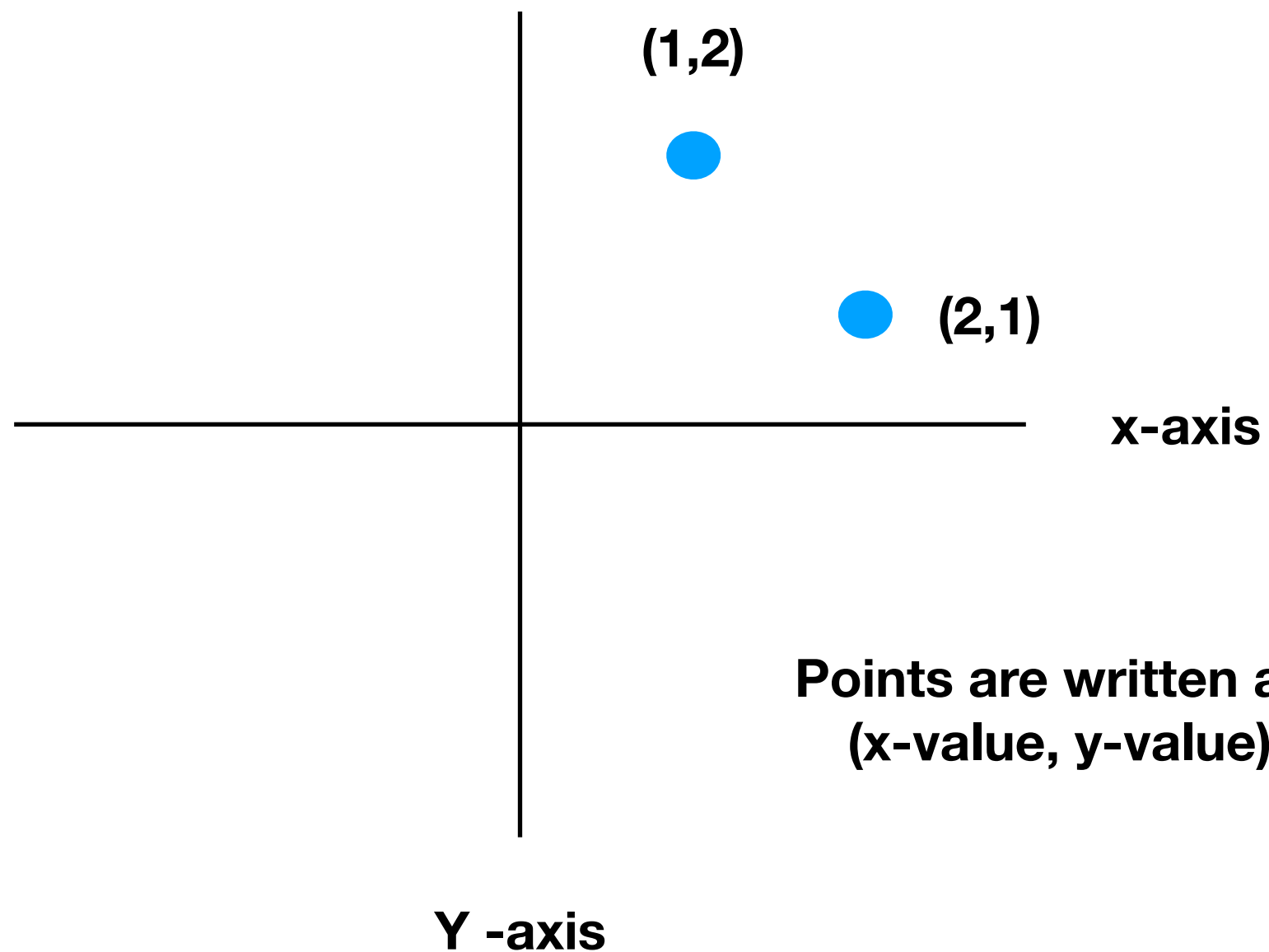


# Correlation and Scatter Plots

how to describe the relationship between two  
variables

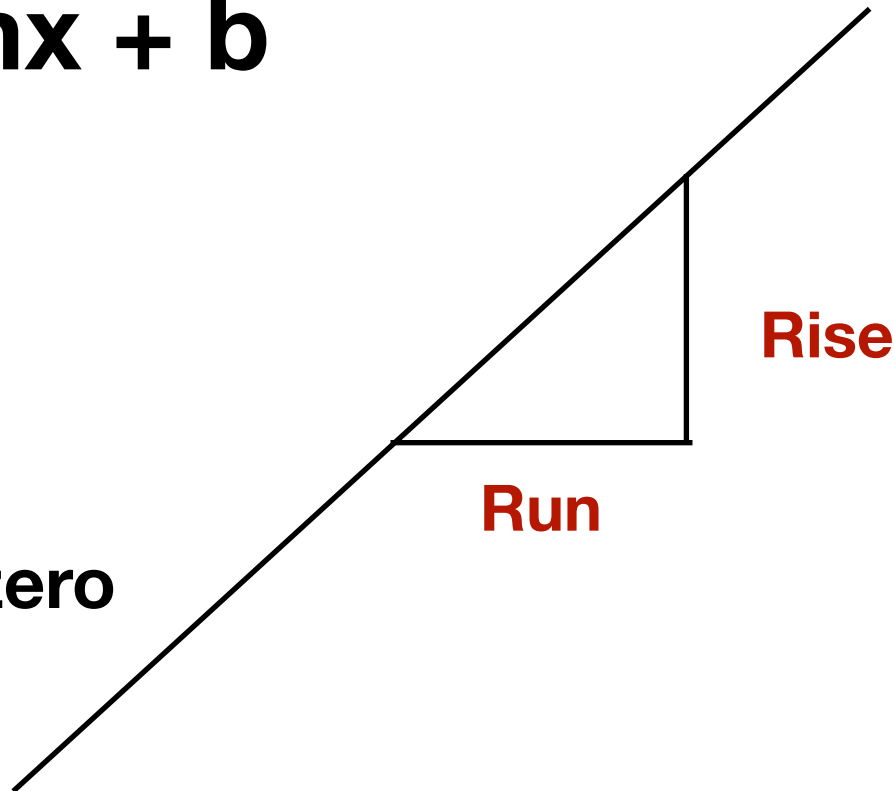
# Coordinate system and Points



# Remember a line

$$y = mx + b$$

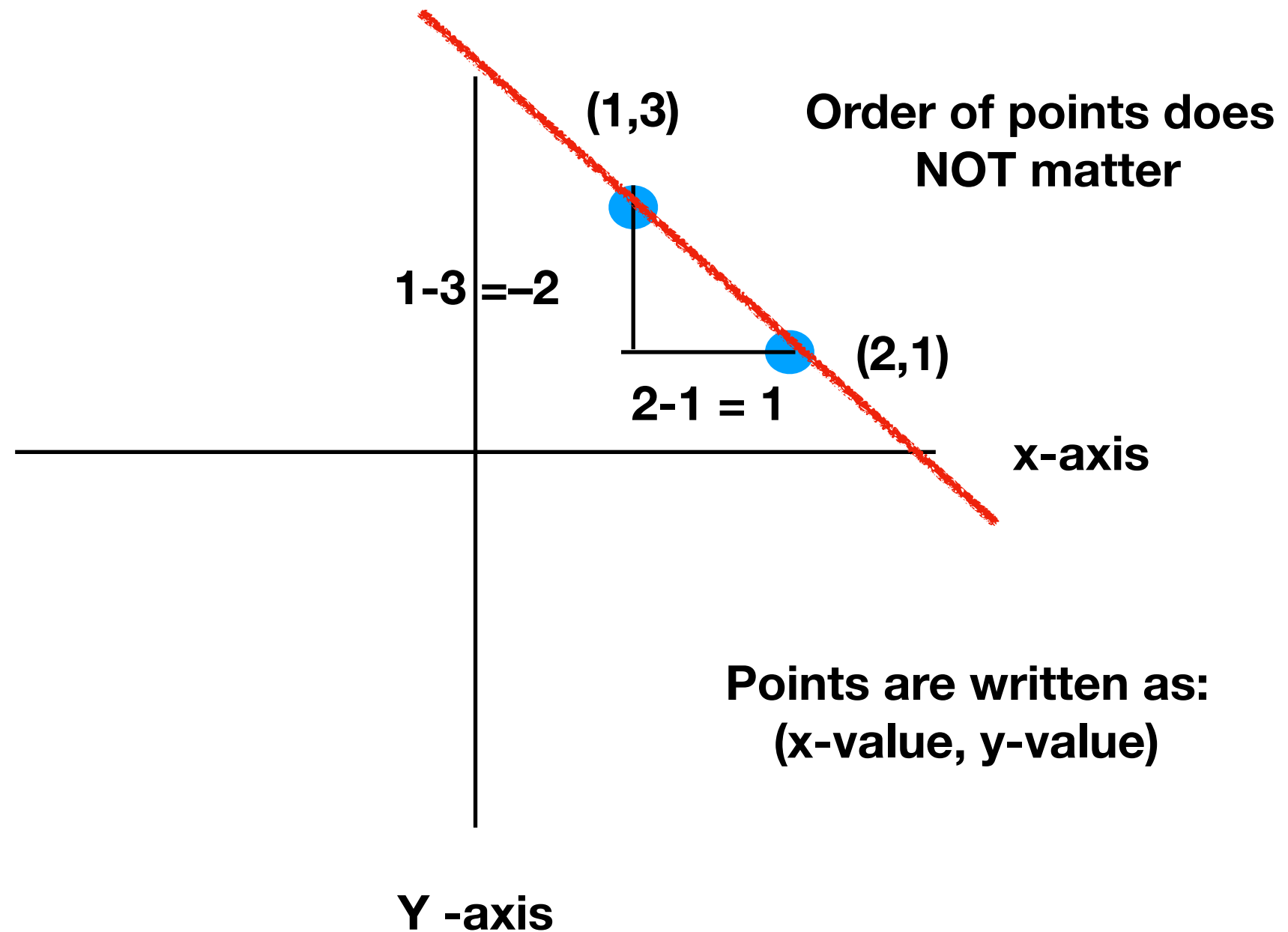
**b** -> intercept  
or value of y when x is zero



**m** -> is the slope of the line  
or “rise over run”

# Finding Formula of line

Find the slope:  $-2/1$



# Finding Formula of line

Find the slope:  $-2/1$

What about the intercept?

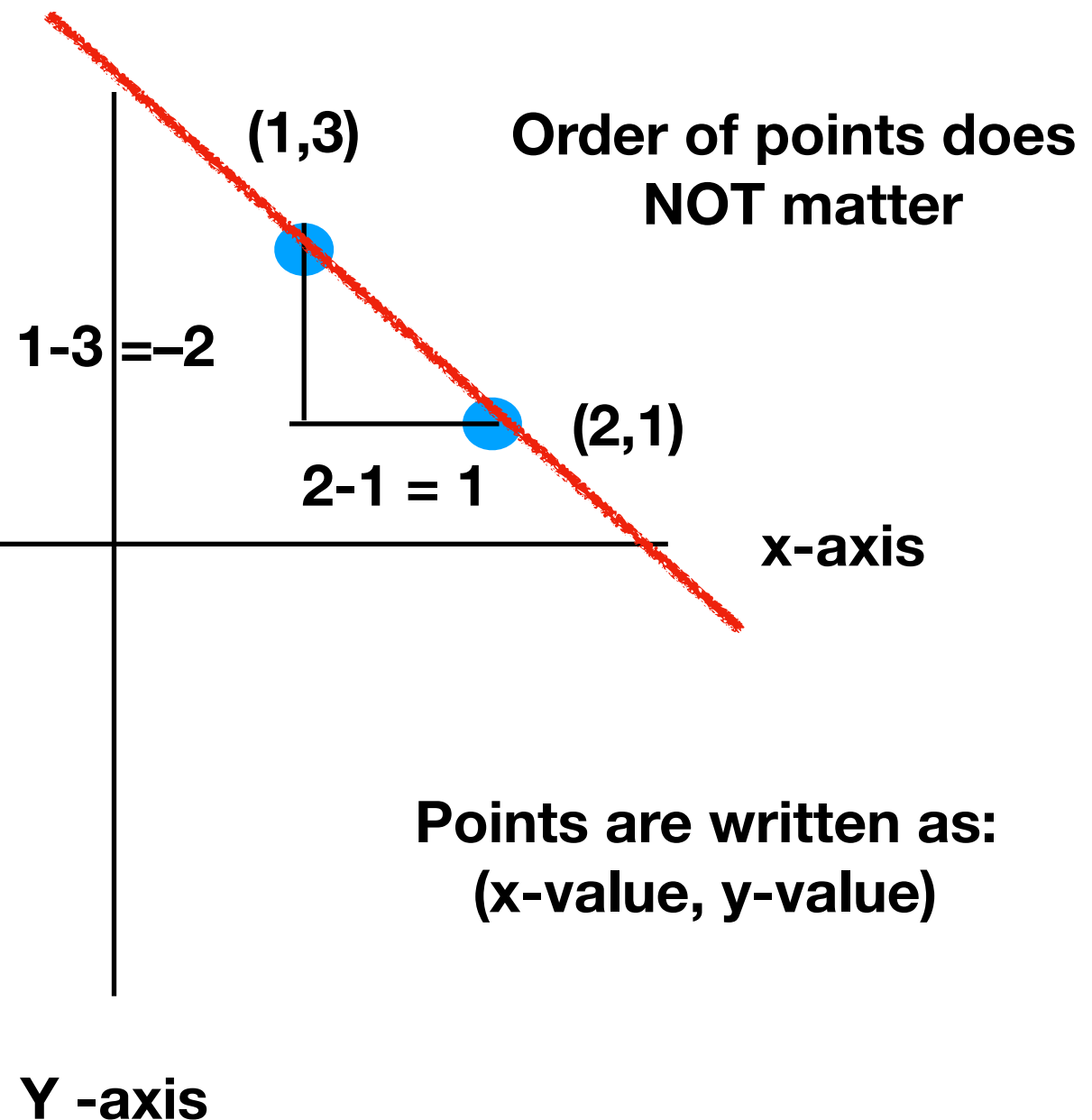
Choose any point on the line:

$$Y - 3 = m(x - 1)$$

$$Y - 3 = -2(x - 1)$$

$$Y - 3 = -2x + 1$$

$$Y = -2x + 4$$



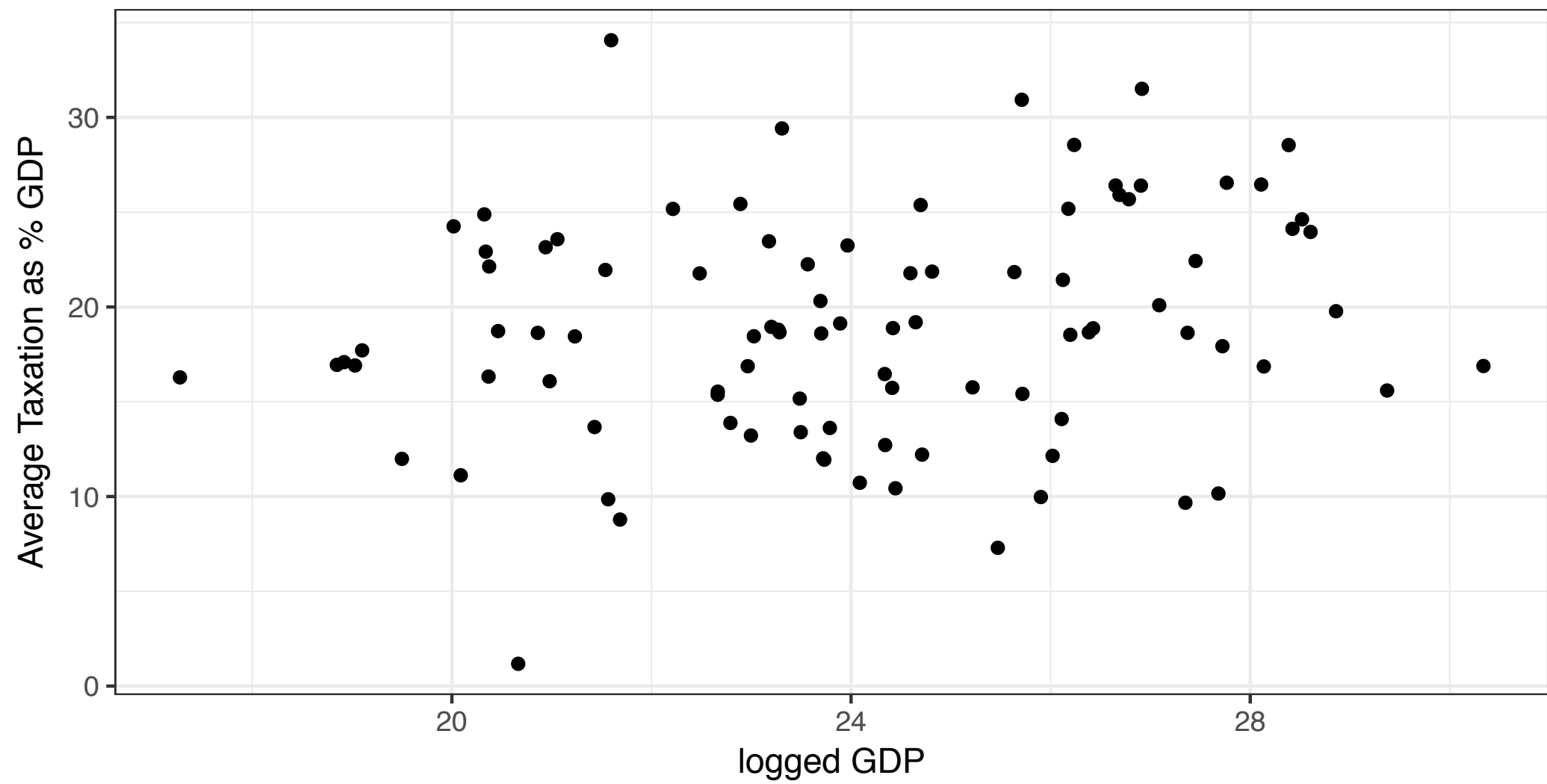
# Scatter plots

- Often we are interested in the relationship between two variables
  - A. Dependent variable:  $Y$
  - B. Independent variable:  $X$
- Each observation has a  $y$ -value and  $x$ -value

# Scatter plots

- We can plot each observation as a point on the coordinate system
- Each plot is drawn as if their y and x values are coordinates
- Scatter the points across the plot

Tax Revenue vs logged GDP





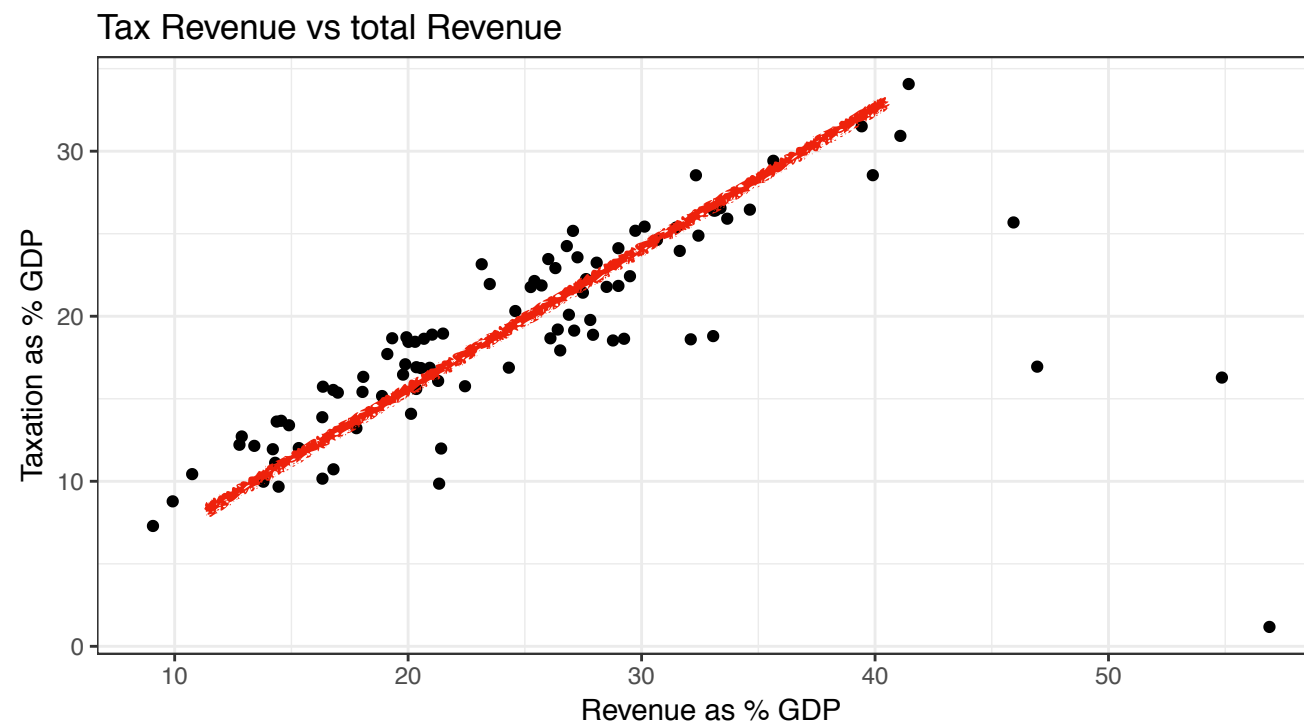
# Correlation

- Correlation is a standardized measure of the co-relationship between two variables
- Or, correlation is a standardized measure of the co-variation between two variables
- I.e how much do two variables move together?

# Correlation

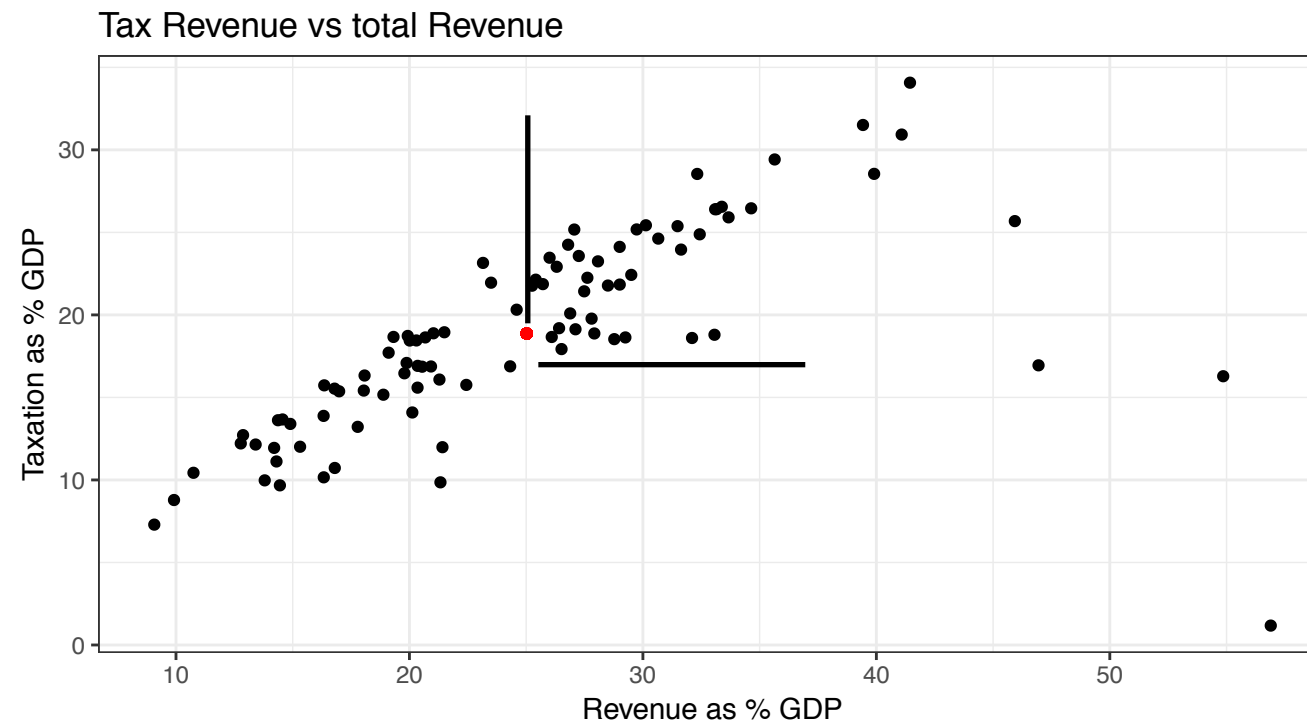
If there is a strong correlation (association) between two variables then knowing one will help predicting the other. If the correlation (association) is weak, then info about one helps little to predict the other.

**Strong correlation looks like a very tight cloud of points that you could easily draw a line through**



**Remember, usually we call the variable we are interested in explaining or predicting the dependent variable (Y)**

**The spread of points in any direction is approximated by mean  $\pm$  2\*SD**  
**Why?**



**y-mean = 18.87**

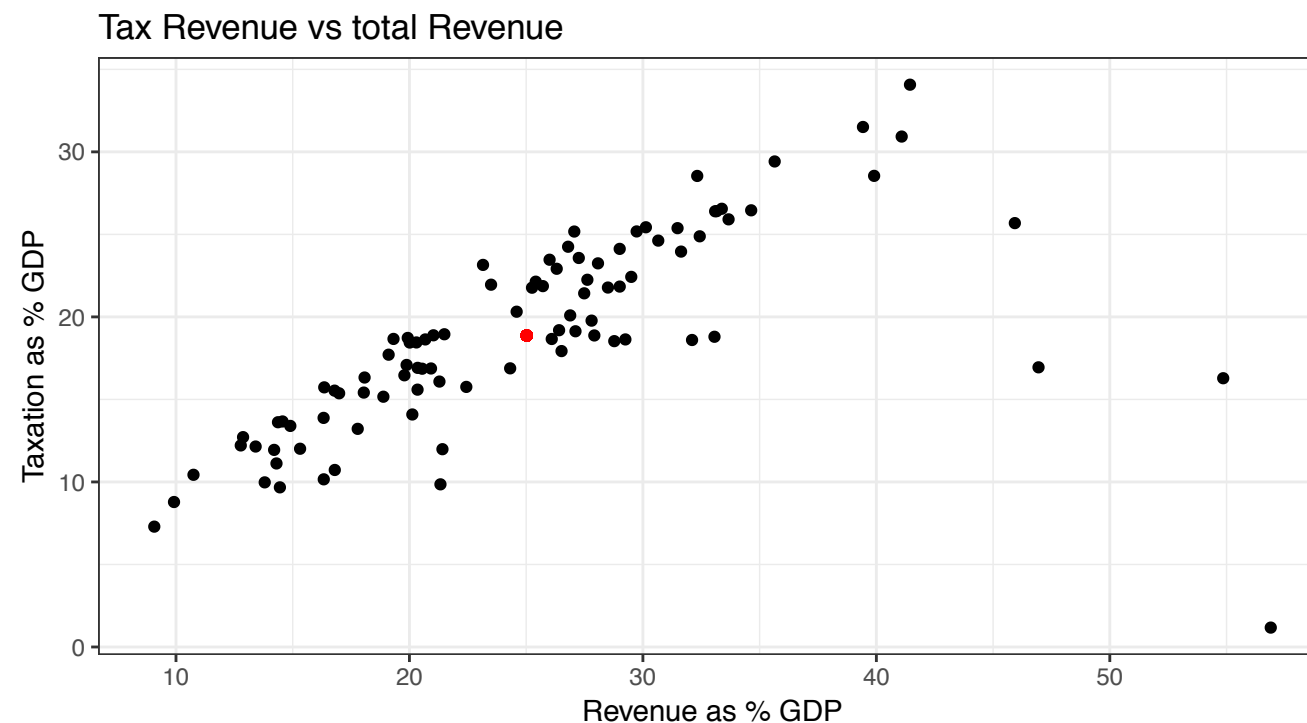
**y-sd = 5.92**

**x-mean = 25.02**

**x-sd = 9.13**

**x ranges approximately from 7 to 43,  
y from 6 to 30**

**The correlation is stronger the less spread out and steeper the imaginary cloud is**



**y-mean = 18.87**

**y-sd = 5.92**

**x-mean = 25.02**

**x-sd = 9.13**

**x ranges approximately from 7 to 43,  
y from 6 to 30**

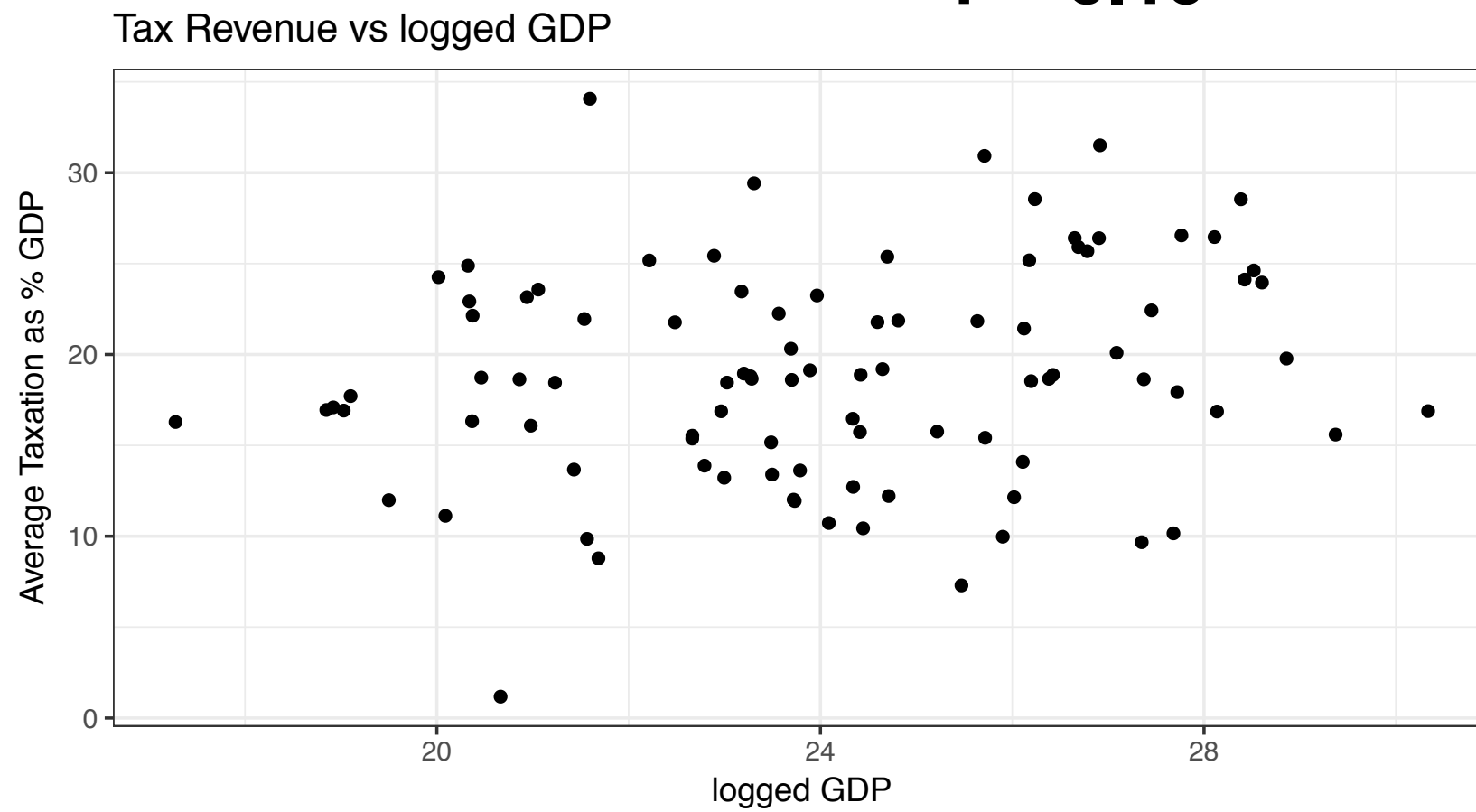
# Correlation Coefficient $r$

- Ranges from -1 to 1
- -1 means strongest possible negative correlation
- 1 means strongest possible positive correlation
- 0 means no relationship

# Correlation Coefficient $r$

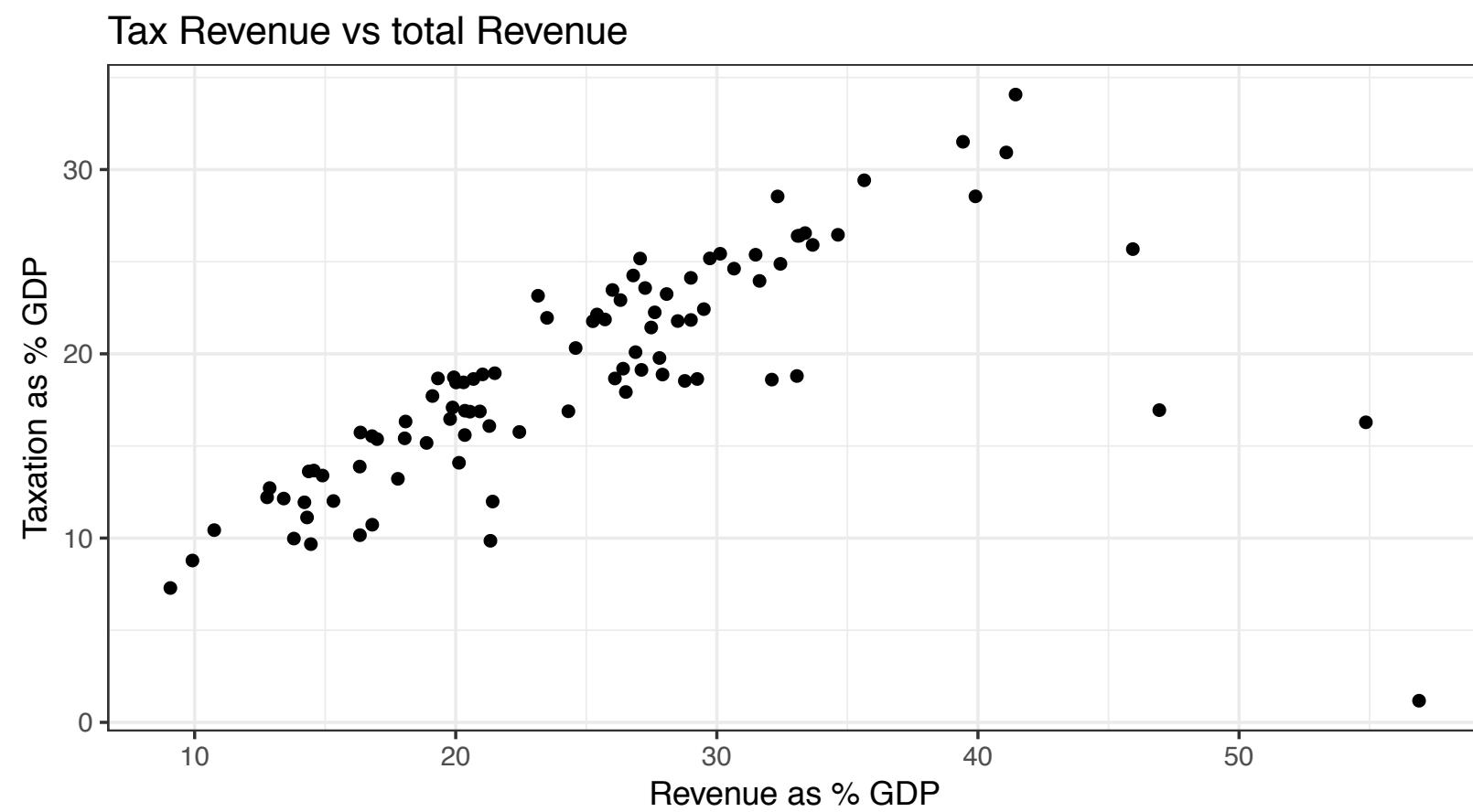
- Negative value means slope is downwards
- Positive value means slope is upwards
- 0 means flat

**$r = 0.19$**

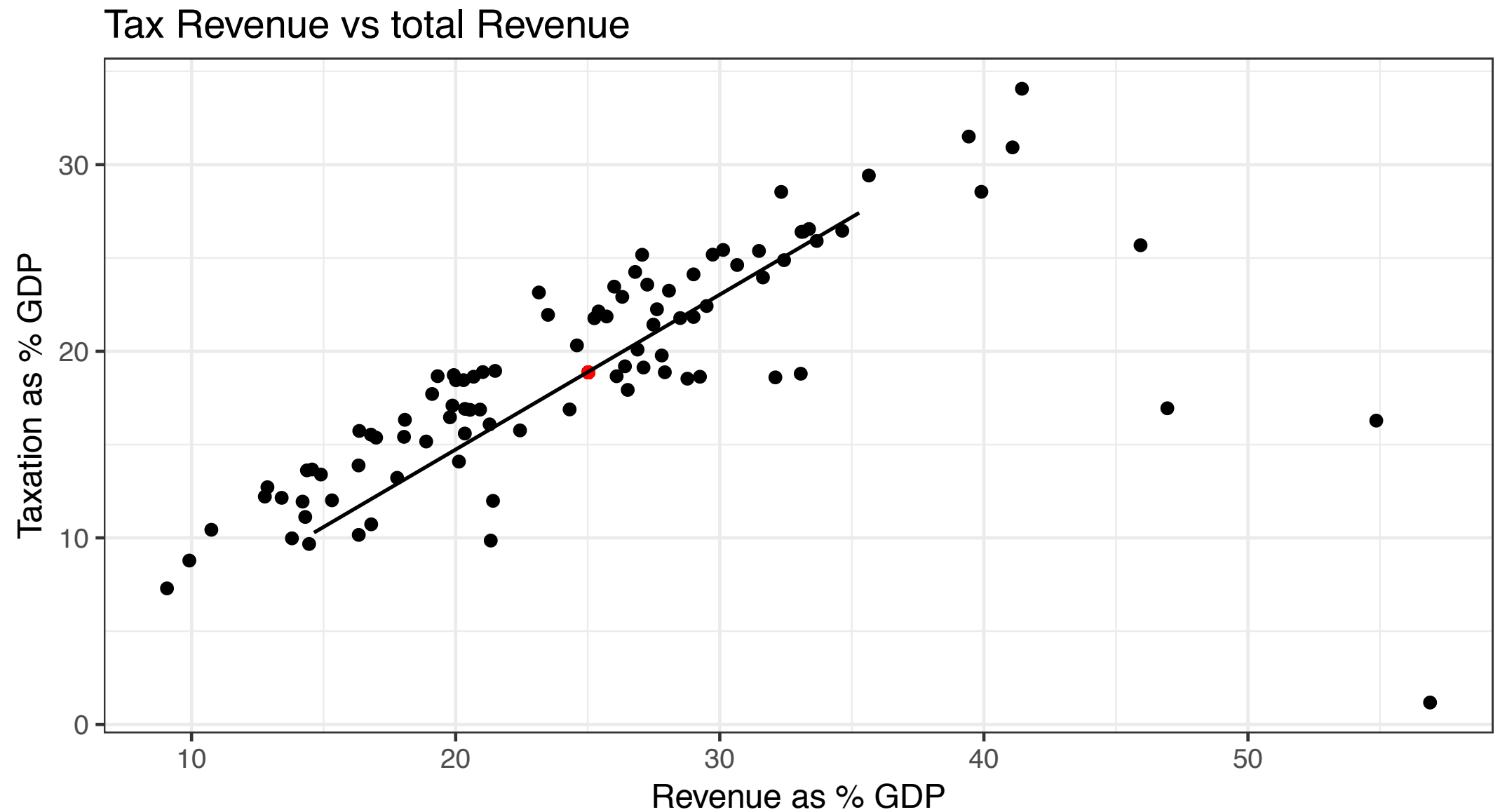




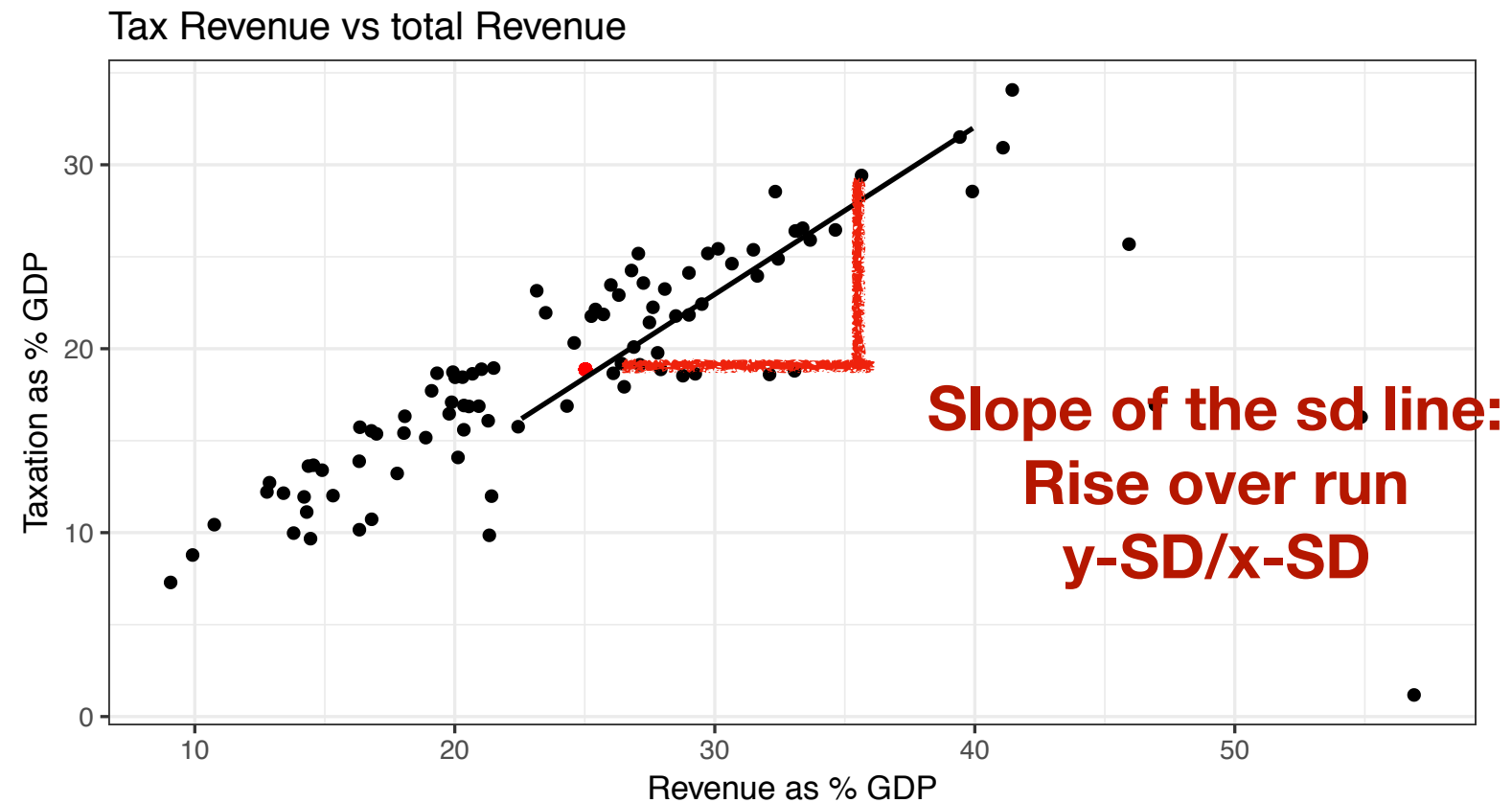
**$r = 0.57$**



**SD line goes through the point of averages and  
all points an equal number of SDs away from the average**



**SD line goes through the point of averages and  
all points an equal number of SDs away from the average**



# Calculating the correlation coefficient

- $r = \text{mean of } (x \text{ in standard units} * y \text{ in standard units})$ 
  - For each  $x$ : subtract mean from value and divide by SD
  - For each  $y$ : subtract mean from value and divide by SD
  - Multiply standard values of both and then divide by  $N$  (number of observations)

# Let's do an example

- $Y = 1, 3, 4, 5, 7$
- $X = 5, 9, 7, 1, 13$

# Let's do an example

- $X = 1, 2, 3, 4, 5, 6, 7$
- $Y = 2, 1, 4, 3, 7, 5, 6$