HIROSHIMA UNIVERSITY

Fundamental Data Science (30104001)

Lecture 6 — Correlation and Regression

Jorge N. Tendeiro
Hiroshima University

Today

Methods to study the association between two quantitative variables:

- Descriptive statistics to quantify the strength of the linear association between the two variables: Covariance, correlation.
- Statistical model used to predict one variable from the other: Simple linear regression.

Covariance

Scatterplot (Review from Lecture 04)

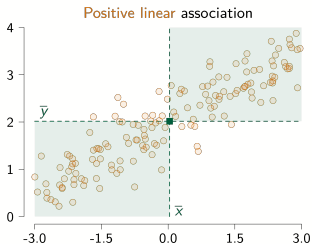
The covariance (and correlation) quantifies the strength of the linear relationship between two variables shown on a scatterplot.

Scatterplot (Review from Lecture 04)

The covariance (and correlation) quantifies the strength of the linear relationship between two variables shown on a scatterplot.

If most points lie in the…

- …top-right and bottom-left panels $\Longrightarrow$ positive linear association.

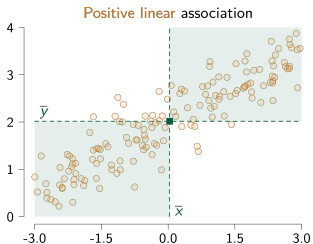


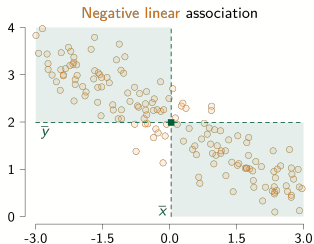
Scatterplot (Review from Lecture 04)

The covariance (and correlation) quantifies the strength of the linear relationship between two variables shown on a scatterplot.

If most points lie in the…

- …top-right and bottom-left panels $\implies$ positive linear association.
- …top-left and bottom-right panels $\implies$ negative linear association.



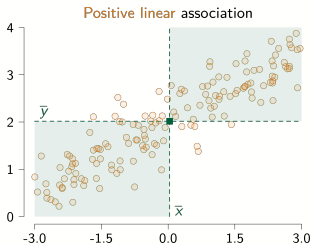


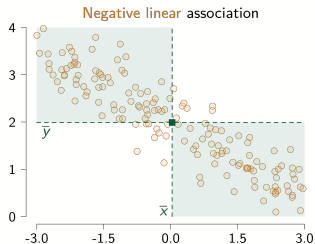
Scatterplot (Review from Lecture 04)

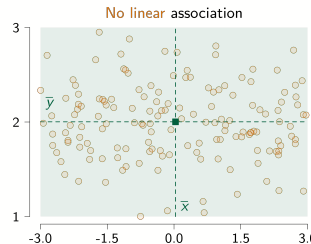
The covariance (and correlation) quantifies the strength of the linear relationship between two variables shown on a scatterplot.

If most points lie in the…

- …top-right and bottom-left panels $\implies$ positive linear association.
- …top-left and bottom-right panels $\implies$ negative linear association.
- …scattered evenly across the four panels $\implies$ no (or very weak) linear association.







Covariance

The covariance between variables x and y, with n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is given by:

$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ and $\overline{y} = \frac{y_1 + \cdots + y_n}{n}$ are the mean of x and y, respectively.

Covariance

The covariance between variables x and y, with n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is given by:

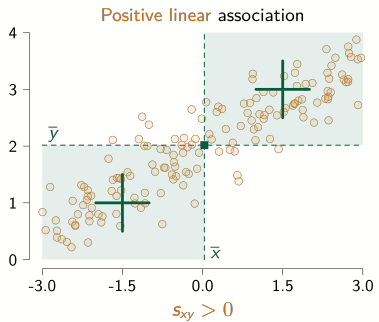
$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n},$$

where $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$ and $\overline{y} = \frac{y_1 + \cdots + y_n}{n}$ are the mean of x and y, respectively.

The covariance describes the $\underbrace{\text{average product of the variables deviation from their means}}_{(x_i - \overline{x})(y_i - \overline{y})}$.

Covariance — Interpretation

$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n}$$



Most point are on the:

- Top-right:

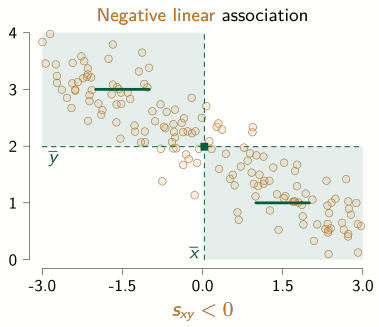
$$\underbrace{(x_i - \overline{x})}_{>0}\underbrace{(y_i - \overline{y})}_{>0} > 0$$

- Bottom-left:

$$\underbrace{(x_i - \overline{x})}_{<0}\underbrace{(y_i - \overline{y})}_{<0} > 0$$

Covariance — Interpretation

$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n}$$



Most point are on the:

- Top-left:

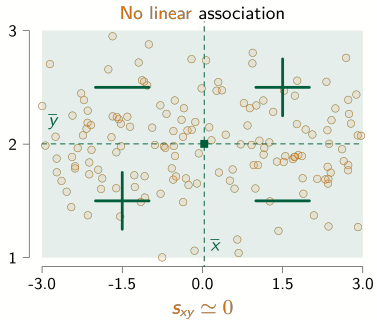
$$\underbrace{(x_i - \overline{x})}_{<0}\underbrace{(y_i - \overline{y})}_{>0} < 0$$

- Bottom-right:

$$\underbrace{(x_i - \overline{x})}_{>0}\underbrace{(y_i - \overline{y})}_{<0} < 0$$

Covariance — Interpretation

$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n}$$



The $(x_i - \overline{x})(y_i - \overline{y})$ terms are scattered across the four panels.

They cancel each other out, thus the covariance is about 0.

To summarize, the covariance between two variables quantifies the strength of their linear relationship.

Problem about covariances

Changing the units of measurement will change the value of the covariance.
This is true even if the data did not change!

Problem about covariances

Changing the units of measurement will change the value of the covariance.
This is true even if the data did not change!

Example:

$s_{xy} = 39.8$

ID	$x =$ Height (cm)	$y =$ Weight (kg)
1	152	44
2	160	49
3	165	54
4	168	59
5	170	61

$s_{xy} = 39800$

ID	$x =$ Height (cm)	$y =$ Weight (g)
1	152	44000
2	160	49000
3	165	54000
4	168	59000
5	170	61000

Problem about covariances

Changing the units of measurement will change the value of the covariance.
This is true even if the data did not change!

Example:

$s_{xy} = 39.8$

ID	$x =$ Height (cm)	$y =$ Weight (kg)
1	152	44
2	160	49
3	165	54
4	168	59
5	170	61

$s_{xy} = 39800$

ID	$x =$ Height (cm)	$y =$ Weight (g)
1	152	44000
2	160	49000
3	165	54000
4	168	59000
5	170	61000

Therefore, the value of the covariance to describe the strength of a linear association is somewhat limited.

'Fixing' unit indeterminacy

We would like to fix this limitation of the covariance.
To understand how this could be done, let's focus on the variance of one of the variables, say weight:

'Fixing' unit indeterminacy

We would like to fix this limitation of the covariance.
To understand how this could be done, let's focus on the variance of one of the variables, say weight:

Variable	y	$(y-\overline{y})^2$	$\left(\frac{y-\overline{y}}{s_y}\right)^2$
Unit	kg	kg^2	unitless
	44	$(44-53.4)^2$	$\left(\frac{44-53.4}{\sqrt{39.44}}\right)^2$
	49	$(49-53.4)^2$	$\left(\frac{49-53.4}{\sqrt{39.44}}\right)^2$
	54	$(54-53.4)^2$	$\left(\frac{54-53.4}{\sqrt{39.44}}\right)^2$
	59	$(59-53.4)^2$	$\left(\frac{59-53.4}{\sqrt{39.44}}\right)^2$
	61	$(61-53.4)^2$	$\left(\frac{61-53.4}{\sqrt{39.44}}\right)^2$
	$\overline{y}=53.4$	$s_y^2=39.44$	**Average** = 1

'Fixing' unit indeterminacy

We would like to fix this limitation of the covariance.
To understand how this could be done, let's focus on the variance of one of the variables, say weight:

Variable	y	$(y-\overline{y})^2$	$\left(\frac{y-\overline{y}}{s_y}\right)^2$
Unit	kg	kg^2	unitless
	44	$(44-53.4)^2$	$\left(\frac{44-53.4}{\sqrt{39.44}}\right)^2$
	49	$(49-53.4)^2$	$\left(\frac{49-53.4}{\sqrt{39.44}}\right)^2$
	54	$(54-53.4)^2$	$\left(\frac{54-53.4}{\sqrt{39.44}}\right)^2$
	59	$(59-53.4)^2$	$\left(\frac{59-53.4}{\sqrt{39.44}}\right)^2$
	61	$(61-53.4)^2$	$\left(\frac{61-53.4}{\sqrt{39.44}}\right)^2$
	$\overline{y}=53.4$	$s_y^2=39.44$	**Average** = 1

Variable	y	$(y-\overline{y})^2$	$\left(\frac{y-\overline{y}}{s_y}\right)^2$
Unit	g	g^2	unitless
	44000	$(44000-53400)^2$	$\left(\frac{44000-53400}{\sqrt{39440000}}\right)^2$
	49000	$(49000-53400)^2$	$\left(\frac{49000-53400}{\sqrt{39440000}}\right)^2$
	54000	$(54000-53400)^2$	$\left(\frac{54000-53400}{\sqrt{39440000}}\right)^2$
	59000	$(59000-53400)^2$	$\left(\frac{59000-53400}{\sqrt{39440000}}\right)^2$
	61000	$(61000-53400)^2$	$\left(\frac{61000-53400}{\sqrt{39440000}}\right)^2$
	$\overline{y}=53400$	$s_y^2=39440000$	**Average** = 1

'Fixing' unit indeterminacy

Here's the 'trick':

$$
\begin{aligned}
1 &= \frac{s_y^2}{s_y^2} \\
&= \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n} \Big/ s_y^2 \\
&= \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{s_y^2} \Big/ n \\
&= \sum_{i=1}^{n}\left(\frac{y_i - \overline{y}}{s_y}\right)^2 \Big/ n \\
&= \text{mean of the elements } \left(\frac{y_i - \overline{y}}{s_y}\right)^2.
\end{aligned}
$$

This works regardless of the units of measurement (e.g., grams, kilograms, tons, etc.).

'Fixing' unit indeterminacy

Thus, instead of using the original scores of a variable, say y:

$$y_1, y_2, \ldots, y_n,$$

we use the so-called standardized scores:

$$\frac{y_1 - \overline{y}}{s_y}, \ldots, \frac{y_n - \overline{y}}{s_y}.$$

'Fixing' unit indeterminacy

Thus, instead of using the original scores of a variable, say y:

$$y_1, y_2, \ldots, y_n,$$

we use the so-called standardized scores:

$$\frac{y_1 - \overline{y}}{s_y}, \ldots, \frac{y_n - \overline{y}}{s_y}.$$

Rules to compute standardized scores:

1. Subtract the mean from each score:

$$y_i \quad \longrightarrow \quad y_i - \overline{y}$$

   This step is known as centering the data.

'Fixing' unit indeterminacy

Thus, instead of using the original scores of a variable, say y:

$$y_1, y_2, \ldots, y_n,$$

we use the so-called standardized scores:

$$\frac{y_1 - \overline{y}}{s_y}, \ldots, \frac{y_n - \overline{y}}{s_y}.$$

Rules to compute standardized scores:

1. Subtract the mean from each score:

$$y_i \quad \longrightarrow \quad y_i - \overline{y}$$

   This step is known as centering the data.

2. Divide by the standard deviation:

$$y_i - \overline{y} \quad \longrightarrow \quad \frac{y_i - \overline{y}}{s_y}$$

   This step (combined with step 1) is known as standardizing the data.

Correlation

Looking at the covariance,

$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n},$$

we can see that the formula is based on the centered scores $(x_i - \overline{x})$ and $(y_i - \overline{y})$, for $i = 1, \ldots, n$.

Correlation

Looking at the covariance,

$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n},$$

we can see that the formula is based on the centered scores $(x_i - \overline{x})$ and $(y_i - \overline{y})$, for $i = 1, \ldots, n$.

We now introduced the so-called correlation coefficient.
The correlation is equal to the covariance, but applied instead to the standardized scores:

$$r_{xy} = \left\{ \left(\frac{x_1 - \overline{x}}{s_x} \right) \left(\frac{y_1 - \overline{y}}{s_y} \right) + \cdots + \left(\frac{x_n - \overline{x}}{s_x} \right) \left(\frac{y_n - \overline{y}}{s_y} \right) \right\} \Big/ n.$$

Correlation

Looking at the covariance,

$$s_{xy} = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (y_1 - \overline{y})(y_1 - \overline{y})}{n},$$

we can see that the formula is based on the centered scores $(x_i - \overline{x})$ and $(y_i - \overline{y})$, for $i = 1, \ldots, n$.

We now introduced the so-called correlation coefficient.
The correlation is equal to the covariance, but applied instead to the standardized scores:

$$r_{xy} = \left\{ \left(\frac{x_1 - \overline{x}}{s_x} \right) \left(\frac{y_1 - \overline{y}}{s_y} \right) + \cdots + \left(\frac{x_n - \overline{x}}{s_x} \right) \left(\frac{y_n - \overline{y}}{s_y} \right) \right\} \Big/ n.$$

Note: In case $s_x = 0$ or $s_y = 0$ (implying that we cannot standardized either variable x or y) then we define $r_{xy} = 0$.

Correlation

$$r_{xy} = \left\{ \left(\frac{x_1 - \overline{x}}{s_x} \right) \left(\frac{y_1 - \overline{y}}{s_y} \right) + \cdots + \left(\frac{x_n - \overline{x}}{s_x} \right) \left(\frac{y_n - \overline{y}}{s_y} \right) \right\} \Big/ n$$

Correlation

$$r_{xy} = \left\{ \left(\frac{x_1 - \overline{x}}{s_x} \right) \left(\frac{y_1 - \overline{y}}{s_y} \right) + \cdots + \left(\frac{x_n - \overline{x}}{s_x} \right) \left(\frac{y_n - \overline{y}}{s_y} \right) \right\} \Big/ n$$

For completeness only (don't worry about the mathematical details):
A little of algebra allows transforming the formula above into this:

$$\begin{aligned} r_{xy} &= \left\{ \frac{(x_1 - \overline{x})(y_1 - \overline{y})}{s_x s_y} + \cdots + \frac{(x_n - \overline{x})(y_n - \overline{y})}{s_x s_y} \right\} \Big/ n \\ &= \left\{ \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})}{n} \right\} \Big/ (s_x s_y) \\ &= \frac{s_{xy}}{s_x s_y}. \end{aligned}$$

Thus, the covariance and the correlation coefficients are related!

Correlation

Example:

ID	$x =$ Height (cm)	$y =$ Weight (kg)
1	152	44
2	160	49
3	165	54
4	168	59
5	170	61

- $s_x = 6.45$
- $s_y = 6.28$
- $s_{xy} = 39.8$

$$r_{xy} = \frac{39.8}{6.45 \times 6.28} = 0.98$$

ID	$x =$ Height (cm)	$y =$ Weight (g)
1	152	44000
2	160	49000
3	165	54000
4	168	59000
5	170	61000

- $s_x = 6.45$
- $s_y = 6280$
- $s_{xy} = 39800$

$$r_{xy} = \frac{39800}{6.45 \times 6280} = 0.98$$

Correlation coefficient — Properties

- r_{xy} is always between -1 and 1:
  - $r = -1$: Perfect decreasing linear association.
    The data fall on a negative slope line.
  - $r = 0$: No linear association.
  - $r = 1$: Perfect increasing linear association:
    The data fall on a positive slope line.

Correlation coefficient — Properties

- r_{xy} is always between -1 and 1:
  - $r = -1$: Perfect decreasing linear association.
    The data fall on a negative slope line.
  - $r = 0$: No linear association.
  - $r = 1$: Perfect increasing linear association:
    The data fall on a positive slope line.

- What is a small, medium, or large correlation depends on the context (research field, problem at hand, etc.).

Correlation coefficient — Properties

- r_{xy} is always between -1 and 1:
  - $r = -1$: Perfect decreasing linear association.
    The data fall on a negative slope line.
  - $r = 0$: No linear association.
  - $r = 1$: Perfect increasing linear association:
    The data fall on a positive slope line.

- What is a small, medium, or large correlation depends on the context (research field, problem at hand, etc.).

- r_{xy} is meant to quantify **linear associations** between two variables.
  Other types of associations (e.g., curvilinear) should **not** be described by r_{xy}!

Correlation coefficient — Properties

- r_{xy} is always between -1 and 1:
  - $r = -1$: Perfect decreasing linear association.
    The data fall on a negative slope line.
  - $r = 0$: No linear association.
  - $r = 1$: Perfect increasing linear association:
    The data fall on a positive slope line.

- What is a small, medium, or large correlation depends on the context (research field, problem at hand, etc.).

- r_{xy} is meant to quantify **linear associations** between two variables.
  Other types of associations (e.g., curvilinear) should **not** be described by r_{xy}!

- **Always** look at the association between two variables through a plot.

Correlation coefficient — Strength of linear association

r_{xy} indicates the strength of the linear association between variables x and y.

The closer r_{xy} is to…

- 1: the stronger the positive linear association.
- -1: the stronger the negative linear association.
- 0: the weaker the linear association.
  When $r_{xy} = 0$ we say that variables x and y are uncorrelated.

Correlation coefficient — Strength of linear association

r_{xy} indicates the strength of the linear association between variables x and y.

The closer r_{xy} is to…

- 1: the stronger the positive linear association.
- -1: the stronger the negative linear association.
- 0: the weaker the linear association.
  When $r_{xy} = 0$ we say that variables x and y are uncorrelated.

What is a small, medium, or large correlation depends on the context
(research field, problem at hand, etc.).

Correlation coefficient — Strength of linear association

r_{xy} indicates the strength of the linear association between variables x and y.

The closer r_{xy} is to…

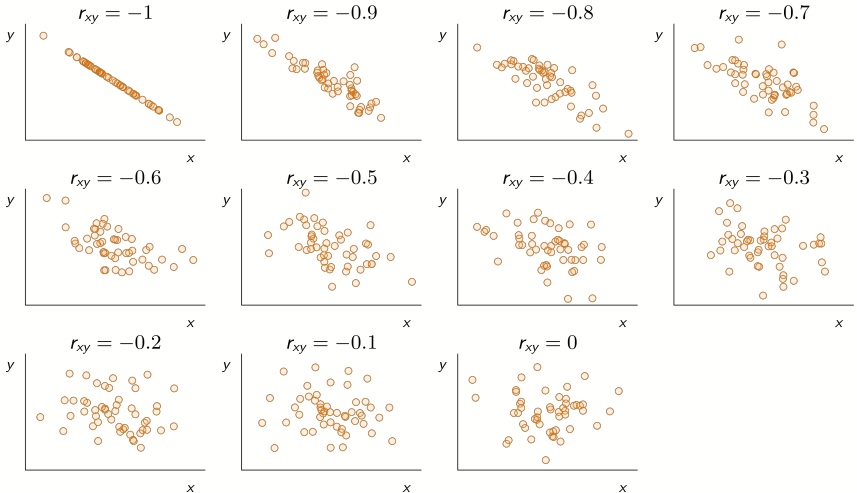
- 1: the stronger the positive linear association.
- -1: the stronger the negative linear association.
- 0: the weaker the linear association.
  When $r_{xy} = 0$ we say that variables x and y are uncorrelated.

What is a small, medium, or large correlation depends on the context
(research field, problem at hand, etc.).

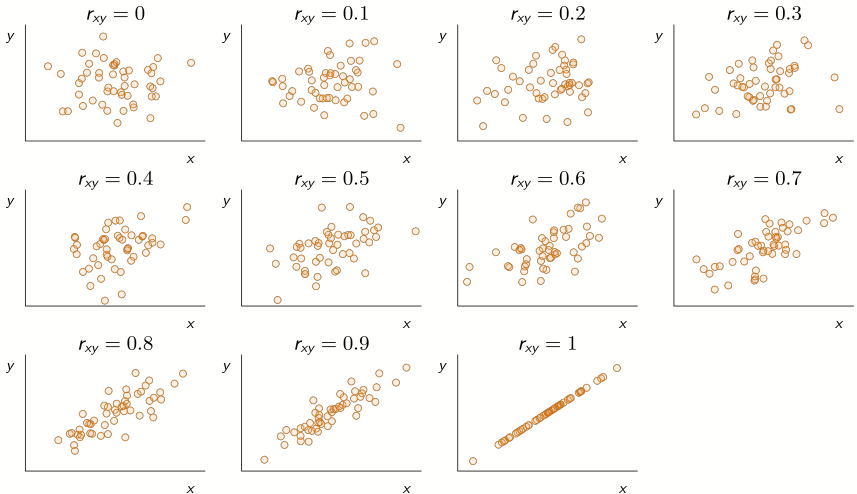
The table on the right should therefore be used just as a reference (**not** as a fixed rule!).

r_{xy}	**Strength**
$-1 \leq \cdot \leq -0.7$	Strong negative
$-0.7 \leq \cdot \leq -0.4$	Somewhat negative
$-0.4 \leq \cdot \leq -0.2$	Weak negative
$-0.2 \leq \cdot \leq 0.2$	Almost uncorrelated
$0.2 \leq \cdot \leq 0.4$	Weak positive
$0.4 \leq \cdot \leq 0.7$	Somewhat positive
$0.7 \leq \cdot \leq 1$	Strong positive

Negative correlation coefficient



Positive correlation coefficient

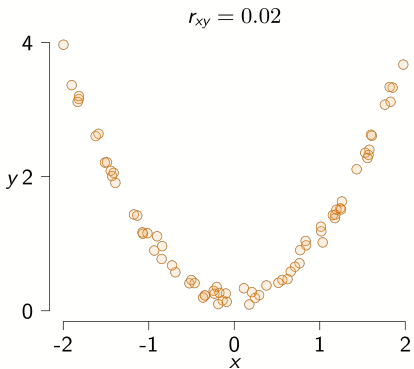


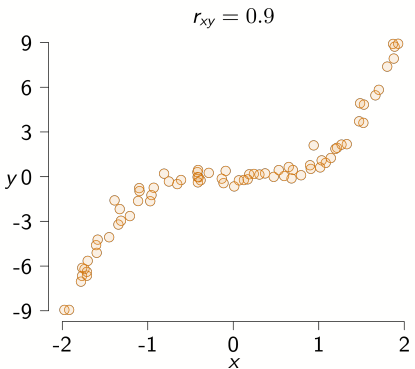
Correlation coefficient — Be careful

The correlative value, by itself, is **not enough** to learn about the association between the two variables.
Always look at the association between two variables through a plot.

Correlation coefficient — Be careful

The correlative value, by itself, is **not enough** to learn about the association between the two variables.
Always look at the association between two variables through a plot.



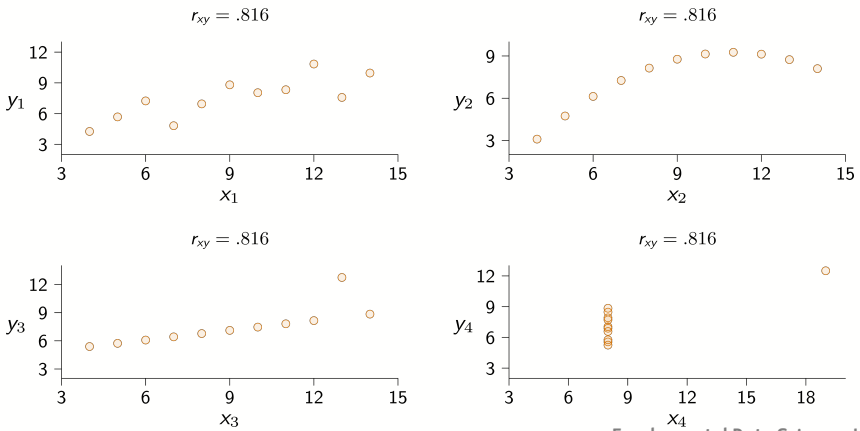


Anscombe's quartet

The value of r_{xy} alone can be misleading, as the famous **Anscombe quartet** illustrates.
Below, r_{xy} is only meaningful for the top-left panel.

Anscombe's quartet

The value of r_{xy} alone can be misleading, as the famous **Anscombe quartet** illustrates.
Below, r_{xy} is only meaningful for the top-left panel.



Linear regression analysis

Linear regression — Motivation

Consider the following dilemma of Student A:

> *I want to increase my test score to 80.*
> *If I study more hours probably my test score improves…*
> *But, how many more hours should I study to get score 80?*
> *I also don't want to put in too many hours...*

Linear regression — Motivation

Consider the following dilemma of Student A:

> *I want to increase my test score to 80.*
> *If I study more hours probably my test score improves…*
> *But, how many more hours should I study to get score 80?*
> *I also don't want to put in too many hours...*

Student A collected data to answer his question:

Student	Study time (h)	Test score
1	8	74
2	5.2	68
⋮	⋮	⋮
30	10	92

Linear regression — Motivation

Consider the following dilemma of Student A:

> *I want to increase my test score to 80.*
> *If I study more hours probably my test score improves…*
> *But, how many more hours should I study to get score 80?*
> *I also don't want to put in too many hours...*

Student A collected data to answer his question:

Student	Study time (h)	Test score
1	8	74
2	5.2	68
⋮	⋮	⋮
30	10	92

By studying how test score varies as a function of the number of study hours, Student A hopes to determine how many hours he should study.

Regression Analysis

Regression analysis is a statistical method that allows modeling a response variable (say, y) as a function of a predictor variable (say, x).

Regression Analysis

Regression analysis is a statistical method that allows modeling a response variable (say, y) as a function of a predictor variable (say, x).

Note. There are various names in use for variables x and y:

- x: Predictor, input, independent variable, explanatory variable.
  Example: $x =$ number of study hours.
- y: Response, output, dependent variable, outcome variable.
  Example: $y =$ test score.

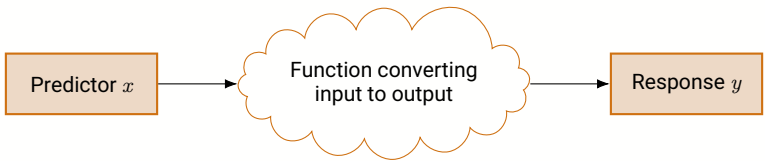
Regression Analysis

Regression analysis is a statistical method that allows modeling a response variable (say, y) as a function of a predictor variable (say, x).

Note. There are various names in use for variables x and y:

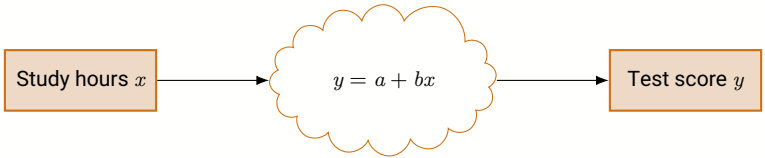
- x: Predictor, input, independent variable, explanatory variable.
  Example: $x =$ number of study hours.
- y: Response, output, dependent variable, outcome variable.
  Example: $y =$ test score.

One of the 'secrets' in regression analysis is to determine the functional relationship between x and y.
This is **required**, otherwise regression analysis won't work.



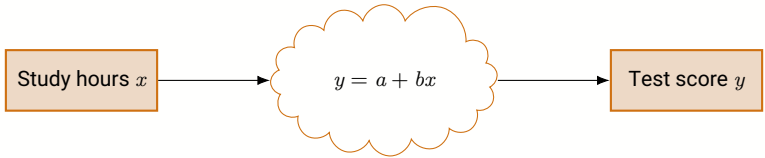
Regression Analysis

Student A decided to use the following model:



Regression Analysis

Student A decided to use the following model:



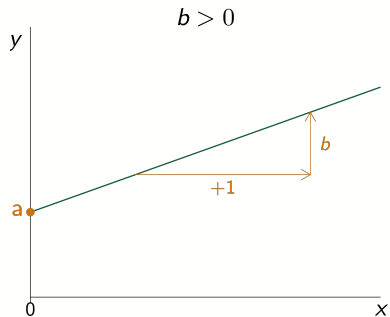
$y = a + bx$ is the equation of a straight line relating x to y.

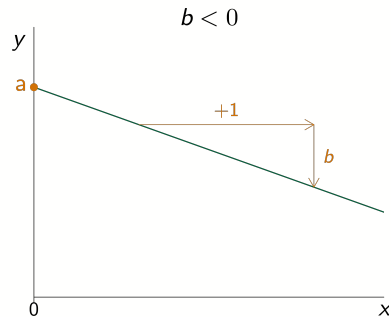
Regression coefficients

$$\boxed{y = a + bx}$$

The values a and b are the model parameters, also known as the regression coefficients.

- a = intercept = predicted value of y when $x = 0$.
- b = slope = amount by which y changes when x increases by 1 unit.





Regression Analysis

$$y = a + bx$$

We don't know the values of a and b.
We will need to estimate them from the data.
We will learn today how to do that.

Regression Analysis

$$y = a + bx$$

We don't know the values of a and b.
We will need to estimate them from the data.
We will learn today how to do that.

The estimated values of a and b will be denoted by $\widehat{a}$ and $\widehat{b}$ (read: "a-hat", "b-hat").

Regression Analysis

$$\boxed{y = a + bx}$$

We don't know the values of a and b.
We will need to estimate them from the data.
We will learn today how to do that.

The estimated values of a and b will be denoted by $\widehat{a}$ and $\widehat{b}$ (read: "a-hat", "b-hat").

Once estimated, we can use the regression model to make predictions.

For example:

> *If I study 7 hours, I predict that I will get a test score equal to $\widehat{a} + \widehat{b} \times 7$.*

Regression Analysis

$$y = a + bx$$

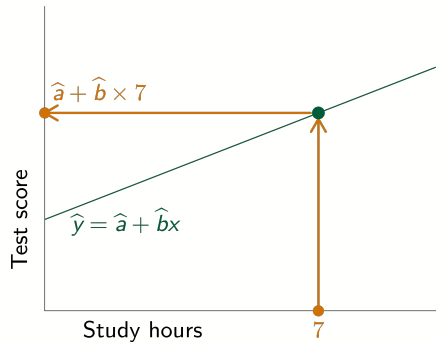
We don't know the values of a and b.
We will need to estimate them from the data.
We will learn today how to do that.

The estimated values of a and b will be denoted by $\widehat{a}$ and $\widehat{b}$ (read: "a-hat", "b-hat").

Once estimated, we can use the regression model to make predictions.

For example:

> *If I study 7 hours, I predict that I will get a test score equal to* $\widehat{a} + \widehat{b} \times 7$.



Versality of the regression model

The regression equation $y = a + bx$ is at the basis of the most famous regression model of all:

the simple linear regression model.

Versality of the regression model

The regression equation $y = a + bx$ is at the basis of the most famous regression model of all:

the simple linear regression model.

For now, just keep in mind the following important idea:

The simple linear regression model is the basis of many other statistical models.

For example:

- Multiple linear regression
- Polynomial regression
- Multivariate regression
- Logistic regression
- Lasso regression
- Deep learning

Versality of the regression model

The regression equation $y = a + bx$ is at the basis of the most famous regression model of all:

the simple linear regression model.

For now, just keep in mind the following important idea:

The simple linear regression model is the basis of many other statistical models.

For example:

- Multiple linear regression
- Polynomial regression
- Multivariate regression
- Logistic regression
- Lasso regression
- Deep learning

We will next focus on the simple linear regression model.

Simple linear regression:

Goals

Simple linear regression

$$\boxed{y = a + bx}$$

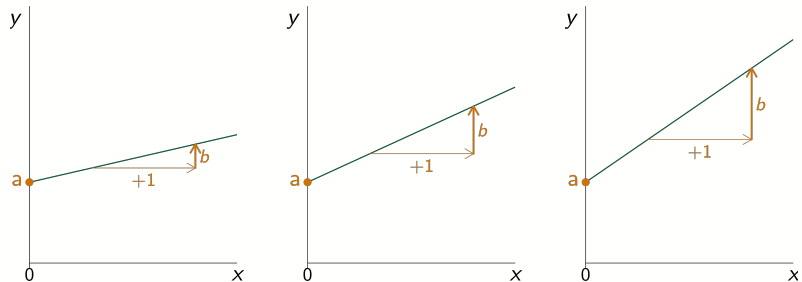
There are two main goals when fitting a simple linear regression model to data.

Simple linear regression

$$y = a + bx$$

There are two main goals when fitting a simple linear regression model to data.

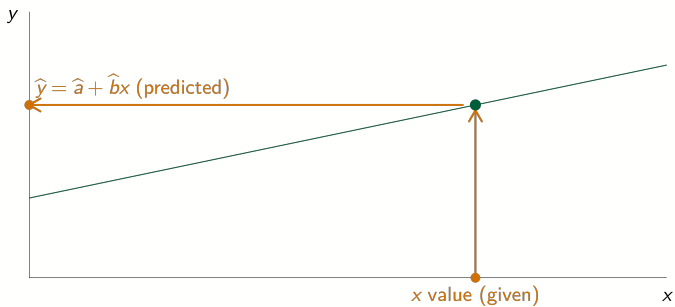
1. Quantify the effect of predictor x on outcome y.
   In other words: Focus on the slope coefficient b.



Simple linear regression

$$\boxed{y = a + bx}$$

2. Predict unknown (future) y values for given x values.
   In other words: Focus on $\widehat{y} = \widehat{a} + \widehat{b}x$ (from estimation).



Simple linear regression:

Estimating the regression coefficients

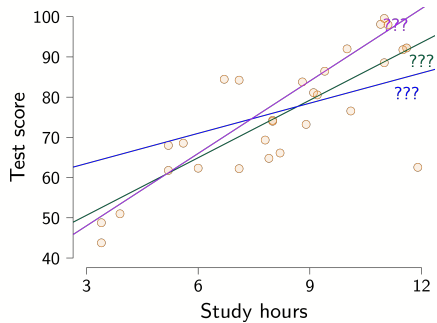
Estimating the regression coefficients

$$y = a + bx$$

How can we determine "the" regression line?

Just by "looking" we cannot determine which line is the "best".

What we can do is use the data (x_i, y_i) for $i = 1, \ldots, n$ to estimate the intercept a and the slope b.



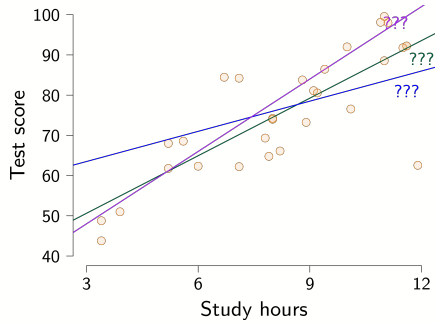
Estimating the regression coefficients

$$y = a + bx$$

How can we determine "the" regression line?

Just by "looking" we cannot determine which line is the "best".

What we can do is use the data (x_i, y_i) for $i = 1, \ldots, n$ to estimate the intercept a and the slope b.



Before performing estimation, we need to set up a criterion (i.e., a rule) that specifies what the "best" regression line should be.

Estimating the regression coefficients a, b

$$y = a + bx$$

We need to consider the so-called residuals.

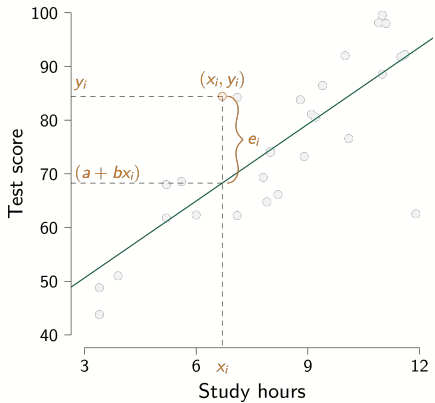
For the i-th observation,

$$\text{residual} = \text{observed} - \text{predicted}$$
$$e_i = y_i - (a + bx_i)$$

The closer each data y_i is to the regression line…

- the closer the residual is to 0;
- the better the prediction from the regression line.

We must consider the residuals for all observations!



Least squares method

$$y = a + bx$$

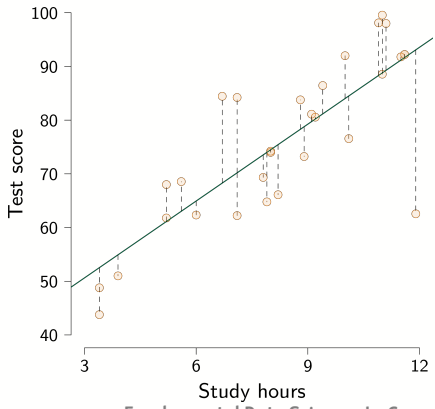
Least squares method:

Find the regression coefficients a and b that minimize the sum of squares of the residuals:

$$\sum_{i=1}^{n} e_i^2$$

$$= \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

$$= [y_1 - (a + bx_1)]^2 + \cdots + [y_n - (a + bx_n)]^2 .$$



Least squares method

$$\boxed{y = a + bx}$$

The mathematical solution that optimizes the least squares criterion is given by

$$\widehat{b} = r_{xy}\frac{s_y}{s_x}, \quad \widehat{a} = \overline{y} - \widehat{b}\overline{x}.$$

$\widehat{a}$ and $\widehat{b}$ are called the least square estimates of parameters a and b.

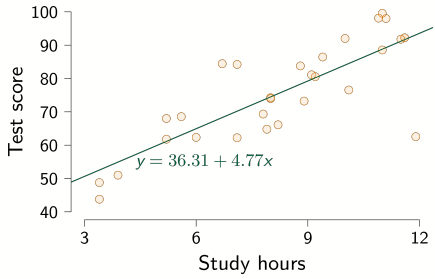
For the running data:

- $r_{xy} = 0.805$
- $\overline{x} = 8.320$, $s_x = 2.551$
- $\overline{y} = 76.007$, $s_y = 15.130$,

thus

$$\widehat{b} = 0.805\left(\frac{15.130}{2.551}\right) = 4.771$$

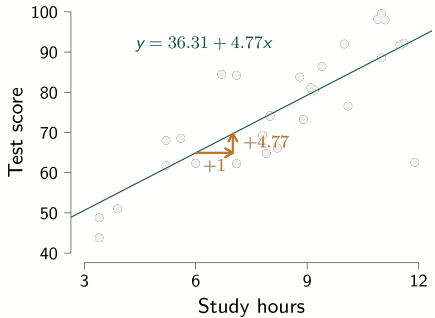
$$\widehat{a} = 76.007 - 4.771(8.320) = 36.311$$



Two main goals in using regression

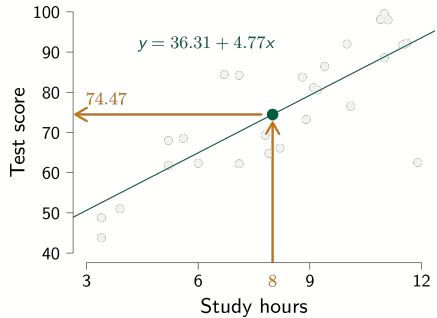
Quantify the effect of x on y:

When x increases by 1 unit, y increases by 4.77 units.



Predict unknown y values for given x values:

When $x = 8$, the predicted test score is $y = 36.31 + 4.77(8) = 74.47$.



Exercise (1)

Suppose we obtained the following values for variables x and y:

$$\overline{x} = 6, \overline{y} = 6, s_x^2 = \frac{26}{5}, s_{xy} = \frac{14}{5}.$$

When simple linear regression analysis is conducted with y as the response variable and x as the predictor, which linear regression line do we obtain with the least squares method?
(*Tip:* Check pages 14 and 36.)

Exercise (1) — ANSWER

Suppose we obtained the following values for variables x and y:

$$\overline{x} = 6, \overline{y} = 6, s_x^2 = \frac{26}{5}, s_{xy} = \frac{14}{5}.$$

When simple linear regression analysis is conducted with y as the response variable and x as the predictor, which linear regression line do we obtain with the least squares method?
(*Tip:* Check pages 14 and 36.)

Answer

$$\widehat{b} = r_{xy}\frac{s_y}{s_x}, \quad \widehat{a} = \overline{y} - \widehat{b}\overline{x}$$

- $\widehat{b} = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} = \frac{14/5}{26/5} = \frac{7}{13}$
- $\widehat{a} = \overline{y} - \widehat{b}\overline{x} = 6 - \left(\frac{7}{13}\right)6 = \frac{36}{13}$

Therefore, the estimated regression line is $y = \frac{36}{13} + \frac{7}{13}x$.

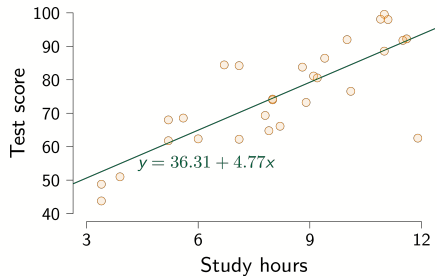
Coefficient of determination

Coefficient of determination

We now know how to find the "best" (in the least squares sense) simple linear regression model relating x to y.

But, does this model fit the data well (enough)?...

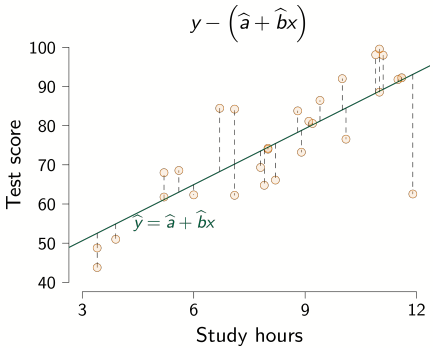
One way to answer this question is to quantify the model fit, by means of the so-called coefficient of determination.



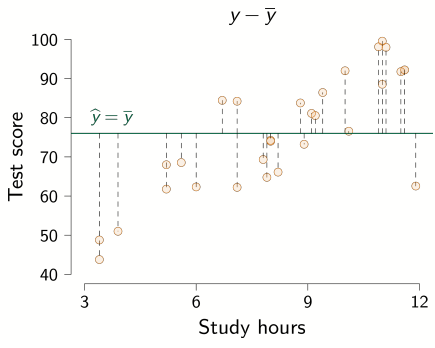
Coefficient of determination

The coefficient of determination is based on comparing two different predictions for y:

y predicted by $\widehat{a} + \widehat{b}x$

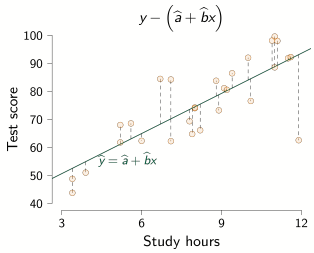


y predicted by $\overline{y}$

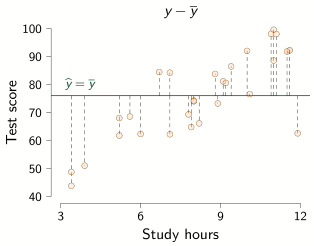


Coefficient of determination

y predicted by $\widehat{a} + \widehat{b}x$



y predicted by $\overline{y}$

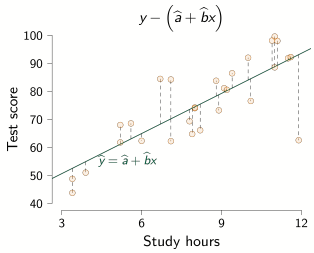


The main idea is this:

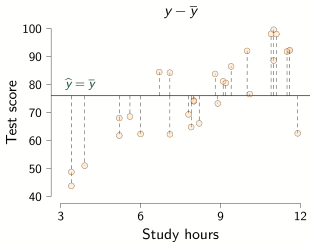
> *Predictor x is helpful in predicting y if and only if $\widehat{y} = a + bx$ (model with x) gives better predictions than $y = \overline{y}$ (model without x), on average.*

Coefficient of determination

y predicted by $\widehat{a} + \widehat{b}x$



y predicted by $\overline{y}$



The main idea is this:

> *Predictor x is helpful in predicting y if and only if $\widehat{y} = a + bx$ (model with x) gives better predictions than $y = \overline{y}$ (model without x), on average.*

The coefficient of determination compares the squared distances between the observations y (the yellow circles) and the predictions from either model's line (in green).

Coefficient of determination

It can be shown that

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\text{total sum of squares } (SS_{\text{total}})} = \underbrace{\sum_{i=1}^{n}\left[y_i-(\widehat{a}+\widehat{b}x_i)\right]^2}_{\text{residual sum of squares } (SS_{\text{res}})} + \underbrace{\sum_{i=1}^{n}\left[(\widehat{a}+\widehat{b}x_i)-\overline{y}\right]^2}_{\text{model sum of squares } (SS_{\text{mod}})}$$

$$\boxed{SS_{\text{total}} = SS_{\text{res}} + SS_{\text{mod}}}$$

Note that:

- SS_{total} is (essentially) the variance of y.
- SS_{res}, which is just $\sum_{i=1}^{n} e_i^2$, is a measure of *how bad* the regression line fits the data:

  The further the predictions $(\widehat{a}+\widehat{b}x_i)$ are from the observed y_i values, the larger SS_{res}.

Coefficient of determination

$$\boxed{\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{res}} + \mathrm{SS}_{\mathrm{mod}}}$$

Dividing both sides by $\mathrm{SS}_{\mathrm{total}}$ we have this:

$$\frac{\mathrm{SS}_{\mathrm{total}}}{\mathrm{SS}_{\mathrm{total}}} = \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{total}}} + \underbrace{\frac{\mathrm{SS}_{\mathrm{mod}}}{\mathrm{SS}_{\mathrm{total}}}}_{R^2}$$

$$1 = (1 - R^2) + R^2$$

R^2 is a measure of goodness of fit.

Coefficient of determination

$$R^2 = \frac{\mathsf{SS}_{\mathsf{mod}}}{\mathsf{SS}_{\mathsf{total}}} = 1 - \frac{\mathsf{SS}_{\mathsf{res}}}{\mathsf{SS}_{\mathsf{total}}}$$

Coefficient of determination

$$R^2 = \frac{\mathsf{SS}_{\mathsf{mod}}}{\mathsf{SS}_{\mathsf{total}}} = 1 - \frac{\mathsf{SS}_{\mathsf{res}}}{\mathsf{SS}_{\mathsf{total}}}$$

R^2 is a real number between 0 and 1:

- $R^2 = 0$ if $\mathsf{SS}_{\mathsf{mod}} = 0$, that is, if all predictions $(\widehat{a} + \widehat{b}x_i)$ are equal to $\overline{y}$.
  $\longrightarrow \widehat{b} = 0$, that is, predictor x is useless.
  Poor model fit.

Coefficient of determination

$$R^2 = \frac{\mathsf{SS}_{\mathsf{mod}}}{\mathsf{SS}_{\mathsf{total}}} = 1 - \frac{\mathsf{SS}_{\mathsf{res}}}{\mathsf{SS}_{\mathsf{total}}}$$

R^2 is a real number between 0 and 1:

- $R^2 = 0$ if $\mathsf{SS}_{\mathsf{mod}} = 0$, that is, if all predictions $(\widehat{a} + \widehat{b}x_i)$ are equal to $\overline{y}$.
  $\longrightarrow \widehat{b} = 0$, that is, predictor x is useless.
  Poor model fit.

- $R^2 = 1$ if $\mathsf{SS}_{\mathsf{res}} = 0$, that is, if all predictions $(\widehat{a} + \widehat{b}x_i)$ coincide with y.
  $\longrightarrow$ all predictions fall perfectly on the regression line; predictor x works great.
  Excellent model fit, but actually *useless* for prediction.

Coefficient of determination

$$R^2 = \frac{\mathsf{SS}_{\mathsf{mod}}}{\mathsf{SS}_{\mathsf{total}}} = 1 - \frac{\mathsf{SS}_{\mathsf{res}}}{\mathsf{SS}_{\mathsf{total}}}$$

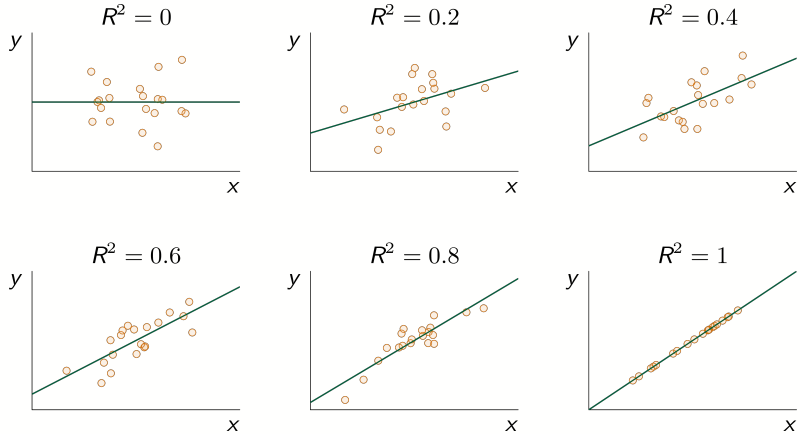
R^2 is a real number between 0 and 1:

- $R^2 = 0$ if $\mathsf{SS}_{\mathsf{mod}} = 0$, that is, if all predictions $(\widehat{a} + \widehat{b}x_i)$ are equal to $\overline{y}$.
  $\longrightarrow \widehat{b} = 0$, that is, predictor x is useless.
  Poor model fit.

- $R^2 = 1$ if $\mathsf{SS}_{\mathsf{res}} = 0$, that is, if all predictions $(\widehat{a} + \widehat{b}x_i)$ coincide with y.
  $\longrightarrow$ all predictions fall perfectly on the regression line; predictor x works great.
  Excellent model fit, but actually *useless* for prediction.

Usually R^2 is between 0 and 1.
We always hope that R^2 is as close to 1 as possible, as long as it can make good predictions.

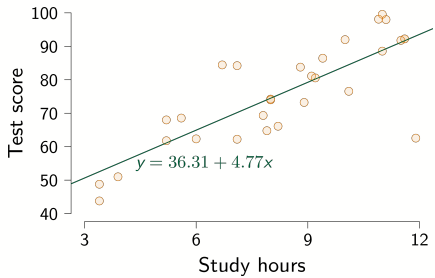
Coefficient of determination



Coefficient of determination

For the running example,

$$R^2 = 0.647.$$



Interpretation:

64.7% *of the total variance of* y *is 'explained' by predictor* x.

Important notes

A small R^2 value may imply that the simple linear regression model is not suitable.

Perhaps:

- We need to include more predictors:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$

  This is known as multiple linear regression.

Important notes

A small R^2 value may imply that the simple linear regression model is not suitable.

Perhaps:

- We need to include more predictors:

  $$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$

  This is known as multiple linear regression.

- Fitting a curve via a polynomial regression model is better suited:

  $$y = a + b_1 x + b_2 x^2 + \cdots + b_k x^k.$$

  This is a special case of multiple linear regression.

Important notes

A small R^2 value may imply that the simple linear regression model is not suitable.

Perhaps:

- We need to include more predictors:

  $$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$

  This is known as multiple linear regression.

- Fitting a curve via a polynomial regression model is better suited:

  $$y = a + b_1 x + b_2 x^2 + \cdots + b_k x^k.$$

  This is a special case of multiple linear regression.

Always remember that simple linear regression is a method only suitable to model linear relationships.

Summary

Follow the following steps in case you are interested in fitting the simple linear regression model:

1. Confirm whether a linear relationship exists between predictor x and response y.
   How: Look at the scatter plot and correlation coefficient r.

Summary

Follow the following steps in case you are interested in fitting the simple linear regression model:

1. Confirm whether a linear relationship exists between predictor x and response y.
   How: Look at the scatter plot and correlation coefficient r.

2. Estimate the regression coefficients a and b from the model $y = a + bx$.
   How: Use the least squared method and the data.

Summary

Follow the following steps in case you are interested in fitting the simple linear regression model:

1. Confirm whether a linear relationship exists between predictor x and response y.
   How: Look at the scatter plot and correlation coefficient r.

2. Estimate the regression coefficients a and b from the model $y = a + bx$.
   How: Use the least squared method and the data.

3. Calculate the coefficient of determination R^2 to evaluate the goodness of fit.

Summary

Follow the following steps in case you are interested in fitting the simple linear regression model:

1. Confirm whether a linear relationship exists between predictor x and response y.
   How: Look at the scatter plot and correlation coefficient r.

2. Estimate the regression coefficients a and b from the model $y = a + bx$.
   How: Use the least squared method and the data.

3. Calculate the coefficient of determination R^2 to evaluate the goodness of fit.

4. Perform the intended analysis:
   - Quantify the effect of x on y.
     How: Look at $\widehat{b}$.
   - Predict unknown y values for given x values.
     How: Use the estimated regression model $y = \widehat{a} + \widehat{b}x$.

To do before the next lecture

Before lecture 7:

- Log in to *Moodle*.
- Go to folder "Lecture 7".
- Download two data files: `HOUSE.csv` and `police.csv`.
- Save the two data files to a folder called `Stat` on your Desktop.

To do before the next lecture

Before lecture 7:

- Log in to *Moodle*.
- Go to folder "Lecture 7".
- Download two data files: `HOUSE.csv` and `police.csv`.
- Save the two data files to a folder called `Stat` on your Desktop.

We will be using Excel in Lecture 7.