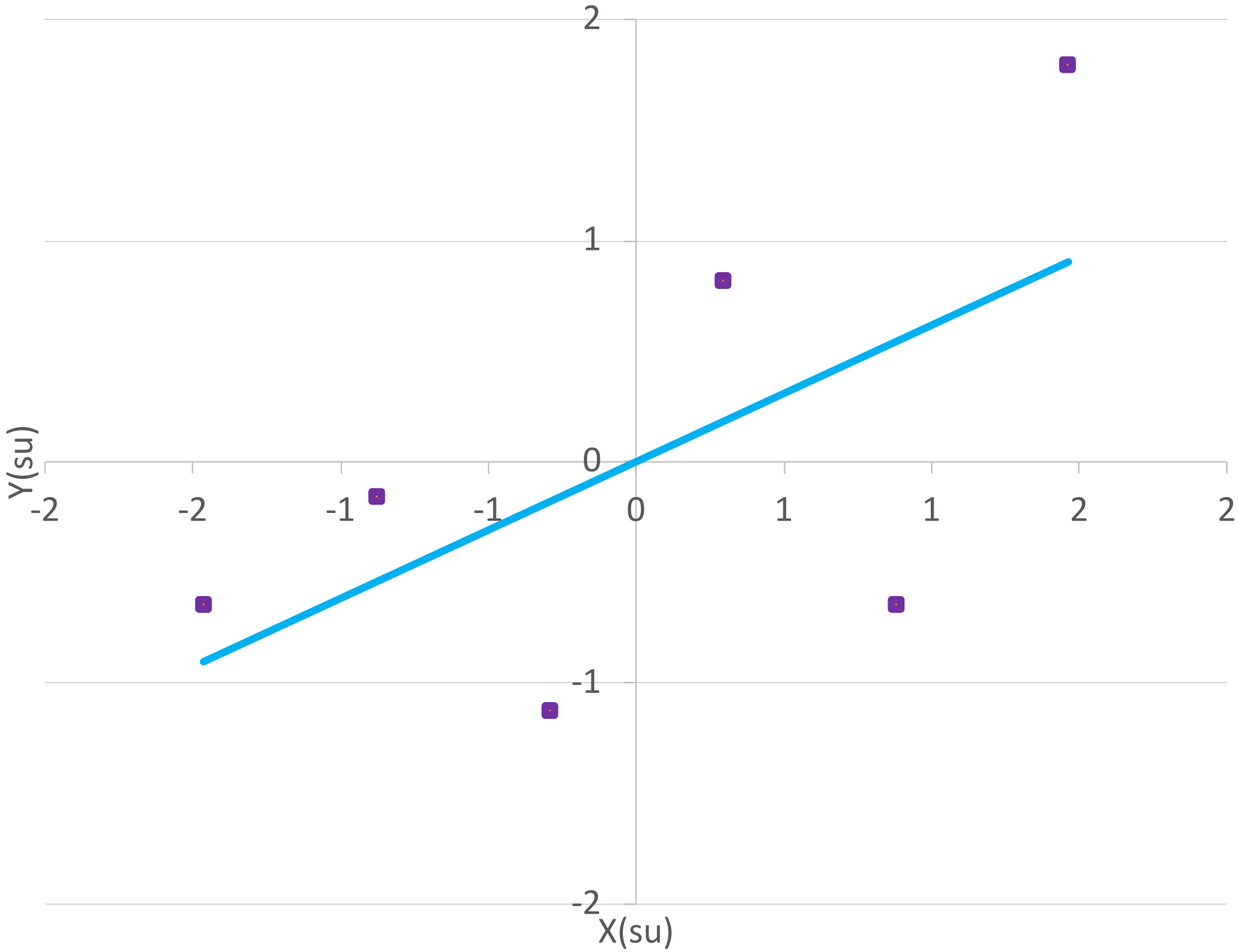
Computing Correlation

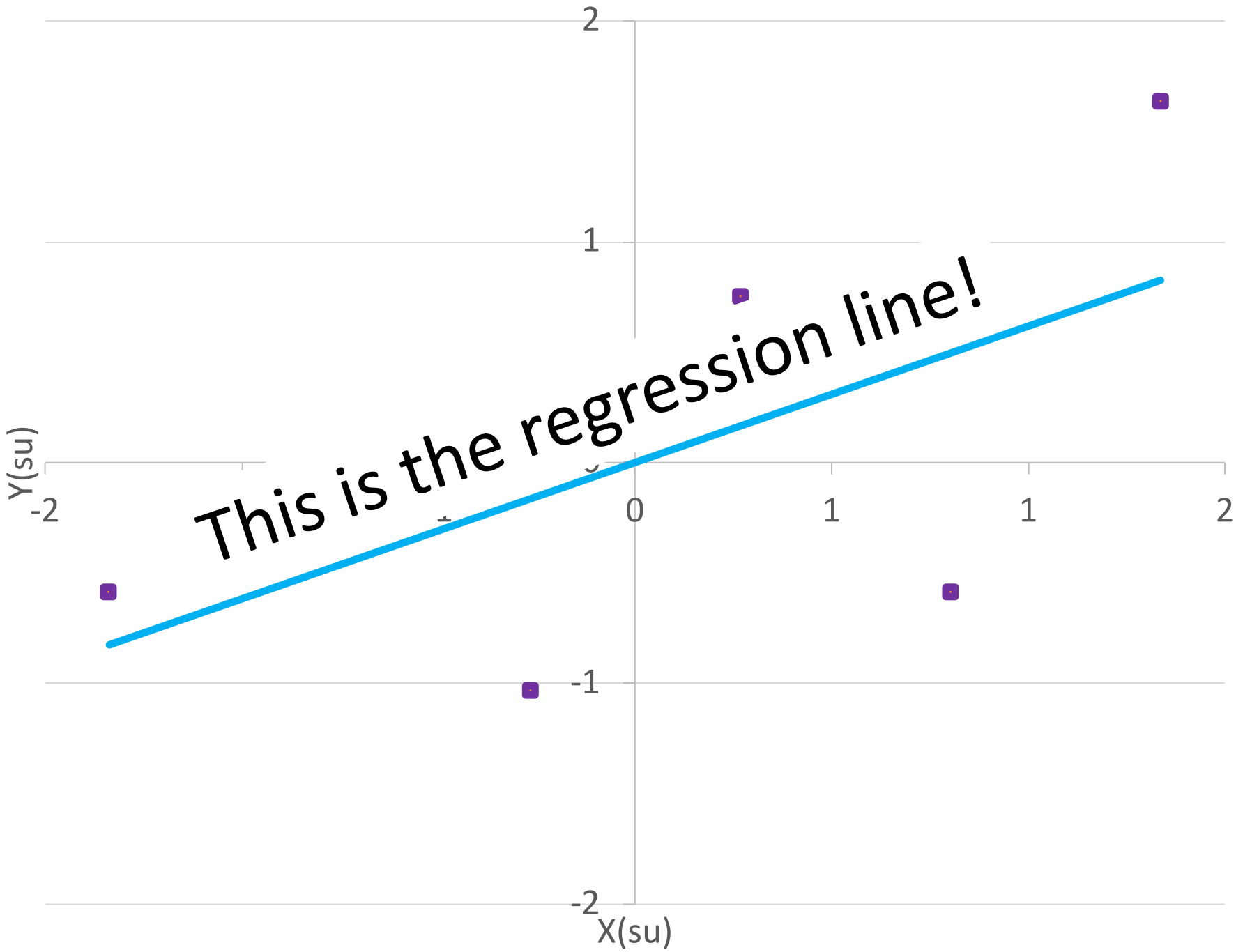
Step 1: Visualize!
Step 2: Convert to standard units
Subtract off the mean and divide by the standard deviation
Step 3: Multiply X(su) * Y(su)
Step 4: Average the products

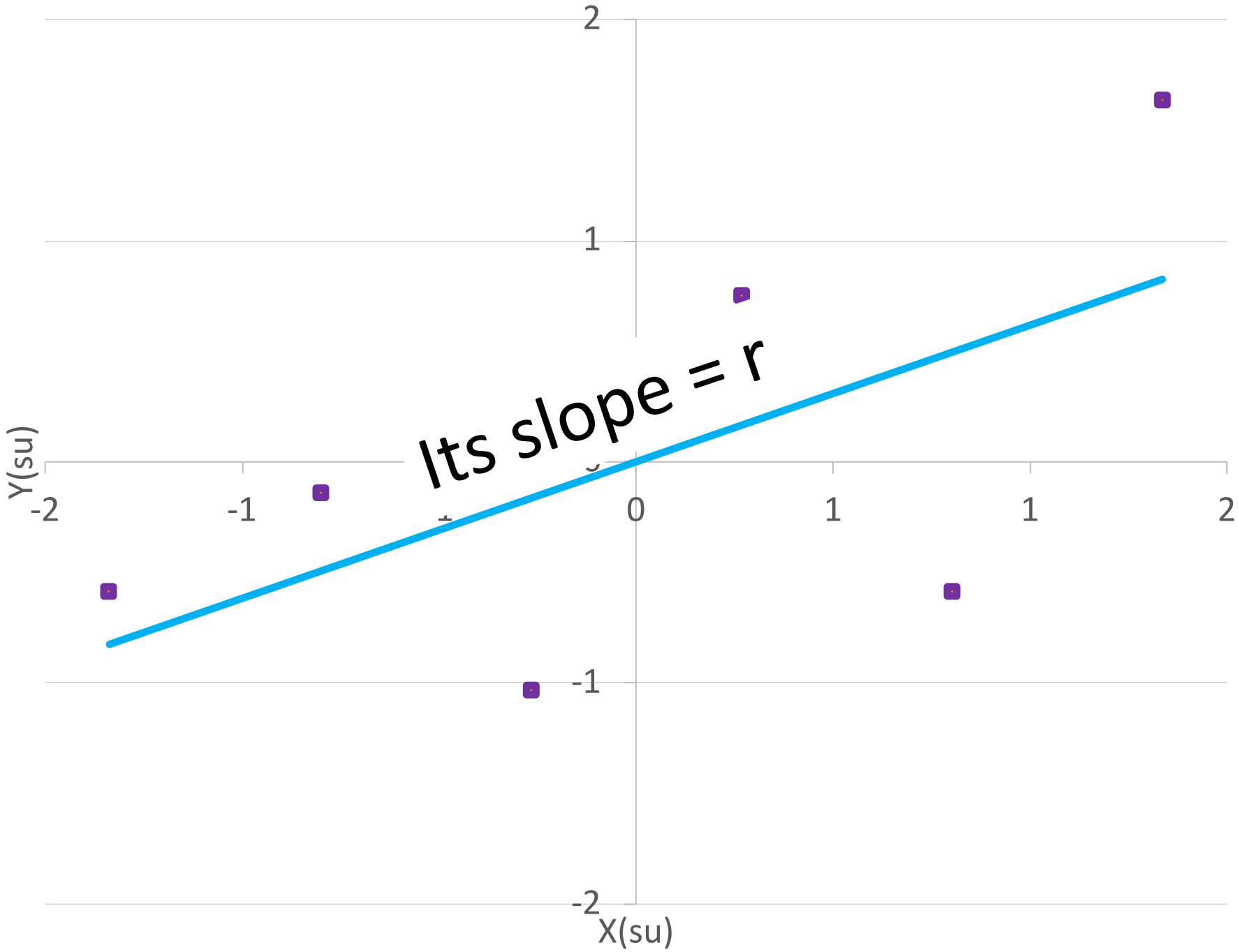
X(su)	Y(su)	Product
-1.34	-0.59	0.79
-0.80	-0.15	0.12
-0.27	-1.04	0.28
0.27	0.74	0.20
0.80	-0.59	-0.47
1.34	1.63	2.18

r = (0.79 + 0.12 + 0.28 + 0.20 + -0.47 + 2.18) / 6

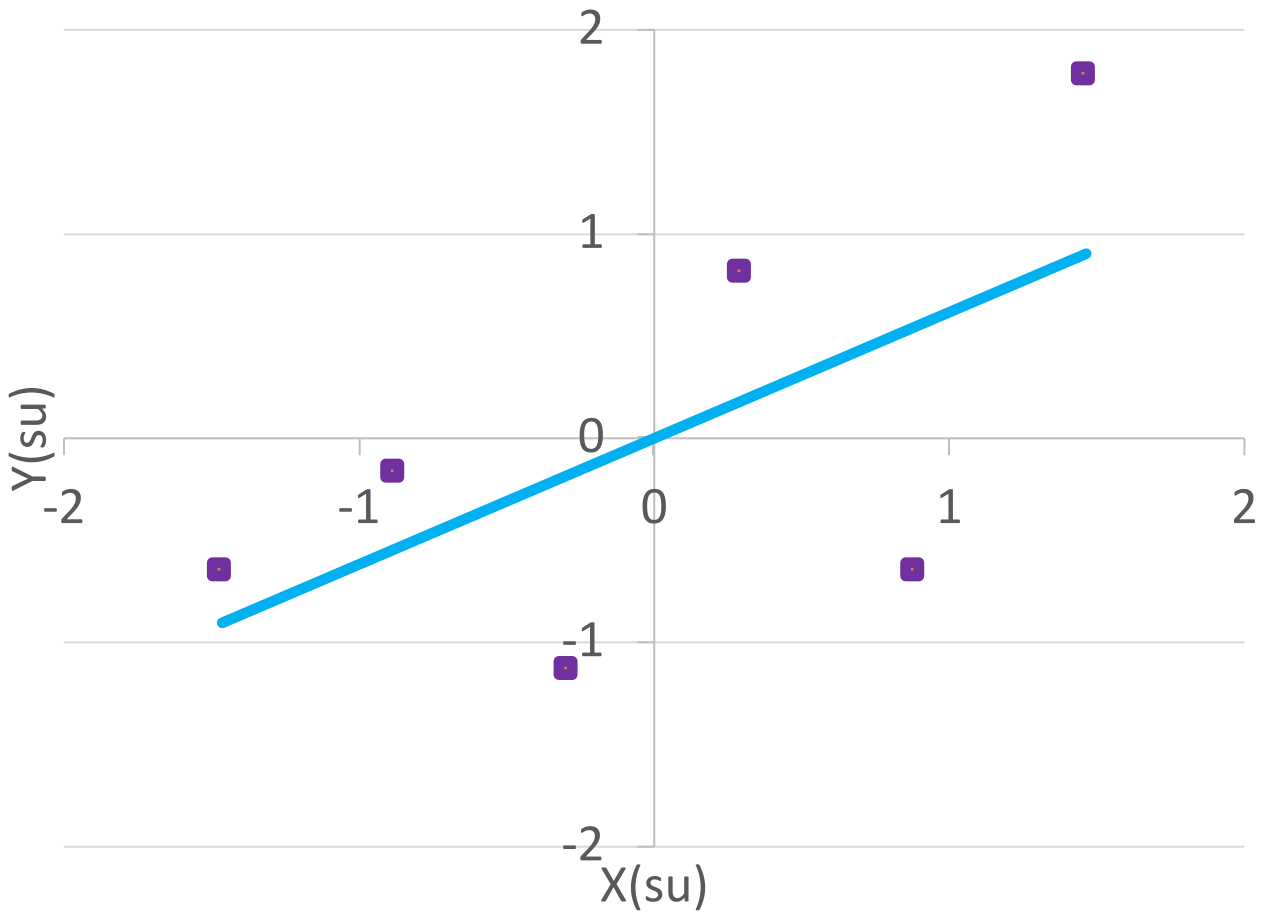
r = 0.51


2
1
0
-1
-2
-2
-1
0
1
2
Y(su)
X(su)


2
1
-1
-2
-2
0
1
1
2
Y(su)
X(su)
This is the regression line!


2
1
-1
-2
-2
-1
0
1
1
2
Y(su)
X(su)
Its slope = r

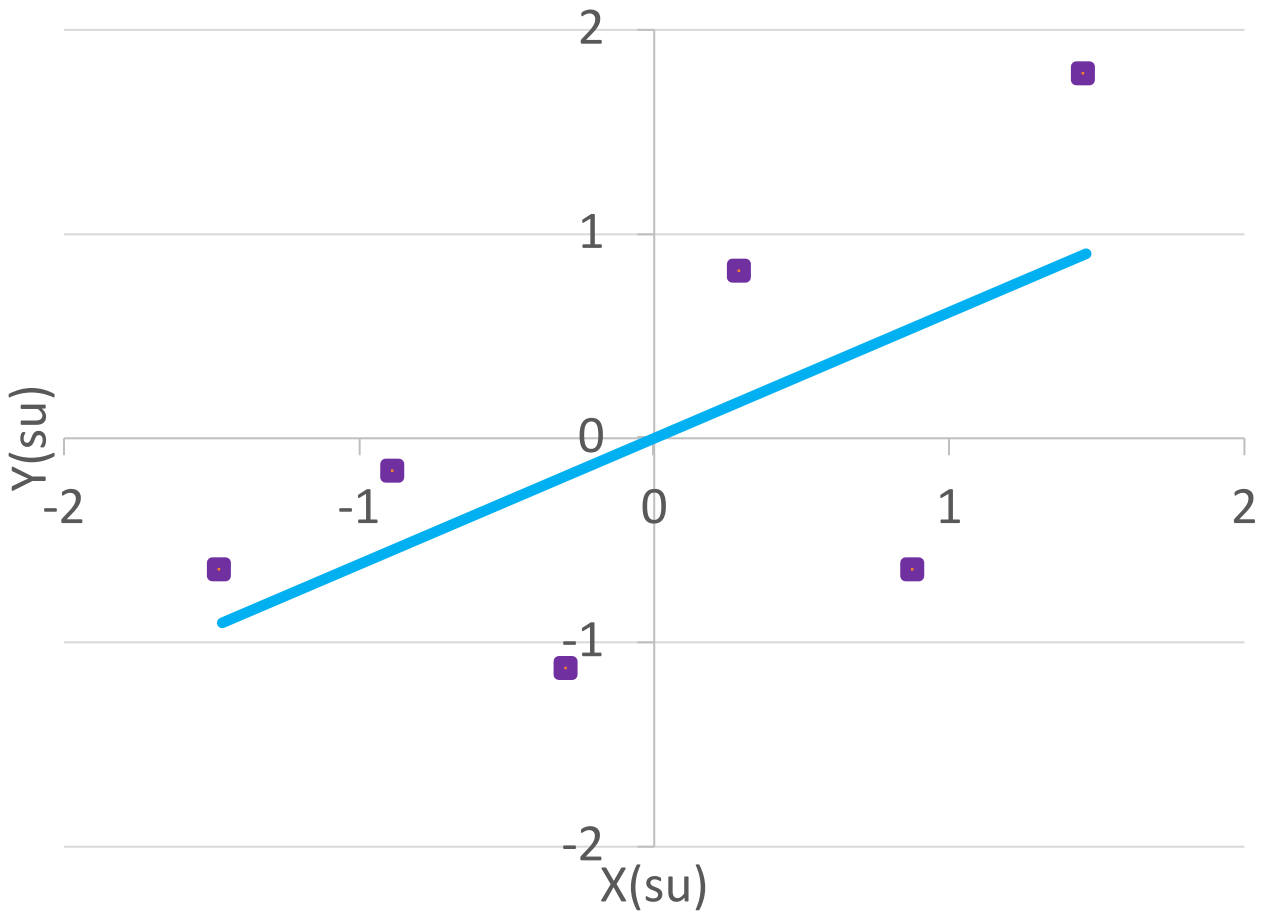
Regression Line Equation



Equation of a line: $y = mx + b$

y: the y-value for a given x-value
x: a given x-value
m: slope of the line (r!)
b: y-intercept

Regression Line Equation



Equation of a line: $y = mx + b$

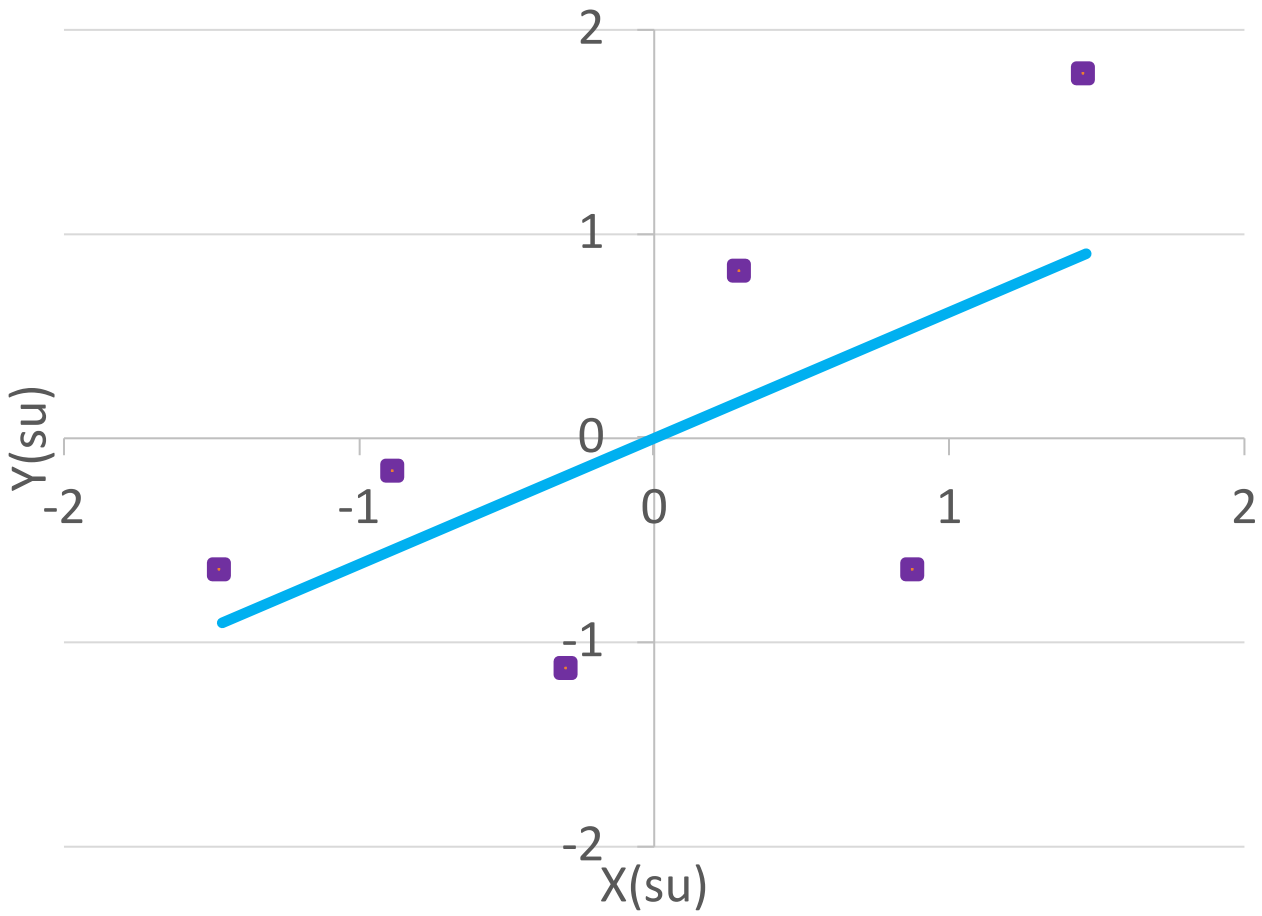
y: the y-value for a given x-value
x: a given x-value
m: slope of the line (r!)
b: y-intercept

In standard units, b = 0

So the equation is just:

$$y = mx$$

Regression Line Equation



Equation of a line: $y = mx + b$

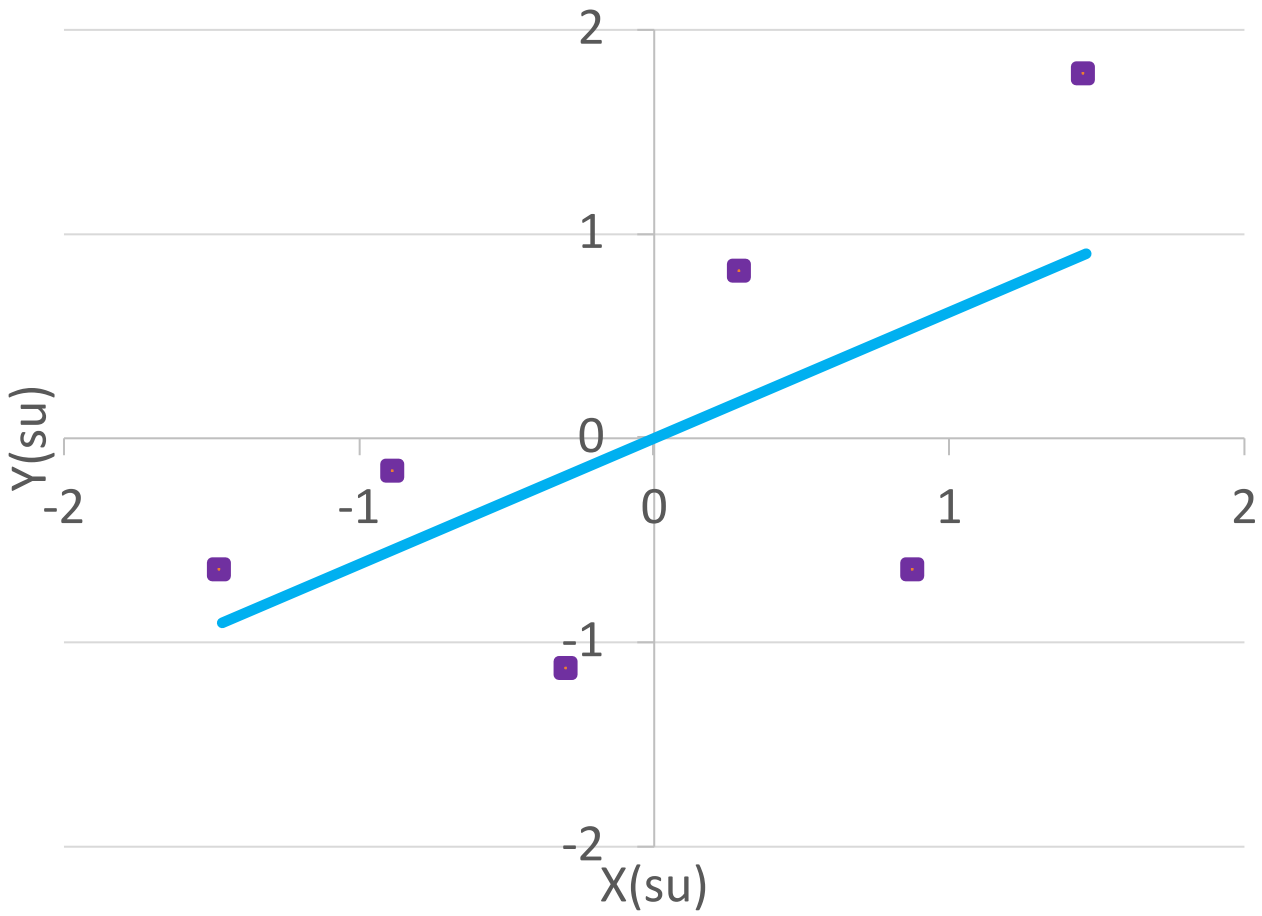
y: the y-value for a given x-value
x: a given x-value
m: slope of the line (r!)
b: y-intercept

In standard units, b = 0

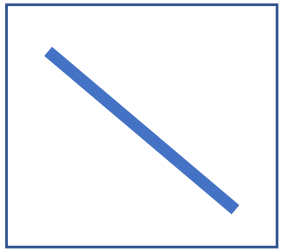
So the equation is just:

$$y = rx$$

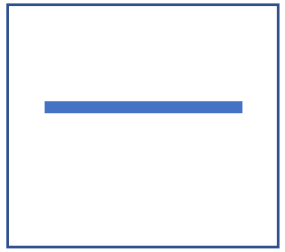
Regression Line Equation



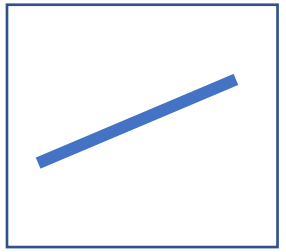
$$y = rx$$



slope < 0

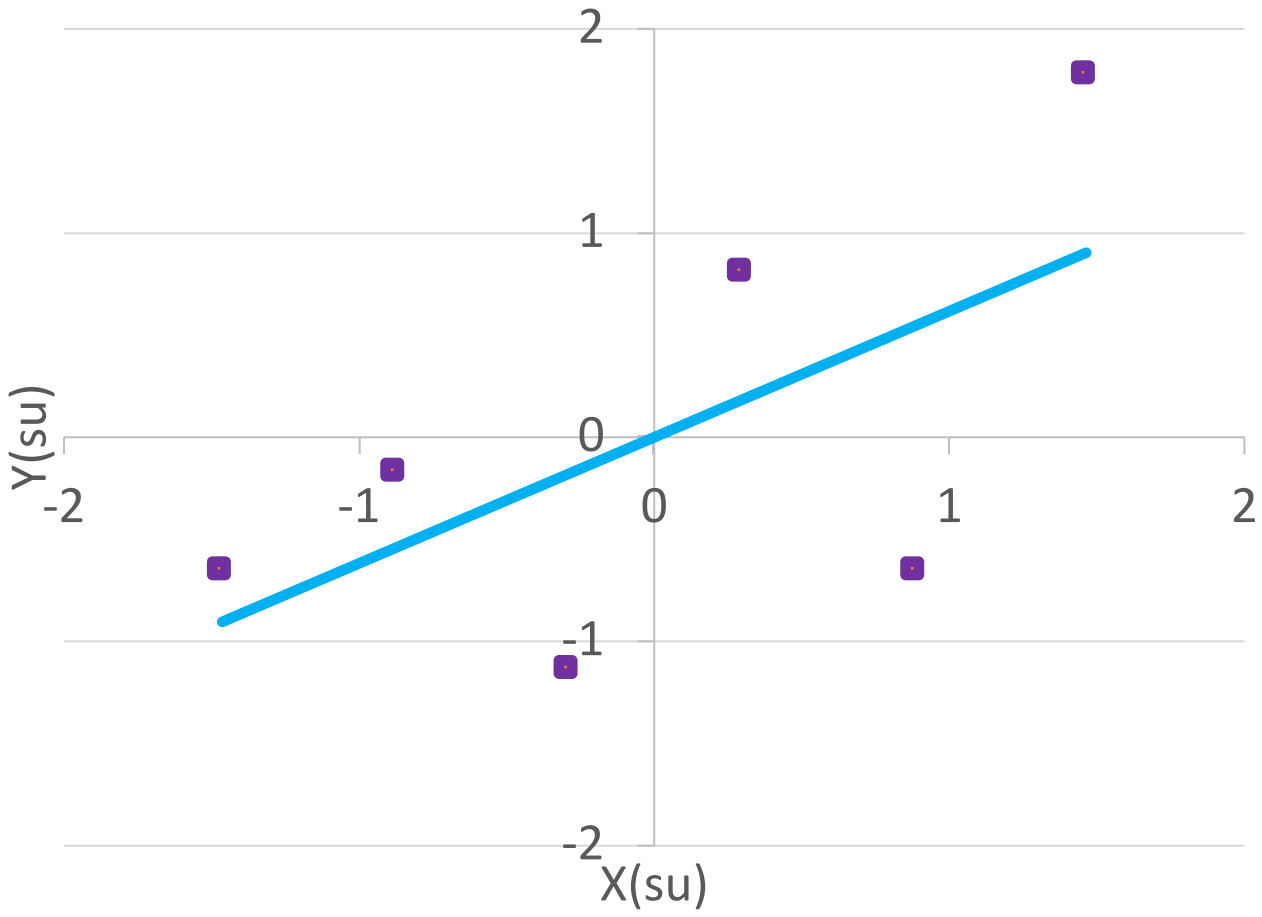


slope = 0

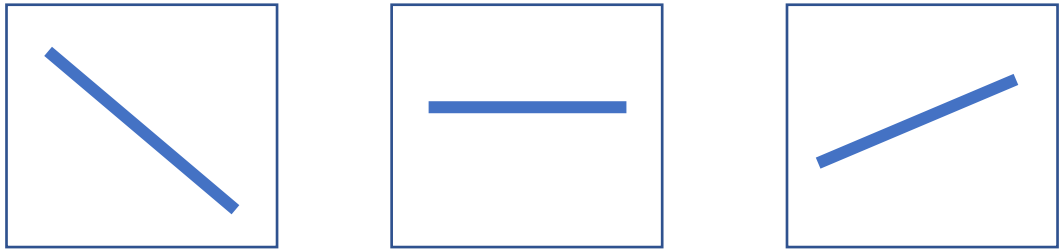


slope > 0

Regression Line Equation

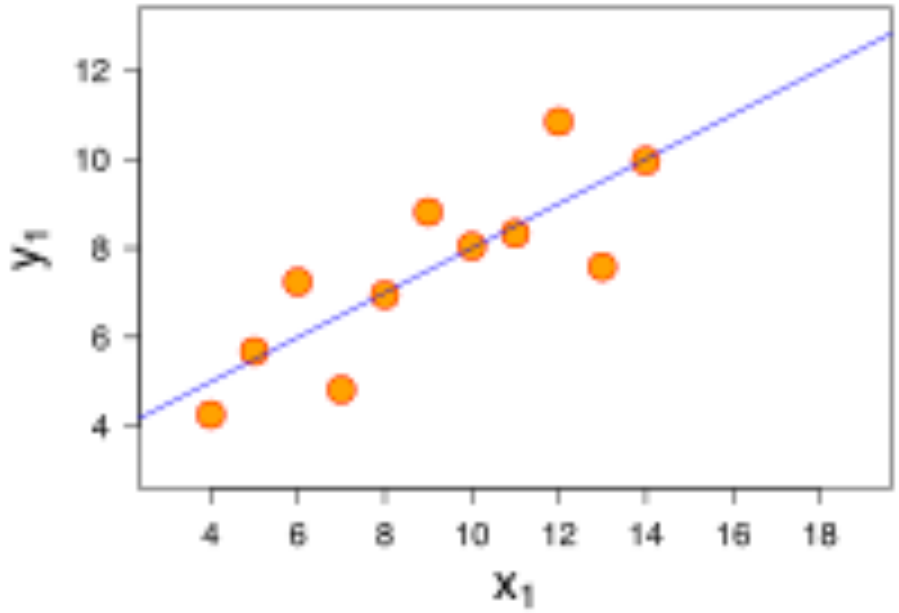


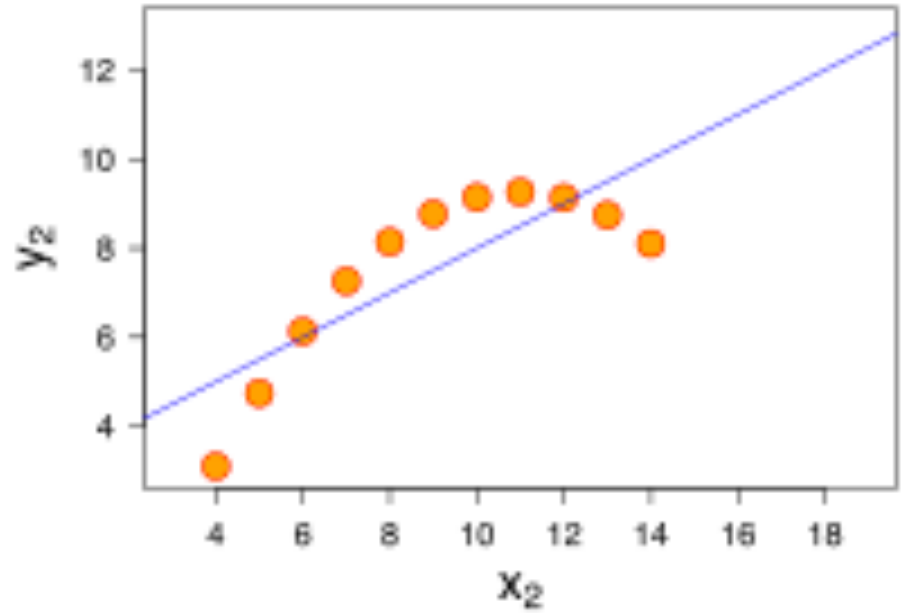
$$y = rx$$

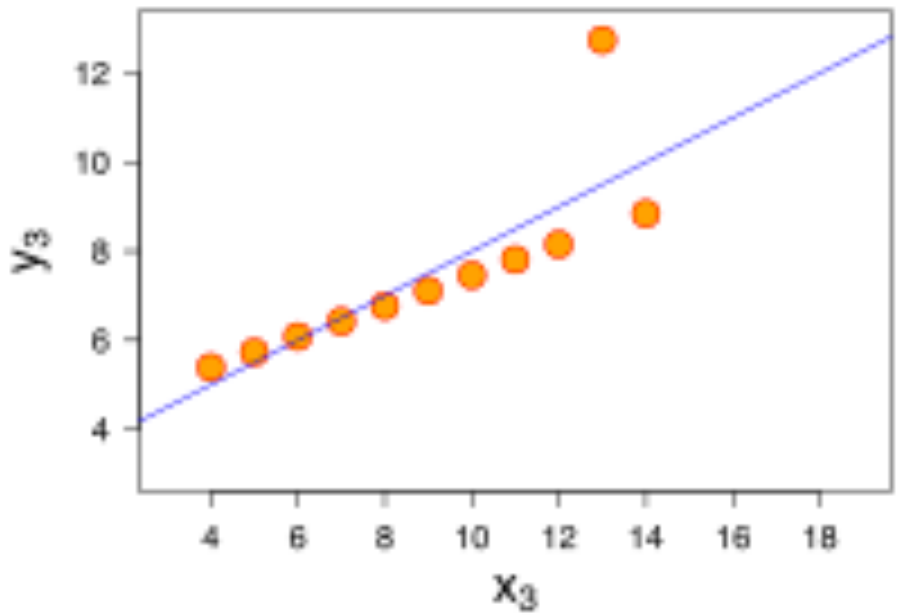


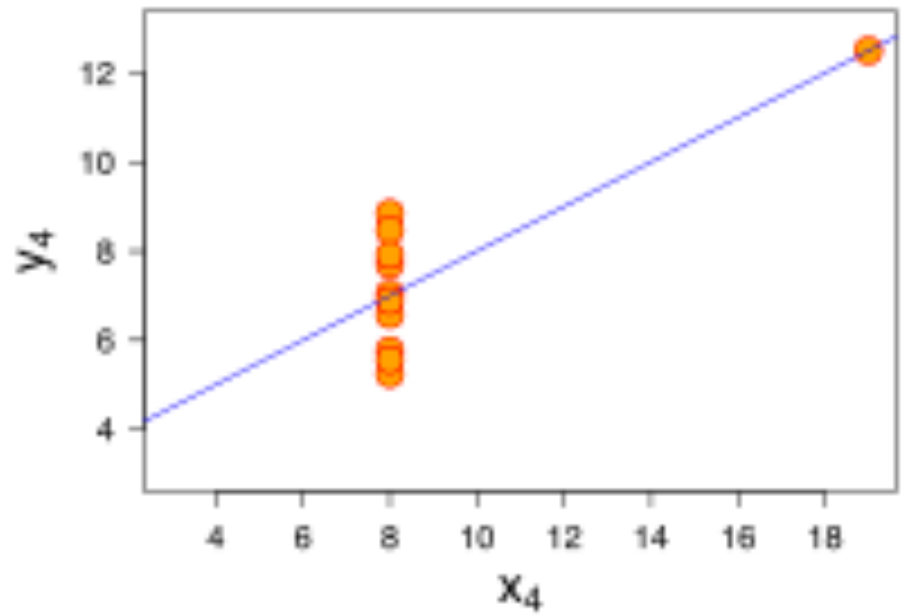
slope < 0 slope = 0 slope > 0

The **slope** of the regression line describes how much we expect *y* to change, on average, for every unit change in *x*.

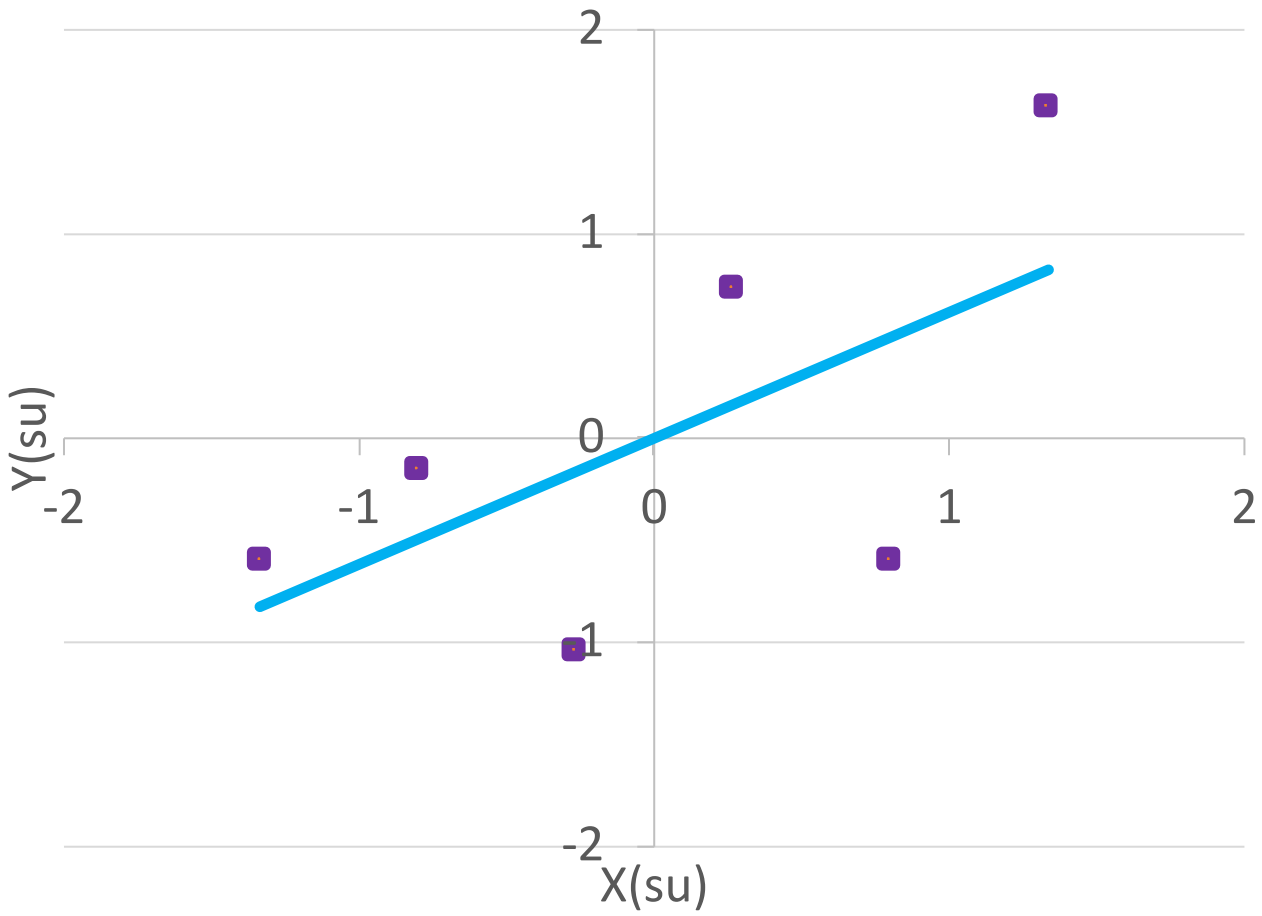

12
10
8
6
4
y_1
4 6 8 10 12 14 16 18
x_1


12
10
8
6
4
y_2
4 6 8 10 12 14 16 18
x_2


12
10
8
6
4
y_3
4 6 8 10 12 14 16 18
x_3


12
10
8
6
4
y_4
4 6 8 10 12 14 16 18
x_4

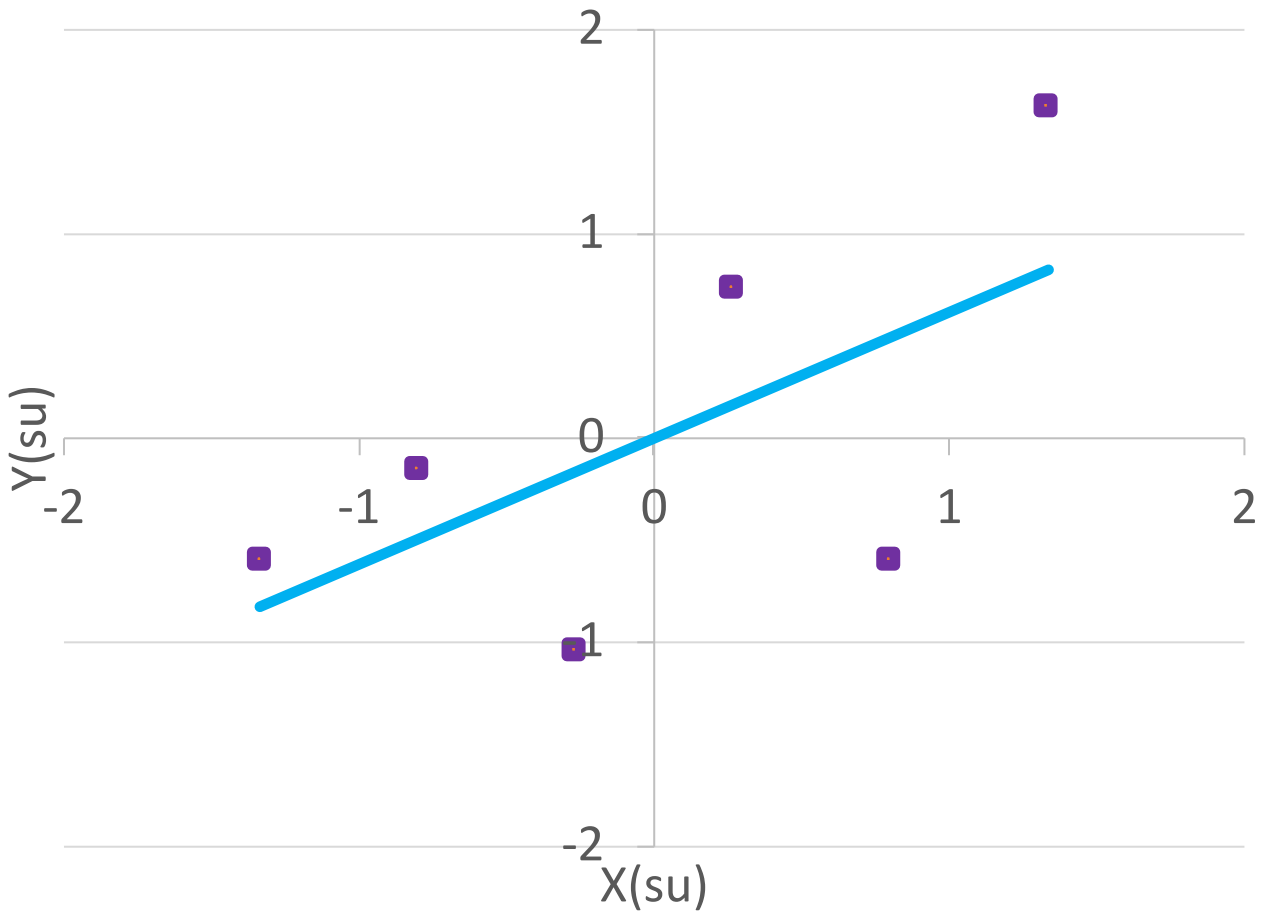
Regression Line Equation



$$y = rx$$

We can use this to predict new y-values

Regression Line Equation



$$y = rx$$

We can use this to predict new y-values

But they'd be in standard units

Regression Line Equation

To get in ***original*** units:

$$y = mx + \mathrm{b}$$

m (slope): $r \cdot \frac{\text{STD of } y}{\text{STD of } x}$

b (intercept): $\text{average of y} - \text{slope} \cdot \text{average of } x$

Regression Line Equation

To get in ***original*** units:

$$y = mx + \mathrm{b}$$

m (slope): $r \cdot \dfrac{\text{STD of } y}{\text{STD of } x}$

b (intercept): $\text{average of y} - \text{slope} \cdot \text{average of } x$

So now you can make your twizzler length estimates in centimeters!