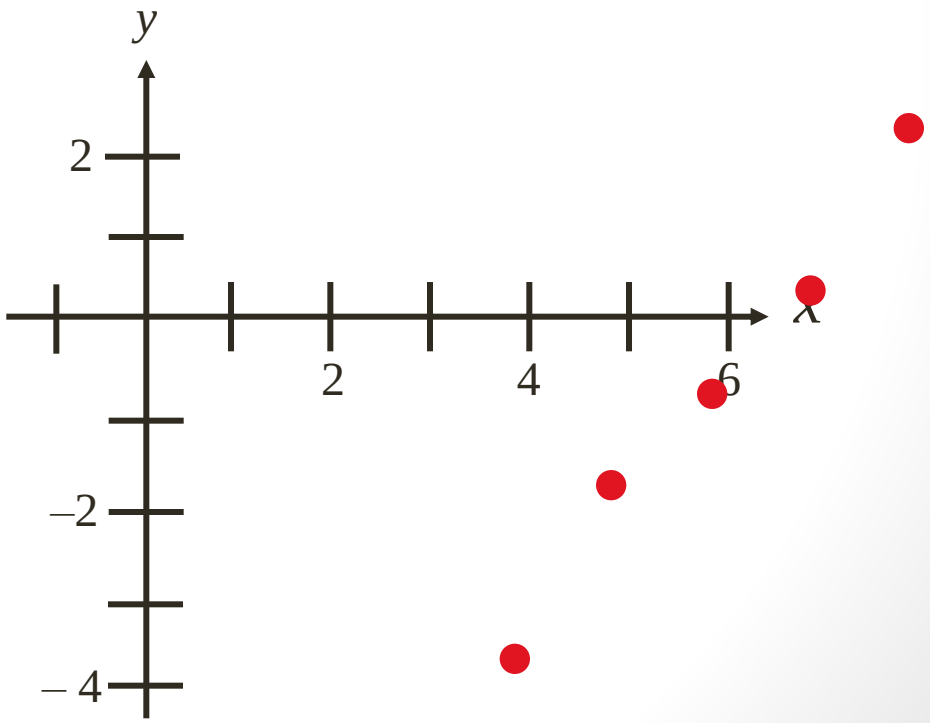
Correlation and Regression

Correlation

A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y) where x is the **independent** (or **explanatory**) **variable**, and y is the **dependent** (or **response**) **variable**.

Example:

x	1	2	3	4	5
y	-4	-2	-1	0	2



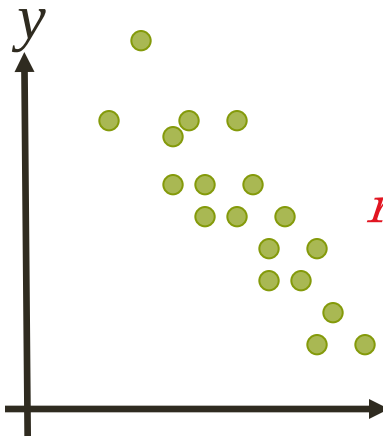
Correlation Coefficient

The **correlation coefficient** is a measure of the strength and the direction of a linear relationship between two variables. The symbol r represents the sample correlation coefficient. The formula for r is

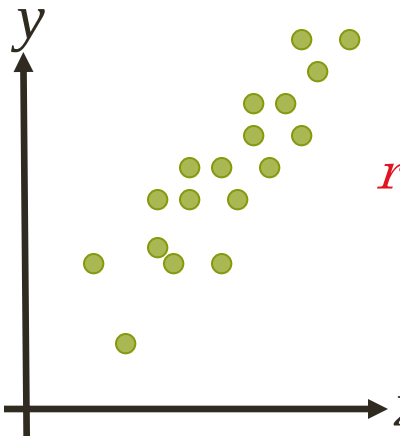
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

The range of the correlation coefficient is –1 to 1. If x and y have a strong positive linear correlation, r is close to 1. If x and y have a strong negative linear correlation, r is close to –1. If there is no linear correlation or a weak linear correlation, r is close to 0.

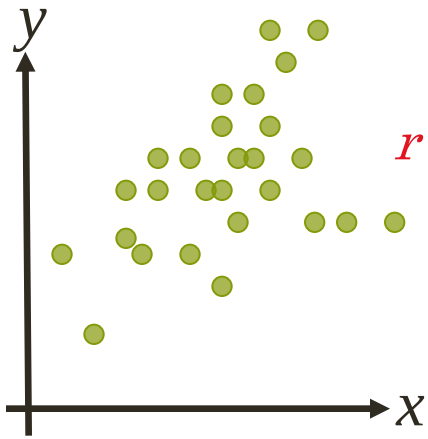
Linear Correlation



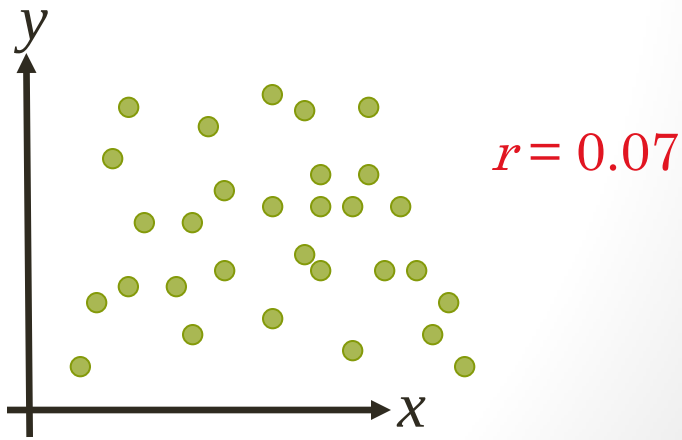
Strong negative correlation



Strong positive correlation



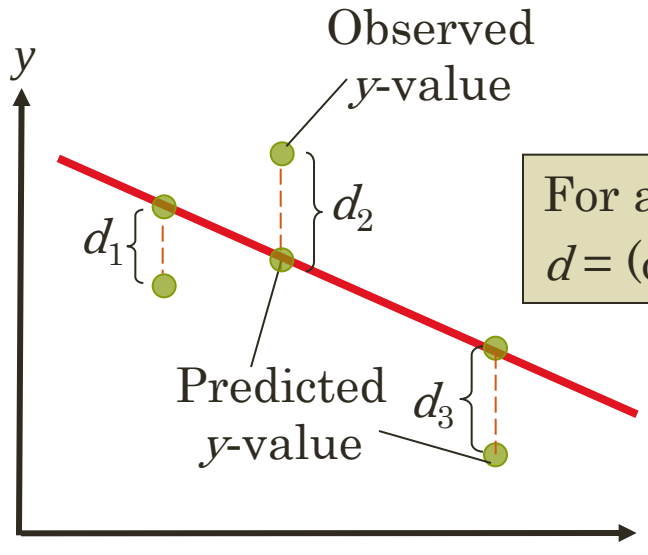
Weak positive correlation



Nonlinear Correlation

Residuals

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of y for a given value of x.



For a given x-value,
d = (observed y-value) – (predicted y-value)

Each data point d_i represents the difference between the observed y-value and the predicted y-value for a given x-value on the line. These differences are called **residuals**.

Regression equation

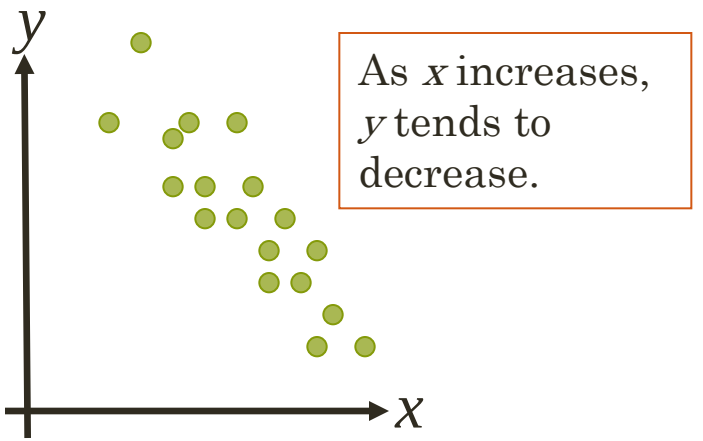
Example continued:

Using the equation $\hat{y} = -4.07x + 93.97$, we can predict the test score for a student who watches 9 hours of TV.

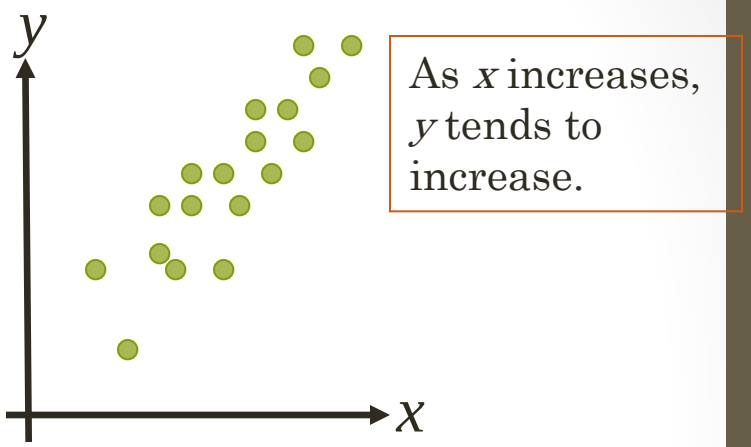
$$\begin{aligned}\hat{y} &= -4.07x + 93.97 \\ &= -4.07(9) + 93.97 \\ &= 57.34\end{aligned}$$

A student who watches 9 hours of TV over the weekend can expect to receive about a 57.34 on Monday's test.

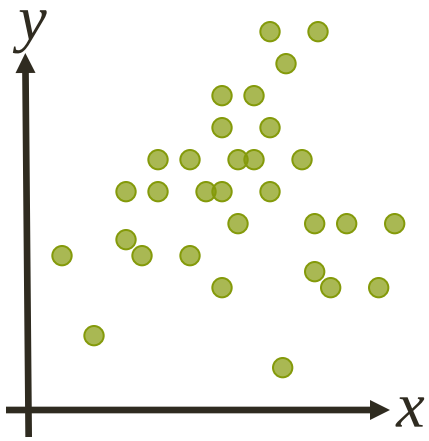
Linear Correlation



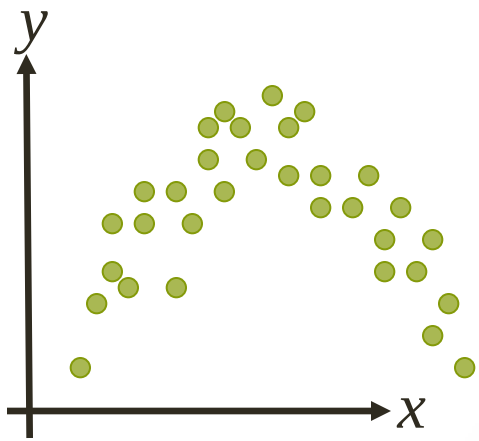
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation