Simple Linear Regression and Correlation

Dependent vs Independent Variables

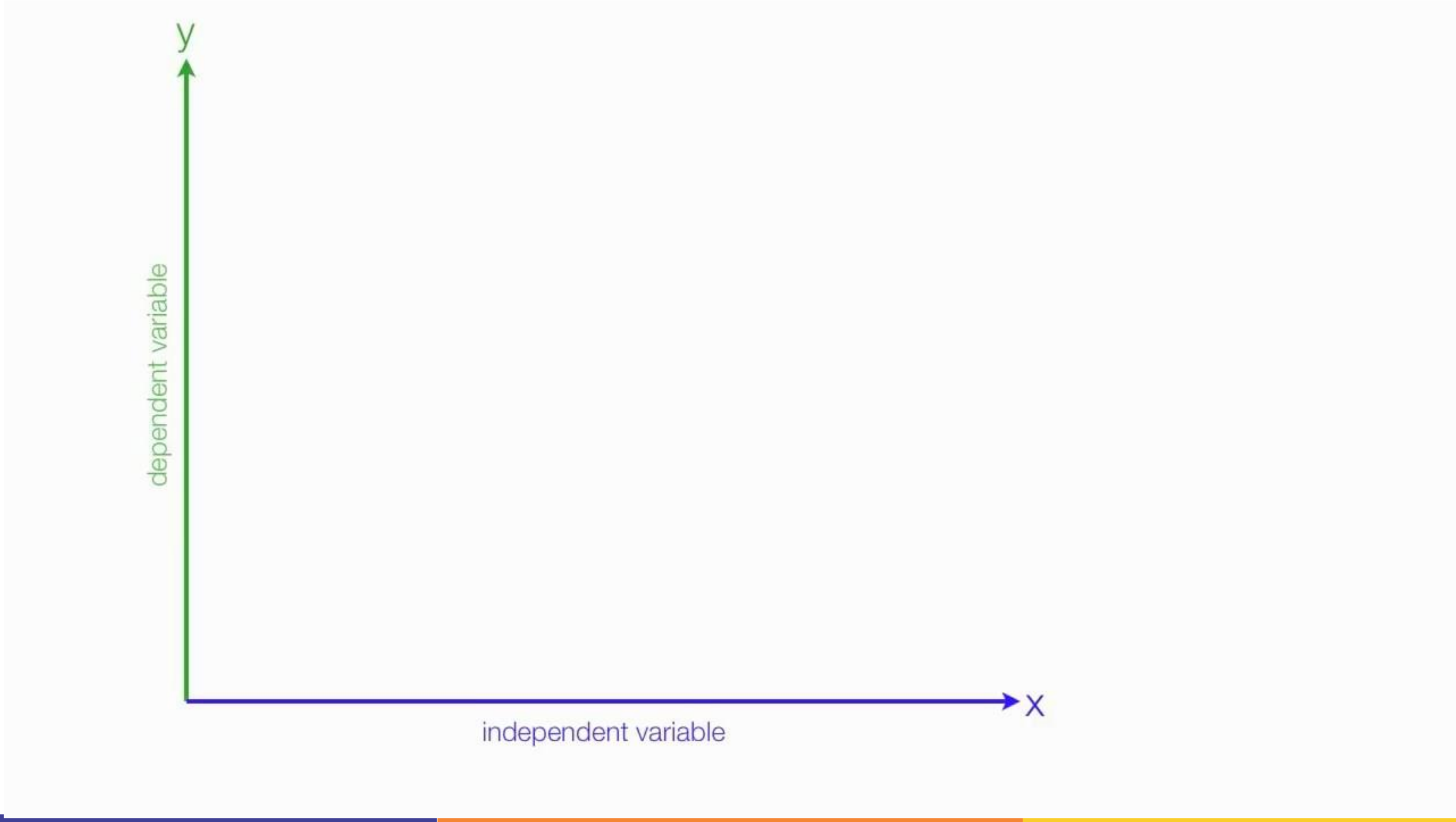
- **Independent Variable**
  - The value does not change due to the impact of any other variable. The researcher manipulates or changes the independent variable to measure its impact on other variables.
- **Dependent Variable**
  - It depends on other variables.
  - It is the variable that is being tested in the experiment.
  - A researcher measures the outcome of the experiment to see how other variables cause changes in the value of a dependent variable.

Dependent vs Independent Variables

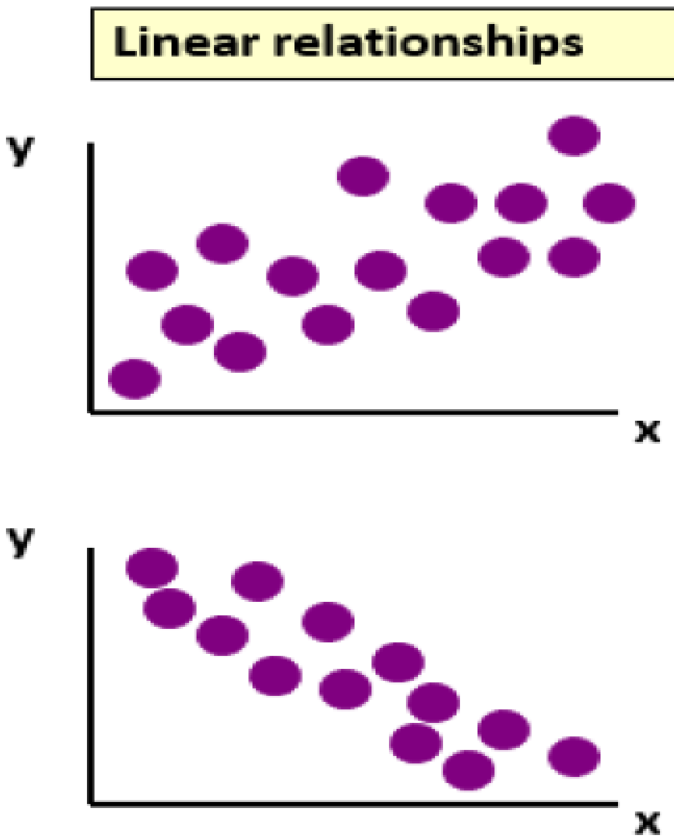
- **Examples:**
- How does the amount of sleep impact test scores?
  - Independent Variable: Time spent on sleeping before the exam
  - Dependent Variable: Test Score
- What is the effect of fast food on blood pressure?
  - Independent Variable: Consumption of fast food
  - Dependent Variable: Blood Pressure
- What is the effect of caffeine on sleep?
  - Independent Variable: the amount of caffeine consumed
  - Dependent Variable: Sleep

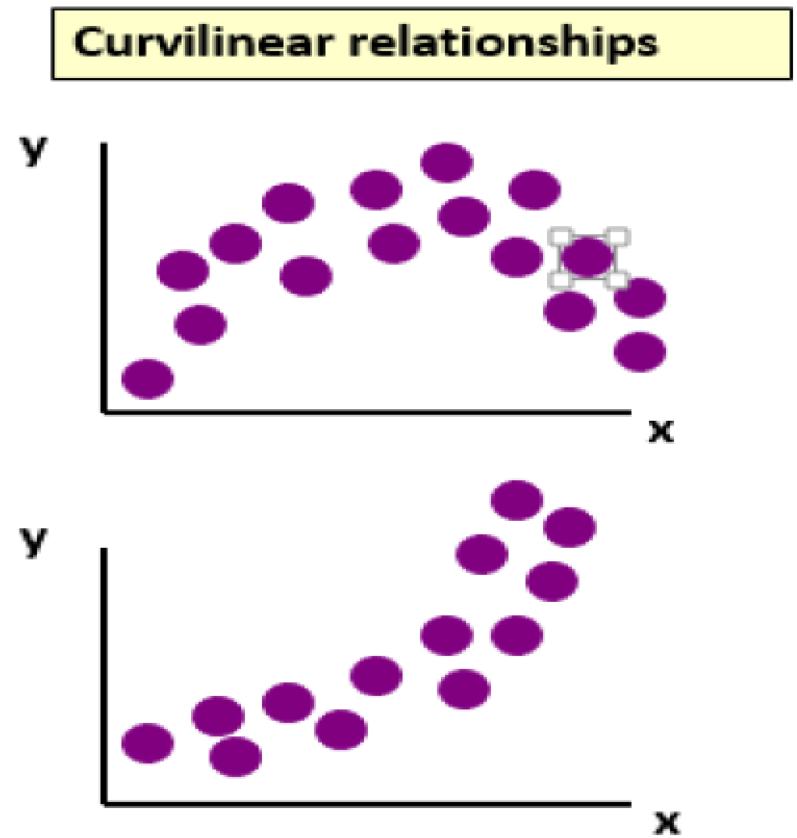
Dependent vs Independent Variables

- We mark x-axis as independent variable and y-axis as dependent variable.

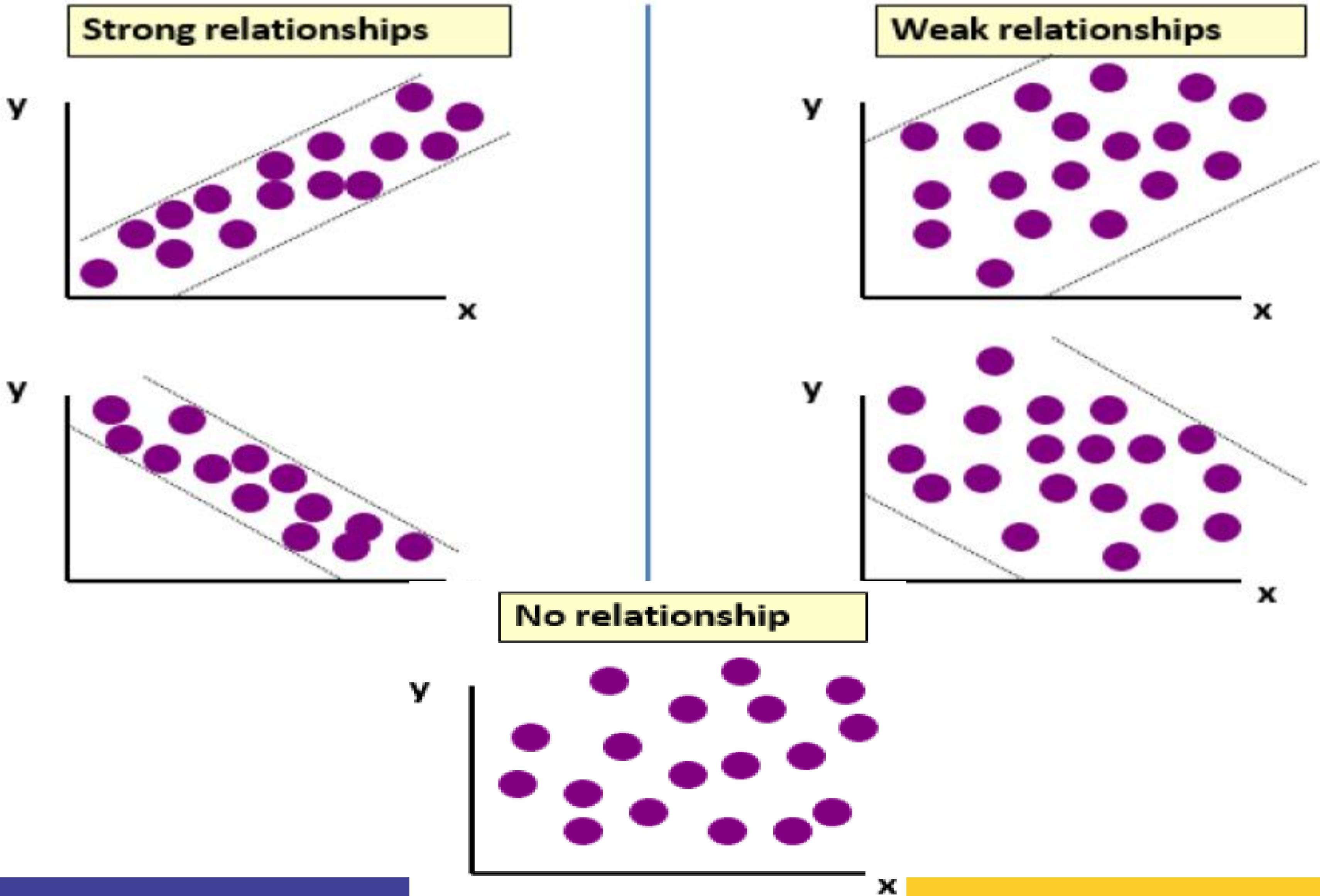


Scatter Plots and Correlation





Scatter Plots and Correlation



Correlation Coefficient

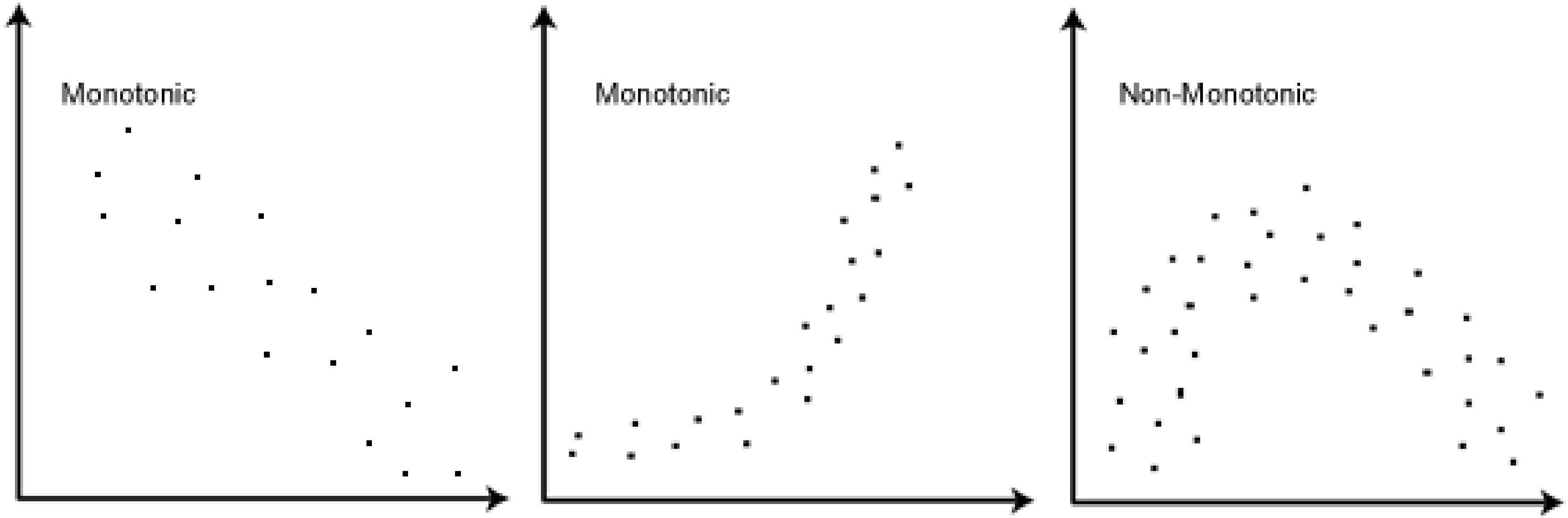
- Correlation coefficients are used to measure how strong a relationship is between two variables.
- There are several types of correlation coefficient, but the most popular are:
  - Pearson Correlation Coefficient (r)
  - Spearman Correlation Coefficient ρ (rho)

Comparison of Pearson and Spearman coefficients

- Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well.
- Pearson works with raw data values of the variables whereas Spearman works with rank-ordered variables.
- The Spearman's rank-order correlation is the nonparametric version of the Pearson.

Monotonic relationship

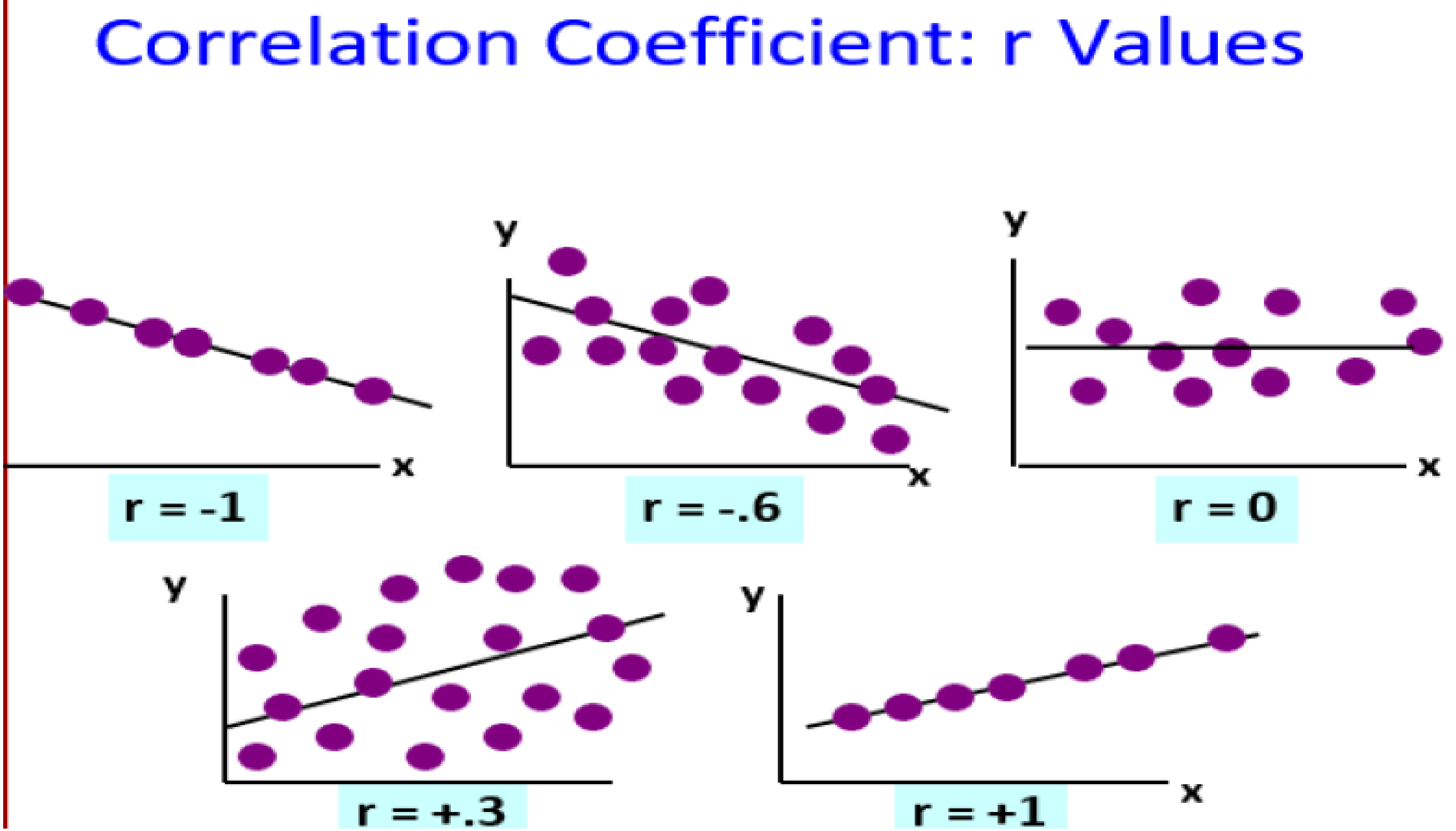
- A monotonic relationship is a relationship that does one of the following:
  - as the value of one variable increases, so does the value of the other variable
  - as the value of one variable increases, the other variable value decreases.



Features of ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Features of ρ and r

- 

Regression

- In this we try to find out the relationship between two variables and form a straight line on the scattered plot called regression line.
- We try to fit this regression line to all the observations.
- Regression line is based upon least squared method.

Estimated Regression Model

The sample regression line provides an estimate of the population regression line

Estimated (or predicted) y value

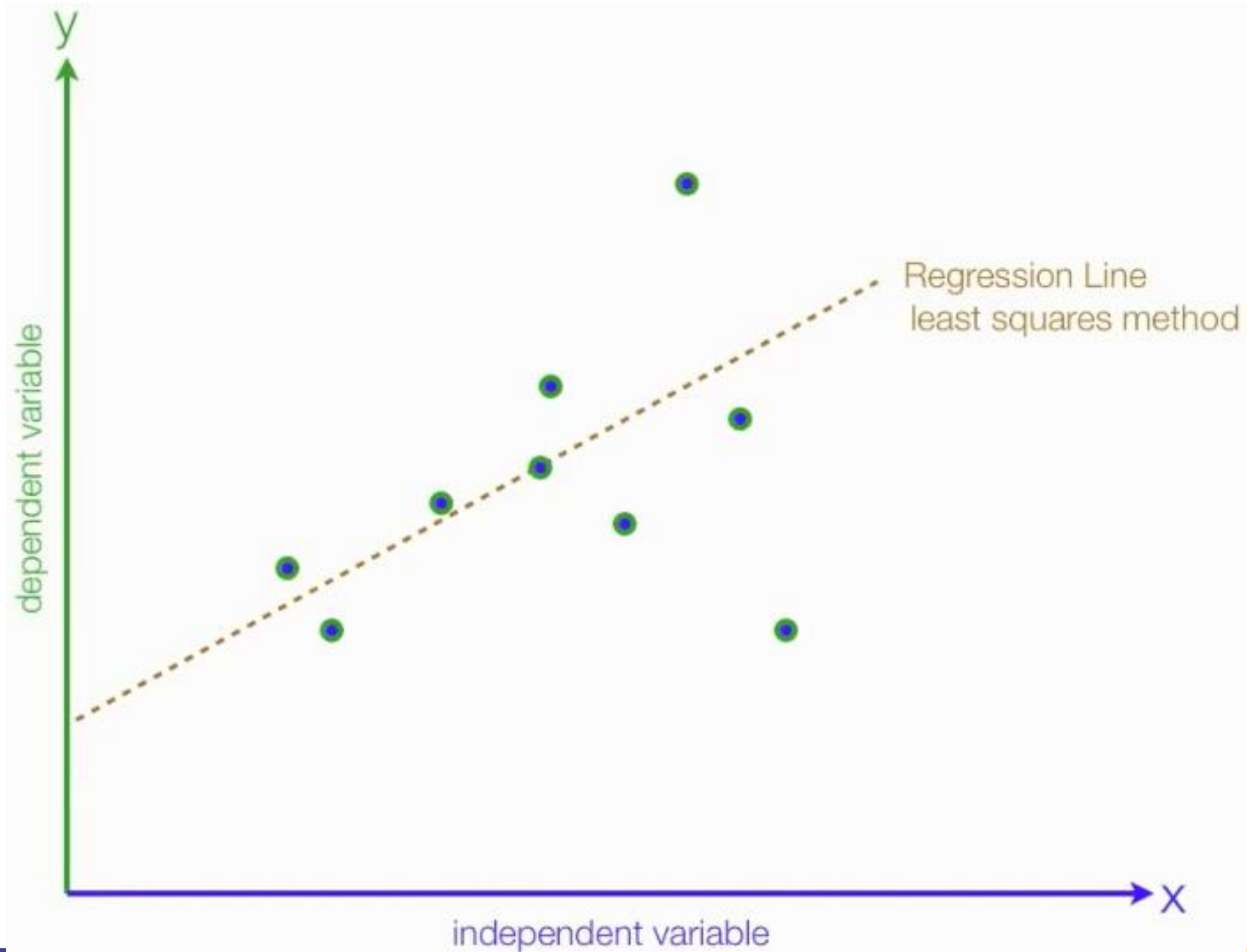
Estimate of the regression intercept

Estimate of the regression slope

Independent variable

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

Regression


y
Regression Line
least squares method
dependent variable
x
independent variable

Least Square Equation

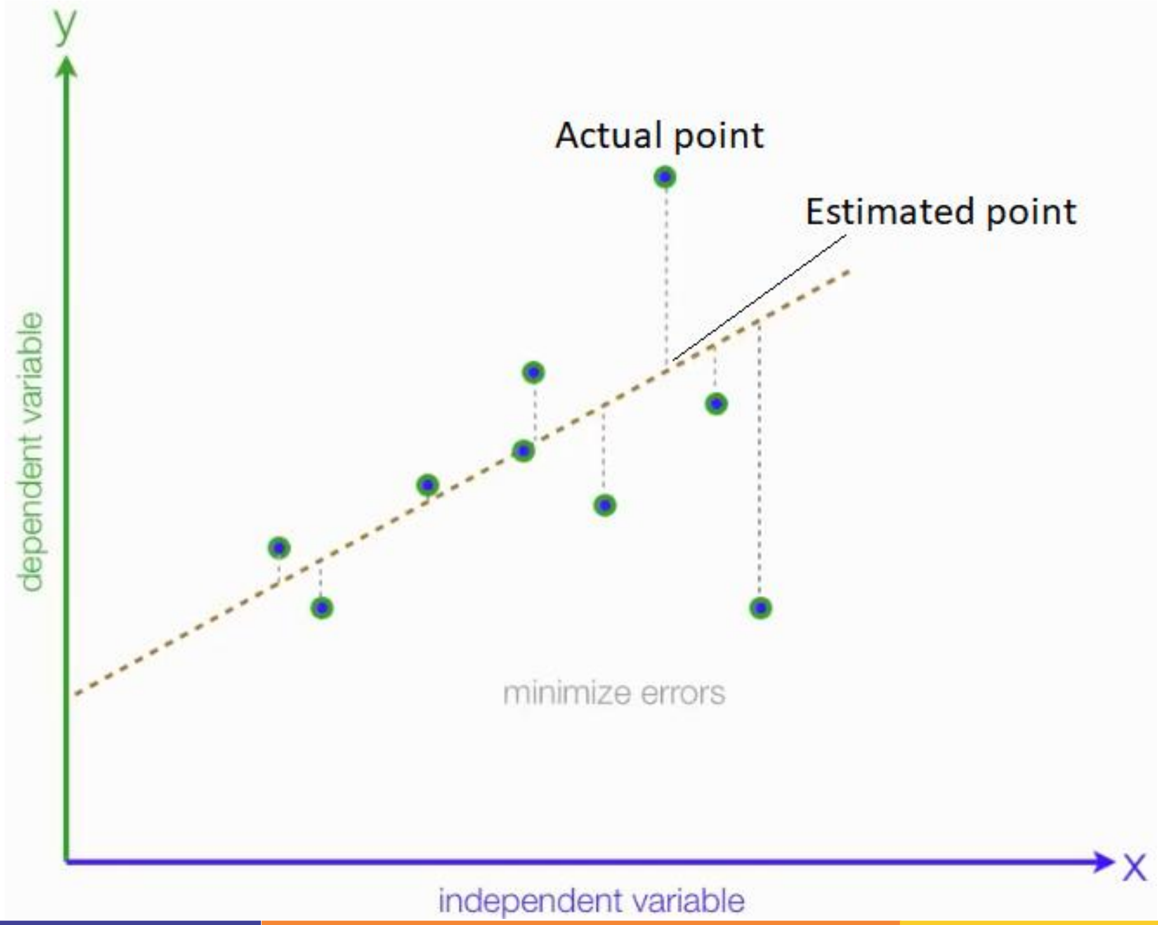
- $\hat{y} = b_0 + b_1 x$

- $b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ $\qquad b_0 = \bar{y} - b_1\bar{x}$

- $\mathbf{b_0}$ is the estimated average value of y when the value of x is zero.
- $\mathbf{b_1}$ is the estimated change in the average value of y as a result of a one-unit change in x.

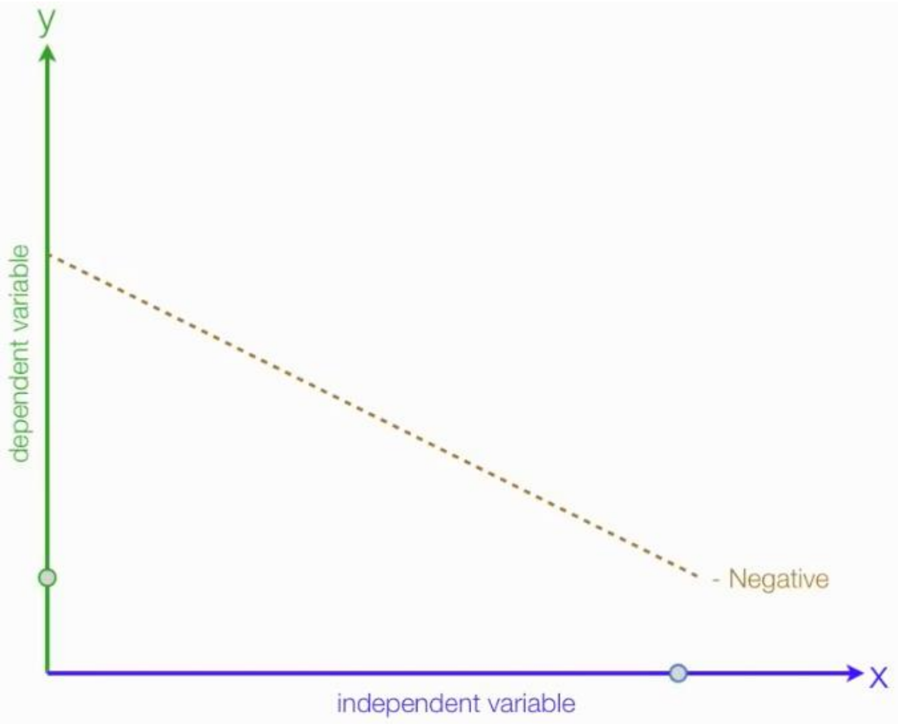
Regression

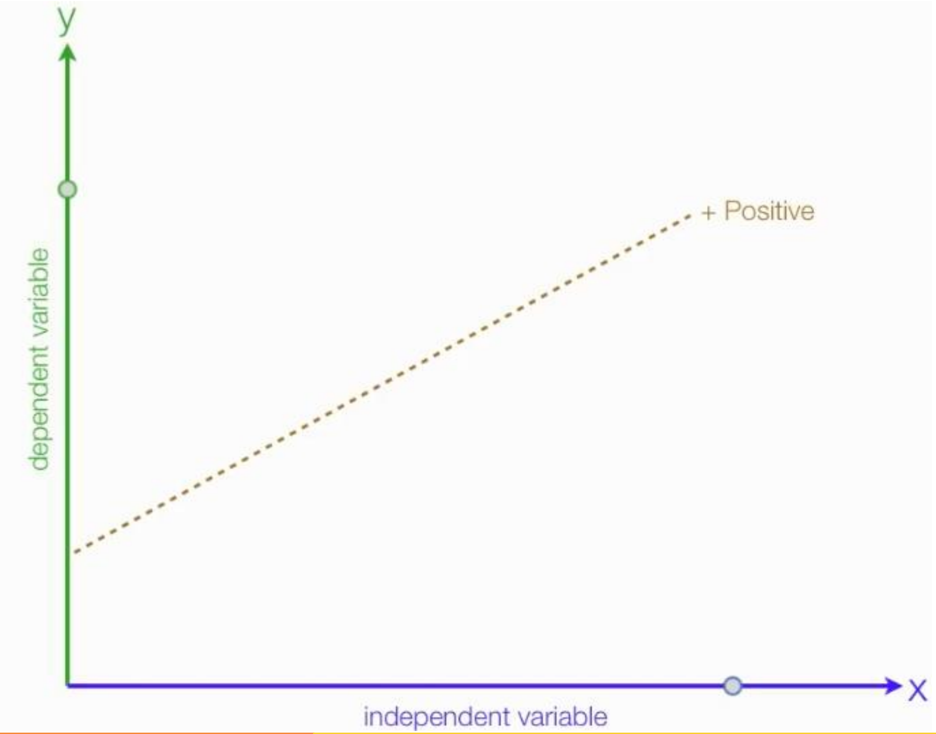
- We try to minimize the errors produced due to the difference between actual and estimated data points.



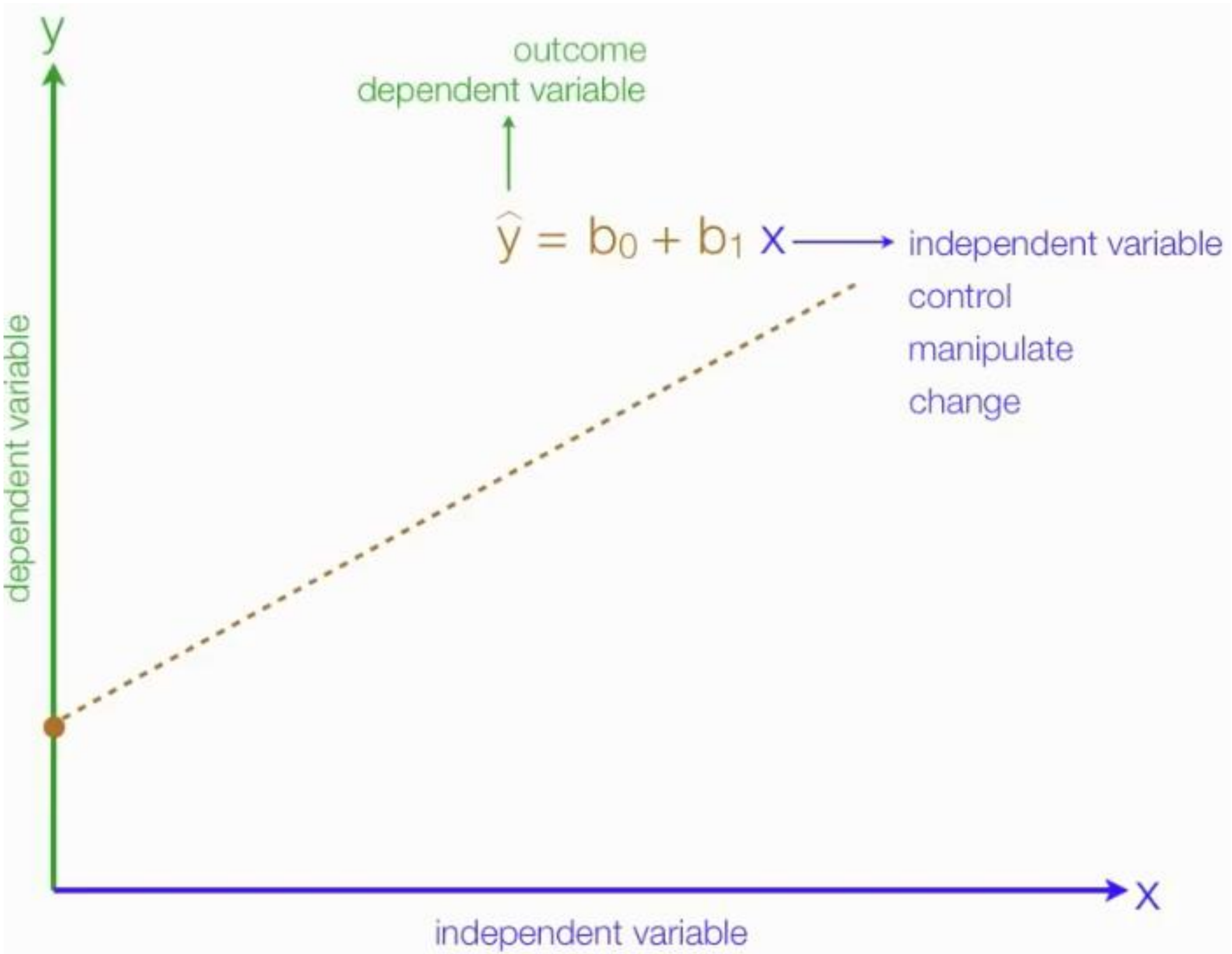
Regression

- We examine the relationship between variables i.e. when one independent variable is changing what is its effect on dependent variable?
  - Positive relationship
  - Negative relationship





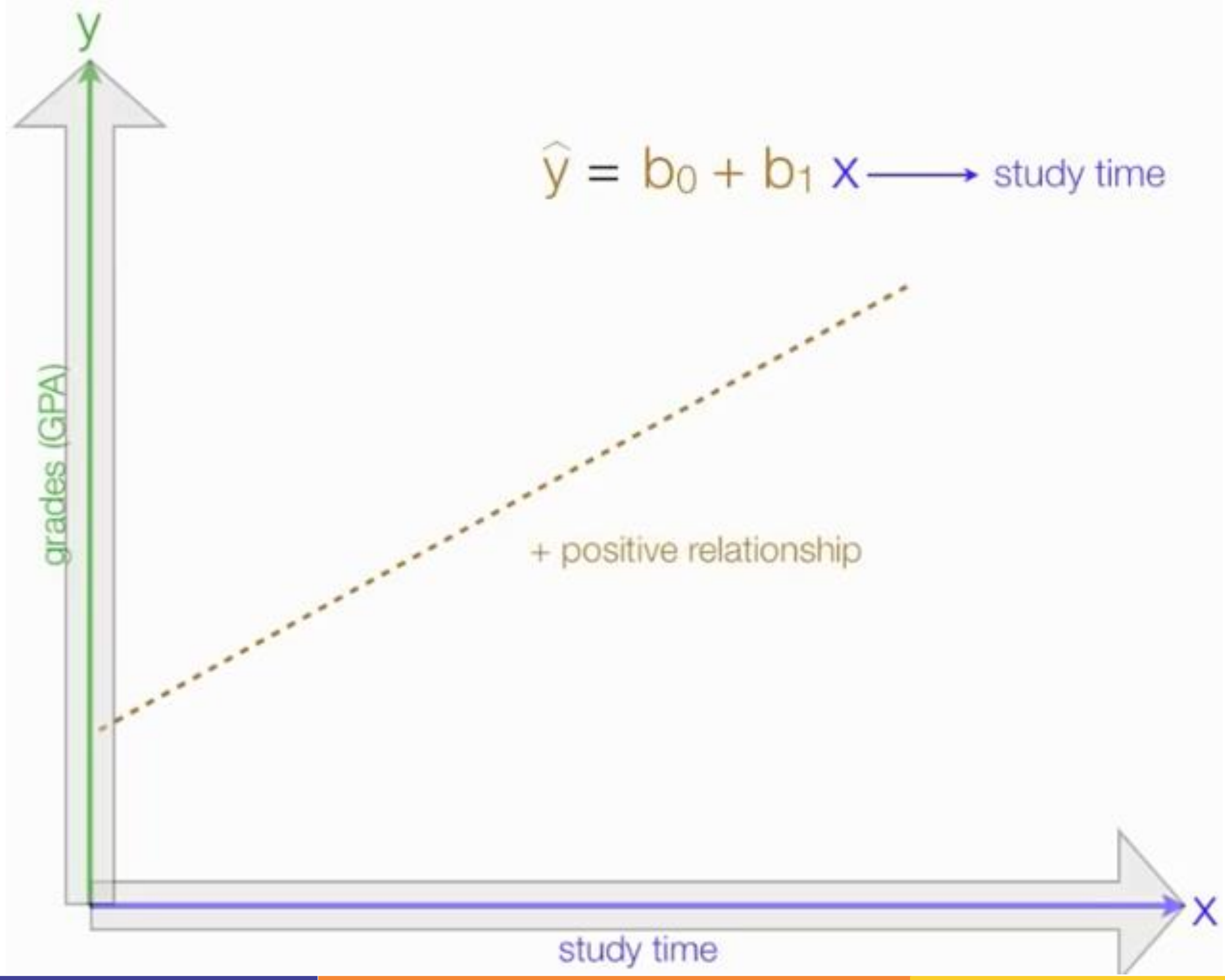
Regression



y
outcome
dependent variable
$\hat{y} = b_0 + b_1 x$
independent variable
control
manipulate
change
dependent variable
x
independent variable

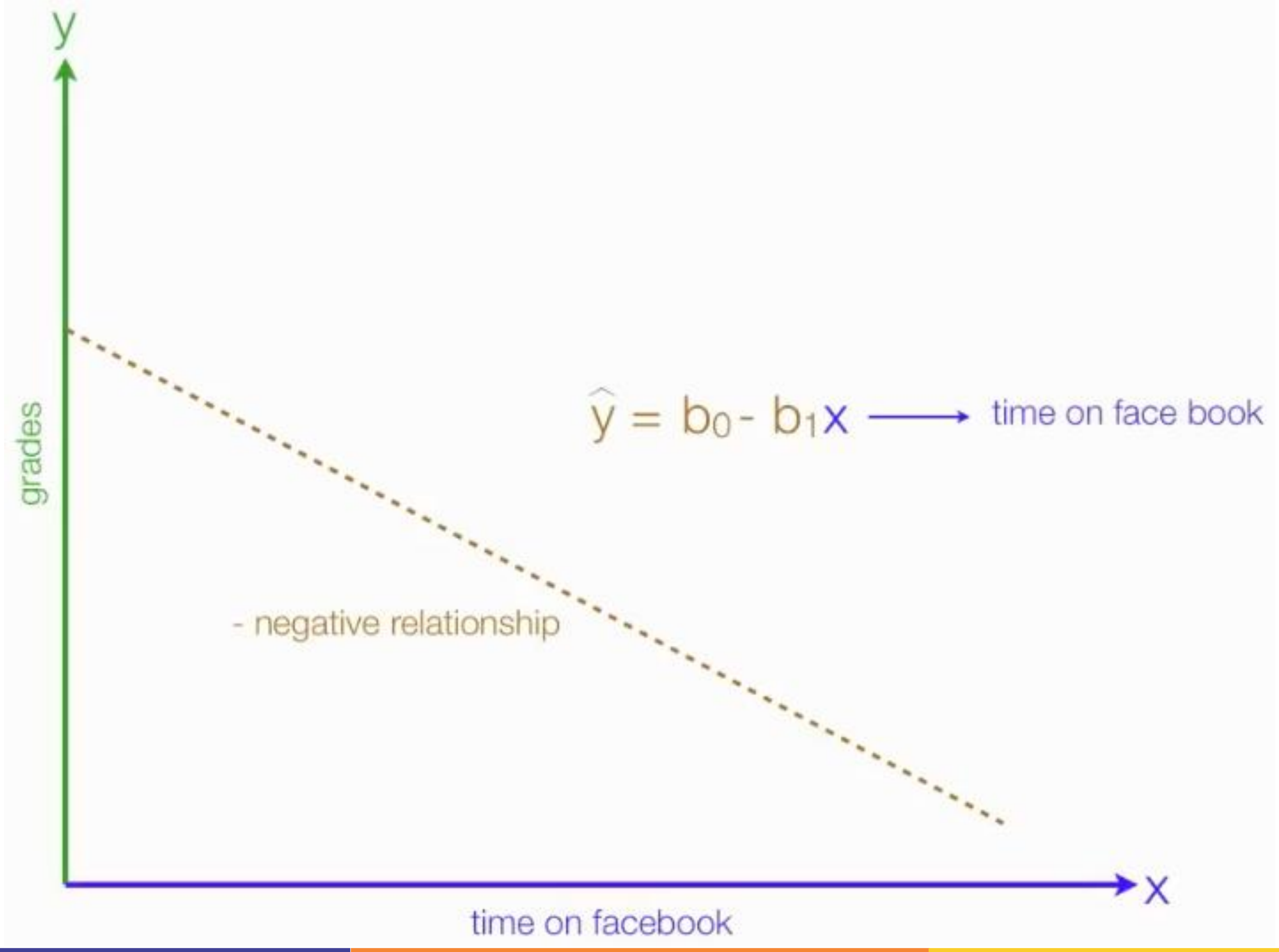
Regression

- Relationship Examples: Positive relationship



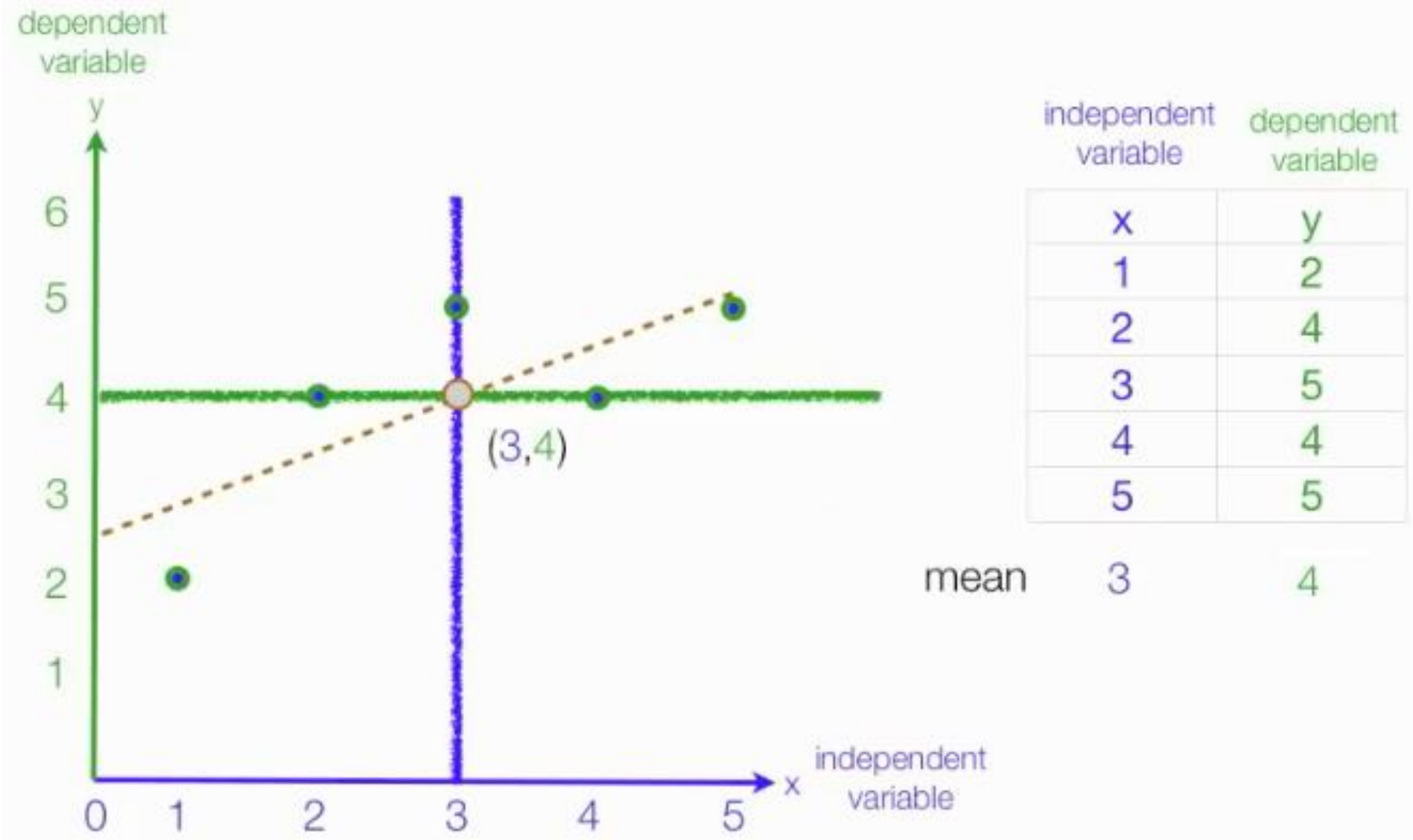
Regression

- Relationship Examples: Negative relationship



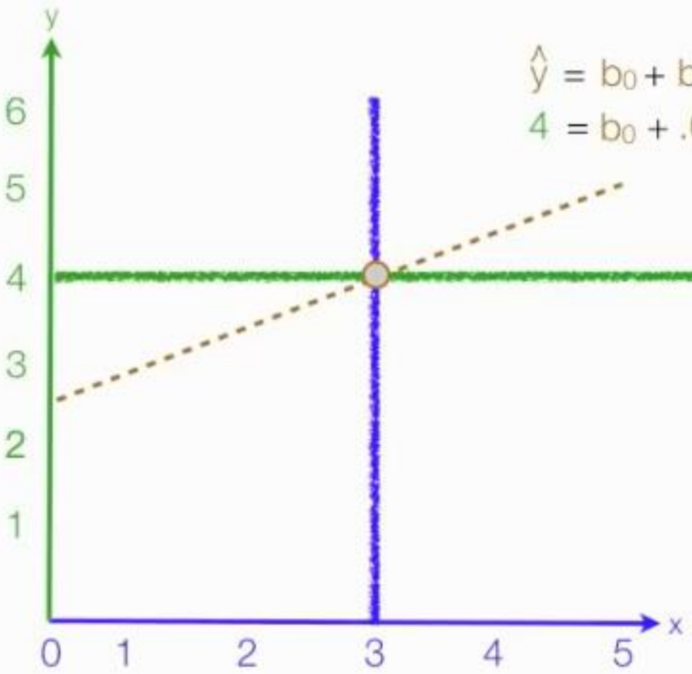
Regression

- Regression line must pass through the point where means of dependent and independent variables crossed.



Regression

$b_0 = 2.2$
$b_1 = .6$
$\hat{y} = 2.2 + .6x$



$\hat{y} = b_0 + b_1 x$
$4 = b_0 + .6(3)$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
mean 3	4			10	6

$4 = b_0 + .6(3)$

$$\begin{aligned} 4 &= b_0 + \cancel{1.8} \\ -1.8 & \quad\;\; \cancel{-1.8} \\ \hline 2.2 &= b_0 \end{aligned}$$

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

Measures of Goodness of fit

- R^2 (R-squared)
  - (Coefficient of Determination)
  - Value ranges from 0(worst fit) to 1(best fit)

$$R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

- Standard Error of Estimate
  - Distance between estimated and actual values

$$= \sqrt{\frac{\sum(\hat{y}-y)^2}{n-2}}$$