Multivariate Expectations and Moments

Probability calculus / Adv Stat I

Prof. Dr. Matei Demetrescu

Getting multivariate

We defined moments and related quantities (like the MGF) for scalar random variables.

We may wonder what happens with random vectors.

Obviously, we may work with the moments of the marginal distributions of each element of the random vector.

- But, just as the joint distribution was more than just the set of marginal distributions,
- ... there is something to learn from joint moments.

Multivariate Expectations and Moments

Outline

The scalar case is not enough...

So far, we considered the expectation of a function of a univariate random variable. But...

Theorem (3.7)

Let $(X_1, ..., X_n)$ be a multivariate random variable with joint pdf $f\,(x_1, ..., x_n)$. Then the expectation of random variable $Y = g(X_1, ..., X_n)$ is given by

$$\mathrm{E}(Y) = \begin{cases} \displaystyle\sum \cdots \sum_{(x_1,...,x_n)\in \mathrm{R}(X)} g\,(x_1, ..., x_n)\, f\,(x_1, ..., x_n) & \textit{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g\,(x_1, ..., x_n)\, f\,(x_1, ..., x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n & \textit{(continuous).} \end{cases}$$

Multivariate results

Theorem (3.8)

$$\mathrm{E}\left(\sum_{i=1}^{k} g_i(X_1,\ldots,X_n)\right) = \sum_{i=1}^{k}\mathrm{E}(g_i(X_1,\ldots,X_n)).$$

Corollary (3.2)

$$\mathrm{E}\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k}\mathrm{E}(X_i).$$

And the much more interesting

Theorem (3.9)

Let $X_1,\ldots X_n$ be independent random variables. Then
$\mathrm{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n}\mathrm{E}(X_i)$.

Joint distributions...

In the case of multivariate random variables, *joint moments* characterize the relationship between the individual variables.

Definition (Joint non-central moment)

Let X and Y be two random variables with joint pdf $f(x,y)$. Then the joint non-central moment of (X,Y) of order (r,s) is defined as

$$\mu'_{r,s} = \mathrm{E}\left(X^r Y^s\right) = \begin{cases} \sum\limits_{x \in \mathrm{R}(X)} \sum\limits_{y \in \mathrm{R}(Y)} x^r y^s f\left(x,y\right) & \text{(discrete)} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^r y^s f\left(x,y\right) \mathrm{d}x \mathrm{d}y & \text{(continuous).} \end{cases}$$

Multivariate MGFs

Definition (Moment-Generating Function; multivariate)

The MGF of a multivariate random variable $\boldsymbol{X} = (X_1, \ldots, X_N)'$ is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}\left(e^{\boldsymbol{t}'\boldsymbol{X}}\right) = \mathrm{E}\left(e^{\sum_{i=1}^{n} t_i X_i}\right), \qquad \text{where} \qquad \boldsymbol{t} = (t_1, ..., t_n)',$$

if the expectation exists for all t_i in some neighborhood of 0, $i = 1, ..., n$. I.e. $\exists\ h > 0$ such that $\mathrm{E}\left(e^{\boldsymbol{t}'\boldsymbol{X}}\right)$ exists $\forall\ t_i \in (-h, h)$, $i = 1, ..., n$.

The rth order non-central moment of X_i obtains from the rth order partial derivative w.r.t. t_i, $\quad \mu_r'(X_i) = \mathrm{E}(X_i^r) = \left.\frac{\partial^r M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_i^r}\right|_{\boldsymbol{t}=0}$.

Cross partial derivatives deliver *joint non-central moments*,

$$\mathrm{E}(X_i^r X_j^s) = \left.\frac{\partial^{r+s} M_{\boldsymbol{X}}(\boldsymbol{t})}{\partial t_i^r \partial t_j^s}\right|_{\boldsymbol{t}=0}.$$

The central version

Definition (Joint central moment)

Let X and Y be two random variables with joint pdf $f(x, y)$. Then the joint central moment of (X, Y) of order (r, s) is defined as

$$\mu_{r,s} = \begin{cases} \displaystyle\sum_{x \in \mathrm{R}(X)} \sum_{y \in \mathrm{R}(Y)} (x - \mathrm{E}(X))^r (y - \mathrm{E}(Y))^s f(x, y) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^r (y - \mathrm{E}(Y))^s f(x, y) \, \mathrm{d}x \mathrm{d}y & \text{(continuous).} \end{cases}$$

The joint moment of order $(1, 1)$, namely $\mu_{1,1}$, is commonly known as the covariance, which measures the 'linear association' between X and Y.

We use it so often that it pays to discuss it in more detail.

Outline

The queen of joint moments

Definition (Covariance)

The **covariance between the random variables** X **and** Y is the joint central moment of the order $(1, 1)$,

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = \mathrm{E}\left((X - \mathrm{E}(X))\,(Y - \mathrm{E}(Y))\right).$$

The covariance can be represented in terms of non-central moments:

$$\begin{aligned} \sigma_{XY} &= \mathrm{E}\left((X - \mathrm{E}(X))\,(Y - \mathrm{E}(Y))\right) \\ &= \mathrm{E}\left(XY - \mathrm{E}(X)Y - \mathrm{E}(Y)X + \mathrm{E}(X)\,\mathrm{E}(Y)\right) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y). \end{aligned}$$

From this relationship we obtain e.g. that

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y) \quad \text{iff} \quad \sigma_{XY} = 0.$$

Some results

Theorem (3.16 (Cauchy-Schwarz Inequality))

$(\mathrm{E}(WZ))^2 \leq \mathrm{E}(W^2)\,\mathrm{E}(Z^2)$.

Theorem (3.17 (Covariance bound))

$|\sigma_{XY}| \leq \sigma_X \sigma_Y$.

Using this upper bound, we can define a normalized version of the covariance, the so-called **correlation**.

Definition (Correlation)

The correlation between the random variables X and Y is defined by

$$\operatorname{corr}(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

More on correlation

Theorem (3.18 (Correlation bound))

$-1 \leq \rho_{XY} \leq 1.$

A fundamental relationship between the covariance and the stochastic (in)dependence is indicated in the next theorem.

Theorem (3.19)

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$.

The converse of the theorem is not true: The fact that $\sigma_{XY} = 0$ *does not necessarily imply* that X and Y are independent:

Let X and Y have the joint pdf $f(x, y) = 1.5\mathbb{I}_{[-1,1]}(x)\mathbb{I}_{[0,x^2]}(y)$. The correlation is zero but the variables are not independent...

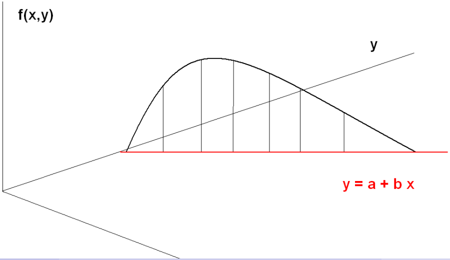
At the other end of the scale

Theorem (3.20)

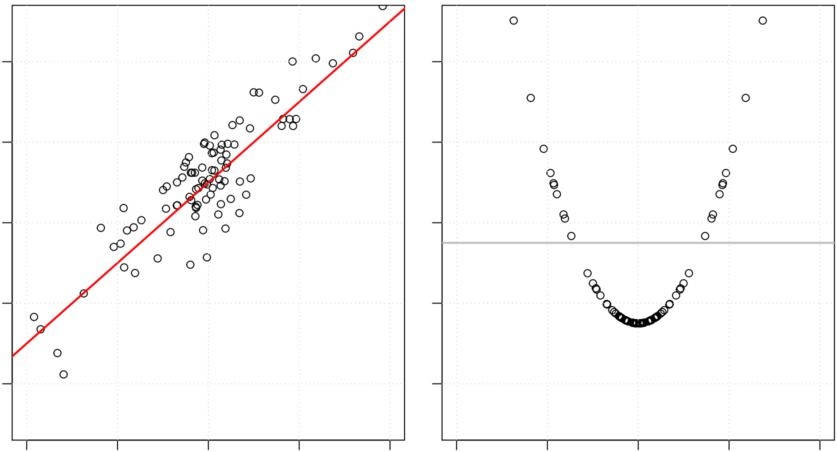
If $\rho_{XY} = 1$ or -1, then $\mathrm{P}(Y = a + bX) = 1$, where $b \neq 0$.

If $\rho_{XY} = 1$ or -1 such that $\mathrm{P}(y = a + bx) = 1$, then the joint pdf $f(x, y)$ is **degenerate**. All the probability mass of $f(x, y)$ is concentrated above the line $y = a + bx$.

This generates a perfect linear relationship between X and Y .



Don't overinterpret!



Left: correlation, imperfect relation; Right: no correlation, perfect dependence.

Mean and variance of linear combinations

Theorem (3.21)

Let $Y = \sum_{i=1}^{n} a_i X_i$*, where* a_i *are constant. Then* $\mathrm{E}(Y) = \sum_{i=1}^{n} a_i \, \mathrm{E}(X_i)$*.*

The matrix representation of this result obtains as follows. Let

$$\boldsymbol{a} = (a_1, \ldots, a_n)' \quad \text{and} \quad \boldsymbol{X} = (X_1, \ldots, X_n)'.$$

Then $\qquad Y = \boldsymbol{a}'\boldsymbol{X} \qquad$ such that $\qquad \mathrm{E}(Y) = \boldsymbol{a}' \, \mathrm{E}(\boldsymbol{X})$.

Theorem (3.22)

Let $Y = \sum_{i=1}^{n} a_i X_i$*, where the* a_i*s are constants. Then*

$$\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2 + 2 \sum_{i<j} a_i a_j \sigma_{X_i X_j}.$$

In order to rewrite this result in matrix notation we shall define the **covariance matrix** of a multivariate random variable.

Covariance matrix

Definition

The covariance matrix of the n-dimensional random vector $\boldsymbol{X} = (X_1, \dots, X_n)'$ is the $n \times n$ symmetric matrix

$$\mathrm{Cov}(\boldsymbol{X}) = \mathrm{E}\left((\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))'\right) = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma^2_{X_n} \end{pmatrix}.$$

- The variance of the ith variable in $\boldsymbol{X}$ is given by the (i, i)th diagonal entry in the covariance matrix.
- A covariance matrix is symmetric, that is $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{X})'$.

The matrix expressions

Let $\boldsymbol{a} = (a_1, \ldots, a_n)'$ and $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Then the variance of $Y = \boldsymbol{a}'\boldsymbol{X}$ given in the theorem can obviously be represented as

$$\sigma_Y^2 = \boldsymbol{a}' \operatorname{Cov}(\boldsymbol{X})\boldsymbol{a}.$$

Note that since a variance is non-negative $\left(\sigma_Y^2 \geq 0\right)$ the expression $\boldsymbol{a}' \operatorname{Cov}(\boldsymbol{X})\boldsymbol{a}$ is also non-negative for any a.

This implies that a covariance matrix is necessarily positive semidefinite!

More matrices

Theorem (3.23)

Let $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X}$, *where* $\mathbf{A} = (a_{hm})$ *is a* $k \times n$ *matrix of real constants, and* $\boldsymbol{X} = (X_i)$ *is an* $n \times 1$ *vector of random variables. Then* $\mathrm{E}(\boldsymbol{Y}) = \mathbf{A}\,\mathrm{E}(\boldsymbol{X})$.

Theorem (3.24)

Let $\boldsymbol{Y} = \mathbf{A}\boldsymbol{X}$, *where* $\mathbf{A} = (a_{hm})$ *is a* $k \times n$ *matrix of real constants, and* $\boldsymbol{X} = (X_i)$ *is a* $n \times 1$ *vector of random variables. Then* $\mathrm{Cov}(\boldsymbol{Y}) = \mathbf{A}\,\mathrm{Cov}(\boldsymbol{X})\mathbf{A}'$.

Outline

The same thing?

- So far, we have considered unconditional expectations, this means the expectations of unconditional/marginal distributions.
- If we take the expectation w.r.t. a conditional distribution, we have the conditional expectation.
- The conditional expectation is one of the most important concepts used in econometrics and empirical economics.
- It is for instance the key element of regression analysis, telling us how a variable reacts on the average to changes in other variables.[1]

[1] This is one interpretation, don't expect uniqueness thereof.

A conditional distribution is just a distribution

Definition (Conditional expectation)

Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be random vectors with joint pdf $f(x_1, \ldots, x_n, y_1, \ldots, y_m)$. The conditional expectation of $g(Y_1, \ldots, Y_m)$, given $(X_1, \ldots, X_n) \in B$, is defined as

(discrete) $\quad \mathrm{E}\left(g\left(Y_1, \ldots, Y_m\right) \mid \left(X_1, \ldots, X_n\right) \in B\right)$

$$= \sum \cdots \sum_{(y_1, \ldots, y_m) \in \mathrm{R}(Y)} g\left(y_1, \ldots, y_m\right) f\left(y_1, \ldots, y_m \mid \left(x_1, \ldots, x_n\right) \in B\right)$$

(continuous) $\quad \mathrm{E}\left(g\left(Y_1, \ldots, Y_m\right) \mid \left(X_1, \ldots, X_n\right) \in B\right)$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g\left(y_1, \ldots, y_m\right) f\left(y_1, \ldots, y_m \mid \left(x_1, \ldots, x_n\right) \in B\right) \mathrm{d}y_1 \cdots \mathrm{d}y_m.$$

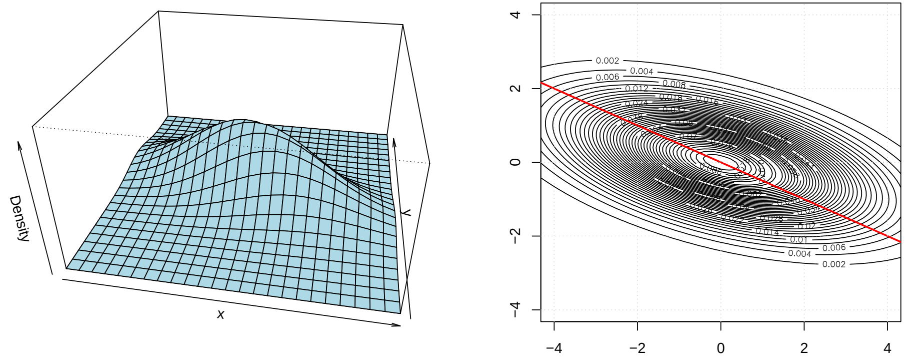
The regression function

An important special case of the definition given above obtains by setting $g(Y_1, ..., Y_n) = Y$, where Y is a univariate random variable, and B is an elementary event.

$$\mathrm{E}\left(Y|\boldsymbol{X}=\boldsymbol{x}\right) = \begin{cases} \displaystyle\sum_{y\in\mathrm{R}(Y)} y\cdot f\left(y \mid \boldsymbol{X}=\boldsymbol{x}\right) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} y\cdot f\left(y \mid \boldsymbol{X}=\boldsymbol{x}\right)\mathrm{d}y & \text{(continuous).} \end{cases}$$

This is a function of $\boldsymbol{x}$; we call it **the regression curve** of Y on $\boldsymbol{X}$.

An example



Left: bivariate pdf, correlation; Right: level curves and regression line

A nonlinear example

Take the bivariate random variable with joint pdf

$$f(x,y) = \frac{1}{96}(x^2 + 2xy + 2y^2)\mathbb{I}_{[0,4]}(x)\mathbb{I}_{[0,2]}(y).$$

The regression function of a regression of Y on X is obtained as

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \int_{-\infty}^{\infty} y \cdot \frac{f(x,y)}{f_X(x)}\mathrm{d}y = \int_0^2 \frac{y \cdot (x^2 + 2xy + 2y^2)\mathbb{I}_{[0,4]}(x)}{(2x^2 + 4x + \frac{16}{3})\mathbb{I}_{[0,4]}(x)}\mathrm{d}y \\ &= \frac{2x^2 + \frac{16}{3}x + 8}{2x^2 + 4x + \frac{16}{3}} \quad \text{for} \quad x \in [0,4]. \end{aligned}$$

For $x \notin [0,4]$, the regression function is not defined.

Getting more random

- The conditional expectation $\mathrm{E}(Y|(X_1, ..., X_n) \in B)$ was introduced as being conditional on a *particular event* B, e.g. $B = ((X_1, ..., X_n) = (x_1, ..., x_n))$.
- Rather than specifying a particular event, we might conceptualize leaving the event for $(X_1, ..., X_n)$ *unspecified* and interpret the conditional expectation of Y as a function of $(X_1, ..., X_n)$ denoted by $\mathrm{E}(Y|X_1, ..., X_n)$.
- Note that $\mathrm{E}(Y|X_1, ..., X_n)$ is then a function of random variables and, therefore, itself a random variable.
- $\mathrm{E}\left(Y|(X_1, ..., X_n) = (x_1, ..., x_n)\right) = \mathrm{E}(Y|x_1, ..., x_n)$ is referred to as the regression function of a regression of Y on the X_is.

Recovering the unconditional

One might ask whether there's any relation between unconditional and conditional expectations. And...

Theorem (3.10)

$\mathrm{E}\left(\mathrm{E}(g(Y)|\boldsymbol{X})\right) = \mathrm{E}(g(Y)).$

For random vectors, we get

$$\mathrm{E}\left(\mathrm{E}(g(Y_1, ...Y_n)|X_1, ..., X_n)\right) = \mathrm{E}(g(Y_1, ..., Y_n)).$$

Final remark: All properties of expectations discussed above also apply analogously to conditional expectations.

Outline

Coming up

Parametric families of distributions