Korreláció és regressziószámítás

Kapcsolatot, összefüggést keresünk két vagy több változó között.

- Van-e összefüggés a 7-10 éves gyermekek testtömeg-indexe (TTI) és a hetente a tv előtt töltött órák száma (TV) között?

- Jól leírja-e a TTI TV-től való függését egy lineáris függvény? Ha igen, mik ennek a függvénynek a paraméterei? Ha nem, milyen más függvényt válasszunk?

- Változik-e a kép, ha a TV mellett még más magyarázó változókat is figyelembe veszünk? (pl. a szülők testtömeg-indexét, a számítógép előtt töltött órák számát stb.)

Korreláció és regressziószámítás

Kapcsolatot, összefüggést keresünk két vagy több változó között.

- Van-e összefüggés a 7-10 éves gyermekek testtömeg-indexe (TTI) és a hetente a tv előtt töltött órák száma (TV) között?

Korrelációszámítás

- Jól leírja-e a TTI TV-től való függését egy lineáris függvény? Ha igen, mik ennek a függvénynek a paraméterei? Ha nem, milyen más függvényt válasszunk?

Regressziószámítás

- Változik-e a kép, ha a TV mellett még más magyarázó változókat is figyelembe veszünk? (pl. a szülők testtömeg-indexét, a számítógép előtt töltött órák számát stb.)

Korreláció és regressziószámítás

Kapcsolatot, összefüggést keresünk két vagy több változó között.

- Van-e összefüggés a 7-10 éves gyermekek testtömeg-indexe (TTI) és a hetente a tv előtt töltött órák száma (TV) között?

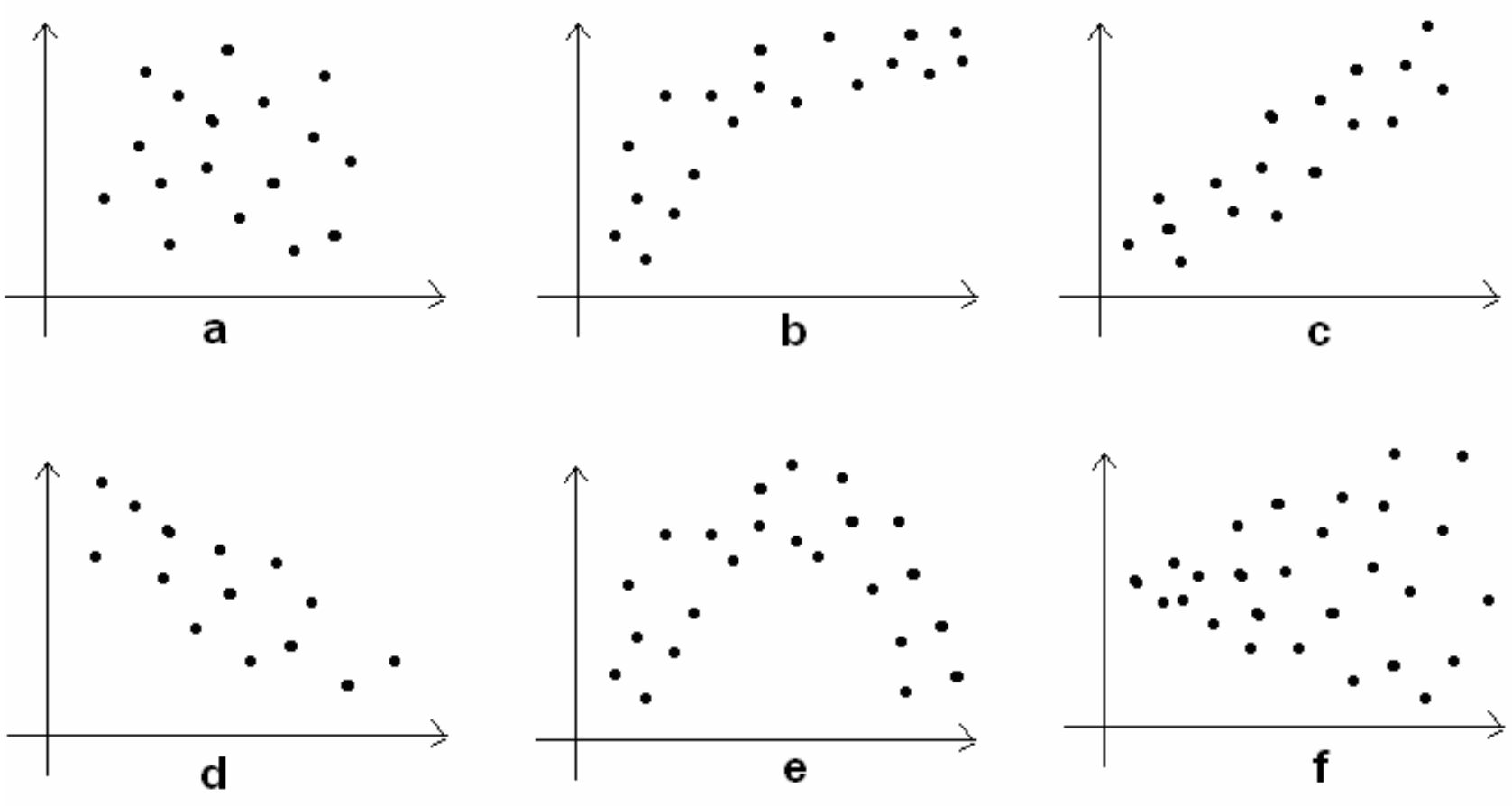
Korrelációszámítás

- Jól leírja-e a TTI TV-től való függését egy lineáris függvény? Ha igen, mik ennek a függvénynek a paraméterei? Ha nem, milyen más függvényt válasszunk?

Hipotézisvizsgálat Regressziószámítás Becslés

- Változik-e a kép, ha a TV mellett még más magyarázó változókat is figyelembe veszünk? (pl. a szülők testtömeg-indexét, a számítógép előtt töltött órák számát stb.)

Figyelem! Nem minden kapcsolat korrelációs jellegű!



a – nincs kapcsolat, b – pozitív nemlineáris korreláció,
c – pozitív lineáris korreláció, d – negatív lineáris korreláció,
e, f – nem korrelációs (nem monoton) kapcsolatok

Általános (bármilyen jellegű) összefüggés: **asszociáció**

Monoton összefüggés: **korreláció**

Pozitív korreláció: nagyobb *x*-hez nagyobb *y* (általában!)

Negatív korreláció: nagyobb *x*-hez kisebb *y* (általában!)

Számszerűsítése a (–1, 1) skálán történik:

–1	0	1
max. negatív korr. (legerősebb)	nincs korr. (leggyengébb)	max. pozitív korr. (legerősebb)

Bár két független változó között mindig 0 a korreláció, a 0 korrelációból nem következik a függetlenség!

(lásd e, f: van kapcsolat, csak nem korrelációs jellegű)

A három leggyakrabban használt korrelációs együttható

Pearson-féle:

- Főleg a lineáris kapcsolatra érzékeny
- A kiugró értékek erősen befolyásolják
- A klasszikus szignifikancia-vizsgálat csak normális változókra megy (de bootstrap-pel eloszlásfüggetlen is végezhető)

Spearman-féle (rang-korrelációs együttható):

- A Pearson-féle a rangszámokra alkalmazva
- Nemlineáris kapcsolatra is érzékeny
- A kiugró értékek kevésbé befolyásolják

Kendall-féle τ *(tau)*:

- Nemlineáris kapcsolatra is érzékeny
- A kiugró értékek még kevésbé befolyásolják

Regressziószámítás

Matematikai **függvénnyel leírható kapcsolat**ot keresünk **egy vagy több magyarázó változó** (x_1, x_2 stb.) és **egy függő változó** (y) között.

Feltételezzük, hogy az y-t az x-ek nem határozzák meg egyértelműen, **az *y* mért értéke a véletlentől is függ.**

Modell: *f a függvény* *ε a „véletlen komponens”*

$$y = f(x_1, x_2, \ldots) + \varepsilon$$

vagy ha a függvény lineáris:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \varepsilon$$

Regressziószámítás

Oksági kapcsolatot keresünk?

Matematikai **függvénnyel leírható kapcsolat**ot keresünk **egy vagy több magyarázó változó** (x_1, x_2 stb.) és **egy függő változó** (y) között.

Feltételezzük, hogy az y-t az x-ek nem határozzák meg egyértelműen, **az *y* mért értéke a véletlentől is függ.**

Modell: *f a függvény* *ε a „véletlen komponens”*

$$y = f(x_1, x_2, \ldots) + \varepsilon$$

vagy ha a függvény lineáris:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \varepsilon$$

Regressziószámítás

Matematikai **függvénnyel leírható kapcsolat**ot keresünk **egy vagy több magyarázó változó** (x_1, x_2 stb.) és **egy függő változó** (y) között.

Feltételezzük, hogy az y-t az x-ek nem határozzák meg egyértelműen, **az *y* mért értéke a véletlentől is függ.**

Modell: *f a függvény* *ε a „véletlen komponens”*

$$y = f(x_1, x_2, \ldots) + \varepsilon$$

vagy ha a függvény lineáris:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \varepsilon$$

Az x-ek ebben a modellben nem véletlen változók!

Regressziós modellezés:

1. **Mit gondolunk**, mi függ mitől, mi a függő változó, hogyan mérjük, mely magyarázó változókat használjuk stb.
2. **Mi a cél**? Csak az összefüggés igazolása, vagy előrejelzés az *x*-(ek)ből az *y*-ra? Esetleg az *y*-ból az *x*-re?
3. **Mit tudunk** már pl. az irodalomból? Lineáris a kapcsolat? Monoton a kapcsolat? Ha monoton, pozitív vagy negatív?
4. **Mit látunk az adatokból** készült grafikonokon? Vannak-e kiugró értékek? Ha igen, értjük-e, hogyan keletkeztek? A grafikonok szerint milyen függvénytípus illeszkedne jól?
5. Válasszunk függvénytípust és **végezzük el az elemzést!** Rendben van-e a modell (**diagnosztika**, reziduumok)? Szignifikáns-e (***p*-értékek**)? Megfelel-e az illeszkedés (R^2)?
6. Ha valami nincs rendben, vissza az 5.-re!

Különbségek a korreláció- és regressziószámítás között

- Míg a korrelációszámítás **szimmetrikus kapcsolat**ot tételez fel az *x* és *y* között, addig a regressziószámítás **egy bizonyos irányú (*x* → *y*) kapcsolat**ot
- Míg a korrelációszámításban **mindkét változó valószínűségi változó**, a regressziószámításban *x* nem feltétlenül az (***x* nem feltétlenül függ a véletlentől**).

A korrelációszámításnak nincs értelme akkor, ha az *x* értékeit (pl. dózist) a kísérletező állítja be!

Két mérési módszer közötti egyezés vizsgálatára sem a korreláció-, sem a regressziószámítás nem megfelelő!

Miért nem mindegy, hogy melyik változót választjuk *x*-nek és melyiket *y*-nak?

Tegyük fel, hogy a valódi összefüggés az *x* és *y* között

$$y = 0.5 \cdot x + 3,$$

amit átírhatunk így is:

$$x = 2 \cdot y - 6.$$

Tegyük fel továbbá, hogy az *y* függ az *x*-től, de emellett egy normális eloszlású véletlen faktort (ε) is tartalmaz, azaz:

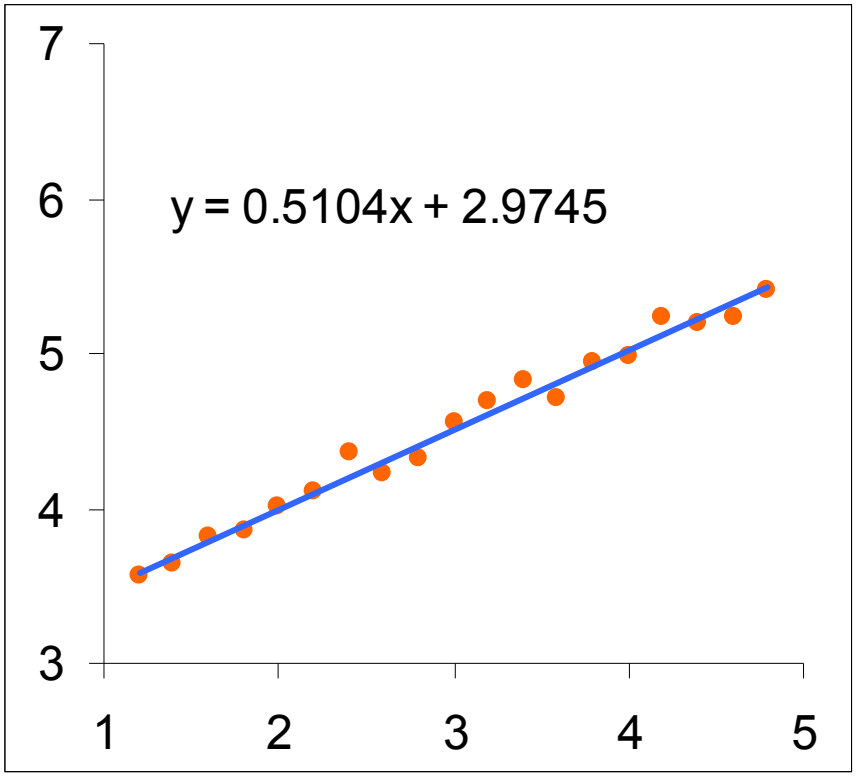
$$y = 0.5 \cdot x + 3 + \varepsilon$$

(tehát az *x* nem véletlen, de az *y* az ε miatt az)

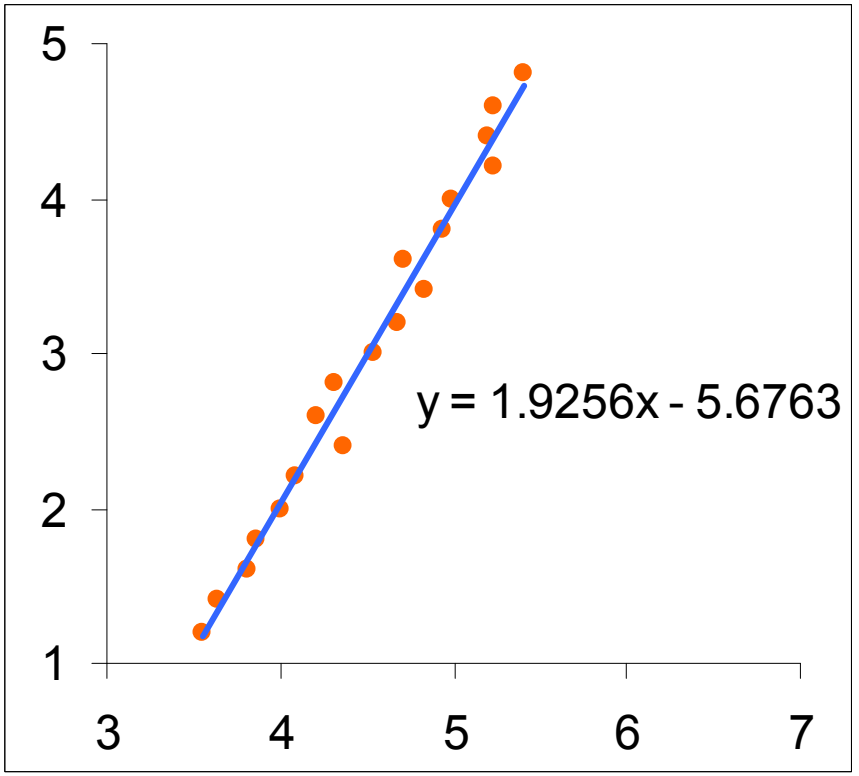
Ha az ε szórása kicsi, mindkét irányú regresszió ugyanazt adja.
Ha az ε szórása nagyobb, a „jó irányú” regresszió továbbra is helyes eredményt ad, míg a másik egyre rosszabbat.

A valódi összefüggés *x* és *y* között:

$$y = 0.5 \cdot x + 3$$



$$x = 2 \cdot y - 6$$

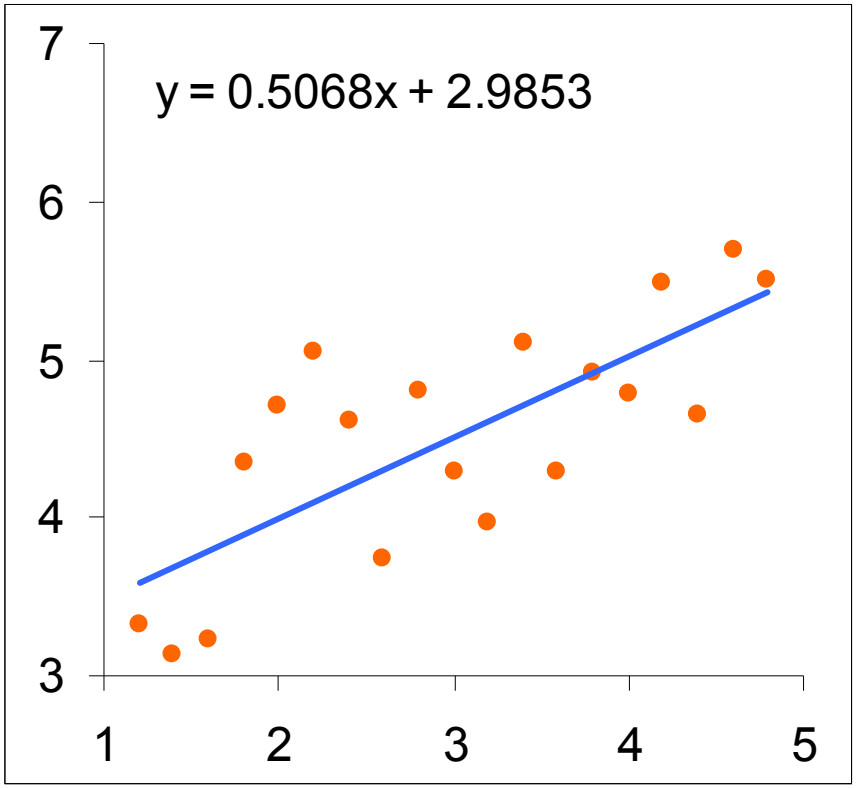


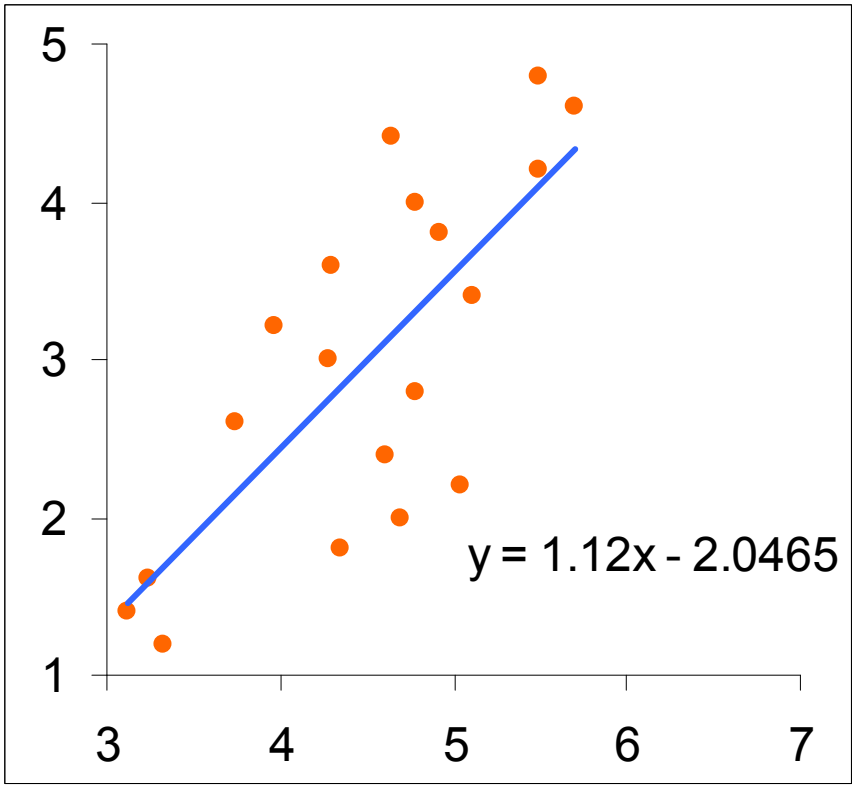
A becsült egyenesek, ha az ε szórása kicsi.

A valódi összefüggés *x* és *y* között:

$$y = 0.5 \cdot x + 3$$

$$x = 2 \cdot y - 6$$

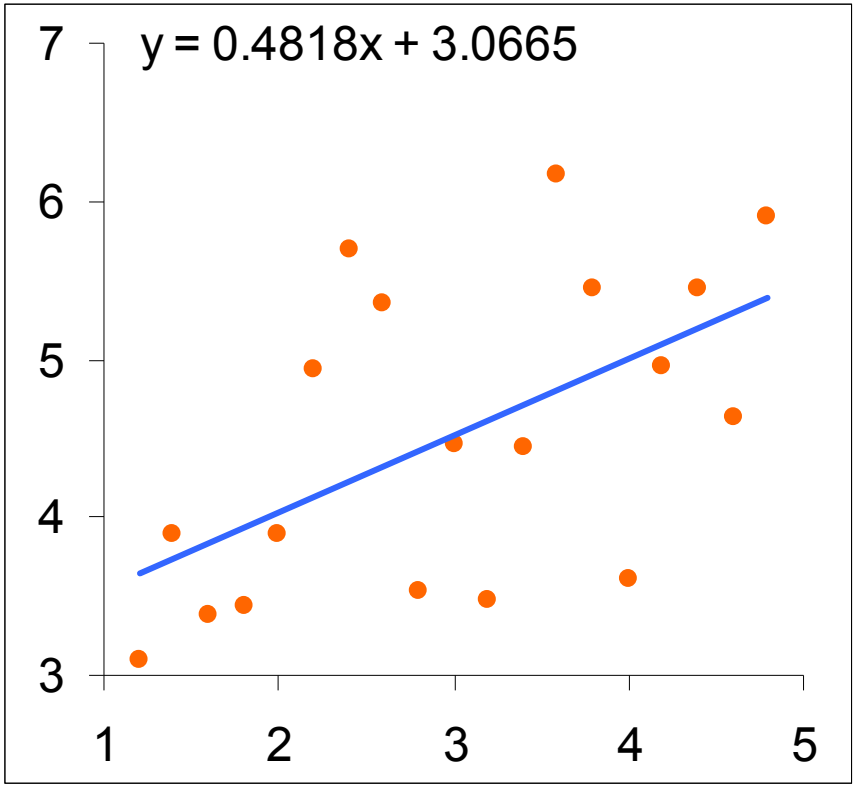


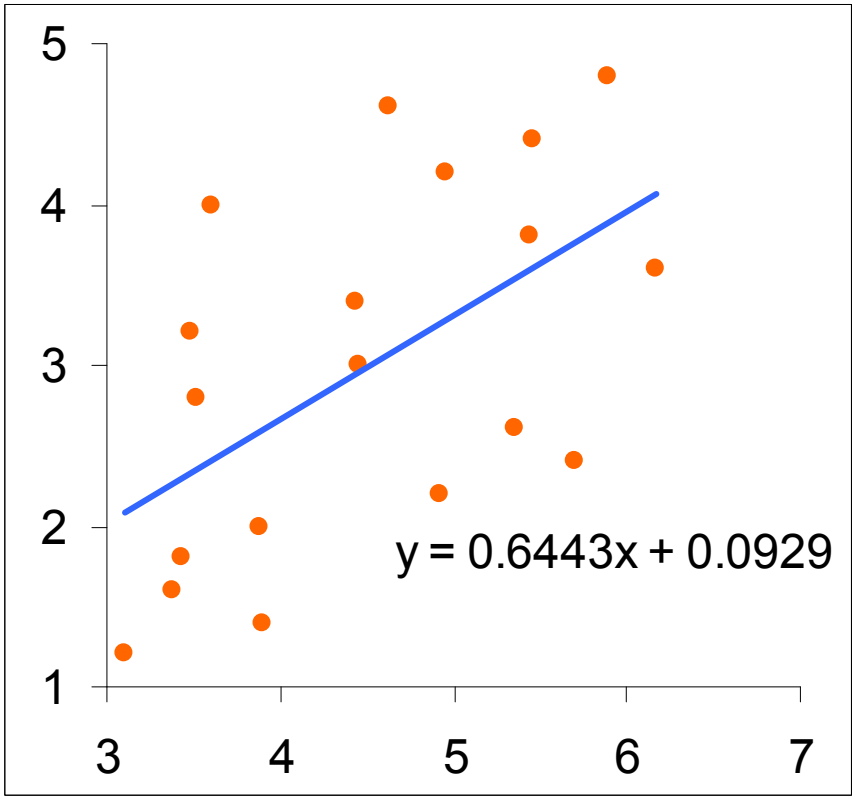


A becsült egyenesek közepes szórású ε esetén.

A valódi összefüggés *x* és *y* között:

$y = 0.5 \cdot x + 3$ $\qquad$ $x = 2 \cdot y - 6$



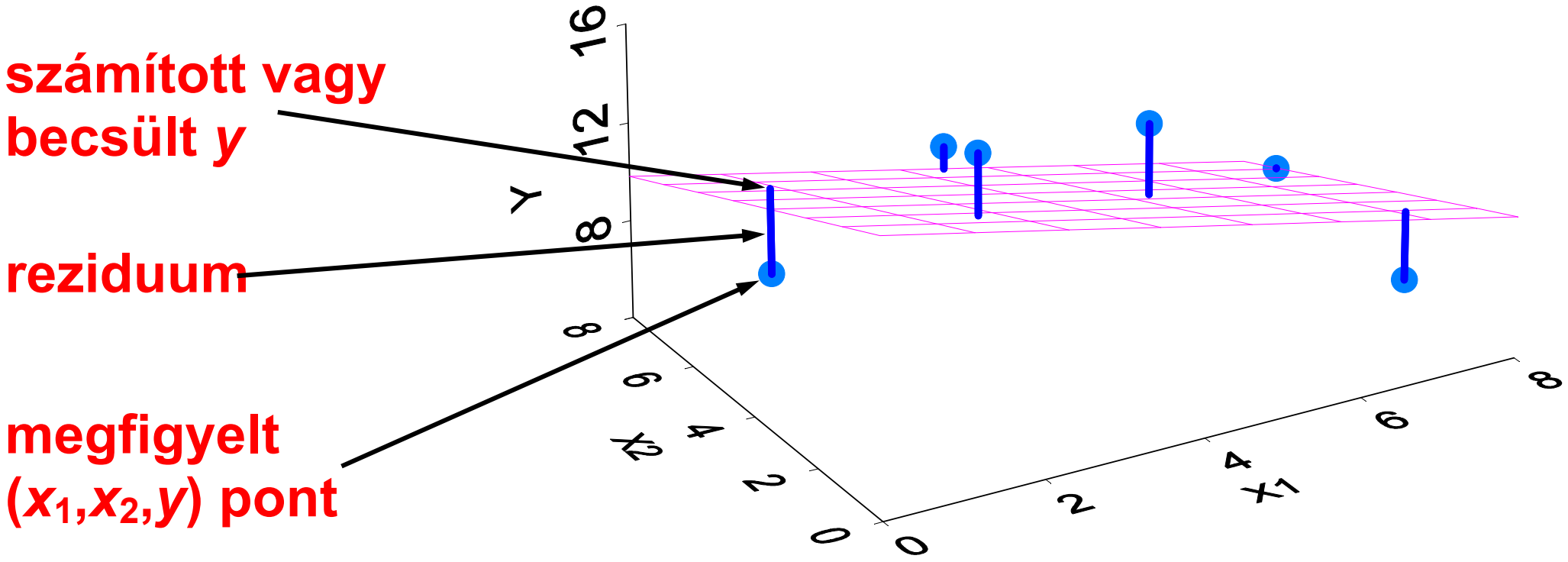


A becsült egyenesek nagy szórású ε esetén.

Megfigyelt és számított (becsült) értékek, reziduumok

Jelölje a két magyarázó változót x_1 és x_2, a függő változót y, és végezzünk **lineáris regressziót**!

A **regressziós felület egy sík** lesz, a megfigyelt pontoknak a síkra való y irányú vetületei a becsült értékek, a síktól való eltéréseik pedig a reziduumok.



A paraméterek becslése: a legkisebb négyzetek elve

Azt a felületet (lineáris esetben síkot, egy magyarázó változó esetén egyenest) **választjuk becslésnek**, azaz mint az adatainkhoz legjobban illeszkedőt, **amelyiktől a megfigyelt pontok eltérés-négyzetösszege** (=a reziduumok négyzet-összege) **a legkisebb**.

A statisztikai programok e függvény (felület, sík, egyenes stb) paramétereit adják meg becslésként. Pl.: lineáris regresszió két magyarázó változóval:

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

Ekkor a paraméterek: b_0, b_1, b_2 .

A programok a becslésekhez általában megadnak SE-t, sőt konfidencia-intervallumot is (95%-ost a $b_i \pm 1.96 \cdot SE$ képlettel).

Hipotézisvizsgálatok

Valóban függ-e…?

H_0: nem függ

H_1: függ (az adatok bizonyító erejűek)

„Szignifikáns a regresszió” = elvetjük a H_0-t!

1. **Valóban függ-e az *y* az x_i–től (egyenként)?**
   - *t*-próbák, minden egyes magyarázó változóra
   - Amelyiktől nem függ szignifikánsan, kihagyhatjuk a modellből
2. **Valóban függ-e az *y* az x_i-ktől (együtt az összestől)?**
   - F-próba

A tesztek szükséges feltétele az *y* véletlen komponensének normalitása, és hogy az ε szórása ne függjön az *x*-ektől!

Mennyire határozzák meg az *x*-ek az *y*-t: az R^2

A determinációs együttható, az R^2 azt mondja meg, hogy mennyire informatívak az *x*-ek az *y*-ra nézve.

Értéke 0 és 1 közötti:

$R^2=0$ $R^2=1$

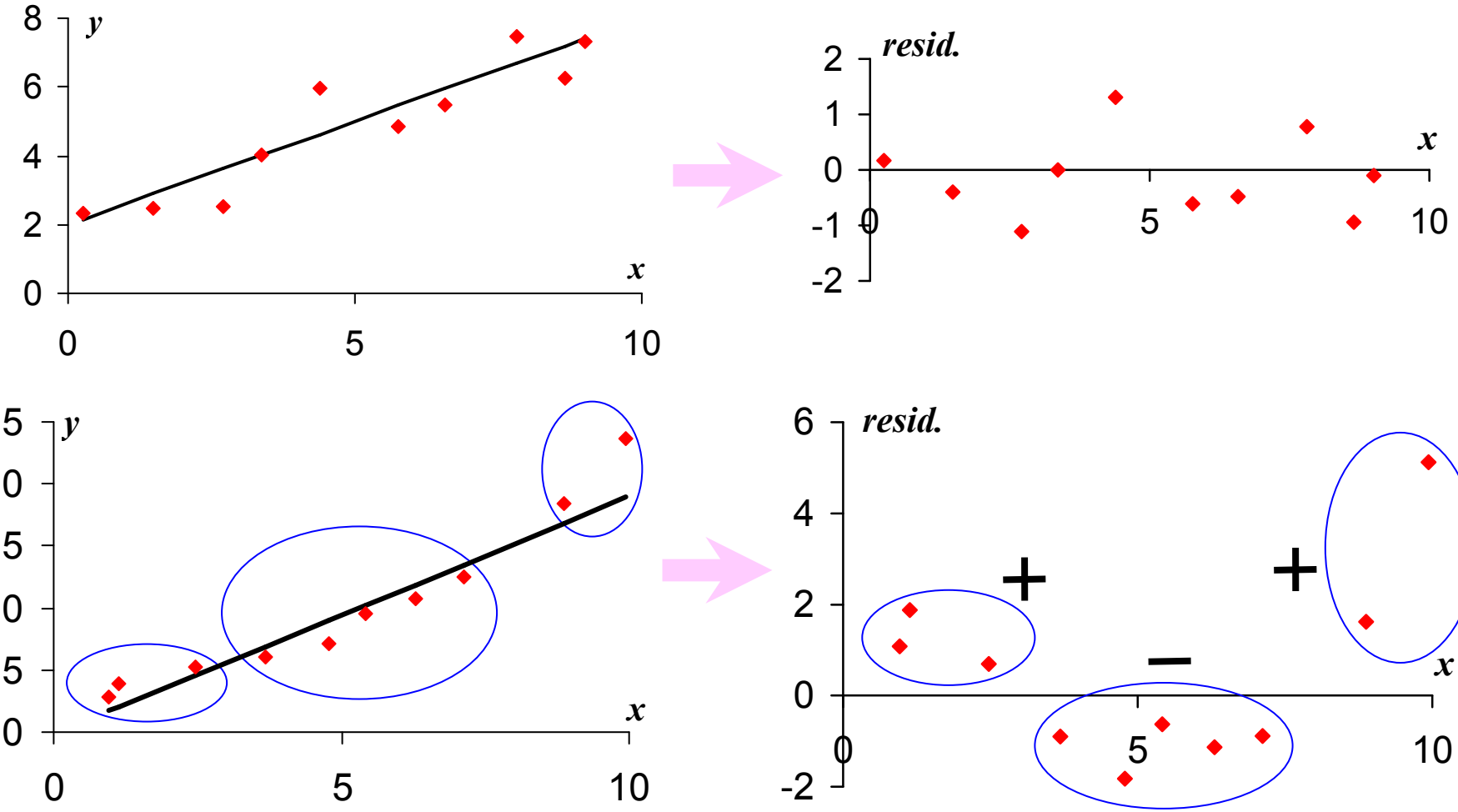
Nem érdemes ismernünk az *x*-eket, úgysem mondanak semmit az *y*-ra nézve (legalábbis a lineáris regresszió modelljében maradva)

Az *x*-ek teljes mértékben meghatározzák az *y*-t, valójában nincs véletlen komponense (tehát az *y* egyszerűen az *x*-ek egy lineáris függvénye)

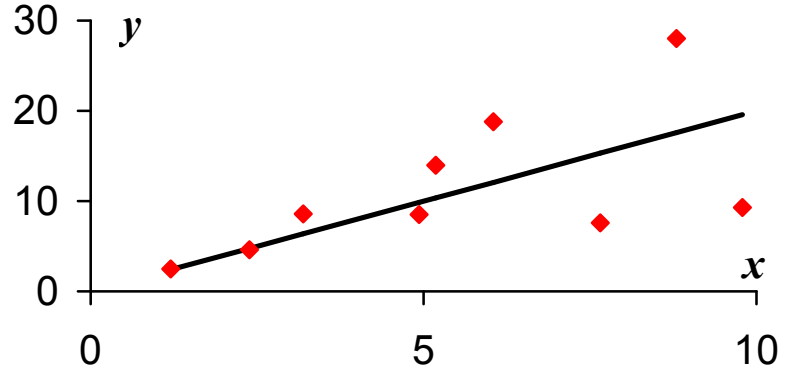
Ne akarjunk minél nagyobb R^2-et! (Inkább értelmes modellt!)

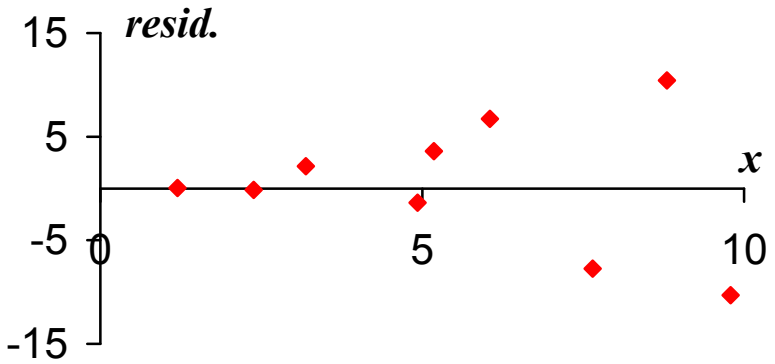
Rendben van-e a modell? „regressziós diagnosztika”

- Ha „rendben van” a modell, akkor a reziduumok véletlenszerű elrendeződést mutatnak.



- A reziduumok normalitását QQ-plottal vizsgálhatjuk (kis mintára a „szemmel” való megítélés jobb, mint a statisztikai teszt).
- Azt, hogy az ε szórása függ-e az x-ektől, a reziduumok szórásából ítélhetjük meg.





- Figyeljünk arra, hogy a magyarázó változók egymással ne legyenek nagyon korreláltak (multikollinearitás)!
- Mindig nézzük meg, hogy az összefüggés nem 1-2 pont műve-e!
- Több magyarázó változó esetén a diagnosztika bonyolultabb (és mindegyik statisztikai programmal másképp kell végezni).

Kitekintés: a lineáris modell

Ha a magyarázó változók nem folytonosak, hanem diszkrétek, kategoriálisak (csoportképző változók), akkor a regresszió-számítás helyett **varianciaelemzést (ANOVA)** végzünk.

Ha a magyarázó változók között csoportképző és folytonos változók is vannak, akkor **variancia-kovariancia-elemzést** végzünk: az *y*-nak a kategoriális változóktól való függését a varianciaelemzés szerint, a folytonosaktól (=kovariánsok) való függését pedig lineáris regresszióval elemezzük.

Nem külön-külön, ez a módszer ezt egyszerre tudja csinálni!

Az általános lineáris modell (general linear model) egységes elméletet kínál e három módszerre.

(A megoldás végül is regressziószámítással történik.)