A Primer in Econometric Theory

Lecture 4: Modelling Dependence

John Stachurski
Lectures by Akshay Shanker

30 сентября 2020 г.

Random Vector

A **random vector** $\mathbf{x}$ in $\mathbb{R}^N$ is a function from Ω to $\mathbb{R}^N$ with the property that

$$\{\omega \in \Omega : \mathbf{x}(\omega) \in B\} \in \mathscr{F} \quad \text{for all } B \in \mathscr{B}(\mathbb{R}^N)$$

We can also define a **random vector** $\mathbf{x}$ in $\mathbb{R}^N$ as a tuple of N random variables $(x_1, \ldots, x_N)$

We write random vectors in rows or columns according to convenience

- during matrix multiplication, random vectors will default to column vectors

Пример. Recall the blindfolded monkey experiment

Sample space is the unit disk $\Omega := \{(h, v) \in \mathbb{R}^2 : \|(h, v)\| \leq 1\}$ and the event space is the Borel sets in Ω

If $\mathbf{x}$ is the identity on Ω, then it simply reports the outcome (h, v) — a random vector

Пример. Consider a random sample listing the income y_n of $n = 1, \ldots, N$ individuals from a given population

The vector $(y_1, \ldots, y_N)$ that reports the outcome of this sampling can be regarded as a random vector in $\mathbb{R}^N$

Measurability

Definition of random vector ensures $\{\mathbf{x} \in B\}$ is a well-defined event for every $B \in \mathscr{B}(\mathbb{R}^N)$

To ensure $\mathbf{y} = f(\mathbf{x})$ is a random vector:

- the function $f\colon \mathbb{R}^N \to \mathbb{R}^M$ must satisfy $f^{-1}(B) \in \mathscr{B}(\mathbb{R}^N)$ for all $B \in \mathscr{B}(\mathbb{R}^M)$

Expectations

Expectations are defined element-by-element

If $\mathbf{x} = (x_1, \ldots, x_N)$ is a random vector in $\mathbb{R}^N$, then

$$\mathbb{E}\,\mathbf{x} = \mathbb{E}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} := \begin{pmatrix} \mathbb{E}\,x_1 \\ \mathbb{E}\,x_2 \\ \vdots \\ \mathbb{E}\,x_N \end{pmatrix}$$

Random Matrix

An $M \times N$ **random matrix** $\mathbf{X}$ is an $M \times N$ array of random variables

Its expectation is defined as

$$\mathbb{E}\,\mathbf{X} := \begin{pmatrix} \mathbb{E}\,x_{11} & \cdots & \mathbb{E}\,x_{1N} \\ \vdots & & \vdots \\ \mathbb{E}\,x_{M1} & \cdots & \mathbb{E}\,x_{MN} \end{pmatrix}$$

From linearity of expectations (fact ??):

Факт. (??) If $\mathbf{X}$ and $\mathbf{Y}$ are random matrices or vectors and $\mathbf{A}$ and $\mathbf{B}$ are constant and conformable, then

$$\mathbb{E}\left[\mathbf{AX}+\mathbf{BY}\right]=\mathbf{A}\mathbb{E}\left[\mathbf{X}\right]+\mathbf{B}\mathbb{E}\left[\mathbf{Y}\right]$$

Variance-Covariance Matrix

The **variance–covariance matrix** of a random vector $\mathbf{x}$ in $\mathbb{R}^N$ with $\boldsymbol{\mu} := \mathbb{E}\,\mathbf{x}$ is the $N \times N$ matrix

$$\operatorname{var}[\mathbf{x}] := \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right]$$

Expanding:

$$\operatorname{var}[\mathbf{x}] = \begin{pmatrix} \mathbb{E}\left[(x_1 - \mu_1)(x_1 - \mu_1)\right] & \cdots & \mathbb{E}\left[(x_1 - \mu_1)(x_N - \mu_N)\right] \\ \vdots & & \vdots \\ \mathbb{E}\left[(x_N - \mu_N)(x_1 - \mu_1)\right] & \cdots & \mathbb{E}\left[(x_N - \mu_N)(x_N - \mu_N)\right] \end{pmatrix}$$

The j,kth term is the scalar covariance between x_j and x_k and the principal diagonal contains the variance of each x_n

Факт. For any random vector $\mathbf{x}$ with $\mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x}] < \infty$,

1. $\mathrm{var}[\mathbf{x}]$ exists and is nonnegative definite,
2. $\mathrm{var}[\mathbf{x}] = \mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}] - \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}$, and
3. $\mathrm{var}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\,\mathrm{var}[\mathbf{x}]\mathbf{A}^\mathsf{T}$ (for any $\mathbf{A}, \mathbf{b}$ constant and conformable).

The cross-covariance between random vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\operatorname{cov}[\mathbf{x}, \mathbf{y}] := \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathsf{T}}\right]$$

Evidently $\operatorname{var}[\mathbf{x}] = \operatorname{cov}[\mathbf{x}, \mathbf{x}]$

Факт. (??) If $\mathbf{z}$ is a random vector in $\mathbb{R}^N$ satisfying $\mathbb{E}[\mathbf{z}\mathbf{z}^{\mathsf{T}}] = \mathbf{I}$ and $\mathbf{A}$ is any constant $N \times N$ matrix, then

$$\mathbb{E}[\mathbf{z}^{\mathsf{T}} \mathbf{A} \mathbf{z}] = \operatorname{trace} \mathbf{A}$$

The proof is a solved exercise (see ex. **??**)

Multivariate Distributions

A **distribution** or **law** P on $\mathbb{R}^N$ is a probability measure over the Borel sets $\mathscr{B}(\mathbb{R}^N)$

By definition, it satisfies $P(\mathbb{R}^N) = 1$ and $P(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n)$ for any disjoint sequence $\{B_n\}$ in $\mathscr{B}(\mathbb{R}^N)$

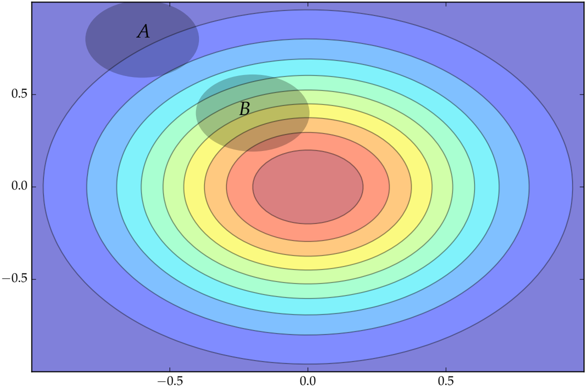

Рис.: Example distribution and events A and B

Any distribution P on $\mathbb{R}^N$ is characterised by the function

$$F(\mathbf{s}) := F(s_1, \ldots, s_N) := P\left(\times_{n=1}^{N}(-\infty, s_n]\right) \qquad (\mathbf{s} \in \mathbb{R}^N)$$

The function F is a **multivariate cumulative distribution function**, which is a function $F\colon \mathbb{R}^N \to [0,1]$ that is

1. right-continuous in each of its arguments,
2. increasing in each of its arguments, and
3. satisfies

$$F(\mathbf{s}_j) \to 1 \text{ as } \mathbf{s}_j \to \infty$$

$$\text{and} \quad F(s_1, \ldots, s_{nj}, \ldots, s_N) \to 0 \text{ as } s_{nj} \to -\infty$$

A distribution P on $\mathbb{R}^N$ is:

- **discrete** if P is supported on a countable subset of $\mathbb{R}^N$
- **absolutely continuous** if $P(B) = 0$ whenever B has zero Lebesgue measure

Again, absolute continuity necessary and sufficient for existence of density representation:

$$P(B) = \int_B p(\mathbf{s})\, \mathrm{d}\mathbf{s} \qquad \text{for all } B \in \mathscr{B}(\mathbb{R}^N)$$

The right-hand is a multivariate integral which we can write as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{1}_B(s_1, \ldots, s_N) p(s_1, \ldots, s_N)\, \mathrm{d}s_1 \cdots \mathrm{d}s_N$$

If p is any density on $\mathbb{R}^N$, then above defines a distribution

Пример. The **multivariate normal density** or **multivariate Gaussian density** on $\mathbb{R}^N$ is a function p of the form

$$p(\mathbf{s}) = (2\pi)^{-N/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{s}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{s}-\boldsymbol{\mu})\right\}$$

where $\boldsymbol{\mu}$ is any $N \times 1$ vector and $\boldsymbol{\Sigma}$ is a positive definite $N \times N$ matrix

We represent this distribution by $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The case $\mathrm{N}(\mathbf{0}, \mathbf{I})$ is called the **multivariate standard normal distribution**

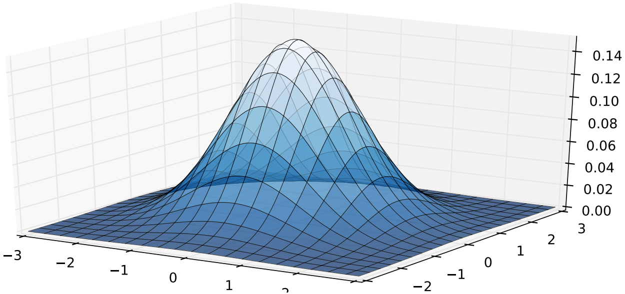

Рис.: Bivariate standard normal density

The **product distribution** of $P_1, \ldots, P_N$ is defined by the next fact:

Факт. (??) Given distributions $P_1, \ldots, P_N$ on $\mathbb{R}$, there exists a unique and well-defined distribution $\mathring{P}$ on $\mathbb{R}^N$ such that

$$\mathring{P}(B_1 \times \cdots \times B_N) = \prod_{n=1}^{N} P_n(B_n) \qquad \text{for all } B_n \in \mathscr{B}(\mathbb{R}),\ n = 1, \ldots, N$$

Unique because the distributions are uniquely pinned down by cylinder sets of $\mathbb{R}^N$ (see page 128 in ET)

Given any distribution P on $\mathbb{R}^N$, the nth **marginal distribution** of P is the distribution on $\mathbb{R}$ defined by

$$P_n(B) = P(\mathbb{R} \times \cdots \times \mathbb{R} \times B \times \mathbb{R} \times \cdots \times \mathbb{R})$$

Here B is the nth element of the Cartesian product

Equivalently,

$$P_n(B) = P\{\mathbf{s} \in \mathbb{R}^N : \mathbf{s}^\mathsf{T}\mathbf{e}_n \in B\}$$

From P_n we can also extract the **marginal** CDF F_n via

$$F_n(s) := P_n((-\infty, s]) \qquad (s \in \mathbb{R})$$

(see page page **??** of ET)

If P_n is absolutely continuous, it has density p_n

When the joint distribution P has a density p, the marginal distribution P_n has a density p_n – "integrate out other variables"

For example, the bivariate case:

$$p_1(s_1) = \int_{-\infty}^{\infty} p(s_1, s_2) \, \mathrm{d}s_2$$

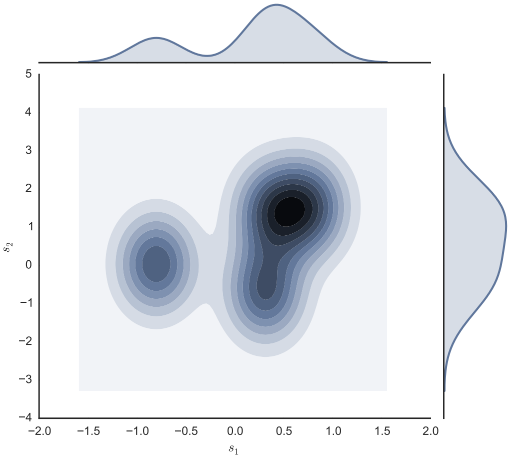

Рис.: Bivariate joint density and its two marginals

Joint distribution cannot be determined from the marginals alone

- marginals do not tell us about the interactions across coordinates

The exception is when there is no interaction – the case for product distributions

Distributions of Random Vectors

Let $\mathbf{x}$ be a random vector in $\mathbb{R}^N$

The **distribution** of $\mathbf{x}$ is the probability measure P on $\mathscr{B}(\mathbb{R}^N)$ defined by

$$P(B) = \mathbb{P}\{\mathbf{x} \in B\} \qquad (B \in \mathscr{B}(\mathbb{R}^N))$$

The P here also called the **joint distribution** of $x_1, \ldots, x_N$, and we write $\mathcal{L}(\mathbf{x}) = P$

Joint distribution represented by the multivariate CDF $F\colon \mathbb{R}^N \to [0,1]$:

$$F(s_1, \ldots, s_N) = \mathbb{P}\{x_1 \leq s_1, \ldots, x_N \leq s_N\}$$

or, in vector notation

$$F(\mathbf{s}) = \mathbb{P}\{\mathbf{x} \leq \mathbf{s}\} \qquad (\mathbf{s} \in \mathbb{R}^N)$$

When the distribution P of $\mathbf{x}$ is absolutely continuous, there exists a non-negative function p on $\mathbb{R}^N$ satisfying

$$\int_B p(\mathbf{s})\,\mathrm{d}\mathbf{s} = \mathbb{P}\{\mathbf{x} \in B\} \qquad (B \in \mathscr{B}(\mathbb{R}^N))$$

The function p is the **joint density** of $\mathbf{x}$

For the above to hold, it suffices that

$$\int_{-\infty}^{s_N} \cdots \int_{-\infty}^{s_1} p(t_1, \ldots, t_N)\,\mathrm{d}t_1 \cdots \mathrm{d}t_N = F(s_1, \ldots, s_N)$$

for all $s_n \in \mathbb{R}$, $n = 1, \ldots, N$

If $\mathbf{x} = (x_1, \dots, x_N)$ is a random vector in $\mathbb{R}^N$, then each x_n is a random variable on $\mathbb{R}$

Let $P_n = \mathcal{L}(x_n)$, so:

$$P_n(B) = \mathbb{P}\{x_n \in B\} \qquad (B \in \mathscr{B}(\mathbb{R}),\ n = 1, \dots, N)$$

P_n is called the **marginal distribution of x_n**

If $P_1 = P_2 = \cdots = P_N$, then $x_1, \dots, x_N$ are **identically distributed**

Gaussian Random Vectors

A random variable x is **normally distributed** if $x = \mu + \sigma z$ for some $\sigma \geq 0$

We write $\mathcal{L}(x) = \mathrm{N}(\mu, \sigma)$

A random vector $\mathbf{x}$ in $\mathbb{R}^N$ is **multivariate normal** or **multivariate Gaussian** if

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{C}\mathbf{z}$$

where the term $\mathbf{z}$ is a $K \times 1$ standard normal random vector, the matrix $\mathbf{C}$ is $N \times K$ and the vector $\boldsymbol{\mu}$ is $N \times 1$

If $\mathbf{x}$ is multivariate normal, then we write $\mathcal{L}(\mathbf{x}) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} := \mathbb{E}\,\mathbf{x} \quad \text{and} \quad \boldsymbol{\Sigma} := \operatorname{var}\mathbf{x}$$

We have $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^\mathsf{T}$ (recall fact 5.1.2 in ET)

$\mathcal{L}(\mathbf{x}) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ does not imply $\mathbf{x}$ has the multivariate normal density

- distribution of $\mathbf{x}$ can fail to be absolutely continuous for e.g. if $\mathbf{C} = \mathbf{0}$

Absolute continuity of the distribution of $\mathbf{x}$ coincides with the setting where $\boldsymbol{\Sigma} := \operatorname{var} \mathbf{x}$ is nonsingular – nonsingularity of $\boldsymbol{\Sigma}$ will be true if and only if $\mathbf{C}^\mathsf{T}$ has full column rank

Факт. (??) Let $\mathbf{x}$ be a random vector in $\mathbb{R}^N$. The following statements are true:

1. The vector $\mathbf{x}$ is multivariate normal if and only if $\mathbf{a}^\top \mathbf{x}$ is normally distributed in $\mathbb{R}$ for every constant $N \times 1$ vector $\mathbf{a}$
2. If $\mathcal{L}(\mathbf{x}) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathcal{L}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathrm{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$$

   for all constant conformable $\mathbf{A}, \mathbf{b}$

Corollary: if $\mathbf{x} = (x_1, \ldots, x_N)$ is multivariate normal, then the marginal distribution of x_n is univariate normal

Is the joint distribution of N univariate normal random variables always multivariate normal?

- Answer: no

Expectations from Distributions

Let $h\colon \mathbb{R}^N \to \mathbb{R}$ be any $\mathscr{B}$-measurable function and let P be a distribution on $\mathbb{R}^N$

The function h now regarded as a random variable on $(\mathbb{R}^N, \mathscr{B}(\mathbb{R}^N), P)$

Expectation of h can be writen as

$$\mathbb{E}_P h :=: \int h(\mathbf{s}) P(\mathrm{d}\mathbf{s}) \tag{1}$$

Факт. (??) Let $h\colon \mathbb{R}^N \to \mathbb{R}$ be $\mathscr{B}$-measurable and let P be a distribution on $\mathbb{R}^N$. If P is discrete, with PMF $\{p_j\}_{j\geq 1}$ and support $\{\mathbf{s}_j\}_{j\geq 1}$, then

$$\int h(\mathbf{s})P(\mathrm{d}\mathbf{s}) = \sum_{j\geq 1} h(\mathbf{s}_j)p_j \tag{2}$$

If P is absolutely continuous with density p, then

$$\int h(\mathbf{s})P(\mathrm{d}\mathbf{s}) = \int h(\mathbf{s})p(\mathbf{s})\,\mathrm{d}\mathbf{s} \tag{3}$$

The right-hand side of (3) should be understood as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(s_1,\ldots,s_N)\, p(s_1,\ldots,s_N)\; \mathrm{d}s_1 \cdots \mathrm{d}s_N$$

As in the univariate case, objects like moments are properties of the distribution

For example, let $\mathbf{x}$ be a random vector in $\mathbb{R}^K$ with $\mathcal{L}(\mathbf{x}) = P$

The variance–covariance matrix $\mathrm{var}[\mathbf{x}]$ of $\mathbf{x}$ has i, jth element $\mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j]$

We can write $\mathrm{var}[\mathbf{x}]$ in terms of P. If

$$\mathbf{\Sigma}_P = (\sigma_{ij}) \quad \text{where} \quad \sigma_{ij} := \int (s_i s_j) P(\mathrm{d}\mathbf{s}) - \int s_i P(\mathrm{d}\mathbf{s}) \cdot \int s_j P(\mathrm{d}\mathbf{s})$$

then $\mathbf{\Sigma}_P = \mathrm{var}[\mathbf{x}]$

Independence of Random Variables

A collection of N random variables $x_1, \ldots, x_N$ is **independent** if

$$\mathbb{P} \bigcap_{n=1}^{N} \{x_n \in B_n\} = \prod_{n=1}^{N} \mathbb{P}\{x_n \in B_n\} \tag{4}$$

for any $B_1, \ldots, B_N$, where each B_n is a Borel subset of $\mathbb{R}$

The random variables $x_1, \ldots, x_N$ are independent when sets of the form $\{x_1 \in B_1\}, \ldots, \{x_N \in B_N\}$ are independent events

An infinite set of random variables $\{x_n\}_{n=1}^{\infty}$ is independent if any finite subset of $\{x_n\}_{n=1}^{\infty}$ is independent

Equivalent definition of independence using distributions

Let P be the joint distribution of $\mathbf{x} = (x_1, \ldots, x_N)$ and P_n be its nth marginal

Since $\cap_{n=1}^{N} \{x_n \in B_n\} = \{(x_1, \ldots, x_N) \in B_1 \times \cdots \times B_N\}$, the random variables $x_1, \ldots, x_N$ are independent if

$$P(B_1 \times \cdots \times B_N) = \prod_{n=1}^{N} P_n(B_n)$$

Elements of a random vector are independent if and only if their joint distribution equals the product distribution formed from their marginals

A necessary and sufficient condition for independence of $x_1, \ldots, x_N$ is:

$$F(s_1, \ldots, s_N) = \prod_{n=1}^{N} F_n(s_n)$$

for all $(s_1, \ldots, s_N) \in \mathbb{R}^N$, where F is the CDF of $\mathbf{x}$ and $F_1, \ldots, F_N$ are the marginal CDFs (why?)

If the distribution of $\mathbf{x}$ is absolutely continuous we can also test independence via its density:

Факт. (??) If $\mathbf{x} = (x_1, \ldots, x_N)$ has joint density p and marginals $p_1, \ldots, p_N$, then $x_1, \ldots, x_N$ are independent if and only if

$$p(s_1, \ldots, s_N) = \prod_{n=1}^{N} p_n(s_n) \qquad \text{for all} \quad (s_1, \ldots, s_N) \in \mathbb{R}^N$$

Пример.

Let $\mathcal{L}(\mathbf{x}) = \mathcal{L}(x_1, \ldots, x_N) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Suppose in addition that $\boldsymbol{\Sigma}$ is diagonal, with nth diagonal component $\sigma_n > 0$, then $x_1, \ldots, x_N$ are independent

To see this, observe for any $\mathbf{s} = (s_1, \ldots, s_N) \in \mathbb{R}^N$, we have

$$p(\mathbf{s}) = (2\pi)^{-N/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{s}-\boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{s}-\boldsymbol{\mu})\right\}$$

$$= \frac{1}{(2\pi)^{N/2} \prod_{n=1}^N \sigma_n} \exp\left\{-\frac{1}{2} \sum_{n=1}^N (s_n - \mu_n)^2 \sigma_n^{-2}\right\}$$

Пример. (cont.) Computation of the determinant and inverse of $\mathbf{\Sigma}$ used facts ?? and ??

The last expression can be factored further

$$p(\mathbf{s}) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{1/2}\sigma_n} \exp\left\{\frac{-(s_n - \mu_n)^2}{2\sigma_n^2}\right\} = \prod_{n=1}^{N} p_n(s_n)$$

where p_n is the density of $\mathrm{N}(\mu_n, \sigma_n^2)$

Факт. (??) If $x_1, \ldots, x_N$ are independent and each x_n is integrable, then

$$\mathbb{E}\left[\prod_{n=1}^{N} x_n\right] = \prod_{n=1}^{N} \mathbb{E}\left[x_n\right]$$

Independence of Random Vectors

Random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ in $\mathbb{R}^K$ are called **independent** if

$$\mathbb{P} \bigcap_{n=1}^{N} \{\mathbf{x}_n \in B_n\} = \prod_{n=1}^{N} \mathbb{P}\{\mathbf{x}_n \in B_n\}$$

for any $B_1, \ldots, B_N$, where each B_n is a Borel subset of $\mathbb{R}^K$

Факт. (??) If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independent random vectors in $\mathbb{R}^K$ and $f_1, \ldots, f_N$ are any $\mathscr{B}$-measurable functions, then $f_1(\mathbf{x}_1), \ldots, f_N(\mathbf{x}_N)$ are also independent.

Доказательство. Observe $f_n(\mathbf{x}_n) \in B_n$ if and only if $\mathbf{x}_n \in f^{-1}(B_n)$. This leads to

$$\bigcap_{n=1}^{N} \{f_n(\mathbf{x}_n) \in B_n\} = \bigcap_{n=1}^{N} \{\mathbf{x}_n \in f^{-1}(B_n)\}$$

Applying independence of $\mathbf{x}_1, \ldots, \mathbf{x}_N$

$$\mathbb{P} \bigcap_{n=1}^{N} \{f_n(\mathbf{x}_n) \in B_n\}$$

$$= \prod_{n=1}^{N} \mathbb{P}\{\mathbf{x}_n \in f^{-1}(B_n)\} = \prod_{n=1}^{N} \mathbb{P}\{f_n(\mathbf{x}_n) \in B_n\}$$

Факт. (??) If $\mathbf{x}$ and $\mathbf{y}$ are independent, then $\text{cov}(\mathbf{x}, \mathbf{y}) = 0$.

Converse not true: one can construct examples of dependent random variables with zero covariance. However,

Факт. (??) If $\mathbf{x}$ is multivariate Gaussian and $\mathbf{A}$ and $\mathbf{B}$ are conformable constant matrices, then $\mathbf{Ax}$ and $\mathbf{Bx}$ are independent if and only if $\text{cov}(\mathbf{Ax}, \mathbf{Bx}) = \mathbf{0}$

Факт. (??) Let S be any linear subspace of $\mathbb{R}^N$, let $\mathbf{P} := \operatorname{proj} S$ and let $\mathbf{M}$ be the residual projection. If $\mathcal{L}(\mathbf{z}) = \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ in $\mathbb{R}^N$ for some $\sigma^2 > 0$, then $\mathbf{Pz}$ and $\mathbf{Mz}$ are independent

Факт. (??) If $w_1, \ldots, w_N$ are independent with $\mathcal{L}(w_n) = \mathrm{N}(\mu_n, \sigma_n^2)$ for all n, then

$$\mathcal{L}\left[\alpha_0 + \sum_{n=1}^{N} \alpha_n w_n\right] = \mathrm{N}\left(\alpha_0 + \sum_{n=1}^{N} \alpha_n \mu_n,\ \sum_{n=1}^{N} \alpha_n^2 \sigma_n^2\right)$$

Sums of arbitrary normals are not always normal — we require a multivariate normal distribution

In fact (**??**) above:

$$\mathcal{L}(w_1, \ldots, w_N) = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\mathbf{e}_n^\mathsf{T} \boldsymbol{\mu} = \mu_n$, and

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_N^2)$$

A **copula** C on $\mathbb{R}^N$ is a multivariate CDF supported on the unit hypercube $[0,1]^N$ with the property that all its marginals are uniform on $[0,1]$

C is a function of the form

$$C(s_1, \ldots, s_N) = \mathbb{P}\{u_1 \leq s_1, \ldots, u_N \leq s_N\} \tag{5}$$

Where $0 \leq s_n \leq 1$ and $\mathcal{L}(u_n) = U[0,1]$ for all n

While each u_n has its marginal distribution pinned down, there are infinitely many ways to specify the joint distribution

Пример. The function $C(s_1, s_2) = s_1 s_2$ on $[0, 1]^2$ is called the **independence copula**

The marginal distributions are $C(s_1, 1) = s_1$ and $C(1, s_2) = s_2$ as required

(These are CDFs for the $U[0, 1]$ distribution.)

Пример. The **Gumbel copulas** are the class of functions on $[0,1]^2$ defined by

$$C(s_1, s_2) = \exp\left\{-\left[(-\ln s_1)^\theta + (-\ln s_2)^\theta\right]^{1/\theta}\right\}, \quad (\theta \geq 1)$$

The **Clayton copulas** are given by

$$C(s_1, s_2) = \left\{\max\left[s_1^{-\theta} + s_2^{-\theta} - 1,\, 0\right]\right\}^{-1/\theta}, \quad (\theta \geq -1,\, \theta \neq 0)$$

Both of these belong to a general class called the **Archimedean copulas**

We can take univariate CDFs $F_1, \ldots, F_N$ and a copula C to create a multivariate CDF on $\mathbb{R}^N$ via

$$F(s_1, \ldots, s_N) = C(F_1(s_1), \ldots, F_N(s_N)) \qquad (s_n \in \mathbb{R}, n = 1, \ldots, N) \quad (6)$$

Benefit: separate out specification of the marginals and specification of the joint distribution

Пример. **bonhomme2009assessing** use copulas to model one component of earnings dynamics in a study based on three-year panels from the French Labor Force Survey

The cross sections are relatively large (around 30,000), allowing for flexible modeling of the marginal distributions via a mixture of normals

However, the time series dimension is short, so a one-parameter family of copulas is used to bind the marginals across time in a parsimonious way

Теорема. (**??**) If F is any CDF on $\mathbb{R}^N$ with marginals $F_1, \ldots, F_N$, then there exists a copula C such that (6) holds. If each F_n is continuous, then this representation is unique.

If $F_1, \ldots, F_N$ are univariate normal, then $C(F_1(s_1), \ldots, F_N(s_N))$ will equal the multivariate normal CDF for one choice of copula, called the Gaussian copula

Other choices lead to different distributions

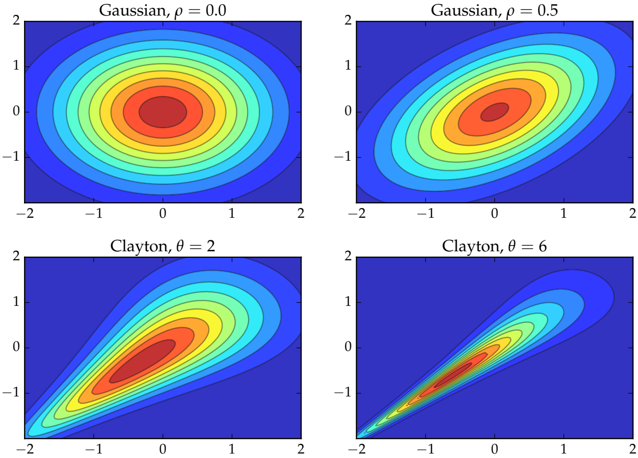

Рис.: Bivariate Gaussian (top) and non-Gaussian (bottom)

Properties of Named Distribution

Факт. (??) If $x_1, \ldots, x_N$ are independent and $\mathcal{L}(x_n) = \chi^2(k_n)$, then $\mathcal{L}(\sum_n x_n) = \chi^2(\sum_n k_n)$

Факт. (??) If z and x are independent with $\mathcal{L}(z) = \mathrm{N}(0,1)$ and $\mathcal{L}(x) = \chi^2(k)$, then

$$z\sqrt{\frac{k}{x}} \quad \text{is } t \text{ distributed with } k \text{ degrees of freedom}$$

Факт. (**??**) If $\mathcal{L}(z_1, \ldots, z_N) = \mathrm{N}(\mathbf{0}, \mathbf{I})$, then $\mathcal{L}(\sum_{n=1}^{N} z_n^2) = \chi^2(N)$.

Факт. (**??**) If $\mathcal{L}(\mathbf{z}) = \mathrm{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{A}$ is symmetric and idempotent, then

$$\mathcal{L}\left(\mathbf{z}^\top \mathbf{A} \mathbf{z}\right) = \chi^2(K) \quad \text{where} \quad K := \operatorname{trace} \mathbf{A}$$

Exercise: obtain Fact (**??**) from Fact (**??**). (See page **??** in eT)

Conditioning and Expectation

Conditional expectation is one of the most important concepts in both economic theory and econometrics

This section gives a construction of expectation based around projection:

- frame conditional expectation as optimal prediction given limited information

Conditional Densities

First some discussion of conditional densities

Let x_1 and x_2 be random variables. The **conditional density** of x_2 given $x_1 = s_1$ is defined as

$$p(s_2 \mid s_1) := \frac{p(s_1, s_2)}{p(s_2)}$$

Here p stands in for either joint, marginal, or conditional density, with the type determined by the argument

The law of total probability extends to the density case as follows: If (x_1, x_2) is a random vector on $\mathbb{R}^2$, then

$$p(s_2) = \int_{-\infty}^{\infty} p(s_2 \mid s_1) p(s_1) \, \mathrm{d}s_1 \qquad (s_2 \in \mathbb{R})$$

Доказательство. To see this, fix $s_2 \in \mathbb{R}$ and integrate the joint density to get the marginal, giving

$$p(s_2) = \int_{-\infty}^{\infty} p(s_1, s_2) \, \mathrm{d}s_1$$

Combine with $p(s_2 \mid s_1) = p(s_1, s_2) / p(s_1)$ to yield the result

Bayes' law also extends to the density case:

$$p(s_2 \mid s_1) = \frac{p(s_1 \mid s_2)p(s_2)}{p(s_1)}$$

The conditional density of $x_{k+1}, \ldots, x_N$ given $x_1 = s_1, \ldots, x_k = s_k$ is defined by

$$p(s_{k+1}, \ldots, s_N \,|\, s_1, \ldots, s_k) = \frac{p(s_1, \ldots, s_N)}{p(s_1, \ldots, s_k)}$$

Rearrange to obtain a useful decomposition of the joint density:

$$p(s_1, \ldots, s_N) = p(s_{k+1}, \ldots, s_N \,|\, s_1, \ldots, s_k) p(s_1, \ldots, s_k)$$

Suppose we want to predict random variable y using another variable x

Choose x such that x and y are expected to be close under most realizations of uncertainty

But what does “expected to be close"mean?

The **mean squared error** (MSE)

$$\mathbb{E}\left[(x-y)^2\right]$$

The **root mean squared error**:

$$\|x-y\| := \sqrt{\mathbb{E}\left[(x-y)^2\right]} \tag{7}$$

There are many parallels between ordinary vector space with the Euclidean norm and the set of random variables combined with the “norm"defined in (7) — we formalise these ideas next

The first geometric concept we defined for vectors was inner product

Analogously, define the **inner product between two random variables** x and y

$$\langle x, y \rangle := \mathbb{E}\left[xy\right]$$

Cauchy–Schwarz inequality for random variables tells us $\mathbb{E}\left[xy\right]$ will be finite and well-defined whenever x and y both have finite second moments

The set of random variables with finite second moments commonly denoted as L_2

$$L_2 := \{ \text{ all random variables } x \text{ on } (\Omega, \mathscr{F}, \mathbb{P}) \text{ with } \mathbb{E}\left[x^2\right] < \infty \}$$

Факт. (??) For any $\alpha, \beta \in \mathbb{R}$ and any $x, y, z \in L_2$, the following statements are true:

1. $\langle x, y \rangle = \langle y, x \rangle$.
2. $\langle \alpha x, \beta y \rangle = \alpha\beta \langle x, y \rangle$.
3. $\langle x, \alpha y + \beta z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$.

Properties follow from the definition of the inner product and linearity of $\mathbb{E}$

Compare above with Fact **??** in ET for vectors in Euclidean space

Define the L_2 **norm** by

$$\|x\| := \sqrt{\langle x, x \rangle} := \sqrt{\mathbb{E}\left[x^2\right]} \qquad (x \in L_2)$$

Norm gives notion of distance $\|x - y\|$ between random variables that agrees with the notion of root MSE

Факт. (??) For any $\alpha \in \mathbb{R}$ and any $x, y \in L_2$, the following statements are true:

1. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$
2. $\|\alpha x\| = |\alpha| \|x\|$
3. $\|x + y\| \leq \|x\| + \|y\|$
4. $|\langle x, y \rangle| \leq \|x\| \|y\|$

Property 2. of above fact is immediate from definition of norm and the linearity of $\mathbb{E}$

Property 3. is called the **triangle inequality**, as in the vector case

Property 4. is just the **Cauchy–Schwarz inequality** for random variables from page **??**

As in the vector case, the triangle inequality can be proved from the Cauchy–Schwarz inequality (see exercise **??**)

Regarding 1., it isn't true that $\|x\| = 0$ implies $x(\omega) = 0$ for all $\omega \in \Omega$

What we can say is that if $\|x\| = 0$, then $\mathbb{P}\{x = 0\} = 1$

In dealing with with L_2, convention to not distinguish between random variables that only differ with zero probability

Linear Subspaces in L_2

Any **linear combination** of random variables with finite variance

$$\alpha_1 x_1 + \cdots \alpha_K x_K, \qquad \alpha_k \in \mathbb{R},\ x_k \in L_2 \tag{8}$$

is again in L_2

When X is a subset of L_2, the set of finite linear combinations that can be formed from elements of X is called the **span** of X, and denoted by $\operatorname{span} X$

Пример. If $x \in L_2$ and $\mathbb{1} := \mathbb{1}_\Omega$ is the constant random variable always equal to 1, then $\operatorname{span}\{\mathbb{1}, x\}$ is the set of random variables

$$\alpha + \beta x := \alpha \mathbb{1} + \beta x \quad \text{for scalars} \quad \alpha, \beta \tag{9}$$

This is the set $\mathcal{L}$ introduced from when we discussed best linear predictors

A subset S of L_2 is called a **linear subspace** of L_2 if it is closed under addition and scalar multiplication

- for each $x, y \in S$ and $\alpha, \beta \in \mathbb{R}$, we have $\alpha x + \beta y \in S$

Пример. The span of any set of elements of L_2 is a linear subspace in L_2

Пример. The set $Z := \{x \in L_2 : \mathbb{E}\, x = 0\}$ is a linear subspace of L_2 because

$$x, y \in Z \text{ and } \alpha, \beta \in \mathbb{R} \implies \mathbb{E}[\alpha x + \beta y] = \alpha \mathbb{E}[x] + \beta \mathbb{E}[y] = 0$$

As in $\mathbb{R}^N$, an **orthonormal basis** of a linear subspace S of L_2 is a set $\{u_1, \ldots, u_K\} \subset S$ with the property

$$\langle u_j, u_k \rangle = \mathbb{1}\{j = k\}$$

$$\text{and} \quad \operatorname{span}\{u_1, \ldots, u_K\} = S$$

Пример. Let $x \in L_2$ so that $S := \mathrm{span}\{\mathbb{1}, x\}$ is the set of random variables

$$\alpha + \beta x := \alpha \mathbb{1} + \beta x \quad \text{for scalars} \quad \alpha, \beta \tag{10}$$

If we define

$$u_1 := \mathbb{1} \quad \text{and} \quad u_2 := \frac{x - \mu}{\sigma_x}$$

Then

$$\langle u_1, u_2 \rangle = \mathbb{E}[u_1 u_2] = \mathbb{E}\left[\frac{x - \mu}{\sigma_x}\right] = 0$$

Clearly, $\|u_1\| = \|u_2\| = 1$, so this pair is orthonormal

Also straightforward to show $\mathrm{span}\{u_1, u_2\} = \mathrm{span}\{\mathbb{1}, x\}$, so $\{u_1, u_2\}$ is an orthonormal basis for S

Projections in L_2

As in the Euclidean case, if $\langle x, y \rangle = 0$, then we say that x and y are **orthogonal**, and write $x \perp y$

Факт. If $x, y \in L_2$ and $\mathbb{E}x = 0$ or $\mathbb{E}y = 0$ then $x \perp y \iff \operatorname{cov}[x, y] = 0$

Given $y \in L_2$ and linear subspace $S \subset L_2$, we seek the closest element $\hat{y}$ of S to y

Closeness is in terms of L_2 norm, so $\hat{y}$ is the minimizer of $\|y - z\|$ over all $z \in S$

We seek

$$\hat{y} = \operatorname*{argmin}_{z \in S} \|y - z\| = \operatorname*{argmin}_{z \in S} \sqrt{\mathbb{E}\left[(y - z)^2\right]} \tag{11}$$

The following theorem mimics the Orthogonal Projection Theorem we have already seen:

Теорема. (??) Let $y \in L_2$ and let S be any nonempty closed linear subspace of L_2

The following statements are true:

1. The optimization problem (11) has exactly one solution
2. $\hat{y} \in L_2$ is the unique solution

The statement S is closed means that $\{x_n\} \subset S$ and $x \in L_2$ with $\|x_n - x\| \to 0$ implies $x \in S$ — condition true for all the linear subspaces we want to work with

Analogous with the case of $\mathbb{R}^N$, the random variable $\hat{y}$ above is called the **orthogonal projection of y onto S**

Holding S fixed, the operation

$$y \mapsto \text{the orthogonal projection of } y \text{ onto } S$$

is a function from L_2 to L_2:

- function called **orthogonal projection onto S**
- function denoted by $\mathbf{P}$
- we write $\mathbf{P} = \operatorname{proj} S$

For each $y \in L_2$, $\mathbf{P}y$ is the image of y under $\mathbf{P}$, which is the orthogonal projection $\hat{y}$

- interpret $\mathbf{P}y$ as the *best predictor of* y *from within the collection of random variables contained in* S

Факт. (??)

If S is any linear subspace of L_2, and $\mathbf{P} = \operatorname{proj} S$, then

1. $\mathbf{P}$ is a linear function.

Moreover, for any $y \in L_2$, we have

2. $\mathbf{P}y \in S$,
3. $y - \mathbf{P}y \perp S$,
4. $\|y\|^2 = \|\mathbf{P}y\|^2 + \|y - \mathbf{P}y\|^2$,
5. $\|\mathbf{P}y\| \leq \|y\|$, and
6. $\mathbf{P}y = y$ if and only if $y \in S$.

In 1, $\mathbf{P}$ is linear means $\mathbf{P}(\alpha x + \beta y) = \alpha \mathbf{P}x + \beta \mathbf{P}y$ for all $x, y \in L_2$ and $\alpha, \beta \in \mathbb{R}$

Факт. (??) Let S_i be a linear subspace of L_2 for $i = 1, 2$ and let $\mathbf{P}_i = \operatorname{proj} S_i$. If $S_1 \subset S_2$, then $\mathbf{P}_1\mathbf{P}_2 y = \mathbf{P}_1 y$ for all $y \in L_2$

Факт. (??) If $\{u_1, \ldots, u_K\}$ is an orthonormal basis of S, then, for all $y \in L_2$,

$$\mathbf{P}y = \sum_{k=1}^{K} \langle y, u_k \rangle \; u_k \tag{12}$$

Example

(**??**) The mean of a random variable x can be thought of as the "best predictor of x within the set of constants."

Let $S := \operatorname{span}\{\mathbb{1}\}$, where $\mathbb{1} := \mathbb{1}_\Omega$, and let $\mathbf{P} := \operatorname{proj} S$

The object $\mathbf{P}x$ is precisely the best predictor of x within the class of constant random variables

Not surprisingly, $\mathbf{P}x = \mu\mathbb{1}$, where $\mu := \mathbb{E}\,x$

The easiest way to check this is to observe that $\{\mathbb{1}\}$ is an orthonormal set spanning S, and hence, by (12),

$$\mathbf{P}x = \langle x, \mathbb{1}\rangle\ \mathbb{1} = \mathbb{E}\,[x\mathbb{1}]\mathbb{1} = \mathbb{E}\,[x]\mathbb{1} = \mu\,\mathbb{1}$$

You can also check the claim that $\mu\mathbb{1}$ is the projection of x onto S by verifying the conditions in (ii) of theorem **??**

Пример.

Fix $x, y \in L_2$ and consider projecting y onto $S := \mathrm{span}\{\mathbb{1}, x\}$

The set S is the set of random variables

$$\alpha + \beta x := \alpha \mathbb{1} + \beta x \quad \text{for scalars} \quad \alpha, \beta$$

The problem of projecting y onto S is equivalent to the best linear prediction problem from §**??**

To implement the projection recall

$$u_1 := \mathbb{1} \quad \text{and} \quad u_2 := \frac{x - \mu}{\sigma_x}$$

form an orthonormal basis for S

Let $\mathbf{P} = \operatorname{proj} S$ and apply fact (**??**) above to give

$$\mathbf{P}y = \langle y, u_1 \rangle u_1 + \langle y, u_2 \rangle u_2 = \mathbb{E}[y] + \frac{\operatorname{cov}[x,y]}{\operatorname{var}[x]}(x - \mathbb{E}[x])$$

Alternatively

$$\mathbf{P}y = \alpha^* + \beta^* x$$

$$\text{where} \quad \beta^* := \frac{\operatorname{cov}[x,y]}{\operatorname{var}[x]} \quad \text{and} \quad \alpha^* := \mathbb{E}[y] - \beta^* \mathbb{E}[x]$$

Population Regression

Consider an extension of the best linear prediction problem above to a setting where the information for predicting y is a random vector $\mathbf{x}$ in $\mathbb{R}^K$

We seek L_2 orthogonal projection of y onto the linear subspace:

$$\text{span}\{\mathbf{x}\} := \text{ random variables of the form } \mathbf{x}^\mathsf{T}\mathbf{b} \text{ for some } \mathbf{b} \in \mathbb{R}^K$$

Assume $\mathbb{E}\left[\mathbf{x}^\mathsf{T}\mathbf{x}\right] < \infty$

Факт. (??) If $\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}]$ is positive definite, then the projection $\mathbf{P}y$ of any $y \in L_2$ onto $\operatorname{span}\{\mathbf{x}\}$ is given by

$$\hat{y} = \mathbf{x}^\mathsf{T}\mathbf{b}^* \quad \text{where} \quad \mathbf{b}^* := \mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}]^{-1}\mathbb{E}[\mathbf{x}y]$$

Exercise **??** asks you to prove the above fact

Positive definiteness of $\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}]$ ensures invertibility, hence $\mathbf{b}^*$ is uniquely defined

By the definition of orthogonal projections, $\mathbf{b}^*$ necessarily satisfies

$$\mathbf{b}^* = \operatorname*{argmin}_{\mathbf{a} \in \mathbb{R}^K} \mathbb{E}[(y - \mathbf{x}^\mathsf{T}\mathbf{a})^2]$$

The linear prediction problem considered also called **population linear regression**

- “population"because we are using the true joint distribution of $(\mathbf{x}, y)$ when we compute expectations

Population regression has a sample counterpart called multivariate linear regression, based on observations of $(\mathbf{x}, y)$ – we will discuss in chapter **??**

Measurability

We don't always want to constrain ourselves to linear predictions

To drop the linearity requirement, change the linear subspaces used for projection from the set of linear functions of $\mathbf{x}$ to the set of arbitrary functions of $\mathbf{x}$

The resulting best predictor is the conditional expectation with respect to $\mathbf{x}$

The subspace of arbitrary real-valued functions of $\mathbf{x}$ is called the $\mathbf{x}$-measurable functions

Let $\mathcal{G} := \{x_1, \ldots, x_D\}$ be any set of random variables and let z be any other random variable

The variable z is **$\mathcal{G}$-measurable** if there exists a $\mathscr{B}$-measurable function $g \colon \mathbb{R}^D \to \mathbb{R}$ such that

$$z = g(x_1, \ldots, x_D)$$

- equality between random variables should be interpreted pointwise

$\mathcal{G}$ sometimes referred to as the **information set**

We'll also write $\mathbf{x} = (x_1, \ldots, x_D)$ and say z is $\mathbf{x}$-measurable

Similar terminology will be used for scalars and matrices

- e.g. if $\mathbf{X}$ is a random matrix, then $\mathbf{X}$-measurability means $\mathcal{G}$-measurability when $\mathcal{G}$ lists all elements of $\mathbf{X}$

Intuition: $\mathcal{G}$-measurability of z means z is completely determined by the elements in $\mathcal{G}$

Пример. Let x, y and z be random variables and let α and β be scalars

If $z = \alpha x + \beta y$, then z is $\{x, y\}$-measurable (take $g(s,t) := \alpha s + \beta t$)

Пример. If $x_1, \ldots, x_N$ are random variables and $\mathcal{G} := \{x_1, \ldots, x_N\}$, then the sample mean $\bar{x}_N := \frac{1}{N} \sum_{n=1}^{N} x_n$ is $\mathcal{G}$-measurable.

Пример. Let $\mathbf{x}$ and y be independent and nondegenerate

Then y is not $\mathbf{x}$-measurable, for if it were, then we would have $y = g(\mathbf{x})$ for some function g, contradicting independence of $\mathbf{x}$ and y

Пример. Let $y = \alpha$, where α is a constant

This degenerate random variable is $\mathcal{G}$-measurable for any information set $\mathcal{G}$, because y is already deterministic

For example, if $\mathcal{G} = \{x_1, \ldots, x_p\}$, then we can take $y = g(x_1, \ldots, x_p) = \alpha + \sum_{i=1}^{p} 0x_i$

Факт. (??) Let α, β be any scalars, and let x and y be random variables. If x and y are both $\mathcal{G}$-measurable, then $u := xy$ and $v := \alpha x + \beta y$ are also $\mathcal{G}$-measurable

Suppose $\mathcal{G} \subset L_2$ and consider the set

$$L_2(\mathcal{G}) := \{\text{all } \mathcal{G}\text{-measurable random variables in } L_2\}$$

In view of fact **??**:

Факт. For any $\mathcal{G} \subset L_2$, the set $L_2(\mathcal{G})$ is a linear subspace of L_2

This furnishes us with a subspace to project onto, allowing us to define conditional expectations

Факт. (??) If $\mathcal{G} \subset \mathcal{H}$ and z is $\mathcal{G}$-measurable, then z is $\mathcal{H}$-measurable.

If z is known once the variables in $\mathcal{G}$ are known, then it is certainly known when the extra information provided by $\mathcal{H}$ is available

Пример. Let x_1, x_2 and y be random variables and let

$$\mathcal{G} := \{x_1\} \subset \{x_1, x_2\} =: \mathcal{H}$$

If y is $\mathcal{G}$-measurable, then $y = g(x_1)$ for some $\mathscr{B}$-measurable g. But then y will also be $\mathcal{H}$-measurable. For example, we can write $y = h(x_1, x_2)$ where $h(x_1, x_2) = g(x_1) + 0x_2$.

Факт. (5.2.12) If $\mathcal{G} \subset \mathcal{H}$, then $L_2(\mathcal{G}) \subset L_2(\mathcal{H})$

Conditional Expectation

Let $\mathcal{G} \subset L_2$ and y be some random variable in L_2

The **conditional expectation** of y given $\mathcal{G}$ is written as $\mathbb{E}\left[y \mid \mathcal{G}\right]$ or $\mathbb{E}^{\mathcal{G}}[y]$ and defined as

$$\mathbb{E}\left[y \mid \mathcal{G}\right] := \operatorname*{argmin}_{z \in L_2(\mathcal{G})} \|y - z\| \tag{13}$$

$\mathbb{E}\left[y \mid \mathcal{G}\right]$ is the best predictor of y given the information contained in $\mathcal{G}$

Does the minimizer generally exist? And is it unique?

- yes and yes

We have

$$\mathbb{E}\,[y \,|\, \mathcal{G}] = \mathbf{P}y \quad \text{when } \mathbf{P} := \operatorname{proj} L_2(\mathcal{G})$$

By the orthogonal projection theorem, the projection exists and is unique

An alternative (and equivalent) definition of conditional expectation

The function $\hat{y}$, where $\hat{y} \in L_2$, is the **conditional expectation** of y given $\mathcal{G}$ if

1. $\hat{y}$ is $\mathcal{G}$-measurable and
2. $\mathbb{E}[\hat{y}\, z] = \mathbb{E}[yz]$ for all $\mathcal{G}$-measurable $z \in L_2$.

When convenient we'll also use symbols like $\mathbb{E}[y \mid x_1, \ldots, x_D]$ or $\mathbb{E}[y \mid \mathbf{x}]$

- same as $\mathbb{E}[y \mid \mathcal{G}]$ when $\mathcal{G}$ is defined as the information set containing the variables we condition on

Пример. If x and u are independent, $\mathbb{E}\, u = 0$ and $y = x + u$, then $\mathbb{E}\,[y \,|\, x] = x$. To prove this we need to show that x satisfies 1–2 above

Clearly, x is x-measurable

For 2. we need to show $\mathbb{E}\,[x\, z] = \mathbb{E}\,[y\, z]$ for all x-measurable z. This translates to the claim

$$\mathbb{E}\,[xg(x)] = \mathbb{E}\,[(x+u)g(x)]$$

for any $\mathscr{B}$-measurable g, which is true from independence and $\mathbb{E}\, u = 0$

Факт. (??) Given $\mathbf{x} \in \mathbb{R}^D$ and y in L_2, there exists a $\mathscr{B}$-measurable function $f^*\colon \mathbb{R}^D \to \mathbb{R}$ such that $\mathbb{E}[y \mid \mathbf{x}] = f^*(\mathbf{x})$

The particular function f^* satisfying $f^*(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$ is called the **regression function** of y given $\mathbf{x}$

Пример. If x and y are random variables and $p(y \mid x)$ is the conditional density of y given x, then

$$\mathbb{E}[y \mid x] = \int t p(t \mid x) \, \mathrm{d}t$$

Proof as exercise **??** in ET

Факт. (??) Let x and y be random variables in L_2, let α and β be scalars, and let $\mathcal{G}$ and $\mathcal{H}$ be subsets of L_2. The following properties hold:

1. Linearity: $\mathbb{E}\left[\alpha x + \beta y \,|\, \mathcal{G}\right] = \alpha \mathbb{E}\left[x \,|\, \mathcal{G}\right] + \beta \mathbb{E}\left[y \,|\, \mathcal{G}\right]$
2. If $\mathcal{G} \subset \mathcal{H}$, then $\mathbb{E}\left[\mathbb{E}\left[y \,|\, \mathcal{H}\right] \,|\, \mathcal{G}\right] \mathbb{E}\left[y \,|\, \mathcal{G}\right]$ and $\mathbb{E}\left[\mathbb{E}\left[y \,|\, \mathcal{G}\right]\right] = \mathbb{E}\left[y\right]$ (**the law of iterated expectations**)
3. If y is independent of the variables in $\mathcal{G}$, then $\mathbb{E}\left[y \,|\, \mathcal{G}\right] = \mathbb{E}\left[y\right]$.
4. If y is $\mathcal{G}$-measurable, then $\mathbb{E}\left[y \,|\, \mathcal{G}\right] = y$
5. If x is $\mathcal{G}$-measurable, then $\mathbb{E}\left[xy \,|\, \mathcal{G}\right] = x\mathbb{E}\left[y \,|\, \mathcal{G}\right]$ (**conditional determinism**)

Recap: given $y \in L_2$ and random vector $\mathbf{x}$ in $\mathbb{R}^D$, the conditional expectation $\mathbb{E}[y \mid \mathbf{x}]$ is a function f^* of $\mathbf{x}$, called the regression function of y given $\mathbf{x}$, such that:

$$f^*(\mathbf{x}) = \operatorname*{argmin}_{g \in G} \mathbb{E}\left[(y - g(\mathbf{x}))^2\right] \tag{14}$$

where G is the set of functions from $\mathbb{R}^D$ to $\mathbb{R}$ with $g(\mathbf{x}) \in L_2$

For any $g \in G$, we also have

$$\mathbb{E}\left[(y - g(\mathbf{x}))^2\right] = \mathbb{E}\left[(y - f^*(\mathbf{x}))^2\right] + \mathbb{E}\left[(f^*(\mathbf{x}) - g(\mathbf{x}))^2\right] \tag{15}$$

This implies (14) because $(f^*(\mathbf{x}) - g(\mathbf{x}))^2 \geq 0$

To prove (15), let f^* be the regression function, pick any $g \in G$ and observe

$$
\begin{aligned}
(y - g(\mathbf{x}))^2 &= (y - f^*(\mathbf{x}) + f^*(\mathbf{x}) - g(\mathbf{x}))^2 \\
&= (y - f^*(\mathbf{x}))^2 + 2(y - f^*(\mathbf{x}))(f^*(\mathbf{x}) - g(\mathbf{x})) \\
&\quad + (f^*(\mathbf{x}) - g(\mathbf{x}))^2
\end{aligned}
$$

Consider the expectation of the cross-product term. From the law of iterated expectations:

$$
\begin{aligned}
&\mathbb{E}\{(y - f^*(\mathbf{x}))(f^*(\mathbf{x}) - g(\mathbf{x}))\} \qquad (16) \\
&= \mathbb{E}\{\,\mathbb{E}[(y - f^*(\mathbf{x}))(f^*(\mathbf{x}) - g(\mathbf{x})) \,|\, \mathbf{x}]\,\}
\end{aligned}
$$

Using conditional determinism, rewrite the term inside the curly brackets on the right-hand side as

$$(f^*(\mathbf{x}) - g(\mathbf{x}))\mathbb{E}\left[(y - f^*(\mathbf{x})) \,|\, \mathbf{x}\right]$$

For the second term in this product

$$\mathbb{E}\left[y - f^*(\mathbf{x}) \,|\, \mathbf{x}\right] = \mathbb{E}\left[y \,|\, \mathbf{x}\right] - \mathbb{E}\left[f^*(\mathbf{x}) \,|\, \mathbf{x}\right] = \mathbb{E}\left[y \,|\, \mathbf{x}\right] - f^*(\mathbf{x}) = 0$$

Hence the expectation in (16) is zero — Equation (15) follows

The Vector Case

Given random matrices $\mathbf{X}$ and $\mathbf{Y}$, we set

$$\mathbb{E}[\mathbf{Y}\,|\,\mathbf{X}] := \begin{pmatrix} \mathbb{E}[y_{11}\,|\,\mathbf{X}] & \cdots & \mathbb{E}[y_{1K}\,|\,\mathbf{X}] \\ \vdots & & \vdots \\ \mathbb{E}[y_{N1}\,|\,\mathbf{X}] & \cdots & \mathbb{E}[y_{NK}\,|\,\mathbf{X}] \end{pmatrix}$$

We also define

1. $\mathrm{cov}[\mathbf{x},\mathbf{y}\,|\,\mathbf{Z}] := \mathbb{E}[\mathbf{x}\mathbf{y}^\top\,|\,\mathbf{Z}] - \mathbb{E}[\mathbf{x}\,|\,\mathbf{Z}]\mathbb{E}[\mathbf{y}\,|\,\mathbf{Z}]^\top$
2. $\mathrm{var}[\mathbf{x}\,|\,\mathbf{Z}] := \mathbb{E}[\mathbf{x}\mathbf{x}^\top\,|\,\mathbf{Z}] - \mathbb{E}[\mathbf{x}\,|\,\mathbf{Z}]\mathbb{E}[\mathbf{x}\,|\,\mathbf{Z}]^\top$

Properties of scalar conditional expectations in fact ?? carry over to the matrix setting

A partial list:

Факт. (??) If $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are random matrices and $\mathbf{A}$ and $\mathbf{B}$ are constant and conformable, then

1. $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]^\intercal = \mathbb{E}[\mathbf{Y}^\intercal \mid \mathbf{Z}]$.
2. $\mathbb{E}[\mathbf{AX} + \mathbf{BY} \mid \mathbf{Z}] = \mathbf{A}\mathbb{E}[\mathbf{X} \mid \mathbf{Z}] + \mathbf{B}\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$.
3. $\mathbb{E}[\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]] = \mathbb{E}[\mathbf{Y}]$ and $\mathbb{E}[\mathbb{E}[\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}] \mid \mathbf{X}] = \mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$.
4. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}] = \mathbb{E}[\mathbf{Y}]$.
5. If $g(\mathbf{X})$ is a matrix depending only on $\mathbf{X}$, then
   5.1 $\mathbb{E}[g(\mathbf{X}) \mid \mathbf{X}] = g(\mathbf{X})$
   5.2 $\mathbb{E}[g(\mathbf{X})\,\mathbf{Y} \mid \mathbf{X}] = g(\mathbf{X})\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$ and $\mathbb{E}[\mathbf{Y}\,g(\mathbf{X}) \mid \mathbf{X}] = \mathbb{E}[\mathbf{Y} \mid \mathbf{X}]\,g(\mathbf{X})$