EMTH211 Statistics

Fabian Dunker

University of Canterbury

2019

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Lecturer

- Fabian Dunker
- fabian.dunker@canterbury.ac.nz
- Office hour Friday 11pm - 12pm
  and by appointment.
- Erskine 703

Outline

Outline

1. Descriptive statistics
2. Simple linear regression
3. Multiple linear regression

Lecture notes

- Lecture notes by Richard Vale
- Slides

are available on Learn.

1. Descriptive statistics

1.1 Random variables and data

Random variables

Statistical methods are useful if we have observations or measurements which cannot be (exactly) predicted.
For example: measurement with measurement errors and perturbations, or events that occur with uncertainty.

Definition

- We call an experiment or measurement whose outcome cannot be predicted a **random experiment**.
- A map that assigns real numbers to the outcomes of a random experiment is called **random variable**.

Remark

You can think of a random variable as a real valued variable whose value is random, i.e. it can be diffent every time we observe the variable.

Examples

- rolling a dice, possible outcomes are 1,2,3,4,5, or 6
- tossing a coin, we assign tail $= 0$, head $= 1$ to get a real valued variable.
- polls and surveys
- daily mean temperature
- electricity consumption
- life time of a machine in sec, min, hours, or years
- precision of a machine (often influenced by run time)
- measurement with measurement errors:
  temperature, pressure, velocity, voltage, current, ...
- bugs in a software
- Randomised software testing has a random variable as input. Hence, the output is also a random variable.

Data

- Observing a random variable generates **data**.
- A collection of one or more observations is called a **sample**.
- The number of observation in a sample is called the **sample size** and is usually denoted by n.
- Statistical methods try to learn properties of the random variable by analysing properties of the sample.
- However, properties of a sample can differ from the properties of the random variable due to random perturbations. We need to be careful!

Example

The average of a sample indicates what the average outcome of a random variable might be.
If we happen to get tail three times in a row when tossing a coin, the average of this sample will be misleading.

- Fair coin 50% head or tail
- 3 times tail $= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$

$= 12.5\%$

Data vector

We store each sample in a vector, e.g.

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top,$$

where $x_1, x_2, \ldots, x_n$ are the individual observations. The sample size is n.

Notation

- $\mathbf{x}$, $\mathbf{y}$, $\mathbf{b}$ vectors on the slides.
- $\underline{x}$, $\underline{y}$, $\underline{b}$ vectors on the white board.
- $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$, $\bar{\mathbf{b}}$ mean of a data vector (next lecture).

- Vectors on the slides <u>bold letters</u>

(←)

$\underset{\sim}{x}, \underset{\sim}{y}$

$$\underset{\sim}{x} = (x_1, x_2 \dots x_n)^T$$

mean on the board $\tilde{x}$

Two ore more random variables

Often two or more random variables are observed simultaneously.

Example Temperature and pressure in a boiler.

Temp (°C)	Pressure (kPa)
0	91
10	95
20	100
30	101
40	107
50	112

$\mathbf{x} = (0, 10, 20, 30, 40, 50)^\top$
$\mathbf{y} = (91, 95, 100, 101, 107, 112)^\top$

Goals Describe and quantify by using linear algebra:

- probabilistic properties of one (or more) random variable,
- how variations in one random variable is linked to another random variable.

Example for random variables that depend on each other

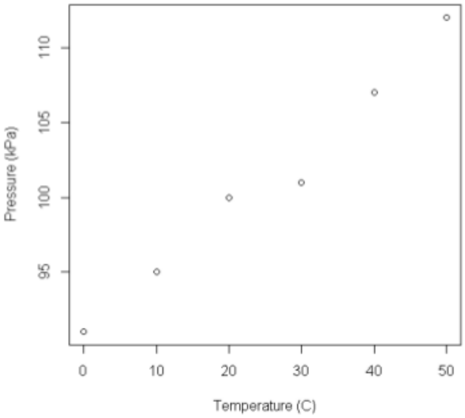
- *temperature* (T) and *pressure* (P)
  Ideal gas law: $P = cT$ with some constant c.
- *voltage* ($\overset{V}{\cancel{U}}$) and *current* (I)
  $\underset{V}{\cancel{U}} = R\,I$ with resistance R.
- *daily mean temperature* and *electricity consumption* in Christchurch
  The lower the temperature the higher the electricity consumption.
  (This will be different in other places, e.g. New York)
- *live time* of a machine and *production costs* of a machine
  A more durable machine is more expensive to produce.
- *Number of bugs* in a software and *number of randomised software tests*.

Scatter plot

Plotting the values of two random variables against each other is called a **scatter plot**. Matlab command: `scatter(x,y)`

A scatter plot can give a rough idea about the data.

Temp ($^\circ$C)	Pressure (kPa)
0	91
10	95
20	100
30	101
40	107
50	112



1.2 Sample mean and centering

Sample mean

Want to find the average outcome

- An important characteristic of a random variable is its average outcome.
- We will use the average of a sample to estimate the average of the random variable.

Definition

For a data vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ the **sample mean** is defined as

$$\bar{\mathbf{x}} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Note

- The average outcome of the random variable is called **expectation** or **population mean**. It is not the same as the sample mean.
- Often the **sample mean** and the **expectation** are both just called the **mean**. Do not confuse these two concepts!

$$\underset{\sim}{x} = (x_1\ x_2\ \dots\ x_n)^T$$ (Proofs)

$$\underset{\sim}{y} = (y_1\ y_2\ \cdots\ y_n)^T$$

(1):

$$\widetilde{x+y} = \frac{1}{n}\sum_{i=1}^{n}(x_i+y_i) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \left(\frac{1}{n}\sum_{j=1}^{n} y_i\right)$$
$$= \tilde{x} + \tilde{y}$$

(2):

$C \in \mathbb{R}$ (C is linear)

$$\widetilde{Cx} = \frac{1}{n}\sum_{i=1}^{n} Cx_i = C\cdot\frac{1}{n}\sum_{i=1}^{n} x_i$$
$$= C\tilde{x}$$

(3)

$$\underset{\sim}{1} = (1 \cdot 1 \dots 1)^T \Big\}\, n \text{ times}$$

$$\underset{\sim}{x} + C \times \underset{\sim}{1} = \left((x_1+C) + (x_2+C) \dots + (x_n+C)\right)^T$$
$$= \underset{\sim}{x} + C$$

$$T = \frac{1}{n}\sum_{i=1}^{n} \cdot 1 = \frac{n}{n} = 1$$

Sample mean

Properties

Let $\mathbf{x}, \mathbf{y}$ be two vectors of the same length n and $c \in \mathbb{R}$.

(1) - $\overline{\mathbf{x} + \mathbf{y}} = \bar{\mathbf{x}} + \bar{\mathbf{y}}$
(2) - $\overline{c\mathbf{x}} = c\bar{\mathbf{x}}$
(3) - Hence, the sample mean is a linear map from $\mathbb{R}^n$ to $\mathbb{R}$.

Example

- $\mathbf{x}$ temperature values in degree Celsius
- $\mathbf{x} + 237.5 \times \mathbf{1}$ temperature values in degree Kelvin
- Hence, the mean temperature in Kelvin is

$$\overline{\mathbf{x} + 237.5 \times \mathbf{1}} = \bar{\mathbf{x}} + \overline{237.5 \times \mathbf{1}} = \bar{\mathbf{x}} + 237.5 \times \bar{\mathbf{1}} = \bar{\mathbf{x}} + 237.5$$

Notation

$\mathbf{1} = (1, 1, \ldots, 1)^\top$ with length n (sample size).

Sample mean

Remark

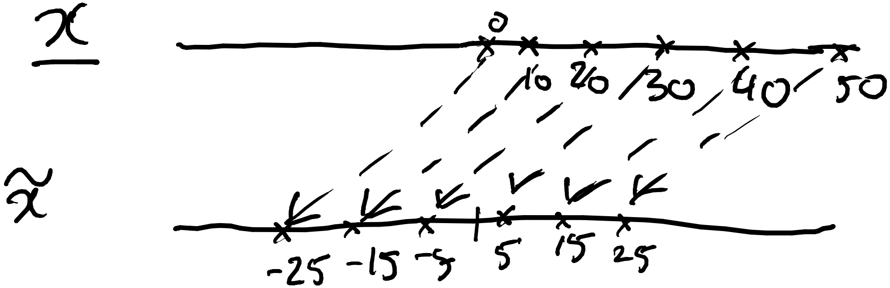
The sample mean is not necessarily a good guess for future observations.

Example

- The sample mean for a fair dice will be about 3.5. This is an impossible result for future observations.
- However, the average daily max temperature in Christchurch is 16.8°C. We have many days with a similar max temperature.

Centering

$$\tilde{x} = \underline{x} - \bar{x}\,\underline{1}$$



Center of gravity is 0

$$\bar{\tilde{x}} = \overline{\underline{x} - \bar{x}\,\underline{1}} = \bar{x} - \overline{\bar{x}\,\underline{1}}$$

$$= \bar{x} - \bar{x}\,\underset{\substack{\uparrow \\ 1}}{\bar{1}} = \bar{x} - \bar{x} = 0$$

• Last Lecture

- random variables
- sample $x_1, x_2, \dots x_n$
- data vector $\underline{x} = (x_1 \dots x_n)^T$
- sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- centred data vector

$\tilde{x} = \underline{x} - \bar{x}\perp$ # $\perp = (1, 1, \dots 1)^T$

$c \in \mathbb{R}$ $\underline{x} \in \mathbb{R}^n$

$\underline{x} + c = \underline{x} + c\perp$

we want to
keep track of th
fact that this is a vector

Centering

It is sometimes more convenient to work with a sample that has mean 0.

Definition

We call

$$\widetilde{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}\mathbf{1} = (x_1 - \overline{\mathbf{x}}, x_2 - \overline{\mathbf{x}}, \ldots, x_n - \overline{\mathbf{x}})^\top$$

the **centred** data vector.

Example

Temperature $\mathbf{x} = (0, 10, 20, 30, 40, 50)^\top$

$$\begin{aligned} \overline{\mathbf{x}} &= \frac{1}{6}(0 + 10 + 20 + 30 + 40 + 50) \\ &= \frac{150}{6} = 25 \\ \widetilde{\mathbf{x}} &= \mathbf{x} - \overline{\mathbf{x}} = (-25, -15, -5, 5, 15, 25)^\top \end{aligned}$$

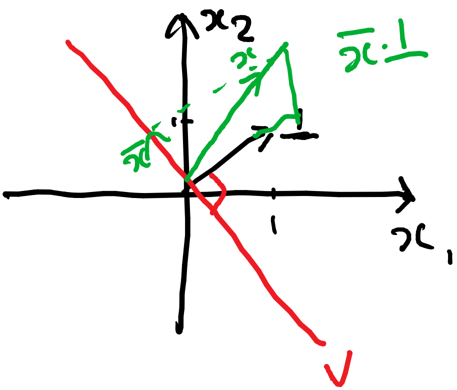
$$\tilde{\underline{x}} \cdot \underline{1} = (\underline{x} - \bar{x}\underline{1}) \cdot \underline{1}$$

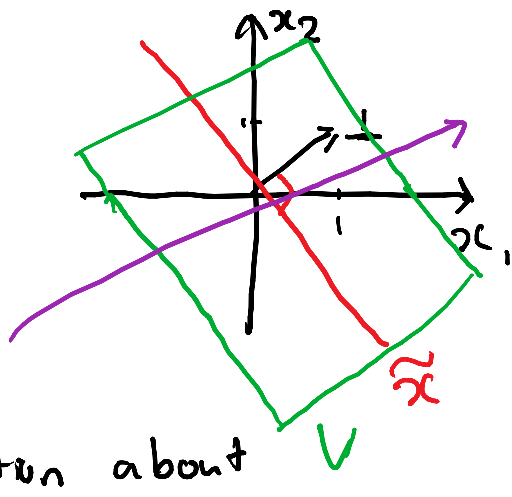
$$= \sum_{i=1}^{n} x_i - \bar{x}$$

$$= n \cdot \frac{1}{n} \sum_{i=1}^{n} x_i - \bar{x}$$

$$= n \bar{\tilde{x}} = 0$$

- Vector diagram





- $\bar{x}$ contains information about the mean.
- $\tilde{x}$ contains all other information.

Centering

A centred vector has two important properties.

Properties

1. Centred vectors have 0 mean

$$\overline{\widetilde{\mathbf{x}}} = 0.$$

2. Centred vectors are orthogonal to $\mathbf{1}$

$$\widetilde{\mathbf{x}} \cdot \mathbf{1} = 0.$$

Orthogonal decomposition

Centering and mean decompose $\mathbb{R}^n$ into two orthogonal subspaces

$$V = \left\{ \mathbf{x} : \sum_{i=1}^{n} x_i = 0 \right\} \qquad \text{and} \qquad V^{\perp} = span\{\mathbf{1}\}$$

with $dim(V) = n - 1$ and $dim(V^{\perp}) = 1$.

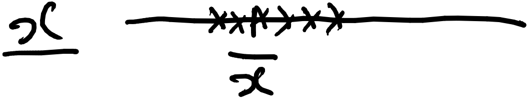
Hence, every $\mathbf{x} \in \mathbb{R}^n$ can be written as

$$\mathbf{x} = \underbrace{\mathbf{x} - \bar{\mathbf{x}}\mathbf{1}}_{= \widetilde{\mathbf{x}} \in V} + \underbrace{\bar{\mathbf{x}}\mathbf{1}}_{\in V^{\perp}}.$$

Degrees of freedom

- All centred vectors of length n form a $n - 1$ dimensional space.
- A centred vector of length n contains the same amount of information as a non-centred vector of length $n - 1$.
- We say $\widetilde{\mathbf{x}}$ has $n - 1$ **degrees of freedom**

- Spread



small spread

Large spread

1.3 Spread

The mean gives a rough idea where the data are. Now we want to describe how much the data vary between measurements, i.e. how they spread around the sample mean.

Sample variance

We measure the spread of $\mathbf{x}$ around the sample mean $\bar{\mathbf{x}}$ by the squared Euclidean norm of $\mathbf{x} - \bar{\mathbf{x}} = \widetilde{\mathbf{x}}$ normalised by its *degrees of freedom*.

Remember $\quad \widetilde{\mathbf{x}} \in V$ and $dim(V) = n - 1$.

variance = average distance from $\bar{x}$

Definition

The **sample variance** of $\mathbf{x}$ is

$\bar{x}$ = sample mean

$$\boxed{var(\mathbf{x}) = \frac{1}{n-1}\|\widetilde{\mathbf{x}}\|^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2.$$

Sample variance

Notation

- We will often just say *variance* instead of *sample variance*.
- A similar concept exists for random variables. It called the *population variance*.

Example

Temperature $\mathbf{x} = (0, 10, 20, 30, 40, 50)^\top$

$$\widetilde{\mathbf{x}} = (-25, -15, -5, 5, 15, 25)^\top$$

$$\|\widetilde{\mathbf{x}}\|^2 = 25^2 + 15^2 + 5^2 + 5^2 + 15^2 + 25^2 = 1750$$

$$var(\mathbf{x}) = \frac{1}{6-1}\|\widetilde{\mathbf{x}}\|^2 = \frac{1750}{5} = 350.$$

- **Check**

$$\mathrm{Var}(\underline{x} + c\underline{1}) = \frac{1}{n-1} \| \widetilde{\underline{x} + c\underline{1}} \|^2$$

$$\widetilde{\underline{x} + c\underline{1}} = \underline{x} + c\underline{1} - (\overline{\underline{x} + c\underline{1}})\underline{1}$$

$$= \underline{x} + c\underline{1} - (\bar{x} + c)\underline{1}$$

$$= \underline{x} - \bar{x}\underline{1} = \tilde{\underline{x}}$$

$$\mathrm{var}(\underline{x} + c\underline{1}) = \mathrm{var}(\underline{x})$$

$$\mathrm{var}(c\underline{x}) = \frac{1}{n-1} \| c\underline{x} \|^2$$

$$= \frac{1}{n-1} c^2 \| \underline{x} \|^2$$

$$= c^2 \,\mathrm{var}(\underline{x})$$

Sample variance

Properties

Let $c \in \mathbb{R}$.

- $var(\mathbf{x} + c\mathbf{1}) = var(\mathbf{x})$ translation invariance
- $var(c\mathbf{x}) = c^2 var(\mathbf{x})$
- The smaller the variance the closer are $x_1, x_2, \ldots, x_n$ to the sample mean $\bar{\mathbf{x}}$.

Example

Assume $\mathbf{x}$ are some distances measured in metre and $\mathbf{y}$ are the same distances measured in centimetre.

- $\bar{\mathbf{y}} = 100\bar{\mathbf{x}}$ # fixed by standard deviation
- $var(\mathbf{y}) = 10000 var(\mathbf{x})$

$$sd(x) = \sqrt{var(c\underline{x})} = \sqrt{\frac{1}{n-1}\|c\underline{x}\|^2}$$

$$= \sqrt{\frac{1}{n-1} c^2 \|\underline{x}\|^2}$$

$$= \sqrt{c^2 \, var(\underline{x})}$$

$$var(\underline{x}) = \sigma^2_x \qquad \sigma = \text{sigma}$$

$$sd(x) = \sigma_x$$

Standard deviation

The quadratic change of the variance when changing the units of measurements is counterintuitive and does sometimes cause problems.

just square root the variance

Definition The **standard deviation** of $\mathbf{x}$ is

$$sd(\mathbf{x}) = \frac{1}{\sqrt{n-1}} \|\widetilde{\mathbf{x}}\| = \sqrt{var(\mathbf{x})}.$$

Properties Let $c \in \mathbb{R}$.

- $sd(\mathbf{x} + c\mathbf{1}) = sd(\mathbf{x})$ translation invariance
- $sd(c\mathbf{x}) = |c| sd(\mathbf{x})$

The standard deviation is in the same units of measurement as the data $\mathbf{x}$. Changing from metre to centimetre results in multiplication by 100.

Sample variance and standard deviation

Notation

Some textbooks use the following notation

- $s_{\mathbf{x}}^2 = var(\mathbf{x})$
- $s_{\mathbf{x}} = sd(\mathbf{x})$.

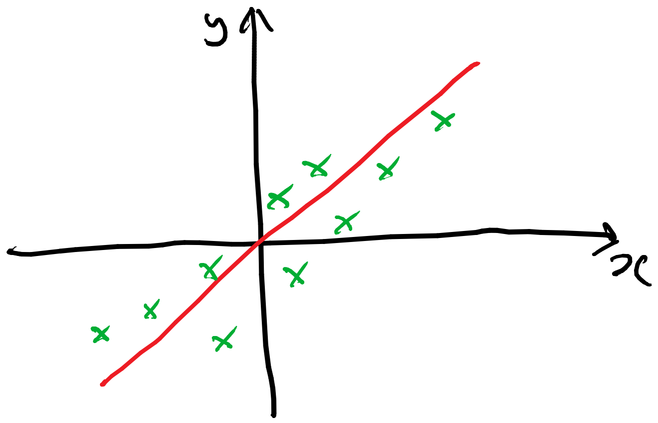
Remark

Other measures for the spread such as

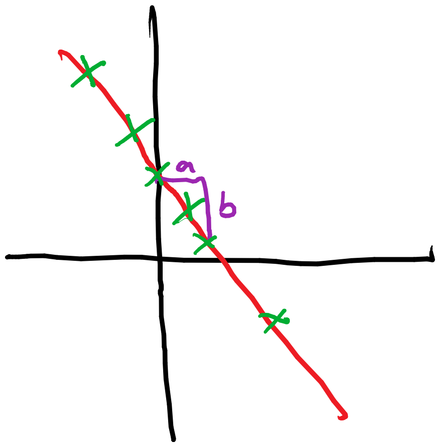
$$\frac{1}{n-1}\sum_{i=1}^{n}|x_i - \overline{\mathbf{x}}| = \frac{1}{n-1}\|\mathbf{x} - \overline{\mathbf{x}}\|_1 = \frac{1}{n-1}\|\widetilde{\mathbf{x}}\|_1$$

are possible. But they would lead to complicated formulas next week! In most fields of statistics *var* and *sd* are used for convenience. However, in image processing and big data other measures of the spread are sometimes useful.

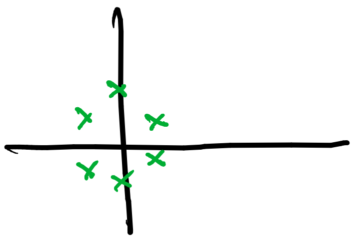
• Scatter plot



• linear relationship



• non-linear relationships



• for centred vectors

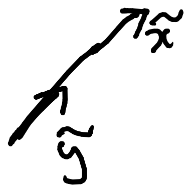
$$\tilde{y} = \underline{y} - \bar{y}\underline{1} = (b\underline{x} + a\underline{1}) - \overline{(b\underline{x} + a\underline{1})}\,\underline{1}$$

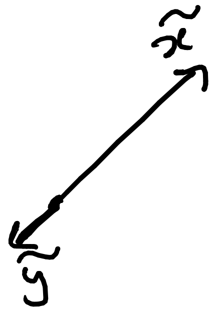
$$= b\underline{x} + \cancel{a\underline{1}} - b\bar{x}\underline{1} - \cancel{a\underline{1}}$$

$$= b\underline{x} - b\bar{x}$$

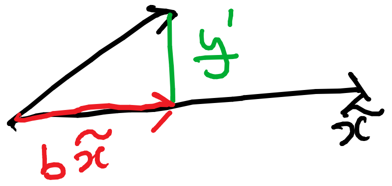
$$= b\tilde{x}$$

• vector diagram





• perfect linear relation



1.4 Covariance and correlation

We want to quantify how two data vectors $\mathbf{x}$ and $\mathbf{y}$ of the same length n behave together. Can $\mathbf{x}$ predict $\mathbf{y}$ in a linear way?

$$\mathbf{y} = b\mathbf{x} + a\mathbf{1}, \qquad a, b \in \mathbb{R}$$

It is easier to work with the centred vectors. If the linear equation holds, then

$$\widetilde{\mathbf{y}} = b\widetilde{\mathbf{x}}.$$

The $a\mathbf{1}$ cancels out.

Such a perfect linear relation is often too much to ask for. But

$$\widetilde{\mathbf{y}} = b\widetilde{\mathbf{x}} + \mathbf{y}' \qquad \text{with } \mathbf{y}' \in span\{\widetilde{\mathbf{x}}\}^{\perp}$$

holds.

Idea The b quantifies the linear dependency while $\mathbf{y}'$ is a component of $\widetilde{\mathbf{y}}$ which is linear independent of $\widetilde{\mathbf{x}}$.

check

$$\mathrm{cov}(\underline{x}, \underline{y}) = \frac{1}{n-1} \tilde{x} \cdot \tilde{y} = \frac{1}{n-1} \tilde{x} \left(b\tilde{x} + y' \right)$$

with $y' \cdot \tilde{x} = 0$

$$= \frac{1}{n-1} \tilde{x} \cdot b\tilde{x} + \underbrace{\tilde{x} \cdot y'}_{\substack{=0 \\ \text{(orthogonal)}}}$$

$$= b \frac{1}{n-1} \tilde{x} \cdot \tilde{x}$$

$$= b \frac{1}{n-1} \|\tilde{x}\|^2 = b \, \mathrm{var}(\underline{x})$$

Sample covariance

Definition

dot product
- projection (orthogonal)

The **sample covariance** of $\mathbf{x}$ and $\mathbf{y}$ is

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \widetilde{\mathbf{x}} \cdot \widetilde{\mathbf{y}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})(y_i - \overline{\mathbf{y}})$$

Notation

- We will often say *covariance* instead of *sample covariance*.
- Some textbooks use the notation $c_{\mathbf{x},\mathbf{y}} = cov(\mathbf{x}, \mathbf{y})$.

Properties Let $c \in \mathbb{R}$

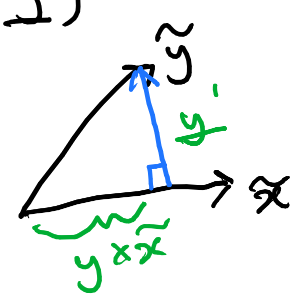
- $cov(\mathbf{x}, \mathbf{y}) = b \times var(\mathbf{x})$
- $cov(\mathbf{x}, \mathbf{x}) = var(\mathbf{x})$
- $cov(\mathbf{x}, \mathbf{y}) = cov(\mathbf{y}, \mathbf{x})$ symmetry
- $cov(c\mathbf{x}, \mathbf{y}) = c \times cov(\mathbf{x}, \mathbf{y}) = cov(\mathbf{x}, c\mathbf{y})$
- $cov(\mathbf{x} + c\mathbf{1}, \mathbf{y}) = cov(\mathbf{x}, \mathbf{y}) = cov(\mathbf{x}, \mathbf{y} + c\mathbf{1})$
  translation invariance

• Last lecture

- $\tilde{x}$ & $\perp$ are orthogonal
- $x = \tilde{x} + \bar{x}\perp$
- $\tilde{x}$ has $n-1$ degrees of freedom
- $\mathrm{var}(\underline{x}) = \frac{1}{n-1} \|\underline{x}\|^2$

  $\mathrm{var}(c\underline{x}) = c^2 \mathrm{var}(\underline{x})$
- $\mathrm{sd}(\underline{x}) = \sqrt{\mathrm{var}(\underline{x})}$

  $\mathrm{sd}(c\underline{x}) = |c|\, \mathrm{sd}(\underline{x})$
- $\mathrm{cov}(\underline{x}\,\underline{y})$



check properties

$$\mathrm{cov}(\underline{x}, \underline{x}) = \frac{1}{n-1}\, \tilde{x} \cdot \tilde{x} = \frac{1}{n-1} \|\tilde{x}\|^x$$

$$= \mathrm{var}(\underline{x})$$

$$\mathrm{cov}(\underline{x}, \underline{y}) = \frac{1}{n-1}\, \tilde{x} \cdot \tilde{y} = \frac{1}{n-1}\, \tilde{y} \cdot \tilde{x} = \mathrm{cov}(\underline{y}, \underline{x})$$

$$\mathrm{cov}(c\underline{x}, \underline{y}) = \frac{1}{n-1}\, c\tilde{x} \cdot \tilde{y}$$

$$= c \frac{1}{n-1}\, \tilde{x} \cdot \tilde{y}$$

$$= c\, \mathrm{cov}(\underline{x}, \underline{y})$$

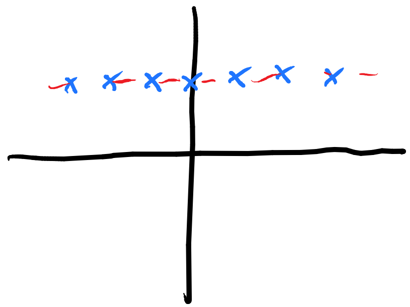
$$= \mathrm{cov}(\underline{x}, c\underline{y})$$

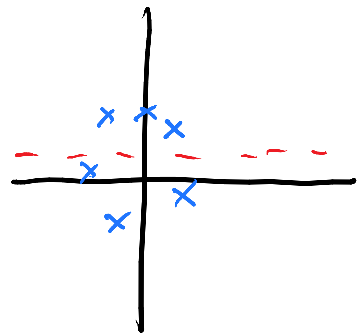
by symmetry

$$\mathrm{cov}(\underline{x} + c\perp + \underline{y}) = \mathrm{cov}(\underline{x}, \underline{y})$$

because $\widetilde{\underline{x} + c\perp} = \tilde{x}$

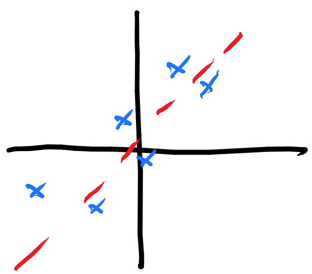
scatter plot $\qquad b = 0$

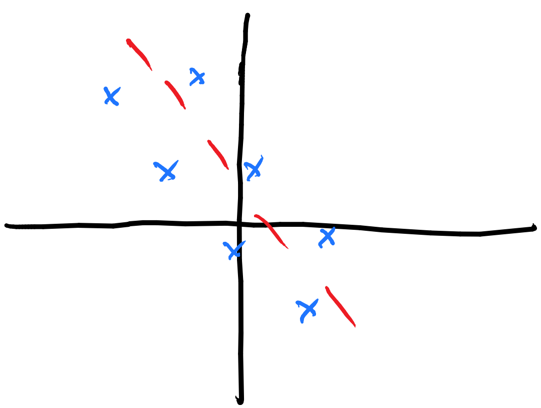




$\mathrm{cov}(\underline{x}, \underline{y}) = 0$

$b > 0 \qquad \mathrm{cov}(\underline{x}, \underline{y}) > 0$



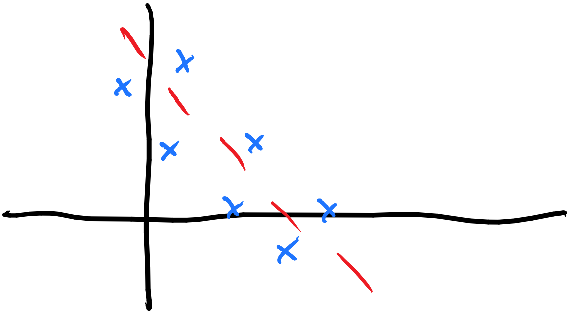


$b < 0 \qquad \mathrm{cov}(\underline{x}, \underline{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

V

$\cancel{V} = RI$

$\mathrm{var}(\underline{u}, \underline{i}) > 0$



$\mathrm{cov}(\underline{t}, \underline{e}) < 0$

Sample covariance

Interpretation

- The covariance quantifies the linear relation between $\mathbf{x}$ and $\mathbf{y}$.
- If they do not even have an approximately linear relation, $cov(\mathbf{x}, \mathbf{y}) = 0$. In this case $\mathbf{x} \cdot \mathbf{y} = 0$.
- If $\mathbf{x}$ and $\mathbf{y}$ tend to be above their respective means at the same time and below their respective means at the same time, then $cov(\mathbf{x}, \mathbf{y}) > 0$.
- If $\mathbf{x}$ and $\mathbf{y}$ tend to go into opposite directions relative to their their respective means, then $cov(\mathbf{x}, \mathbf{y}) < 0$.
- The absolute value of $cov(\mathbf{x}, \mathbf{y})$ depends on $var(\mathbf{x})$ and by symmetry also on $var(\mathbf{y})$.

Example

- Voltage and current in a wire will have positive covariance.
- Electricity consumption and temperature in Christchurch will have negative covariance.

Sample correlation

Let us remove the dependency on $var(\mathbf{x})$ and $var(\mathbf{y})$ from the covariance.

Definition The **sample correlation** of $\mathbf{x}$ and $\mathbf{y}$ is

$$corr(\mathbf{x}, \mathbf{y}) = \frac{cov(\mathbf{x}, \mathbf{y})}{sd(\mathbf{x})sd(\mathbf{y})} = \frac{\widetilde{\mathbf{x}} \cdot \widetilde{\mathbf{y}}}{\|\widetilde{\mathbf{x}}\| \|\widetilde{\mathbf{y}}\|} \; .$$

Notation

- We will often just say *correlation* instead of *sample correlation*.
- This concept is also called **Pearson correlation**.
- Some textbooks use the notation $r_{\mathbf{x},\mathbf{y}} = corr(\mathbf{x}, \mathbf{y})$.

- $\text{cov}(\underline{x},\underline{y})$ & $\text{corr}(\underline{x},\underline{y})$ have the same sign because we divide by $sd(\underline{x})sd(\underline{y})>0$
- corr is symmetric because cov is symmetric.
- $\text{corr}(\underline{x} + c\underline{1}, \underline{y}) = \text{corr}(\underline{x},\underline{y})$

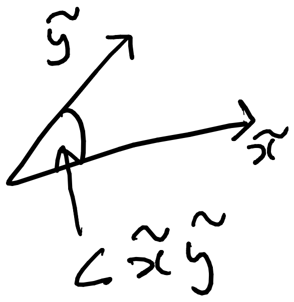
  because $\widetilde{\underline{x} + c\underline{1}} = \tilde{x}$

$$\text{corr}(c\underline{x},\underline{y}) = \frac{c\tilde{x}\cdot\tilde{y}}{\|c\tilde{x}\|\,\|\tilde{y}\|}$$

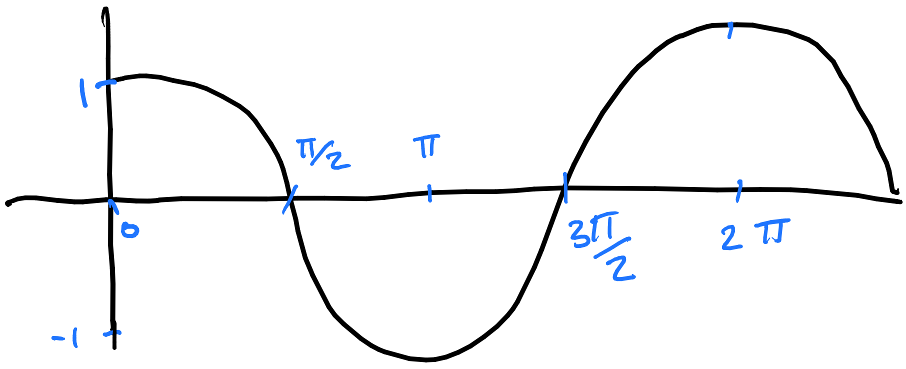
$$= \frac{c}{|c|}\,\frac{\tilde{x}\cdot\tilde{y}}{\|\tilde{x}\|\,\|\tilde{y}\|}$$

$$= \text{sign}(c)$$

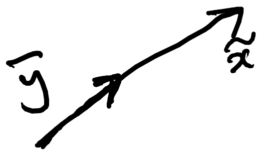
- $\tilde{x}\cdot\tilde{y} = \|\tilde{x}\|\,\|y\| \quad \cos(\angle\tilde{x}\tilde{y})$

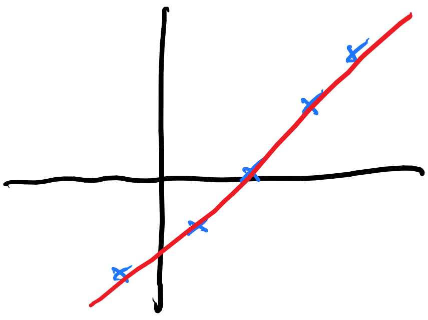


Hence,
$\text{corr}(\underline{x},\underline{y}) = \cos(\angle\tilde{x}\tilde{y})$



- $\text{corr}(\underline{x},\underline{y}) = 1 \rightarrow \angle\tilde{x},\tilde{y} = 0$

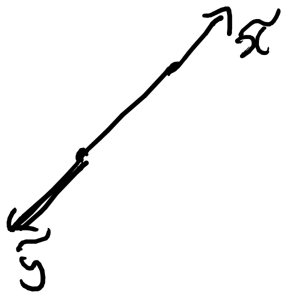


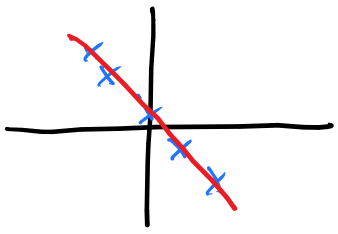


perfect linear relation
$\tilde{y} = b\tilde{x} \quad b>0$

- $\text{corr}(\underline{x},\underline{y}) = -1$

$\angle\tilde{x}\;\tilde{y} = \pi$ perfect linear relation
$\tilde{y} = b\tilde{x}$
$b<0$





Sample correlation

Properties Let $c \in \mathbb{R}$.

- $corr(\mathbf{x}, \mathbf{y})$ and $cov(\mathbf{x}, \mathbf{y})$ have the same sign. $corr(\mathbf{x}, \mathbf{y}) = 0$ if and only if $cov(\mathbf{x}, \mathbf{y}) = 0$.
- $corr(\mathbf{x}, \mathbf{y}) = corr(\mathbf{y}, \mathbf{x})$ symmetry
- $corr(\mathbf{x} + c\mathbf{1}, \mathbf{y}) = corr(\mathbf{x}, \mathbf{y}) = corr(\mathbf{x}, \mathbf{y} + c\mathbf{1})$ translation invariance
- $corr(c\mathbf{x}, \mathbf{y}) = sign(c) corr(\mathbf{x}, \mathbf{y}) = corr(\mathbf{x}, c\mathbf{y})$ scale invariance
- $corr(\mathbf{x}, \mathbf{y}) = \cos(\angle \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$
- $-1 \leq corr(\mathbf{x}, \mathbf{y}) \leq 1$.

Here $\angle \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}$ is the angle between $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{y}}$ and

$$sign(c) = \begin{cases} -1 & \text{if } c < 0 \\ 0 & \text{if } c = 0 \\ 1 & \text{if } c > 0. \end{cases}$$

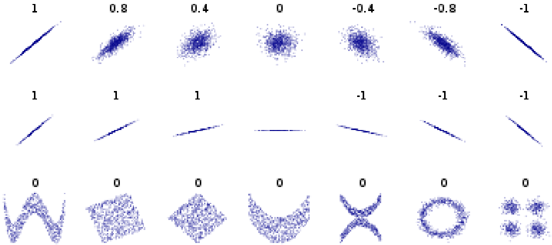
Sample correlation

Note

- $corr(\mathbf{x}, \mathbf{y}) = 1$ if $\widetilde{\mathbf{y}} = b\widetilde{\mathbf{x}}$ with $b > 0$, i.e. $\angle\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}} = 0$
- $corr(\mathbf{x}, \mathbf{y}) = -1$ if $\widetilde{\mathbf{y}} = b\widetilde{\mathbf{x}}$ with $b < 0$, i.e. $\angle\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}} = \pi$ (180°).

Interpretation

The correlation measures how perfect the linear relation $\widetilde{\mathbf{y}} = b\widetilde{\mathbf{x}} + \mathbf{y}'$ is.



Be careful when interpreting *cov* and *corr*! Only approximately linear relations are detected.

Covariance and correlation

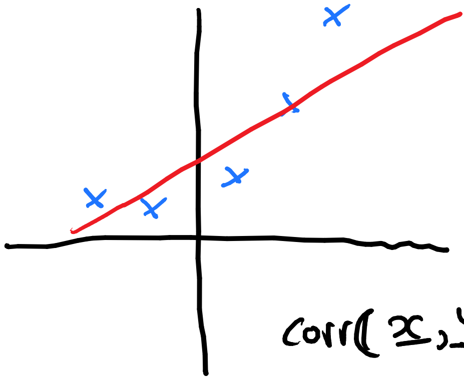
Example

Temperature and pressure in a boiler.

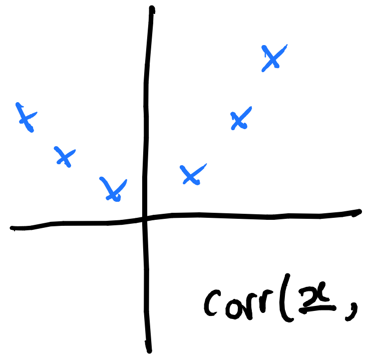
$$\widetilde{\mathbf{x}} = (-25, -15, -5, 5, 15, 25)^\top$$
$$\widetilde{\mathbf{y}} = (-10, -6, -1, 0, 6, 11)^\top$$
$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{6-1}\Big((-25)(-10) + (-15)(-6) + (-5)(-1) + (5)(0) + (15)(6) + (25)(11)\Big)$$
$$= \frac{710}{5} = 142$$

$$corr(\mathbf{x}, \mathbf{y}) = \frac{(-25)(-10) + (-15)(-6) + (-5)(-1) + (5)(0) + (15)(6) + (25)(11)}{\sqrt{25^2 + 15^2 + 5^2 + 5^2 + 15^2 + 25^2}\sqrt{10^2 + 6^2 + 1^2 + 0^2 + 6^2 + 11^2}}$$
$$= \frac{710}{\sqrt{1750}\sqrt{294}} \approx 0.9898$$

- Last Lecture
- $\text{cov}(\underline{x}, \underline{y}) = \dfrac{\tilde{x} \cdot \tilde{y}}{n-1}$
  - translation invariance
  - Symmetric
  - $\text{cov}(c\,\underline{x}, \underline{y}) = c\,\text{cov}(\underline{x}, \underline{y})$
- $\text{corr}(\underline{x}, \underline{y}) = \dfrac{\tilde{x} \cdot \tilde{y}}{\|\tilde{x}\|\|\tilde{y}\|} = \cos(\angle \tilde{x}, \tilde{y})$
  - translation invariance
  - Scale invariance
  - Symmetric



$\text{corr}(\underline{x}, \underline{y}) > 0$



$\text{corr}(\underline{x}, \underline{y}) = 0$

el - monthly electricity in kWh

temp - monthly mean temperature in °F

house in upstate NY

covariance matrix

$\text{cov}(el, tmp)$

$$\text{cov}(\underline{x}, \underline{y}) = \begin{pmatrix} \text{var}(\underline{x}) & \text{cov}(\underline{x}, \underline{y}) \\ \text{cov}(\underline{x}, \underline{y}) & \text{var}(\underline{y}) \end{pmatrix}$$

$°C = \frac{5}{9}(°F - 32)$ # Scaling