Valószínűségszámítás

2021. november 17.

Mészáros Szabolcs

Tárgyhonlap:
cs.bme.hu/valszam

A prezentáció anyagát és az abból készült videofelvételt a tárgy hallgatói jogosultak használni, kizárólag saját célra. A felvétel másolása, videómegosztókra való feltöltése részben vagy egészben tilos, illetve csak a tantárgyfelelős előzetes engedélyével történhet.

Ismétlés: kovariancia

Definíció: Legyenek X és Y valószínűségi változók. Ekkor

$$\operatorname{cov}(X,Y) = \mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Kérdés: Hogyan számoljuk ki $\mathbb{E}(XY)$-t, ha csak az $f_{X,Y}$ együttes sűrűségfüggvényt ismerjük?

Ötlet:
Szorzat sűrűségfüggvényével? De azt hogy kapjuk meg?
(Főleg ha esetleg nem is függetlenek.)

Többdim. transzformált várható értéke

Állítás: Legyen $\underline{X} = (X_1, \dots, X_n)$ folytonos valószínűségi vektorváltozó, és legyen $g : \mathbb{R}^n \to \mathbb{R}$ olyan, amire $\mathbb{E}(g(\underline{X}))$létezik. Ekkor

$$\mathbb{E}(g(\underline{X})) = \int_{\mathbb{R}^n} g(\underline{x}) f_{\underline{X}}(\underline{x}) \mathrm{d}\underline{x}$$

Ha g folytonos és nemnegatív, akkor $\mathbb{E}(g(\underline{X}))$ létezik.

Többdim. transzformált várható értéke

Speciális eset: $g : \mathbb{R}^2 \to \mathbb{R} \qquad g(x, y) = x \cdot y$

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x, y) \, \mathrm{d}x \mathrm{d}y$$

Példa: X éves csapadékmennyiség (/1000 mm),

Y idén eladott esernyők száma (/1000 db).

Mennyi a kovarianciájuk, ha az együttes sűrűségfüggvényük:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{5}(4 - 2x^2 + xy - y^2) & \text{ha } 0 < x < 1 \text{ és } 0 < y < 2, \\ 0 & \text{egyébként.} \end{cases}$$

Többdim. transzformált várható értéke, példa

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot y\cdot f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y$$

$$= \int_{0}^{2}\int_{0}^{1} xy\cdot\frac{1}{5}(4-2x^2+xy-y^2)\mathrm{d}x\mathrm{d}y$$

$$= \frac{1}{5}\int_{0}^{2}\int_{0}^{1} (4xy-2x^3y+x^2y^2-xy^3)\mathrm{d}x\mathrm{d}y$$

$$= \frac{1}{5}\int_{0}^{2}\left[2x^2y-\frac{1}{2}x^4y+\frac{1}{3}x^3y^2-\frac{1}{2}x^2y^3\right]_{x=0}^{1}\mathrm{d}y$$

Többdim. transzformált várható értéke, példa

$$= \frac{1}{5}\int_0^2 \left(\frac{3}{2}y + \frac{1}{3}y^2 - \frac{1}{2}y^3\right)\mathrm{d}y = \frac{1}{5}\left[\frac{3}{4}y^2 + \frac{1}{9}y^3 - \frac{1}{8}y^4\right]_0^2 = \frac{17}{45}$$

$$\mathbb{E}(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \qquad \mathbb{E}(Y) = \frac{4}{5}$$

$g(x,y) = x$

$$= \int_0^2 \int_0^1 x \cdot \frac{1}{5}(4 - 2x^2 + xy - y^2)\mathrm{d}x\mathrm{d}y = \frac{7}{15}$$

$$\mathrm{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{17}{45} - \frac{7}{15}\cdot\frac{4}{5} = \frac{1}{225}$$

Kovariancia tulajdonságai

Lemma: Legyen (X, Y, Z) valószínűségi vektorváltozó. Ha az alábbi mennyiségek léteznek, akkor a következők teljesülnek:

1. Ha $c \in \mathbb{R}$, akkor $\mathbb{D}(X + c) = \mathbb{D}(X)$ és $\mathbb{D}(cX) = |c|\mathbb{D}(X)$
2. $\mathbb{D}^2(X+Y) = \mathbb{D}^2(X)+\mathbb{D}^2(Y)+2\mathrm{cov}(X, Y)$
3. $\mathbb{D}^2(X) = 0$ pontosan akkor, ha $\mathbb{P}(X = c) = 1$ valamilyen c-re.
4. Ha X és Y függetlenek, akkor $\mathrm{cov}(X, Y) = 0$.
5. Ha $b, c \in \mathbb{R}$, akkor $\mathrm{cov}(X, bY + cZ) = b \cdot \mathrm{cov}(X, Y) + c \cdot \mathrm{cov}(X, Z)$

Kovariancia tulajdonságai, példa

Lemma: Ha X és Y független valószínűségi változók, g és h folytonos $\mathbb{R} \to \mathbb{R}$ függvények, akkor $g(X)$ és $h(Y)$ is függetlenek.

Példa: A fenti függetlenségi feltétel esetén

$$\operatorname{cov}\left(\frac{2}{X}+Y^2, \frac{2}{Y}-X^2\right) = \underbrace{\operatorname{cov}\left(\frac{2}{X}, \frac{2}{Y}\right)}_{=0} - \operatorname{cov}\left(\frac{2}{X}, X^2\right)$$
$$+\operatorname{cov}\left(Y^2, \frac{2}{Y}\right) - \underbrace{\operatorname{cov}\left(Y^2, X^2\right)}_{=0}$$

Kovariancia-mátrix

Definíció: Az $\underline{X} = (X_1, \dots, X_n)$ val. vektorváltozó *kovarianciamátrix*a:

$$\operatorname{cov}(\underline{X}) = \begin{pmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \dots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & & \vdots \\ \vdots & & \ddots & \\ \operatorname{cov}(X_n, X_1) & \dots & & \operatorname{cov}(X_n, X_n) \end{pmatrix}$$

azaz $\operatorname{cov}(\underline{X})_{i,j} = \operatorname{cov}(X_i, X_j)$

Megj.: $\operatorname{cov}(X_i, X_i) = \mathbb{D}^2(X_i)$

Kovariancia-mátrix, példa

Példa: az előző.

$$\operatorname{cov}\big((X,Y)\big) = \begin{pmatrix} ? & \frac{1}{225} \\ \frac{1}{225} & ? \end{pmatrix}$$

$$\mathbb{D}^2(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 =$$

$$= \int_0^2 \int_0^1 x^2 \cdot \frac{1}{5}(4 - 2x^2 + xy - y^2)\,\mathrm{d}x\,\mathrm{d}y$$

$$-\left(\int_0^2 \int_0^1 x \cdot \frac{1}{5}(4 - 2x^2 + xy - y^2)\,\mathrm{d}x\,\mathrm{d}y\right)^2 = \frac{7}{90}$$

$$\mathbb{D}^2(Y) = \frac{58}{225}$$

Kovariancia-mátrix tulajdonságai

Állítás: Legyen $\underline{X} = (X_1, \ldots, X_n)$ valószínűségi vektorváltozó. Ekkor

1. $\operatorname{cov}(\underline{X})$ szimmetrikus mátrix,
2. $\operatorname{cov}(\underline{X})$ pozitív szemidefinit mátrix

   (azaz $\underline{a}^T \operatorname{cov}(\underline{X})\underline{a} \geq 0, \quad \forall \underline{a} \in \mathbb{R}^n$).

Példa: az $\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$ szimmetrikus mátrix, de nem pozitív szemidefinit, mert

$$\begin{bmatrix} -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \end{bmatrix} = -1$$

Kovariancia-mátrix tulajdonságai, biz.

Biz.: Szimmetrikus, azaz $\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_j, X_i)$

Pozitív szemidefinit: Legyen $Z = \sum_{i=1}^{n} a_i X_i$

$$\mathbb{D}^2(Z) = \mathbb{D}^2\Big(\sum_{i=1}^{n} a_i X_i\Big) = \operatorname{cov}\Big(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} a_j X_j\Big)$$

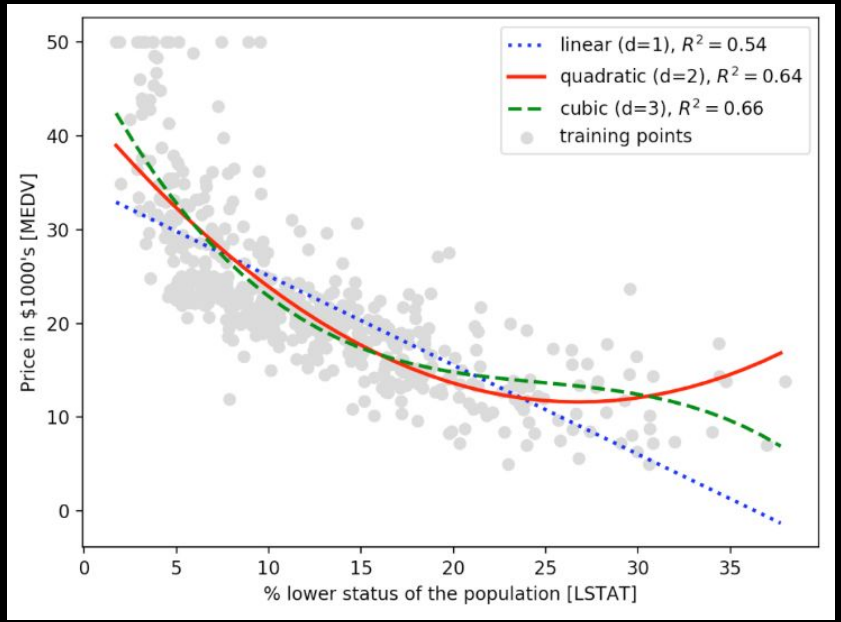
≥ 0

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \operatorname{cov}(X_i, X_j) a_j = \underline{a}^T \operatorname{cov}(\underline{X})\underline{a}$$

Lineáris regresszió, fogalmak

Regresszió változatai:

- lineáris regresszió
  - egyszerű lineáris regresszió,
    avagy legkisebb négyzetek módszere
  - regularizált lineáris regresszió
    (ridge, lasso, ...)
  - ...
- logisztikus regresszió
- lokális regresszió
- ...



Forrás:
https://charleshsliao.wordpress.com/2017/06/16/ransac-and-nonlinear-regression-in-python/

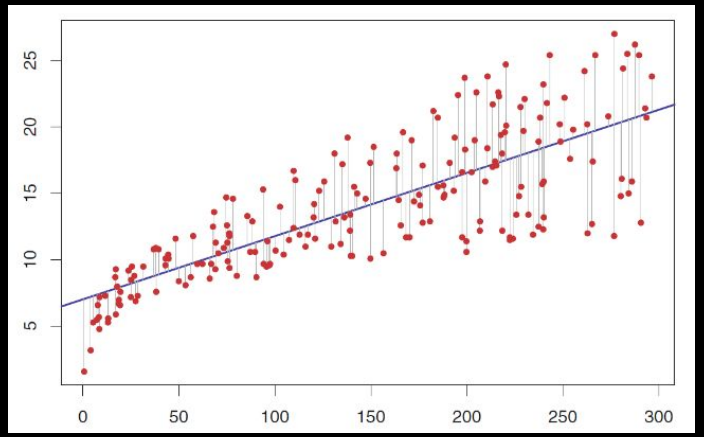
Lineáris regresszió, definíció

Definíció: Legyenek X és Y val. változók. Ekkor az Y-nak az X-re vett *lineáris regresszió*ján azt a $\beta X + \alpha$ val. változót értjük, amire $\alpha, \beta \in \mathbb{R}$, és az

$$\mathbb{E}\Big(\big(Y - (\beta X + \alpha)\big)^2 \Big)$$

mennyiség minimális.

Értelmezés: Megpróbáljuk megtippelni Y-t az X-nek egy lineáris függvényét használva, a lehető legkisebb átlagos négyzetes hibával.



Lineáris regresszió, állítás

Állítás: Legyenek X és Y olyan valószínűségi változók, amire $\mathrm{cov}(X,Y)$, $\mathbb{D}^2(X), \mathbb{D}^2(Y)$ véges, és $\mathbb{D}^2(X) \neq 0$. Ekkor az Y-nak az X-re vett *lineáris regressziós együtthatói:*

$$\beta = \frac{\mathrm{cov}(X,Y)}{\mathbb{D}^2(X)} \qquad \alpha = \mathbb{E}(Y) - \frac{\mathrm{cov}(X,Y)}{\mathbb{D}^2(X)}\mathbb{E}(X)$$

Lineáris regresszió, példa

Példa: az előző, csapadékmennyiséges.

$$\beta = \frac{\mathrm{cov}(X,Y)}{\mathbb{D}^2(X)} = \frac{1/225}{7/90} = \frac{2}{35}$$

$$\alpha = \mathbb{E}(Y) - \frac{\mathrm{cov}(X,Y)}{\mathbb{D}^2(X)}\mathbb{E}(X) = \frac{4}{5} - \frac{2}{35}\frac{7}{15} = \frac{58}{75}$$

Tehát a lineáris regresszió: $\frac{2}{35}X + \frac{58}{75}$

Lineáris regresszió, alternatív felírás

Definíció: Az Y val. változó X -re vett regressziós egyenese az

$$\{(x, y) \in \mathbb{R}^2 \mid y = \beta x + \alpha\}$$

egyenes a síkon, ahol α és β a lineáris regressziós együtthatók.

Kérdés: Hogy a szöszbe jegyzem meg az együtthatók formuláit?

Lehetséges válasz: Ha Y-t próbálja közelíteni a $\beta X + \alpha$, akkor logikus lenne, ha az alábbiak teljesülnének:

$$\mathbb{E}(Y) = \mathbb{E}(\beta X + \alpha) \qquad \operatorname{cov}(X, Y) = \operatorname{cov}(X, \beta X + \alpha)$$

Ebből a két egyenletből átrendezéssel adódik a korábbi két egyenlet.

Lineáris regresszió, korrelációval

Definíció: Legyenek X és Y valószínűségi változók. Ekkor

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\mathbb{D}(X)\mathbb{D}(Y)}$$

ami egy -1 és +1 közti szám (ha létezik).

Kérdés: felírható a lineáris regresszió korrelációval (egyszerűen)?

Válasz: A regressziós együtthatók éppen azok, amikre teljesül, hogy

$$\frac{(\beta X + \alpha) - \mathbb{E}(Y)}{\mathbb{D}(Y)} = \frac{X - \mathbb{E}(X)}{\mathbb{D}(X)} \cdot \operatorname{corr}(X,Y)$$

Lineáris regresszió, bizonyítás

Biz.: Amit minimalizálnunk kell:

$$h(\alpha,\beta) = \mathbb{E}\Big(\big(Y-(\beta X+\alpha)\big)^2\Big)$$
$$= \mathbb{E}\Big(Y^2+\beta^2X^2+\alpha^2-2\beta XY-2\alpha Y+2\alpha\beta X\Big)$$
$$= \mathbb{E}(Y^2)+\beta^2\mathbb{E}(X^2)+\alpha^2-2\beta\mathbb{E}(XY)-2\alpha\mathbb{E}(Y)+2\alpha\beta\mathbb{E}(X)$$

Deriválással kereshető a minimumhelye:

$$\alpha \text{ szerint: } 2\alpha-2\mathbb{E}(Y)+2\beta\mathbb{E}(X)$$

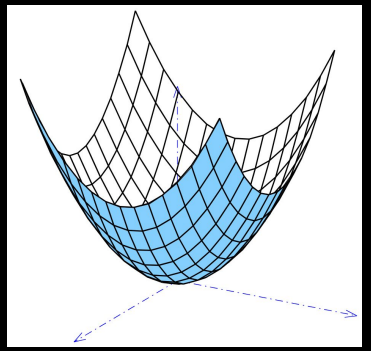
$$\beta \text{ szerint: } 2\beta\mathbb{E}(X^2)-2\mathbb{E}(XY)+2\alpha\mathbb{E}(X)$$

Lineáris regresszió, bizonyítás

Biz. (folyt.): A parciális deriváltak pontosan akkor nullák, ha

$$\alpha + \beta\mathbb{E}(X) = \mathbb{E}(Y)$$

$$\alpha\mathbb{E}(X) + \beta\mathbb{E}(X^2) = \mathbb{E}(XY)$$



Ennek a megoldása:

$$\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X) = \mathbb{E}(Y) - \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2(X)}\mathbb{E}(X)$$

$$\beta = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - \mathbb{E}(X)^2} = \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2(X)}$$

Lineáris regresszió hibája

Állítás: Legyen az Y val. változó X-re vett lineáris regressziója $\beta X + \alpha$. Ekkor

$$\mathbb{D}^2\Big(Y-(\beta X+\alpha)\Big) = \mathbb{D}^2(Y)-\frac{\operatorname{cov}(X,Y)^2}{\mathbb{D}^2(X)}$$

Megj.: Korrelációval felírva

$$\mathbb{D}^2\Big(Y-(\beta X+\alpha)\Big)= \mathbb{D}^2(Y)\cdot\big(1-\operatorname{corr}(X,Y)^2\big)$$

Lineáris regresszió hibája, példa

Példa: az előző.

$$\mathbb{D}^2\Big(Y-(\beta X+\alpha)\Big) = \mathbb{D}^2(Y)-\frac{\operatorname{cov}(X,Y)^2}{\mathbb{D}^2(X)}$$

$$= \frac{58}{225} - \frac{(1/225)^2}{7/90} \approx 0{,}2575$$

Lineáris regresszió hibája, biz.

Biz.: $$\mathbb{D}^2\Big(Y-(\beta X+\alpha)\Big)=$$

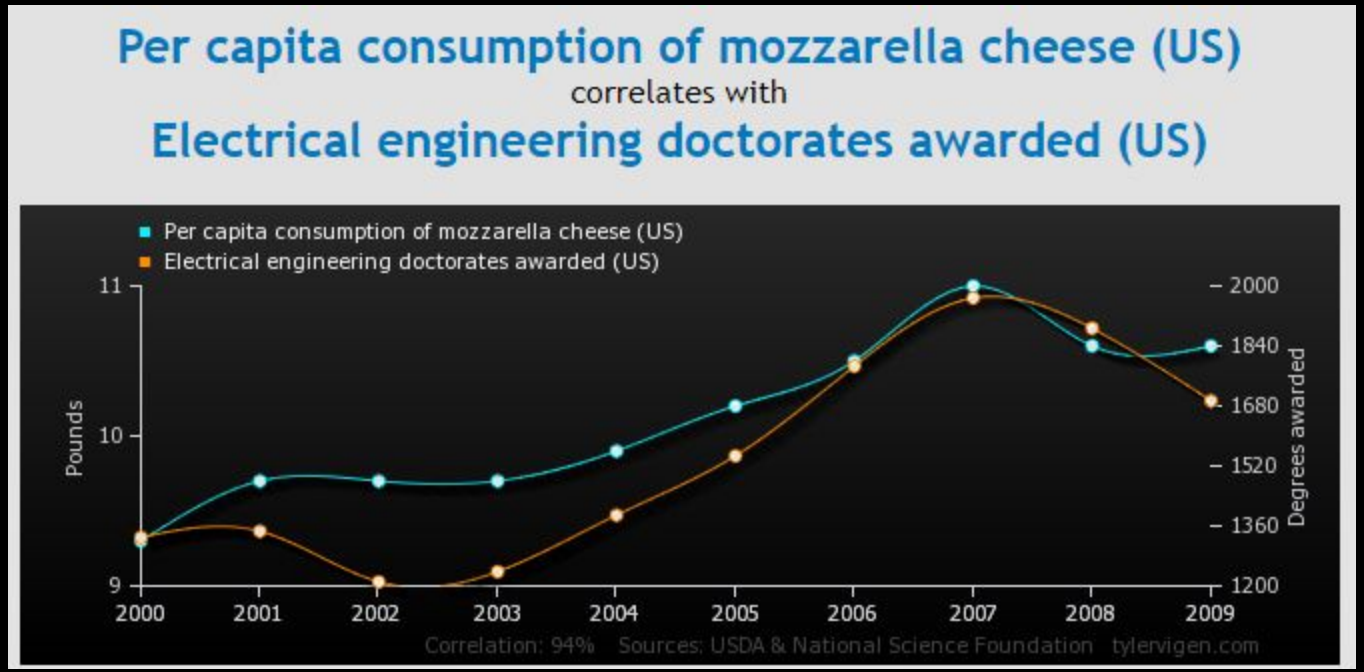
$$=\mathbb{D}^2(Y-\beta X)$$

$$=\mathbb{D}^2(Y)+\beta^2\mathbb{D}^2(X)-2\mathrm{cov}(Y,\beta X)$$

$$=\mathbb{D}^2(Y)+\frac{\mathrm{cov}(X,Y)^2}{\big(\mathbb{D}^2(X)\big)^2}\mathbb{D}^2(X)-2\frac{\mathrm{cov}(X,Y)}{\mathbb{D}^2(X)}\mathrm{cov}(Y,X)$$

$$=\mathbb{D}^2(Y)-\frac{\mathrm{cov}(X,Y)^2}{\mathbb{D}^2(X)}$$

Hamis korrelációk

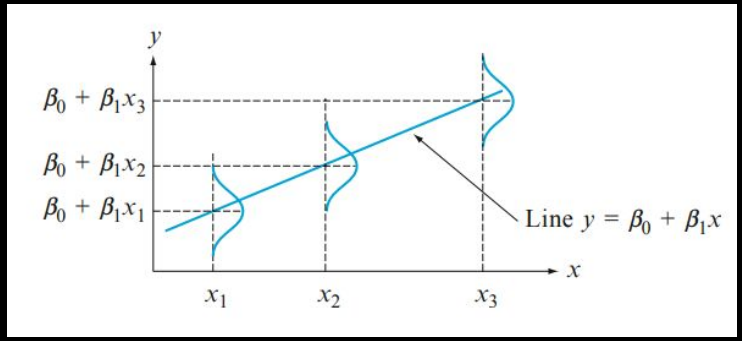


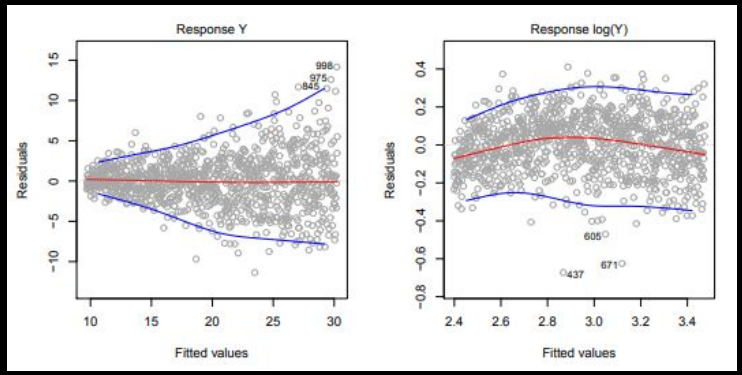
Forrás: www.tylervigen.com/spurious-correlations

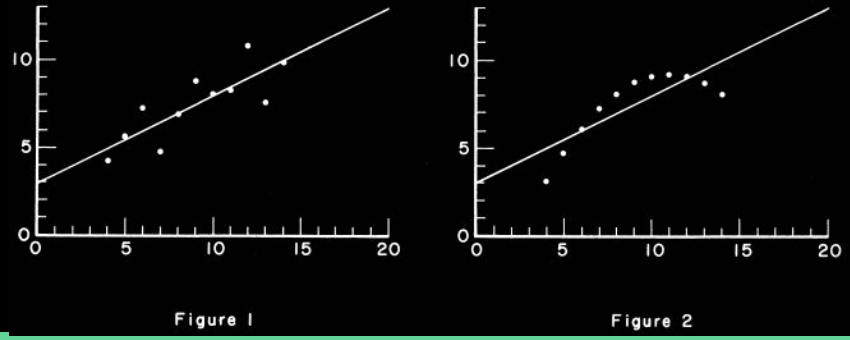
Továbbá, “a korreláció nem implikált kauzalitást.”

További olvasnivaló

- Devore, Berk - Modern mathematical statistics with applications, Ch. 12.1.
- James, Witten, Hastie, Tibshirani - An introduction to statistical learning, Ch. 3.
- Simpson’s paradox
- Anscombe’s quartet







Köszönöm a figyelmet!