A sztochasztika alapjai
MBNXK262

8. előadás: Véletlen vektorváltozók, függetlenség, kovariancia, korreláció

Kevei Péter

2023/24 tavasz

Véletlen vektorváltozók

Definíció

$\xi = (\xi_1, \ldots, \xi_n) : \Omega \to \mathbb{R}^n$ véletlen vektorváltozó, ha minden komponense véletlen változó. Eloszlásfüggvénye

$$F(x_1, \ldots, x_n) = \mathbf{P}(\xi_1 < x_1, \ldots, \xi_n < x_n).$$

$(\xi_1, \ldots, \xi_n)$ véletlen vektorváltozó diszkrét, ha értékkészlete megszámlálható.
ξ_i, $i = 1, 2, \ldots, n$, eloszlása a peremeloszlás, vagy marginális eloszlás

Trinomiális eloszlás

Szabályos dobókockával n-szer dobunk. ξ a hatosok, η egyesek száma
peremeloszlások: mind a hatosok, mind az egyesek száma *binomiális eloszlású* véletlen változó $(n, 1/6)$ paraméterrel:

$$\mathbf{P}(\xi = k) = \mathbf{P}(\eta = k) = \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Így,

$$\mathbf{E}(\xi) = \frac{n}{6}, \quad \mathbf{D}^2(\xi) = n\frac{5}{36}.$$

Trinomiális eloszlás

Szabályos dobókockával n-szer dobunk. ξ a hatosok, η egyesek száma
(ξ, η) lehetséges értékei: olyan (k, ℓ) párok, melyre $0 \le k, \ell \le n$, és $k + \ell \le n$. A megfelelő valószínűségek

$$\mathbf{P}(\xi = k, \eta = \ell) = \binom{n}{k} \binom{n-k}{\ell} \left(\frac{1}{6}\right)^{k+\ell} \left(\frac{4}{6}\right)^{n-k-\ell},$$

Függetlenség

Definíció

$\xi_1, \ldots, \xi_n$ függetlenek, ha minden $x_1, \ldots, x_n \in \mathbb{R}$ esetén

$$\mathbf{P}(\xi_1 < x_1, \ldots, \xi_n < x_n) = \mathbf{P}(\xi_1 < x_1) \ldots \mathbf{P}(\xi_n < x_n).$$

Állítás

$\xi_1, \ldots, \xi_n$ diszkrét véletlen változók úgy, hogy ξ_i lehetséges értékei $x_1^{(i)}, x_2^{(i)}, \ldots$, $i = 1, 2, \ldots, n$. Ekkor $\xi_1, \ldots, \xi_n$ pontosan akkor függetlenek, ha

$$\mathbf{P}(\xi_1 = x_{i_1}^{(1)}, \ldots, \xi_n = x_{i_n}^{(n)}) = \mathbf{P}(\xi_1 = x_{i_1}^{(1)}) \ldots \mathbf{P}(\xi_n = x_{i_n}^{(n)})$$

tetszőleges $i_i, \ldots, i_n$ indexekre.

Kockadobás

Példa

Egy szabályos dobókockával kétszer dobunk. Jelölje ξ az első, η a második dobás eredményét. Ekkor tetszőleges $k, \ell \in \{1, 2, \ldots, 6\}$ esetén

$$\mathbf{P}(\xi = k, \eta = \ell) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathbf{P}(\xi = k) \cdot \mathbf{P}(\eta = \ell),$$

azaz ξ és η függetlenek.

Állítás

ξ, η független diszkrét véletlen változók.

$$\mathbf{E}\left(g_1(\xi)g_2(\eta)\right) = \mathbf{E}\left(g_1(\xi)\right)\mathbf{E}\left(g_2(\eta)\right).$$

Speciálisan, ha ξ és η függetlenek, akkor $\mathbf{E}(\xi\eta) = \mathbf{E}(\xi)\mathbf{E}(\eta)$.

$$\begin{aligned}
&\mathbf{E}(g_1(\xi)g_2(\eta)) \\
&= \sum_i \sum_j g_1(x_i)g_2(y_j)\mathbf{P}(\xi = x_i, \eta = y_j) && \mathbf{E}(h(\xi,\eta))=\sum\sum h(x_i,y_j)\mathbf{P}(\xi=x_i,\eta=y_j) \\
&= \sum_i \sum_j g_1(x_i)g_2(y_j)\mathbf{P}(\xi = x_i)\mathbf{P}(\eta = y_j) && \text{függetlenség} \\
&= \sum_i g_1(x_i)\mathbf{P}(\xi = x_i) \sum_j g_2(y_j)\mathbf{P}(\eta = y_j) \\
&= \mathbf{E}(g_1(\xi))\mathbf{E}(g_2(\eta)).
\end{aligned}$$

Kovariancia, korreláció

Definíció

Az ξ és η véletlen változók *kovarianciája*

$$\mathbf{Cov}(\xi, \eta) = \mathbf{E}[(\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))],$$

korrelációja

$$\rho(\xi, \eta) = \frac{\mathbf{Cov}(\xi, \eta)}{\mathbf{D}(\xi)\mathbf{D}(\eta)}.$$

Tulajdonságok

Állítás

Tetszőleges $\xi, \xi_1, \ldots, \xi_n, \eta, \eta_1, \ldots, \eta_m$ *véletlen változók és* a, b *valós számok esetén igazak az alábbiak.*

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

(ii) $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi)$;

(iii) $\mathbf{Cov}(\xi, \eta) = \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta)$;

(iv) $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta)$;

(v) $\mathbf{Cov}\left(\sum_{i=1}^{n} \xi_i, \sum_{j=1}^{m} \eta_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{Cov}(\xi_i, \eta_j)$;

(vi) *ha* ξ *és* η *függetlenek, akkor* $\mathbf{Cov}(\xi, \eta) = 0$.

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

(ii) $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi)$;

A definíció azonnali következménye.

(iii) $\mathbf{Cov}(\xi, \eta) = \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta)$;

$$\begin{aligned}\mathbf{Cov}(\xi, \eta) &= \mathbf{E}\left((\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))\right) \\ &= \mathbf{E}\left(\xi\eta - \xi\mathbf{E}(\eta) - \mathbf{E}(\xi)\eta + \mathbf{E}(\xi)\mathbf{E}(\eta)\right) \\ &= \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta).\end{aligned}$$

(iv) $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta)$;

$\mathbf{Cov}(a(\xi + c), b(\xi + d))$

$= \mathbf{E}\left[(a(\xi + c) - \mathbf{E}(a(\xi + c)))(b(\eta + d) - \mathbf{E}(b(\eta + d)))\right]$

$= ab\mathbf{E}((\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))) = ab\mathbf{Cov}(\xi, \eta).$

(v) $\mathbf{Cov}\left(\sum_{i=1}^{n} \xi_i, \sum_{j=1}^{m} \eta_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{Cov}(\xi_i, \eta_j)$;

$$\mathbf{Cov}\left(\sum_{i=1}^{n} \xi_i, \sum_{j=1}^{m} \eta_j\right) = \mathbf{E}\left(\sum_{i=1}^{n} (\xi_i - \mathbf{E}(\xi_i)) \sum_{j=1}^{m} (\eta_j - \mathbf{E}(\eta_j))\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{E}((\xi_i - \mathbf{E}(\xi_i))(\eta_j - \mathbf{E}(\eta_j)))$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{Cov}(\xi_i, \eta_j).$$

(vi) ha ξ és η függetlenek, akkor $\mathbf{Cov}(\xi, \eta) = 0$.

A függetlenségből következik, hogy $\mathbf{E}(\xi\eta) = \mathbf{E}(\xi)\mathbf{E}(\eta)$, így (iii) szerint $\mathbf{Cov}(\xi, \eta) = 0$.

Állítás

(i) *Bunyakovszkij–Cauchy–Schwarz-egyenlőtlenség:*

$$|\mathbf{Cov}(\xi,\eta)| \le \mathbf{D}(\xi)\mathbf{D}(\eta).$$

Innen adódik, hogy $\rho(\xi,\eta) \in [-1,1]$*;*

(ii) *ha* $\rho(\xi,\eta) = 1$*, akkor*

$$\xi = \mathbf{E}(\xi) + \frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta - \mathbf{E}(\eta));$$

(iii) *ha* $\rho(\xi,\eta) = -1$*, akkor*

$$\xi = \mathbf{E}(\xi) - \frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta - \mathbf{E}(\eta)).$$

Bizonyítás

U, V véletlen változók, $t \in \mathbb{R}$: $\mathbf{E}[(U+tV)^2] \geq 0$, ezért a

$$p(t) = \mathbf{E}[(U+tV)^2] = t^2\mathbf{E}(V^2) + 2t\mathbf{E}(UV) + \mathbf{E}(U^2)$$

t-ben másodfokú polinom diszkriminánsa nempozitív. Azaz

$$4[\mathbf{E}(UV)]^2 \leq 4\mathbf{E}(U^2)\mathbf{E}(V^2),$$

így

$$|\mathbf{E}(UV)| \leq \sqrt{\mathbf{E}(U^2)\mathbf{E}(V^2)}.$$

Az $U = \xi - \mathbf{E}(\xi)$ és $V = \eta - \mathbf{E}(\eta)$ választással ez éppen az állítás.

Bizonyítás

Ha $|\rho(\xi,\eta)| = 1$, akkor a másodfokú p polinom diszkriminánsa 0. Így

$$t_0 = -\frac{\mathbf{E}[(\xi - \mathbf{E}\xi)(\eta - \mathbf{E}\eta)]}{\mathbf{E}[(\eta - \mathbf{E}\eta)^2]} = -\rho(\xi,\eta)\frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}$$

zérushely, vagyis

$$\xi - \mathbf{E}(\xi) + t_0(\eta - \mathbf{E}(\eta)) = 0,$$

ami éppen a bizonyítandó.

Összeg szórásnégyzete

Legyenek ξ, η véletlen változók, ρ a korrelációjuk. Ekkor

$$\begin{aligned}\mathbf{D}^2(\xi+\eta) &= \mathbf{Cov}(\xi+\eta, \xi+\eta) \\ &= \mathbf{Cov}(\xi,\xi) + 2\mathbf{Cov}(\xi,\eta) + \mathbf{Cov}(\eta,\eta) \\ &= \mathbf{D}^2(\xi) + 2\mathbf{D}(\xi)\mathbf{D}(\eta)\rho + \mathbf{D}^2(\eta).\end{aligned}$$

Összeg szórásnégyzete

Ha ξ és η függetlenek, akkor $\rho = 0$, így

$$\mathbf{D}^2(\xi + \eta) = \mathbf{D}^2(\xi) + \mathbf{D}^2(\eta).$$

Indukcióval, ha $\xi_1, \xi_2, \ldots, \xi_n$ páronként független (korrelálatlan) véletlen változók, akkor

$$\mathbf{D}^2\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mathbf{D}^2(\xi_i).$$

Kovariancia

Példa

Egy szabályos dobókockával n-szer dobunk. Jelölje ξ a hatosok, η egyesek számát! Adjuk meg a várható értéket, szórást, kovarianciát, korrelációt!

A kovariancia tulajdonságai szerint

$$\mathbf{D}^2(\xi+\eta) = \mathbf{D}^2(\xi) + 2\mathbf{Cov}(\xi,\eta) + \mathbf{D}^2(\eta).$$

Mivel ξ, η binomiális $(n, 1/6)$ paraméterekkel, így $\mathbf{D}^2(\xi) = \mathbf{D}^2(\eta) = n\frac{5}{6}\cdot\frac{1}{6}$. Továbbá, $\xi+\eta$ binomiális $(n, 2/6)$ paraméterekkel, így $\mathbf{D}^2(\xi+\eta) = n\frac{1}{3}\cdot\frac{2}{3}$. Rendezve adódik, hogy

$$\mathbf{Cov}(\xi,\eta) = -\frac{n}{36}, \quad \rho(\xi,\eta) = -\frac{1}{5}.$$

Lineáris regresszió felé

Példa

3, külsőre egyforma érmével a fejdobás valószínűsége $1/4, 2/4, 3/4$. Véletlenszerűen választunk egy érmét, és azzal kétszer dobunk. Legyen η a választott érmével dobva a fej valószínűsége, ξ a dobott fejek száma. ξ-ből szeretnénk η értékére következtetni.

ξ lehet 0,1,2, míg η lehet $1/4$, $1/2$, $3/4$. A szorzási szabállyal kapjuk

$$\mathbf{P}\left(\eta = \frac{1}{4}, \xi = 0\right) = \mathbf{P}\left(\eta = \frac{1}{4}\right) \cdot \mathbf{P}\left(\xi = 0 | \eta = \frac{1}{4}\right)$$
$$= \frac{1}{3} \cdot \left(\frac{3}{4}\right)^2 = \frac{9}{48}.$$

Lineáris regresszió felé

Hasonlóan

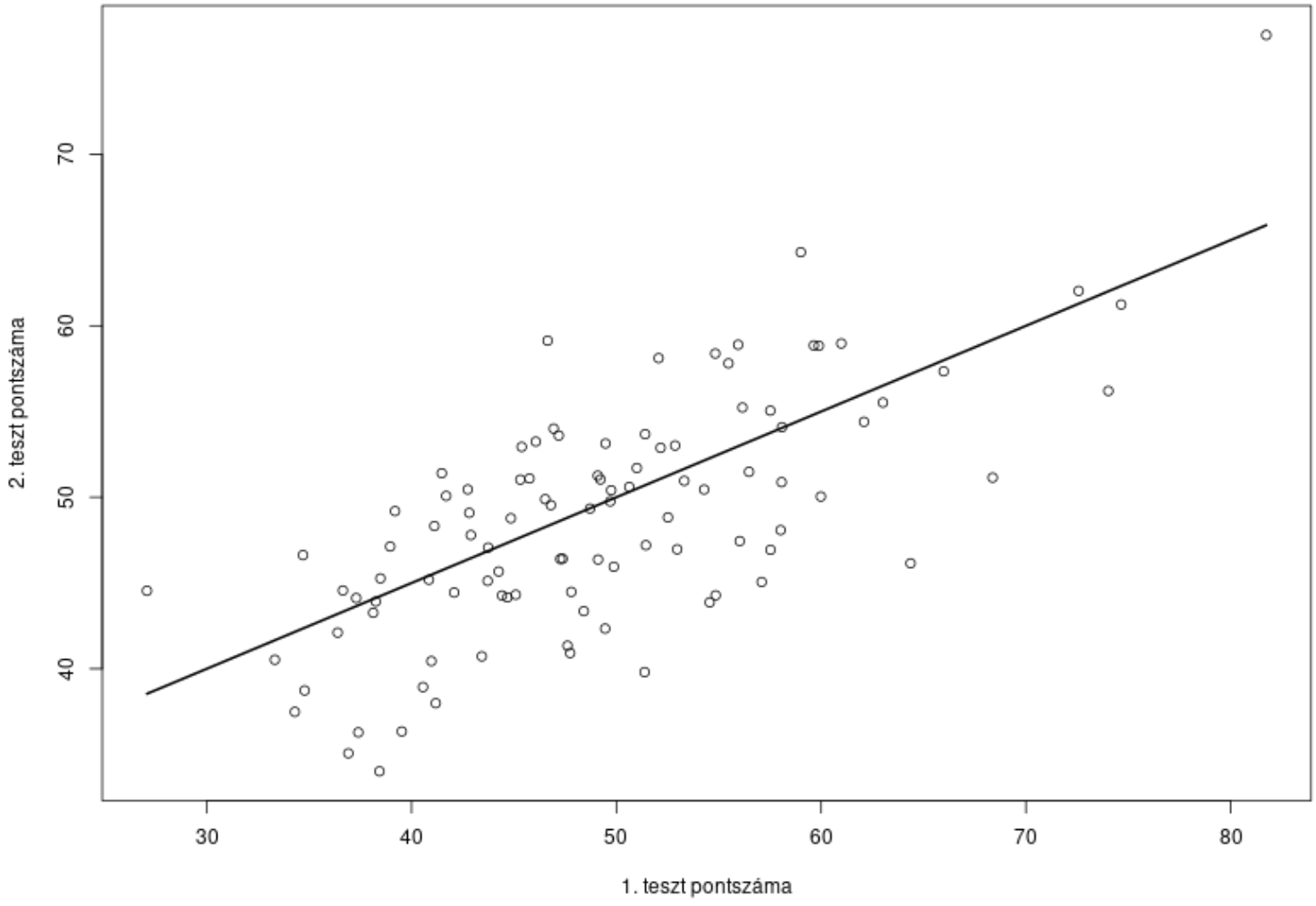
$$\mathbf{P}\left(\eta = \frac{1}{2}, \xi = 1\right) = \frac{1}{3} \cdot 2\left(\frac{1}{2}\right)^2 = \frac{1}{6}.$$

(ξ, η) együttes eloszlása:

$\eta \backslash \xi$	0	1	2	$\sum$
$\frac{1}{4}$	$\frac{9}{48}$	$\frac{6}{48}$	$\frac{1}{48}$	$\frac{1}{3}$
$\frac{1}{2}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{3}$
$\frac{3}{4}$	$\frac{1}{48}$	$\frac{6}{48}$	$\frac{9}{48}$	$\frac{1}{3}$
$\sum$	$\frac{14}{48}$	$\frac{20}{48}$	$\frac{14}{48}$	1

Lineáris regresszió felé

A sztochasztika alapjai kurzus elején a hallgatók kitöltenek egy tesztet, mely az eddigi matematika tudásukat méri. A kurzus teljesítéséhez a félév végén is kitöltenek egy tesztet a kurzus anyagából. A korábbi évek tapasztalatai alapján feltehető, hogy az i-edik hallgató első teszten elért pontszáma ξ_i, a másodikon $\eta_i = \frac{\xi_i + \xi_i'}{2}$, ahol $\xi_1, \xi_2, \ldots$ és $\xi_1', \xi_2', \ldots$ független normális eloszlású változók $\mu = 50$ várható értékkel és $\sigma = 10$ szórással, $i = 1, 2, \ldots, N$, ahol N a hallgatók száma.



ábra: 100 hallgató pontszámának alakulása

Számítsuk ki a két változó, ξ és η korrelációs együtthatóját! A kovariancia tulajdonságai szerint

$$\mathbf{Cov}(\xi,\eta)=\mathbf{Cov}\left(\xi,\frac{\xi+\xi'}{2}\right)=\frac{1}{2}\mathbf{Cov}(\xi,\xi)+\frac{1}{2}\mathbf{Cov}(\xi,\xi')=\frac{1}{2}\mathbf{D}^2(\xi).$$

A szórások $\mathbf{D}^2(\xi)=100$, és a függetlenség miatt

$$\mathbf{D}^2(\eta)=\mathbf{D}^2\left(\frac{\xi+\xi'}{2}\right)=\frac{1}{4}\left(\mathbf{D}^2(\xi)+\mathbf{D}^2(\xi')\right)=50,$$

ahonnan a korrelációs együttható

$$\rho(\xi,\eta)=\frac{\mathbf{Cov}(\xi,\eta)}{\mathbf{D}(\xi)\mathbf{D}(\eta)}=\frac{1}{2}\frac{\mathbf{D}^2(\xi)}{\mathbf{D}(\xi)\mathbf{D}(\eta)}=\frac{1}{\sqrt{2}}.$$

Ha ismerjük ξ értékét, akkor

$$\mathbf{E}[\eta|\xi] = \mathbf{E}\left[\frac{\xi + \xi'}{2}|\xi\right] = 25 + \frac{\xi}{2},$$

hiszen ξ értékét pontosan tudom, ξ' várható értéke pedig 50, és ξ független ξ'-től. Ez éppen a regressziós egyenes.

Lineáris regresszió

(ξ, η) véletlen vektorváltozó. Az η változót tekintem *függő* változónak, ennek az értékére szeretnék következtetni a ξ *független változó* értékéből. Vagyis ismert ξ esetén szeretném megmondani η-t. Keressük azokat az a, b valós számokat, melyre a $\eta - (a\xi + b)$ változó kicsi. A kicsiséget négyzetes hibában mérve, keressük az

$$h(a, b) = \mathbf{E}\left[(\eta - (a\xi + b))^2\right]$$

függvény minimumhelyét, azaz a legjobb a, b választást.

Lineáris regresszió

$$
\begin{aligned}
&\mathbf{E}\left[(\eta-(a\xi+b))^2\right]\\
&=\mathbf{E}\left[((\eta-a\xi)-\mathbf{E}(\eta-a\xi)+\mathbf{E}(\eta-a\xi)-b)^2\right]\\
&=\mathbf{E}\left[((\eta-a\xi)-\mathbf{E}(\eta-a\xi))^2\right]+[\mathbf{E}(\eta-a\xi)-b]^2\\
&\quad+2\mathbf{E}\left[((\eta-a\xi)-\mathbf{E}(\eta-a\xi))\right](\mathbf{E}(\eta-a\xi)-b)\\
&=\mathbf{E}\left[((\eta-a\xi)-\mathbf{E}(\eta-a\xi))^2\right]+[\mathbf{E}(\eta-a\xi)-b]^2
\end{aligned}
$$

látjuk, hogy $b=\mathbf{E}(\eta)-a\mathbf{E}(\xi)$ választás adja b-ben a minimumhelyet. Tehát a $\mathbf{D}^2(\eta-a\xi)$ mennyiség minimuma kell a-ban. A szórásnégyzetre vonatkozó formulák szerint

$$
\begin{aligned}
\mathbf{D}^2(\eta-a\xi)&=\mathbf{Cov}(\eta-a\xi,\eta-a\xi)\\
&=\mathbf{D}^2(\eta)+a^2\mathbf{D}^2(\xi)-2a\mathbf{Cov}(\eta,\xi).
\end{aligned}
$$

Lineáris regresszió

$$\mathbf{D}^2(\eta - a\xi) = \mathbf{D}^2(\eta) + a^2\mathbf{D}^2(\xi) - 2a\mathbf{Cov}(\eta, \xi).$$

Ez a-ban másodfokú polinom, főegyütthatója pozitív. Ennek minimumhelye

$$a = \frac{2\mathbf{Cov}(\eta, \xi)}{2\mathbf{D}^2(\xi)} = \frac{\mathbf{Cov}(\eta, \xi)}{\mathbf{D}^2(\xi)}.$$

Ezek szerint a legjobb lineáris közelítést a

$$g(x) = \frac{\mathbf{Cov}(\eta, \xi)}{\mathbf{D}^2(\xi)}(x - \mathbf{E}(\xi)) + \mathbf{E}(\eta)$$

függvény adja. Ő a *regressziós egyenes*.
Ha ξ és η korrelálatlanok, azaz $\mathbf{Cov}(\xi, \eta) = 0$, akkor a legjobb közelítés $\mathbf{E}(\eta)$, vagyis ξ semmi információt nem ad η értékéről.