# Quantitative Finance

Lectures in Quantitative Finance

Spring Term 2022

*1. Basic probability theory & regression*

Prof. Dr. Erich Walter Farkas
walter.farkas@bf.uzh.ch

# Basic probability theory & regression

1. Basic probability theory
   - Probability spaces
   - Real-valued random variables
   - Moments
   - Joint cumulative distribution function
   - Independence
2. Appendix
   - Maximum Likelihood methods
   - The linear model
     - The concept of regression
     - Simple linear regression model
     - General linear regression model
     - Ordinary least squares
     - Maximum Likelihood
   - Complements
3. References

# Basic probability theory

# Probability space

The very base of our quantitative studies is a *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of

- a *sample space* $\Omega \neq \emptyset$,
- a *$\sigma$-algebra* $\mathcal{F} \subset \mathcal{P}(\Omega)$, a subset of the power set of $\Omega$, and
- a *probability measure* $\mathbb{P} : \mathcal{F} \to [0, 1]$.

# The sample space

The sample space $\Omega$ is an arbitrary non-empty set. The only assumption here is *non-emptiness*, but it is important and needs to be stated!

The elements of $\Omega$ are usually denoted by $\omega$ for the simple reason that ‘small omegas’ live in the ‘big omega’. We may give them subscripts, bars, hats, and so on, $\omega_1, \omega_k, \bar{\omega}, \hat{\omega}$.

In principle, you are free to use your imagination at the cost of transparency: $\odot, \star,$ 😐 $\in \Omega$. However, we strongly advise you against doing that. Some conventions are so firmly established that any deviation causes confusion.

# The power set

By $\mathcal{P}(\Omega)$ or $2^{\Omega}$ we denote the *power set* of $\Omega$. This is the set of all subsets of $\Omega$, including the empty set $\emptyset$ and $\Omega$.

**Example.** Consider the set $S = \{a, b, c\}$. All possible strict subsets are $\{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}$. Including the empty set and $S$ itself, we get the power set of $S$,

$$\mathcal{P}(S) = \big\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\big\}\,.$$

Notice that $S$ has $|S| = 3$ elements and $|\mathcal{P}(S)| = 8 = 2^3$. This is true in general for any *finite set*: if $|S| = n \in \mathbb{N}$, then $|\mathcal{P}(S)| = 2^n$ (exercise). That is the reasoning behind the notation $\mathcal{P}(S) = 2^S$.

# A $\sigma$-algebra

Let $\Omega$ be some non-empty set. A *$\sigma$-algebra* or *$\sigma$-field* $\mathcal{F}$ over $\Omega$ is a subset of the power set $\mathcal{P}(\Omega)$ which satisfies the following three properties:

- $\Omega \in \mathcal{F}$,
- if $A \in \mathcal{F}$, then $A^{\complement} := \Omega \setminus A := \{\omega \in \Omega \mid \omega \notin A\} \in \mathcal{F}$ (we say $\mathcal{F}$ is closed under complementation),
- if $A_1, A_2, \ldots \in \mathcal{F}$, then $\cup_{k=1}^{\infty} A_k \in \mathcal{F}$ (we say $\mathcal{F}$ is closed under countable unions).

**Exercise.** Using the defining properties above, show that

- $\emptyset \in \mathcal{F}$, and
- if $A_1, A_2, \ldots \in \mathcal{F}$, then $\cap_{k=1}^{\infty} A_k \in \mathcal{F}$ (use De Morgan's laws).

# $\sigma$-algebra : Intuition

The $\sigma$-algebra is just a definition of which sets may be considered as events.

Elements not in $\mathcal{F}$ simply have no defined probability measure.

Why are they important?

- It is not always possible to assign a measure to **all** subsets of $\Omega$,
- Keeping track of $\sigma$-algebras allows us to formulate some key concepts in probability very elegantly!

# (General) measures

Let $\Omega$ be a non-empty set and let $\mathcal{F}$ be a $\sigma$-algebra over $\Omega$.

We call a function $\mu : \mathcal{F} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ a *measure* if it satisfies

- $\mu(\emptyset) = 0$,
- *non-negativity*: for all $F \in \mathcal{F}$ we have $\mu(F) \geq 0$, and
- *countable additivity ($\sigma$-additivity)*: for any *countable collection* $\{F_k\}_{k=1}^{\infty} \subset \mathcal{F}$ of *pairwise disjoint sets* we have $\mu(\cup_{k=1}^{\infty} F_k) = \sum_{k=1}^{\infty} \mu(F_k)$.

In general, we call a set together with a $\sigma$-algebra over that set, in our case the pair $(\Omega, \mathcal{F})$, a *measurable space*, since we can define a measure on it.

Elements of $\mathcal{F}$ are called *measurable sets*.

The triplet $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*.

# Probability measures

Let $\Omega$ be a non-empty set and let $\mathcal{F}$ be a $\sigma$-algebra over $\Omega$.

We call a function $\mathbb{P} : \mathcal{F} \to [0,1]$ a *probability measure* if it satisfies

- ▷ $\mathbb{P}[\emptyset] = 0$ and $\mathbb{P}[\Omega] = 1$, and
- ▷ $\mathbb{P}$ is *countably additive*.

Notice that a probability measure is a special kind of measure with total measure of one.

For a probability measure $\mathbb{P}$, we call the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space*.

# Real-valued random variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{B}(\mathbb{R})$ be the *Borel $\sigma$-algebra* on $\mathbb{R}$. This is (without going into much detail here) the smallest $\sigma$-algebra on $\mathbb{R}$ which *contains all the intervals*.

In particular, sets of the form $(-\infty, x]$ for some $x \in \mathbb{R}$ are elements of $\mathcal{B}(\mathbb{R})$.

A *real-valued random variable* is a *measurable* function $X \colon \Omega \to \mathbb{R}$. This means that for every set $B \in \mathcal{B}(\mathbb{R})$ the preimage $X^{-1}(B) := \{\omega \in \Omega \mid X(\omega) \in B\}$ is an element of $\mathcal{F}$.

To emphasise the dependence of measurability on the underlying $\sigma$-algebra, authors also write $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

# Real-valued random variables

The *probability* that $X$ takes on a value in a specified measurable set $B \in \mathcal{B}(\mathbb{R})$ is written as

$$\mathbb{P}[\{\omega \in \Omega \mid X(\omega) \in B\}] .$$

Most of the time, we will abbreviated the set $\{\omega \in \Omega \mid X(\omega) \in B\}$ by $\{X \in B\}$ and hence, we will write $\mathbb{P}[X \in B]$.

Notice that $X^{-1}(B) = \{X \in B\}$ and that already here the notion of measurability is essential: the set $\{X \in B\}$ has to be an element of $\mathcal{F}$ so that we can measure it!

# Cumulative distribution function (cdf)

Every *real-valued* random variable $X$ can be described by its *cumulative distribution function (cdf)* $F_X : \mathbb{R} \to [0,1]$ defined as

$$F_X(x) = \mathbb{P}[X \leq x]\,.$$

Notice that, as mentioned above, $\{X \leq x\} = X^{-1}((-\infty, x]) \in \mathcal{F}$.

**Exercise.** Argue that any cdf $F$ is *increasing* and that

$$\lim_{x\to-\infty} F(x) = 0 \quad \text{and} \quad \lim_{x\to+\infty} F(x) = 1\,.$$

# Discrete random variables

We call a real-valued random variable $X : \Omega \to S$ for some $S \subset \mathbb{R}$ *discrete* if its image is *finite* or *countably infinite*.

The function $f_X : S \to [0,1]$ which gives the probability that $X$ is exactly equal to a certain value in its image is called *probability mass function (pmf)* (not pdf!),

$$f_X(x) = \mathbb{P}[X = x] \text{ for } x \in S\,.$$

Notice that we can assume without loss of generality (w.l.o.g.) that these probabilities are non-zero if we restrict $S$ to the image of $X$.

Since the image of $X$ is at most countably infinite, we can also enumerate the corresponding probabilities. That means for every $x_k$ in the image of $X$ we can define $p_k = \mathbb{P}[X = x_k] = f_X(x_k)$.

**Exercise.** Argue that $\sum_{x \in S} f_X(x) = \sum_{k=1}^{\infty} p_k = 1$.

# Discrete random variables

**Exercise.** Let $\Omega = \{\omega_1, \ldots, \omega_6\}$ describe all possible states after rolling a six faced dice, that is $\omega_1 = \{$After rolling the dice, it shows '1'$\}$, ..., $\omega_6 = \{$After rolling the dice, it shows '6'$\}$.

The discrete random variable which gives us the *value of the dice* is defined as $X(\omega_1) = 1, \ldots, X(\omega_6) = 6$.

We assume that the dice is perfect, that means every face has the same probability, $\mathbb{P}[\omega_1] = \ldots = \mathbb{P}[\omega_6] = \frac{1}{6}$.

Write down the cumulative distribution function $F_X(x) = \mathbb{P}[X \leq x]$ of $X$. Recall that $\{X \leq x\} = \{\omega \in \Omega \mid X(\omega) \leq x\}$.

# Continuous random variables

We say $X$ is a *continuous random variable* if its cdf is continuous. In particular, this implies that the image of $X$ is *uncountably infinite*.

If $X$ is *absolutely continuous*, then $X$ admits a *probability density function (pdf)* $f_X$ which assigns probabilities to intervals.

We have

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f_X(x)\mathrm{d}x\,,$$

in particular, we have

$$F_X(b) = \int_{-\infty}^b f_X(x)\mathrm{d}x\,.$$

# Continuous random variables

**Example.** Let $X$ be (continuous) uniformly distributed on the closed interval $[a,b]$ for $-\infty < a < b < \infty$. We write $X \sim \mathcal{U}(a,b)$. By definition, the pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b\,, \\ 0 & \text{else.} \end{cases}$$

Hence, the cdf is given as

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t = \begin{cases} 0 & \text{for } x < a\,, \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b\,, \\ 1 & \text{for } x > b\,. \end{cases}$$

# Continuous random variables

**Exercise.** Let $X$ be a absolutely continuous random variable with cdf $F_X$ and pdf $f_X$.

- ▷ Show that $F'_X(x) = f_X(x)$ (use the *first fundamental theorem of calculus*).
- ▷ Argue that $\mathbb{P}[X = x] = 0$ for any $x \in \mathbb{R}$.

# Expected value

The *expected value (or expectation, mean, average, first moment)* of a random variable $X$ intuitively characterises the central tendency or long-run average value.

For absolutely continuous random variables it is defined via the integral

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, \mathrm{d}x\,.$$

In the discrete case, the integral becomes a sum

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} x_k \mathbb{P}[X = x_k] = \sum_{k=1}^{\infty} x_k p_k\,.$$

# Expected value

**Example.** Let $X \sim \mathcal{U}(a,b)$, then the expected value of $X$ is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\mathrm{d}x = \frac{x^2}{2(b-a)}\bigg|_{x=a}^{b}$$
$$= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}\,.$$

Now let $X$ denote the value of a perfect dice as in the example above. The expected value of $X$ is

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} x_k \mathbb{P}[X = x_k] = \sum_{k=1}^{6} k\frac{1}{6} = \frac{21}{6} = 3.5\,.$$

# Expected value: Linearity

**Exercise.** Show that the expectation is a *linear functional*. This means that for any random variables $X, Y$ and any real number $c \in \mathbb{R}$ we have

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and
- $\mathbb{E}[cX] = c\mathbb{E}[X]$.

# Variance and standard deviation

The *variance* of a random variable is the expectation of the squared deviation from its mean. It is defined as

$$\mathrm{Var}(X) = \mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \,.$$

The *standard deviation (std)* $\sigma(X)$ of a random variable $X$ is the square root of its variance,

$$\sigma(X) = \sqrt{\mathbb{V}[X]} \,.$$

# Variance and standard deviation

**Exercise.** Let $X$ be a random variable and let $c \in \mathbb{R}$. Show that

- ▷ $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ (use linearity of the expected value),
- ▷ $\mathbb{V}[X + c] = \mathbb{V}[X]$,
- ▷ $\mathbb{V}[cX] = c^2\mathbb{V}[X]$.

Find an example of two random variables $X, Y$ such that

$$\mathbb{V}[X + Y] \neq \mathbb{V}[X] + \mathbb{V}[Y] .$$

*Hint*: Make it simple by taking for example $\Omega = \{\omega_1, \omega_2\}$ and defining $X(\omega_1) = 1$ and $X(\omega_2) = 0$. What could be a possible definition for $Y$?

# Variance and standard deviation

**Example.** Let $X \sim \mathcal{U}(a, b)$, then the variance of $X$ is

$$\begin{aligned}\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \int_a^b \frac{x^2}{b-a}\mathrm{d}x - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{x^3}{3(b-a)}\Big|_{x=a}^{b} - \left(\frac{a+b}{2}\right)^2 = \frac{b^3-a^3}{3(b-a)} - \left(\frac{a+b}{2}\right)^2 .\end{aligned}$$

We can now expand the fraction so that we have the same denominator and get

$$\begin{aligned}\mathbb{V}[X] &= \frac{4(b^3-a^3) - 3(b-a)(a+b)^2}{12(b-a)} \\ &= \frac{b^3+3ba^2-3ab^2-a^3}{12(b-a)} = \frac{(b-a)^3}{12(b-a)} = \frac{(b-a)^2}{12} .\end{aligned}$$

# Variance and standard deviation

**Example.** Let $X$ denote the value of a dice again. The variance of $X$ is given by

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sum_{k=1}^{6} k^2 \frac{1}{6} - \left(\frac{21}{6}\right)^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2$$

$$= \frac{35}{12} = 2.91\bar{6}\,.$$

# Normal distribution

**Example.** One of the most prominent continuous probability distributions is the *normal or Gauss distribution*. We say a random variable $X$ is *normally distributed* with *mean (or location parameter)* $\mu$ and *variance (or scale parameter)* $\sigma^2$ and we write $X \sim \mathcal{N}(\mu, \sigma^2)$ if its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) .$$

We call an $X \sim \mathcal{N}(0, 1)$ a *standard* normal random variable.

For the density function of a standard normal random variable, by convention, we often use the symbol $\varphi$,

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) ,$$

and for its cumulative distribution function we use $\Phi$.

# Normal distribution

Notice that

$$f_X(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right).$$

**Exercise.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$. We call the location parameter $\mu$ the mean and the scale parameter $\sigma^2$ the variance of $X$. But are they really what their names suggest? Let's find out.

First of all, you have to either accept the fact, that

$$\int_{-\infty}^{\infty} e^{\frac{-x^2}{2\sigma^2}}\,\mathrm{d}x = \sqrt{2\pi\sigma^2},$$

or ask in the break or after the lecture how you can calculate this.

# Normal distribution - expected value

So we want to show that

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \mu\,.$$

We start with the substitution $z = x - \mu$ with $\mathrm{d}z = \mathrm{d}x$. We notice that the limits stay the same.

Hence, we get

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (z+\mu) e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \left( \underbrace{\int_{-\infty}^{\infty} z e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{=0} + \mu \underbrace{\int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{=\sqrt{2\pi\sigma^2}} \right) = \mu\,.\end{aligned}$$

Notice that the first integral evaluates to 0, since $ze^{-\frac{z^2}{2\sigma^2}}$ is an *odd function*.

# Normal distribution - variance

Next we want to show that

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu^2 = \sigma^2 \,.$$

So we have to calculate $\mathbb{E}[X^2]$. Using the same substitution as above, we get

$$\begin{aligned}
\mathbb{E}[X^2] &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (z+\mu)^2 e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \left( \mu^2 \underbrace{\int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{=\sqrt{2\pi\sigma^2}} + 2\mu \underbrace{\int_{-\infty}^{\infty} z e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{=0} + \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z \right) \\
&= \mu^2 + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z \,.
\end{aligned}$$

Hence, we can already intuit that the second term must evaluate to $\sigma^2$.

# Normal distribution - variance

To show this, we want use a second substitution $x = z^2$ with $\mathrm{d}x = 2z\mathrm{d}z$. However, since we are integrating from $-\infty$, we cannot directly do that. Instead, we notice that the integrand is an *even function* and thus, we can split the integral in two equal parts at 0. We get

$$\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z = 2\int_0^{\infty} z^2 e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z = 2\int_0^{\infty} x e^{-\frac{x}{2\sigma^2}} \frac{1}{2\sqrt{x}} \mathrm{d}x$$

$$= \int_0^{\infty} \sqrt{x} e^{-\frac{x}{2\sigma^2}} \mathrm{d}x \,.$$

Next, we use integration by parts to get

$$\int_0^{\infty} \sqrt{x} e^{-\frac{x}{2\sigma^2}} \mathrm{d}x = -2\sigma^2 e^{-\frac{x}{2\sigma^2}} \sqrt{x} \Big|_{x=0}^{\infty} + \sigma^2 \int_0^{\infty} \frac{1}{\sqrt{x}} e^{-\frac{x}{2\sigma^2}} \mathrm{d}x \,.$$

Notice that the first term is equal to 0.

# Normal distribution - variance

Finally, we substitute again with $x = z^2$, $\mathrm{d}x = 2z\mathrm{d}z$ and arrive at

$$\sigma^2 \int_0^\infty \frac{1}{\sqrt{x}} e^{-\frac{x}{2\sigma^2}} \mathrm{d}x = \sigma^2 \int_0^\infty \frac{1}{z} 2z e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z = \sigma^2 \sqrt{2\pi\sigma^2}\,.$$

Plugging everything in, we get

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2\,.$$

We conclude that calling the location parameter $\mu$ 'mean' and the scale parameter $\sigma^2$ 'variance' is justified.

# Standard normal distribution

With the above results we can easily show that any normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ can be written in terms of a standard normally distributed random variable $Z \sim \mathcal{N}(0, 1)$ as

$$X = \mu + \sigma Z\,.$$

Indeed, we have

$$\mathbb{E}[\mu + \sigma Z] = \mu + \sigma \mathbb{E}[Z] = \mu = \mathbb{E}[X]$$

and

$$\mathbb{V}[\mu + \sigma Z] = \sigma^2 \mathbb{V}[Z] = \sigma^2 = \mathbb{V}[X]\,.$$

# Covariance

The *covariance* between two real-valued random variables $X$ and $Y$ (with finite second moment) is the expected value of the product of their deviations from their expected values,

$$\operatorname{cov}(X, Y) = \sigma_{XY} = \sigma(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \,.$$

Using the linearity of the expected value, we can simplify it to

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \,.\end{aligned}$$

# Pearson's correlation coefficient

*Pearson's correlation coefficient* is a measure of the *linear* correlation between two (non-constant) random variables $X$ and $Y$, and it is defined as

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

If two random variables are independent, then they are also (linearly) uncorrelated. It is a common misconception that the reverse is also true. *Uncorrelatedness does not imply independence!* Bear this in mind, we will define independence below.

**Exercise.** Let $X$ and $Y$ be two random variables. Show that Pearson's correlation coefficient always lies in the interval $[-1, 1]$. Use the Cauchy-Schwarz inequality.

# Variance: sum of random variables

We have already seen, that the expected value is linear and that the variance, in general, is not. Indeed, the variance of the sum of two random variables is given by

$$\begin{aligned}\mathbb{V}[X+Y] &= \mathbb{E}[(X+Y)^2]-\mathbb{E}[X+Y]^2\\ &= \mathbb{E}[X^2+2XY+Y^2]-\mathbb{E}[X]^2-2\mathbb{E}[X]\mathbb{E}[Y]-\mathbb{E}[Y]^2\\ &= \mathbb{E}[X^2]-\mathbb{E}[X^2]+\mathbb{E}[Y^2]-\mathbb{E}[Y]^2+2(\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y])\\ &= \mathbb{V}[X]+\mathbb{V}[Y]+2\mathrm{cov}(X,Y)\,.\end{aligned}$$

Note that the variance of a sum is the sum of variances only if the variables are uncorrelated.

# Expectation and variance of a portfolio

We can extend these results to a general weighted sum of random variables. Let $\{X_k\}_{k=1}^n$ be a family of random variables. Define

$$Y = \sum_{k=1}^{n} w_k X_k \,.$$

This could represent a portfolio of assets, where $X_k$ represents, for example, the return of asset $k$. We assume that $\sum_{k=1}^n w_k = 1$, that means we invest all our wealth in the portfolio.

By linearity of the expected value, we immediately get that the expected value of the weighted sum is the weighted sum of the expected values,

$$\mathbb{E}[Y] = \mathbb{E}\Big[\sum_{k=1}^{n} w_k X_k\Big] = \sum_{k=1}^{n} \mathbb{E}[w_k X_k] = \sum_{k=1}^{n} w_k \mathbb{E}[X_k]\,.$$

# Expectation and variance of a portfolio

Calculating the variance, we need to be more careful.

$$\begin{aligned}
\mathbb{V}[Y] = \mathbb{V}\Big[\sum_{k=1}^{n} w_k X_k\Big] &= \mathbb{E}\Big[\Big(\sum_{k=1}^{n} w_k X_k\Big)^2\Big] - \mathbb{E}\Big[\sum_{k=1}^{n} w_k X_k\Big]^2 \\
&= \mathbb{E}\Big[\sum_{k=1}^{n}\sum_{\ell=1}^{n} w_k w_\ell X_k X_\ell\Big] - \sum_{k=1}^{n}\sum_{\ell=1}^{n} w_k w_\ell \mathbb{E}[X_k]\mathbb{E}[X_\ell] \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} w_k w_\ell \big(\mathbb{E}[X_k X_\ell] - \mathbb{E}[X_k]\mathbb{E}[X_\ell]\big)\,. \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} w_k w_\ell \operatorname{cov}(X_k, X_\ell) \\
&= \sum_{k=1}^{n} w_k^2 \mathbb{V}[X_k] + 2\sum_{k<\ell} w_k w_\ell \operatorname{cov}(X_k, X_\ell)\,.
\end{aligned}$$

# Expectation and variance of a portfolio

Since this is a lot to write, we want to abbreviate it and use matrix notation. For this we define the *weight vector* $w = (w_1, \ldots, w_n)^\intercal$ and the *covariance matrix*

$$\Sigma = \begin{pmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & \cdots & \operatorname{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots & \operatorname{cov}(X_n, X_n) \end{pmatrix}$$

Now we can compactly write the variance of the sum as

$$\mathbb{V}[Y] = w^\intercal \Sigma w\,.$$

**Exercise.** Make sure that $w^\intercal \Sigma w$ is indeed equal to $\sum_{k=1}^n \sum_{\ell=1}^n w_k w_\ell \operatorname{cov}(X_k, X_\ell)$.

# Skewness

The *skewness* of a real-valued random variable is a measure of the asymmetry of the distribution about its mean. It is defined as the third standardised moment

$$\gamma_1(X) = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}}\right)^3\right] = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right].$$

We differentiate between

- negative skewness: long left tail (high probability of observing large negative values), and
- positive skewness: long right tail (high probability of observing large positive values).

# Kurtosis

The *kurtosis* of a real-valued random variable is (roughly speaking) a measure of the width of its tails. It is defined as the fourth standardised moment,

$$\mathrm{Kurt}[X] = \gamma_2(X) = \kappa(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

- ▷ A distribution with kurtosis of 3 is considered average and it is called *mesokurtic*. Any normally distributed random variable has kurtosis of 3. In fact, *excess kurtosis* is defined as kurtosis minus 3.
- ▷ A distribution with kurtosis greater than 3 is called *leptokurtic*. It indicates fatter tails.
- ▷ A distribution with kurtosis smaller than 3 is called *platykurtic*. It indicates thinner tails.

# Kurtosis of a normal random variable

**Exercise.** As mentioned above, the kurtosis of any normally random variable is equal to 3. Using the fact that $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ and the 'tricks' we used to calculate the expectation and variance, prove this result.

*Hint:* First use the substitution $z = x^2$, then use integration by parts, and finally substitute back.

# Importance of skewness and kurtosis

Two types of risks emerge:

- ▷ **Skewness risk**: risk that results when observations are not spread symmetrically around the average value. Ignoring this will cause the model to understate the risk of variables with high skewness.
- ▷ **Kurtosis risk**: kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns.

Not taking them into account can thus severely impare our model.

# A gambler and his two dice

**Example.** A gambler wants to know the expected value of the outcome of throwing two dice simultaneously.

- ▷ He writes down all possible outcomes: $x_k \in \{2, 3, \dots, 12\}$.
- ▷ He determines the probabilities $f(x_k)$ as all possibilities to achieve $x_k$ divided by the total number of outcomes, which is $6^2 = 36$. We get

$$\begin{array}{ll}
f(2) = \dfrac{1}{36} = 0.02\overline{7} & f(7) = \dfrac{6}{36} = 0.1\overline{6} \\
f(3) = \dfrac{2}{36} = 0.0\overline{5} & f(8) = \dfrac{5}{36} = 0.13\overline{8} \\
f(4) = \dfrac{3}{36} = 0.08\overline{3} & \quad\quad\vdots \\
f(5) = \dfrac{4}{36} = 0.\overline{1} & f(11) = \dfrac{2}{36} = 0.0\overline{5} \\
f(6) = \dfrac{5}{36} = 0.13\overline{8} & f(12) = \dfrac{1}{36} = 0.02\overline{7}
\end{array}$$

# A gambler and his two dice

- ▷ Then he computes all values $x_k f(x_k)$. The first values are:

$$\begin{aligned} x_1 f(x_1) &= 2 \cdot f(2) = 0.0\overline{5} \\ x_2 f(x_2) &= 3 \cdot f(3) = 0.1\overline{6} \\ x_3 f(x_3) &= 4 \cdot f(4) = 0.\overline{3}\,, \end{aligned}$$

  and so on.

- ▷ He sums up across all events and calculates the mean as

$$\mu = \sum_{k=1}^{12} x_k f(x_k) = 7\,.$$

  He notes, that this is also the median, since the distribution is symmetrical.

# A gambler and his two dice

**Exercise.** By now, he is so intrigued by all these calculations, that he does not want to stop there. He decides to also calculate the variance, the skewness, and the kurtosis. He wonders if you can check and verify his results :)

- ▷ All terms $(x_k - \mu)^2 f(x_k)$ sum up to $\sum(x_k - \mu)^2 f(x_k) = 5.833$ or, taking the square root, $\sigma = 2.4152$.
- ▷ All terms $(x_k - \mu)^3 f(x_k)$ sum up to zero, since for each entry with a positive deviation $(x_k - \mu)^3$ there is an identical one with a negative sign and with the same probability. Hence, the skewness is zero.
- ▷ All terms $(x_k - \mu)^4 f(x_k)$ sum up to 80.5. dividing by $\sigma^4 = 34.0278$, he calculates the kurtosis of 2.3657.

# Joint cumulative distribution function

We now extend the notion of a cumulative distribution function to multiple random variables.

For two random variables $X$ and $Y$ we define the *joint (bivariate) distribution function* as

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}[X \leq x, Y \leq y] = \mathbb{P}[\{X \leq x\} \cap \{Y \leq y\}] \\ &= \mathbb{P}[\{\omega \in \Omega \mid X(\omega) \leq x \text{ and } Y(\omega) \leq y\}]\,. \end{aligned}$$

The *joint probability density function* $f_{X,Y}$ for two absolutely continuous random variables is defined as the derivative with respect to both arguments,

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}\,.$$

Notice that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)\,\mathrm{d}s\mathrm{d}t$$

# Joint cumulative distribution function

Naturally, we can define the joint cdf for any finite number of random variables $X_1, \dots, X_n$,

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}[X_1 \le x_1, \dots, X_n \le x_n] \, .$$

The joint density is similarly defined as

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \frac{\partial^n F_{X_1,\dots,X_n}(x_1, \dots, x_n)}{\partial x_1 \cdots \partial x_n} \, ,$$

and again we have

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\dots,X_n}(t_1, \dots, t_n) \, \mathrm{d}t_1 \cdots \mathrm{d}t_n \, .$$

# Independence of events

We say two elements $A$ and $B$ of a $\sigma$-algebra are *independent* and we write $A \perp B$ if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]\,.$$

To see why we call this property independence, we recall the definition of the conditional probability and use 'independence',

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\mathbb{P}[A]\mathbb{P}[B]}{\mathbb{P}[B]} = \mathbb{P}[A]\,.$$

So the probability of $A$ stays the same independent of the occurrence of $B$. The same is true the other way around

$$\mathbb{P}[B \mid A] = \frac{\mathbb{P}[B \cap A]}{\mathbb{P}[A]} = \mathbb{P}[B]\,.$$

# Independence of two random variables

This notion of independence can now easily be translated to the independence of two random variables.

We say two *random variables $X$ and $Y$ are independent* if for all $x, y \in \mathbb{R}$ the sets $\{X \leq x\}$ and $\{Y \leq y\}$ are independent.

From here we directly get the equivalent property that the joint cdf is the product of the individual cdf's. That is, for all $x, y \in \mathbb{R}$ we have

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)\,.$$

This is also equivalent (if all densities exists) that for all $x, y \in \mathbb{R}$, we have

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)\,.$$

# Independence of random variables

For the independence of *more than two random variables* we differentiate between *pairwise independence* and *mutual independence*.

We say a set $\{X_1, \ldots, X_n\}$ of random variables is *pairwise independent* if every pair of random variables is independent (as defined above).

However, we say they are *mutually independent* if for all $x_1, \ldots, x_n \in \mathbb{R}$ we have

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n)\,.$$

# Independence implies uncorrelatedness

As already mentioned, if two random variables $X$ and $Y$ are independent, then they are also uncorrelated. A short calculation shows that,

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x)f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty} xf_X(x)\left(\int_{-\infty}^{\infty} yf_Y(y)\,\mathrm{d}y\right)\mathrm{d}x \\
&= \left(\int_{-\infty}^{\infty} xf_X(x)\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} yf_Y(y)\,\mathrm{d}y\right) \\
&= \mathbb{E}[X]\mathbb{E}[Y]\,.
\end{aligned}$$

Hence, $\mathrm{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$ and hence $\rho_{X,Y} = 0$.

# Uncorrelated but not independent!

However, if two random variables $X$ and $Y$ are uncorrelated, then they do not have to be independent.

One of the standard example is the following. Consider $X \sim \mathcal{U}(-1,1)$ and define $Y = X^2$. Clearly, $X$ and $Y$ are *not independent!* But there correlation is 0, so there uncorrelated,

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] \\ &= \mathbb{E}[X^3] = \int_{-1}^{1} \frac{1}{2} x^3 \mathrm{d}x = 0\,. \end{aligned}$$

**Exercise.** Come up with an own example of uncorrelated but dependent random variables.

# Lognormal distribution

**Exercise.** We want to practice working with distributions by means of the *log-normal distribution.* The log-normal distribution is often use to model returns of stocks.

We say $X$ is *log-normally distributed* and write $X \sim \log \mathcal{N}(\mu, \sigma^2)$ if $Y = \log(X) \sim \mathcal{N}(\mu, \sigma^2)$.

- ▷ First of all, show that $X \sim \log \mathcal{N}(\mu, \sigma^2)$ can be written as $X = \exp(\mu + \sigma Z)$, where $Z \sim \mathcal{N}(0, 1)$.
- ▷ Now show that $\mathbb{E}[X] = \exp(\mu + \frac{\sigma^2}{2})$. You will have to calculate $\mathbb{E}[e^{\sigma Z}]$. For this write down the integral and *complete the square*. Try using $(z - \sigma)^2$.
- ▷ Derive the density function of $X$. Notice that it is only defined on $(0, +\infty)$. You will get

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right]$$

# Appendix

# Maximum Likelihood (ML) methods

How are the parameters estimated from historical data in the models we have been considering?

A commonly applied approach is known as the *maximum likelihood (ML) method*.

It involves choosing values for the parameter that maximise the chance (or likelihood) of the data occurring.

# In General

Suppose we have a sample $x_1, x_2, \dots, x_N$ of $N$ i.i.d. random variables, coming from a parametric model.

- ▷ The joint density function of the observations is

$$f(x_1, x_2, \dots, x_N|\theta) = f_1(x_1|\theta) \cdot f_2(x_2|\theta) \cdot \dots \cdot f_n(x_N|\theta),$$

  where $\theta$ summarises the model parameters.

- ▷ The idea of the maximum likelihood (ML) method is to *choose $\theta$ such that the joint density function is maximised*, given the observed sample of data.

- ▷ A natural tool to this end is the *likelihood function*, which we define as

$$\mathcal{L}(\theta|x_1, x_2, \dots, x_N) := f(x_1, x_2, \dots, x_N|\theta) = \prod_{i=1}^{N} f_i(x_i|\theta).$$

# In General

To convert the product to summation (which is easier to handle on a computer), we take the logarithm. The result is called the *log-likelihood*:

$$\log \mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log(f_i(x_i|\theta)).$$

The ML method estimates $\theta$ by finding a value for $\theta$ that maximises $\log \mathcal{L}(\theta|x_1, x_2, \ldots, x_n)$, that is,

$$\hat{\theta}_{\text{mle}} := \arg\max_{\theta} \log \mathcal{L}(\theta|x_1, x_2, \ldots, x_n).$$

# Example: Estimating a constant variance

Estimate the variance of a variable $X$ from $m$ observations on $X$ when the underlying distribution is *normal with zero mean*.

Let $u_1, u_2, \ldots$ denote the sample of $m$ observations and denote the unknown variance parameter by $v$.

The likelihood of $u_i$ being observed is the probability density function for $X$ when $X = u_i$

$$\frac{1}{\sqrt{2\pi v}} \exp\left(\frac{-u_i^2}{2v}\right).$$

The likelihood of $m$ observations occurring in order in which they are observed is

$$\prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi v}} \exp\left(\frac{-u_i^2}{2v}\right)\right].$$

Using the maximum likelihood method, the best estimate of $v$ is the value that maximises this expression.

# Estimating a constant variance

Maximising an expression is equivalent to maximising the logarithm of that expression, since the logarithm is strictly increasing.

Taking logarithms and ignoring constant multiplicative factors, it can be seen that we wish to maximise

$$\sum_{i=1}^{m}\left[-\log v - \frac{u_i^2}{v}\right].$$

Differentiating this expression with respect to $v$ and setting the result equation to zero, we see that the maximum likelihood estimator of $v$ is

$$\frac{1}{m}\sum_{i=1}^{m} u_i^2 .$$

# The linear model

The natural starting point for learning about statistical data analysis is with a sample of *independent identically distributed (i.i.d.)* data, say $\mathrm{Y} = (Y_1, \dots, Y_n)$, as for example in an idealized experiment of randomly drawing a marble from an urn.

The *linear regression model* relaxes both assumptions

- ▷ allowing the means of the $Y_i$ to depend, in a linear way, on other additional variables,
- ▷ allowing for the $Y_i$ to have different variances, and
- ▷ allowing for correlation between the $Y_i$.

The linear regression model:

- ▷ is of fundamental importance in a large variety of quantitative disciplines, and
- ▷ it forms the basis of many complex and seemingly unrelated models.

# Regressors

The *univariate linear regression model* relates the scalar random variable $Y$ to $k$ other (possibly random) variables, or *regressors*, $x_1, \ldots, x_k$ as

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \text{ where, typically, } \epsilon \sim \mathcal{N}\left(0, \sigma^2\right).$$

Values $\beta_1, \ldots, \beta_k$ and $\sigma^2$ are unknown, constant parameters to be estimated from the data.

To emphasise that the means of the $Y_i$ are not constant, we write

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

The $\epsilon_i$ represent the difference between the values of $Y_i$ and the model $\sum_{j=1}^{k} \beta_j x_{i,j}$, and so are referred to as the *error terms*.

# Example

Notice that the error terms are i.i.d., but the $Y_i$ are not. However, if we take $k = 1$ and $x_{i,1} \equiv 1$, then the above equation reduces to

$$Y_i = \beta_1 + \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\beta_1, \sigma^2\right) .$$

In fact, it is usually the case that $x_{i,1} \equiv 1$ for any $k \geq 1$, in which case the model is said to *include a constant* or *have an intercept term*.

In certain applications, the ordering of the dependent variable and the regressors is important, because they are observed in time, usually equally spaced. For this we use the notation $Y_t$, we get

$$Y_t = \beta_1 x_{t,1} + \beta_2 x_{t,2} + \cdots + \beta_k x_{t,k} + \epsilon_t, \quad t = 1, 2, \ldots, T.$$

# Example

An important special case is with $k = 2$, $x_{t,1} = 1$, and $x_{t,2}$ not constant.

Then

$$Y_t = \beta_1 + \beta_2 X_t + \epsilon_t, \quad t = 1, \ldots, T\,,$$

is referred to as the *simple linear regression model*.

The goal is to find $\beta_1$ and $\beta_2$ which provide a '*best*' fit. One possibility is to minimise the square-error, that is minimise the function

$$S(\beta_1, \beta_2) = \sum_{t=1}^{T} \epsilon_t^2 = \sum_{t=1}^{T} (Y_t - \beta_1 - \beta_2 X_t)^2\,.$$

# Example

For this, we set the partial derivatives equal zero, and solve for $\beta_1$ and $\beta_2$,

$$\partial_{\beta_1} S(\beta_1, \beta_2) = -2 \sum_{t=1}^{T} (Y_t - \beta_1 - \beta_2 X_t) \overset{!}{=} 0\,, \quad \text{hence, } \hat{\beta}_1 = \bar{Y} - \beta_2 \bar{X}\,.$$

Substituting $\hat{\beta}_1$ in $S$ and differentiating with respect to $\beta_2$ gives us

$$\begin{aligned} \partial_{\beta_2} S(\hat{\beta}_1, \beta_2) &= \partial_{\beta_2} \left( \sum_{t=1}^{T} (Y_t - \bar{Y} - \beta_2 (X_t - \bar{X}))^2 \right) \\ &= -2 \sum_{t=1}^{T} (Y_t - \bar{Y} - \beta_2 (X_t - \bar{X}))(X_t - \bar{X}) \overset{!}{=} 0\,, \end{aligned}$$

which leads us to

$$\hat{\beta}_2 = \frac{\sum_{t=1}^{T} (Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1}^{T} (X_t - \bar{X})^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}\,.$$

# Example (contn'd)

So we have

$$\hat{\beta}_1 = \bar{Y} - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \bar{X} \quad \text{and} \quad \hat{\beta}_2 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \, .$$

In order to verify that this indeed provides a minimum, we consider the second partial derivatives,

$$\partial_{\beta_1 \beta_1} S = 2T > 0 \, , \;\; \partial_{\beta_2 \beta_2} S = 2 \sum_{t=1}^{T} X_t^2 > 0 \, , \text{ and}$$

$$\partial_{\beta_1 \beta_2} S = \partial_{\beta_2 \beta_1} S = 2 \sum_{t=1}^{T} X_t \, .$$

Hence, $(\partial_{\beta_1 \beta_1} S)(\partial_{\beta_2 \beta_2} S) - (\partial_{\beta_1 \beta_2} S)^2 = 4(T \sum_{t=1}^{T} X_t^2 - (\sum_{t=1}^{T} X_t)^2) > 0$, confirming a minimum.

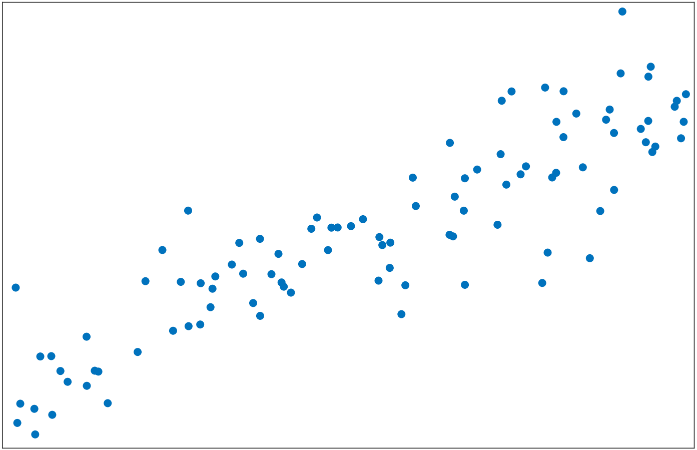

**Figure 1:** Linearly correlated data and the simple linear regression model

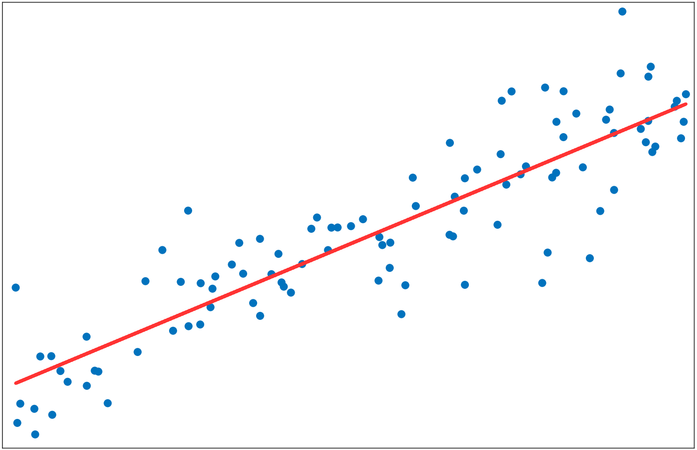

**Figure 2:** Linearly correlated data and the simple linear regression model

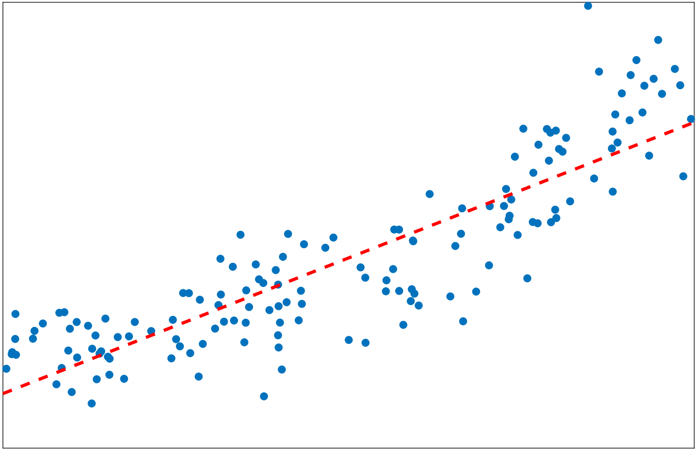

**Figure 3:** Nonlinearly correlated data and the simple linear regression model

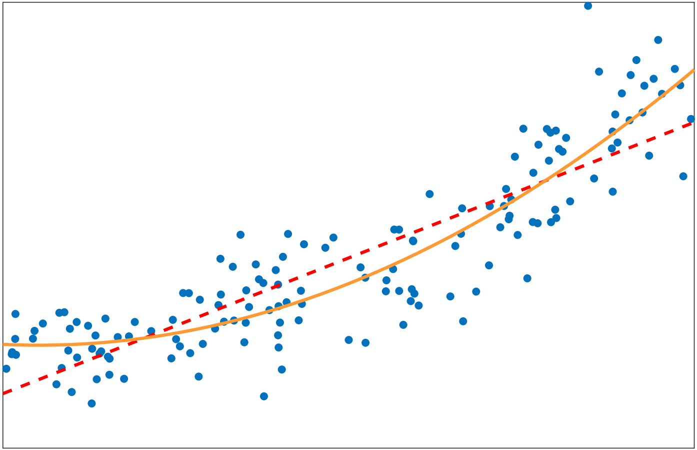

**Figure 4:** Nonlinearly correlated data and linear regression with three regressors

# General linear regression model

We see that the simple linear regression model quickly reaches its limits. So we want to consider the general case.

Using standard matrix notation, with $\mathsf{Y} = (Y_1, \dots Y_T)'$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$, $\mathsf{x}_t = (x_{t,1}, \dots, x_{t,k})'$,

$$\mathsf{X} = \begin{bmatrix} \mathsf{x}_1' \\ \vdots \\ \mathsf{x}_T' \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ x_{T,1} & x_{T,2} & & x_{T,k} \end{bmatrix}, \quad \text{and } \boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2 \mathsf{I}_T\right),$$

we can compactly express the model as

$$\mathsf{Y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Notice that $Y \sim \mathcal{N}\left(\mathsf{X}\boldsymbol{\beta}, \sigma^2 \mathsf{I}_T\right)$.

# Ordinary least squares (OLS)

The method we used before is the most popular way of estimating the $k$ parameters in $\boldsymbol{\beta}$, namely, the *method of least squares*, which takes $\widehat{\boldsymbol{\beta}} = \arg\min \mathsf{S}(\boldsymbol{\beta})$, where

$$\mathsf{S}(\boldsymbol{\beta}) = \mathsf{S}(\boldsymbol{\beta}; \mathsf{Y}, \mathsf{X}) = (\mathsf{Y} - \mathsf{X}\boldsymbol{\beta})' (\mathsf{Y} - \mathsf{X}\boldsymbol{\beta}) = \sum_{t=1}^{T} \left(Y_t - \mathsf{x}_t' \boldsymbol{\beta}\right)^2 .$$

This is referred to as *ordinary* least squares.

Assume that $\mathsf{X}$ is of full rank $k$. Using matrix calculus, we get

$$\partial\, \mathsf{S}(\boldsymbol{\beta}) \,/ \partial\boldsymbol{\beta} = -2\mathsf{X}' \left(\mathsf{Y} - \mathsf{X}\boldsymbol{\beta}\right),$$

and setting this to zero yields the solution

$$\widehat{\boldsymbol{\beta}} = \left(\mathsf{X}'\mathsf{X}\right)^{-1} \mathsf{X}'\mathsf{Y}.$$

This is referred to as the *ordinary least squares (o.l.s.) estimator* of $\boldsymbol{\beta}$.

# Ordinary least squares (OLS)

To verify that $\widehat{\boldsymbol{\beta}} = (\mathsf{X}'\mathsf{X})^{-1}\,\mathsf{X}'\mathsf{Y}$ indeed corresponds to the minimum of $\mathsf{S}(\boldsymbol{\beta})$, the second derivative is checked for positive definiteness, yielding

$$\partial^2\,\mathsf{S}(\boldsymbol{\beta})\,/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = 2\mathsf{X}'\mathsf{X},$$

which is necessarily positive definite when X is full rank.

# Example

Observe that, if $\mathsf{X} = 1$, then

$$\widehat{\boldsymbol{\beta}} = (1'1)^{-1}1'\mathsf{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t = \bar{Y}, \text{ the mean of the } Y_t.$$

Also, if $k = T$ (and $\mathsf{X}$ is full rank), then

$$\widehat{\boldsymbol{\beta}} = \left(\mathsf{X}'\mathsf{X}\right)^{-1}\mathsf{X}'\mathsf{Y} = \mathsf{X}^{-1}\left(\mathsf{X}'\right)^{-1}\mathsf{X}'\mathsf{Y} = \mathsf{X}^{-1}\mathsf{Y}$$

and using this we get $\mathsf{S}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\mathsf{Y} - \mathsf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathsf{Y} - \mathsf{X}\widehat{\boldsymbol{\beta}}\right) = 0.$

# Example

Notice that the derivation of $\widehat{\boldsymbol{\beta}}$ did not involve any explicit distributional assumptions. Consequently, the estimator may not have any meaning if the maximally existing moment of the $\{\epsilon_t\}$ is too low.

- ▷ Take for example $\mathsf{X} = 1$ and let $\epsilon_t \overset{\text{i.i.d.}}{\sim} \text{Cauchy}$. Then $\widehat{\beta} = \bar{Y}$ is a useless estimator.
- ▷ Assume that the first moment of the $\{\epsilon_t\}$ exists and is zero. Writing

  $$\widehat{\boldsymbol{\beta}} = \left(\mathsf{X}'\mathsf{X}\right)^{-1}\mathsf{X}'\left(\mathsf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) = \boldsymbol{\beta} + \left(\mathsf{X}'\mathsf{X}\right)^{-1}\mathsf{X}'\boldsymbol{\epsilon},$$

  we see that $\widehat{\boldsymbol{\beta}}$ is *unbiased*, which means that

  $$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + \left(\mathsf{X}'\mathsf{X}\right)^{-1}\mathsf{X}'\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{\beta}.$$

- Next, if we have existence of second moments, and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathsf{I}$, then

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\beta}} \mid \sigma^2) &= \mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \sigma^2] \\ &= \left(\mathsf{X}'\mathsf{X}\right)^{-1} \mathsf{X}'\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] \mathsf{X} \left(\mathsf{X}'\mathsf{X}\right)^{-1} = \sigma^2 \left(\mathsf{X}'\mathsf{X}\right)^{-1}.\end{aligned}$$

The *Gauss-Markov Theorem* tells us that $\widehat{\boldsymbol{\beta}}$ has the *smallest variance among all linear unbiased estimators*. We call such an estimator *BLUE* – **b**est **l**inear **u**nbiased **e**stimator.

- Next, if we have existence of second moments, and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathsf{I}$, then

$$\begin{aligned}\mathrm{Var}(\widehat{\boldsymbol{\beta}} \mid \sigma^2) &= \mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \sigma^2] \\ &= \left(\mathsf{X}'\mathsf{X}\right)^{-1} \mathsf{X}'\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] \mathsf{X} \left(\mathsf{X}'\mathsf{X}\right)^{-1} = \sigma^2 \left(\mathsf{X}'\mathsf{X}\right)^{-1}.\end{aligned}$$

The *Gauss-Markov Theorem* tells us that $\widehat{\boldsymbol{\beta}}$ has the *smallest variance among all linear unbiased estimators*. We call such an estimator *BLUE* – **b**est **l**inear **u**nbiased **e**stimator.

To see this, consider another linear estimator $\widehat{\boldsymbol{\beta}}^* = \mathsf{A}'\mathsf{Y}$ and let $\mathsf{D} = \mathsf{A} - \mathsf{X}(\mathsf{X}'\mathsf{X})^{-1}$. The unbiased property $\mathbb{E}\big[\widehat{\boldsymbol{\beta}}^*\big] \overset{!}{=} \boldsymbol{\beta}$ implies that $\mathsf{D}'\mathsf{X} = 0$.

Next, we calculate $\mathrm{Var}\big(\widehat{\boldsymbol{\beta}}^* \mid \sigma^2\big) = \mathrm{Var}(\widehat{\boldsymbol{\beta}} \mid \sigma^2) + \sigma^2\mathsf{D}'\mathsf{D}$. The result follows because $\mathsf{D}'\mathsf{D}$ is positive semi-definite and the variance is minimized when $\mathsf{D} = 0$.

# Maximum Likelihood

In many situations, it is reasonable to assume normality for the $\{\epsilon_t\}$. In this case, we can estimate $\sigma^2$ and $\beta_i$, $i = 1, \ldots, k$, by *maximum likelihood*.

We start with the density function

$$f_{\mathsf{Y}}(\mathsf{y}) = (2\pi\sigma^2)^{-T/2} \exp\Big\{ -\frac{1}{2\sigma^2} \underbrace{(\mathsf{y} - \mathsf{X}\boldsymbol{\beta})'(\mathsf{y} - \mathsf{X}\boldsymbol{\beta})}_{=\mathsf{S}(\boldsymbol{\beta};\mathsf{y},\mathsf{X})} \Big\},$$

and the log-likelihood

$$\ell\left(\boldsymbol{\beta}, \sigma^2; \mathsf{Y}\right) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\mathsf{S}(\boldsymbol{\beta}).$$

# Maximum Likelihood

To get the *Maximum Likelihood Estimator (m.l.e.)* we have to solve

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\frac{2}{2\sigma^2}\mathsf{X}'\left(\mathsf{Y} - \mathsf{X}\boldsymbol{\beta}\right) \overset{!}{=} 0$$

and

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\,\mathsf{S}(\boldsymbol{\beta}) \overset{!}{=} 0.$$

This yields the same estimator for $\boldsymbol{\beta}$ and the m.l.e. of $\sigma^2$, namely

$$\widehat{\boldsymbol{\beta}} = \left(\mathsf{X}'\mathsf{X}\right)^{-1}\mathsf{X}'\mathsf{Y} \quad \text{and} \quad \tilde{\sigma}^2 = \mathsf{S}(\widehat{\boldsymbol{\beta}})/T.$$

# Appendix A: Cauchy-Schwarz inequality

Let $X, Y$ real valued random variables with finite second moment. Then:

$$\mathbb{E}[XY] \leq \mathbb{E}[X]^2\mathbb{E}[Y]^2,$$

with equality if $X = \alpha Y$ for some $\alpha \in \mathbb{R}$.

# Appendix B: Matrix Calculus

Let $x$ be a column vector. Then the following rules apply:

$$\begin{aligned} \frac{a'x}{\partial x} = \frac{\partial x'a}{\partial x} &= a', \\ \frac{\partial b'Ax}{\partial x} &= b'A, \\ \frac{\partial x'Ax}{\partial x} &= x'(A+A'), \\ \frac{\partial g(u)}{\partial x} &= \frac{\partial g(u)}{\partial u}\frac{\partial u}{\partial x}, \end{aligned}$$

where $a, b$ are vectors that are not functions of $x$, $A$ is a squared matrix that is not a function of $x$ and $u = u(x)$.

# References

# References

▷ Marc S. Paolella, *Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH*

# Acknowledgements

Thanks goes to Alexander Smirnow for his support in preparing these slides and for challenging an earlier version of them!