Regressziós vizsgálatok

Regresszió (regression)

- **Általános jelentése:** visszaesés, hanyatlás, visszafelé mozgás, visszavezetés.
- **Orvosi területen:** visszafejlődés, involúció. A betegség tünetei, vagy maga a betegség javulása.

  Regrediál.

Regressziószámítás

Francis Galton angol természettudós, aki biológiai vizsgálatai során fogalmazta meg az átlaghoz való visszatérés (**regression to mean**) elvét, melyet apák és fiaik testmagasságának kapcsolatára alkalmazott.

Lényege, hogy magasabb apák fiai tendenciaszerűen („átlagosan”) magasabbaknak bizonyultak, azaz a fiúk „visszatértek” az apáikhoz.

Ekkor számolt először függvényt a két megfigyelés-sorozat között, és ezt nevezte el **regressziós függvénynek**.

Regressziószámítás

Regresszió:

a változók közötti kapcsolat elemzésének elterjedt eszköze.

Vizsgálja: egy kitüntetett, a vizsgálat tárgyát képező változó, amelyet *eredményváltozónak* (vagy függő változónak, response) nevezünk, hogyan függ egy vagy több ún. *magyarázó* (vagy független, prediktor) *változótól*.

Regressziós felület

- Egy változó esetén: (elsőfajú) regressziós egyenes.
- Két változó esetén: (elsőfajú) regressziós sík.
- Több változó esetén: regressziós felület.

Nemlineáris regresszió

- Polinomos regresszió:
- Ebben az esetben a regressziós függvényt

$$\hat{y} = a_0 + a_1x + a_2x^2 + \quad + a_nx^n$$

alakban keressük.

- Hatványkitevős regresszió:

$$\hat{y} = ax^b$$

- Ez azzal ekvivalens: $\ln\hat{y} = \ln a + b\ln x$

Nemlineáris regresszió

- Exponenciális regresszió:

$$\hat{y} = ab^x$$

Ezzel ekvivalens: $\ln \hat{y} = \ln a + (\ln b)x$

- Logaritmikus regresszió:

$$\hat{y} = a + b\ln x$$

Logisztikus regresszió

- **A függő változó kategórikus** :

  - bináris (a megfigyelt eseménynek csak két állapota van),

  - polychotomus (a megfigyelt esemény több állapotú).

- **Többváltozós módszer, amelyben**
  - **Több tényező (jellemző, tünet) alapján valamely betegség előfordulásának valószínűségét becsüljük.**
  - **A független változók eloszlására nincs feltétel.**

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}}$$

Pearson féle lineáris korreláció

(bivariate correlation)

Hipotézisek

- H_0: nincs korrelációs kapcsolat

  $r = 0$

- H_1: van korrelációs kapcsolat

  $r \neq 0$

Feltételek

- Y: normális eloszlású legyen.
- X: normális eloszlású legyen.

$$\operatorname{cov}(x, y) = \frac{\sum d_x d_y}{n}$$

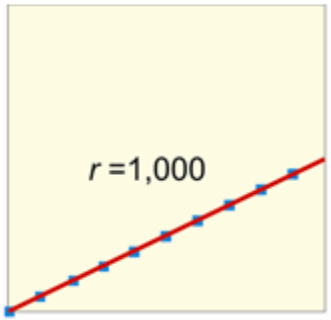
kovariancia

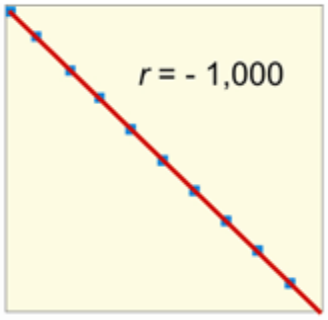
$$r = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \sum d_y^2}}$$

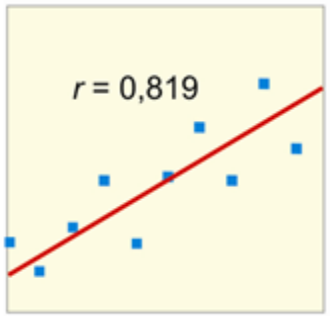
Az *r* korrelációs együttható olyan -1 és +1 között elhelyezkedő mutatószám, amelyik 1-hez közeli abszolút értékei szoros, közel lineáris függvényszerű kapcsolatot, 0 körüli értékei a lineáris kapcsolat hiányát (az ún. korrelálatlanságot) jelentik.

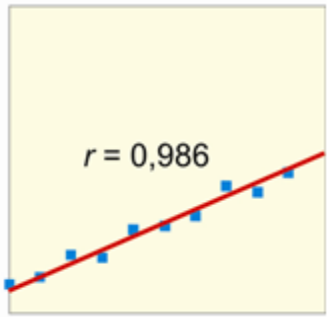
Az *r* a két változó kapcsolat szorosságának mérőszáma.

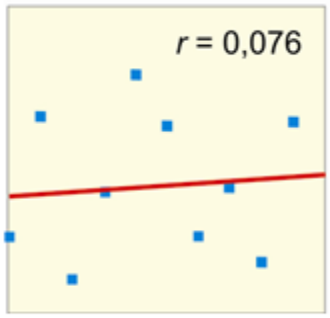

Correlation r = 0
Correlation r = −0.3
Correlation r = 0.5
Correlation r = −0.7
Correlation r = 0.9
Correlation r = −0.99

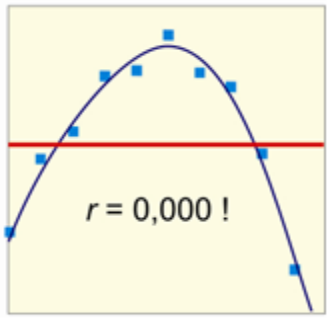

r =1,000


r = - 1,000


r = 0,819


r = 0,986


r = 0,076


r = 0,000 !

Szignifikancia számítása

A korrelációs együttható szignifikanciájának vizsgálatához a H_0: r = 0 hipotézist fogalmazzuk meg. Döntésünk alapja egy N elemű mintában kiszámított korrelációs együttható (r). A H_0 elutasíthatósága függ az r együttható nagyságától és a df szabadságfok nagyságától (df = N-2).
A szignifikancia kiszámításához t eloszlású statisztikát használunk. Ennek képlete:

$$t = r * \sqrt{\frac{n-2}{1-r^2}}$$

Ha $|t| > t_{table}$, elvetjük H_0-t és azt mondjuk, hogy a populáció korrelációs együtthatója különbözik 0-tól. Tehát, ha a kapott eredményünk abszolút értéke nagyobb, mint a táblázatban az adott szabadságfokhoz és szignifikanciaszinthez (ez általában 0,95) tartozó szám, akkor 95%-os bizonyossággal elutasíthatjuk a nullhipotézist.

Rangkorreláció

A *rangkorrelációs együtthatók* azt mérik, hogy két sorozat együtt változik-e. Ha az egyik sorozat nő, a másik csökken, akkor a rangkorrelációk negatívak lesznek. Rangkorrelációt minimum ordinális változók között számíthatunk.
Egyik fajtája a Spearman-féle rangkorreláció, ami egy Pearson-féle korreláció a rangszámok között.
Spearman-rangkorreláció:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

Egy másik lehetőség a Kendall-féle rangkorreláció, ami a pozitív és a negatív kapcsolatok arányának a különbségét számolja ki

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

ahol n_c a megfelelő, és n_d az eltérő párok száma.

A korrelációhoz hasonlóan értékeik a [-1,1] intervallumba esnek. Értékük 1, ha a két rangsor ugyanaz; 0, ha a két rangsor egymástól független, és -1, ha egymás megfordításai.
A rangkorrelációkat sokszor a korrelációs együttható könnyen számítható és kevésbé eloszlásérzékeny alternatíváiként kezelik. Ennek azonban nincs sok matematikai alapja: a rangkorrelációkkal más összefüggéseket lehet kimutatni, mint a korrelációs együtthatóval.

A korreláció (a két változó közötti kapcsolat) erősségének megítélése.

- **Leegyszerűsített megoldás a kapcsolat erősségére:**

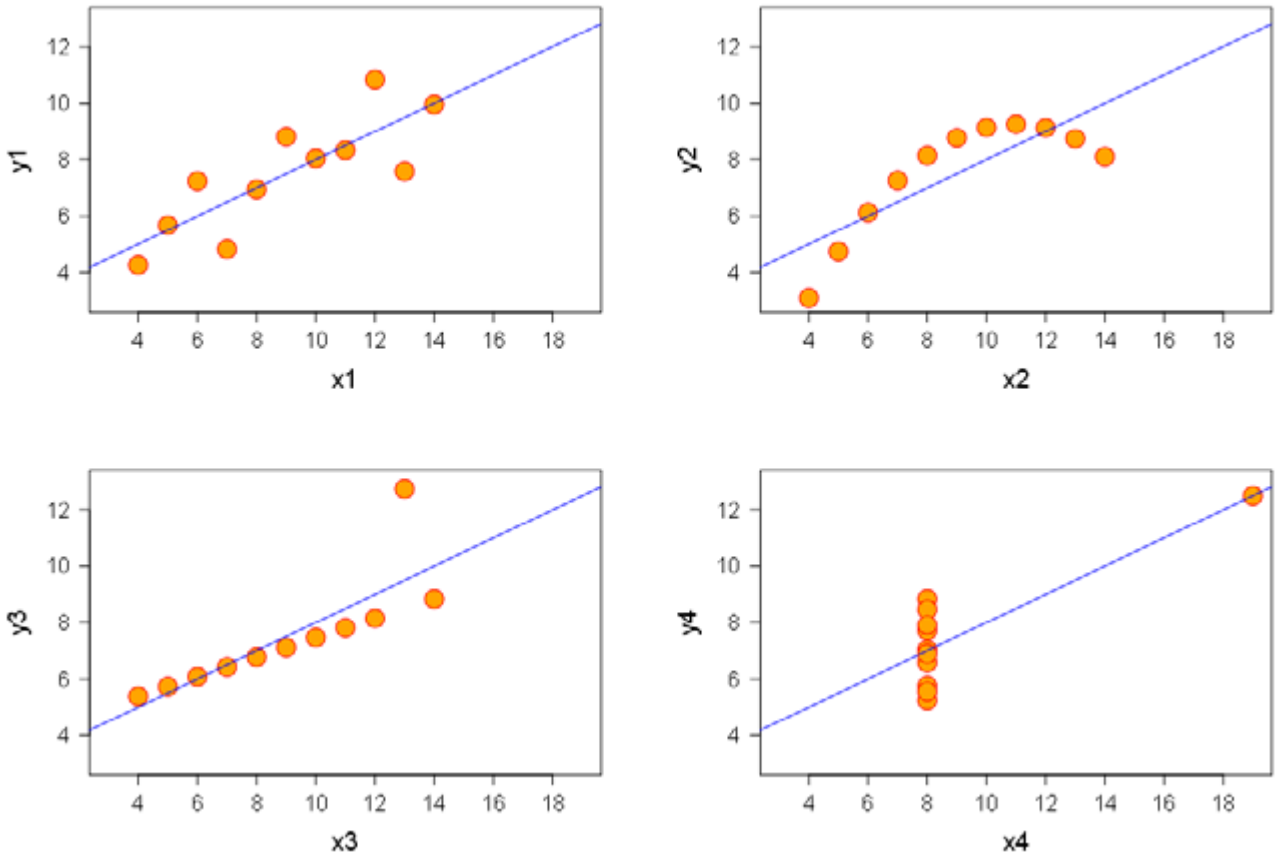
  0-0,25: Nincs vagy igen gyenge

  0,25-0,50: Gyenge

  0,50-0,70: Mérsékelten erős vagy erős

  0,70-1,00: Igen erős

Extrém példa: *r = 0.816*



Parciális korreláció

- Egymás hatásától „megtisztított” érték.
- X, Y, Z változók esetén:

pl. X, Y korrelációja: levesszük Z hatását a kapcsolatból.

Point biserial correlation coefficient (r_{pb})

- Ha az egyik változó (általában Y) természetes **dichotom változó.**

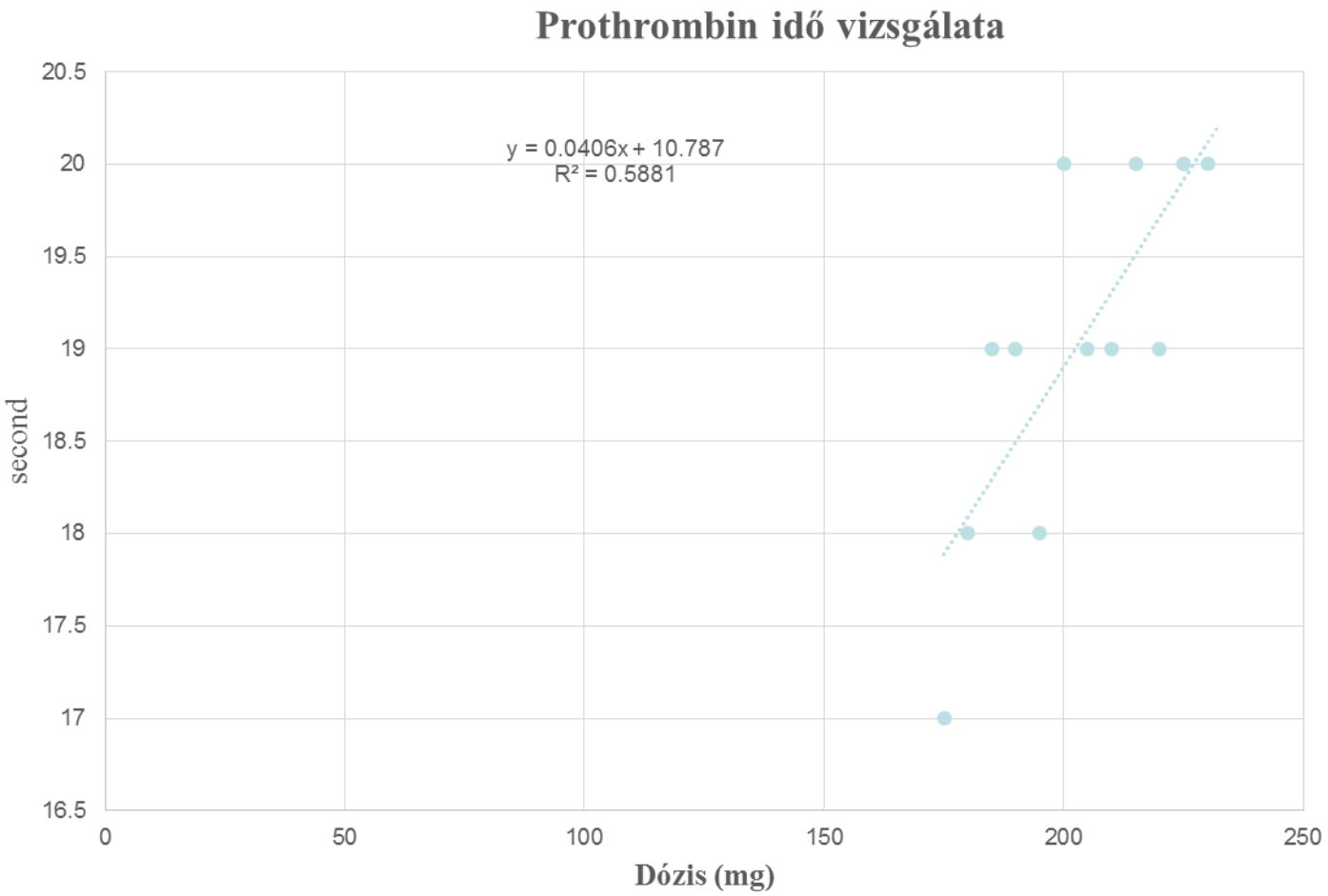
$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} = \frac{M_1 - M_0}{s_{n-1}}\sqrt{\frac{n_1 n_0}{n(n-1)}}.$$

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Lineáris regresszió

Véralvadási időt vizsgáljuk különböző dózisok esetén:

Beteg	Dózis / mg (X)	Prothrombin idő / seconds (Y)
1	200	20
2	180	18
3	225	20
4	205	19
5	190	19
6	195	18
7	220	19
8	175	17
9	215	20
10	185	19
11	210	19
12	230	20


Prothrombin idő vizsgálata
y = 0.0406x + 10.787
R² = 0.5881
second
Dózis (mg)
20.5
20
19.5
19
18.5
18
17.5
17
16.5
0
50
100
150
200
250

Lineáris regressziós modell

$$\hat{y} = a \pm b \cdot x + \varepsilon$$

- ahol:
- $E(\varepsilon) = 0$
- $Var(\varepsilon) = \sigma^2$
- $\varepsilon \sim N(0,\sigma)$

Annyit feltételezünk, hogy:

- Van két változónk (legalább intervallumskála),
- Közöttük kvázi lineáris kapcsolat van.

$$\hat{y} = a \pm b \cdot x + \varepsilon$$

Hiba (0 átlagú)

Eredményváltozó
(prediktandusz)

Magyarázó változó
(prediktor)

Egyváltozós lineáris modell

Hipotézisek

- H_0: nincs regressziós kapcsolat

  $b = 0$

- H_1: van regressziós kapcsolat

  $b \neq 0$

Feltételek

- Y: normális eloszlású legyen.
- X: hibamentesen mérjük és legalább 3 értéke legyen.

Klasszikus legkisebb négyzetek módszerrel készült becslés (KLN) – Ordinary Least Squares (OLS)

A kapott paraméterek az adott megfigyelésekből számított, becsült regressziós együtthatók.

A fontosabb b jelentése az, hogy a magyarázó változó egységnyi növekedése átlagosan hány egységnyi növekedéssel / csökkenéssel jár együtt a becsült eredményváltozóban.

a a konstans együttható, vagy tengelymetszet paraméter (intercept) jelentése az, hogy ha a magyarázó változó 0 értéket vesz fel, a modell szerint mekkora lesz az eredményváltozó értéke.

- A hibák négyzetösszegét akarjuk minimalizálni.
- e_i véletlen zaj minimalizálnunk kell:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{N} \left(y_i - a - b \cdot x_i \right)^2$$

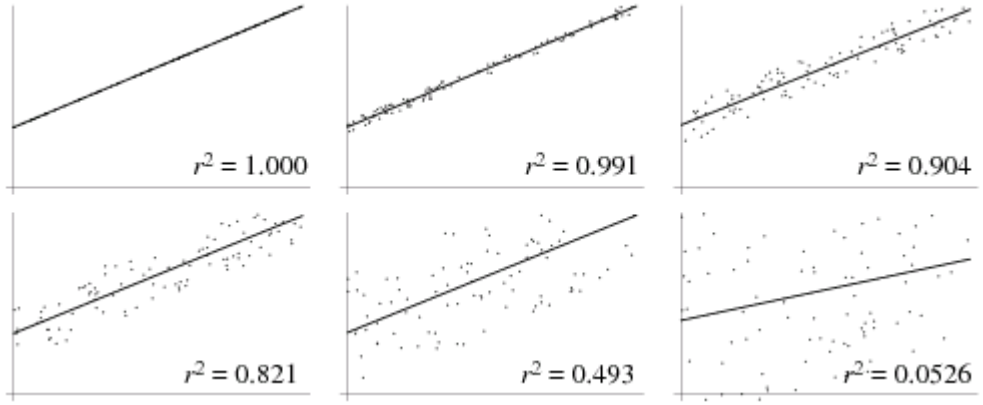
- Ez egy kvadratikus függvény, pontosan ott lesz minimális, ahol az *a* illetve a *b* szerinti deriváltak nullák.
- Eredmény:

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Determinációs együttható - az eredményváltozónak a regresszió által magyarázott és teljes eltérésnégyzetösszegei hányadosaként számítható mutatószám. Jelölése R^2 .

Az R^2 százszorosa megmutatja, hogy a regressziós modellel az y adatokban meglévő variancia (bizonytalanság) hány százaléka szüntethető meg.

- Az r^2 érték megmutatja a lineáris kapcsolat mértékét.



- Linearizálás módszere nem lineáris függés esetén segíthet:
  - Az $y = a + bx$ helyett az $y' = a + bx$ kapcsolatot keressük, ahol $y' = \ln y$. (Vagy, a logaritmikus skálára áttérést az x változóban is megtehetjük.

A becsült regressziós függvény segítségével a megfigyelési pontokban meghatározhatjuk a **reziduumok** értékeit:

$$e_i = y_i - \hat{y}_i$$

A megfigyelések és a becsült függvényértékek különbségét adják meg.

kis $|e_i|$ értékek= jó illeszkedés

nagy $|e_i|$ értékek = gyenge illeszkedés

$$SSE = \sum_{i=1}^{n} e_i^2$$

Sum of Squares of the Errors

reziduális szórás:

$$s_e^* = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}$$

Elaszticitás - rugalmasság

- Egy függvény (*f*(*x*)) **rugalmassága** azt mutatja meg, hogy hány százalékkal változik meg a függvény értéke, ha *x* értéke 1%-kal nő.
- A rugalmasság, amit *ε*-nal (epszilon) vagy **E**-vel jelölünk (***E****lasticity*) szó alapján, *f*(*x*) *x* hatására történő százalékos változásának és *x* „nagyon kicsi” százalékos változásának hányadosa:

$$E_{(X,Y)} = \lim_{\Delta X \to 0} \frac{\frac{\Delta Y}{Y}}{\frac{\Delta X}{X}} = \frac{dY}{dX} \cdot \frac{X}{Y}$$

$$E(X,Y) = \frac{b \cdot x}{a + b \cdot x} = \frac{b \cdot x}{\hat{y}}$$

Elaszticitás - rugalmasság

- Minél nagyobb ε abszolútértéke, az f függvényt annál rugalmasabbnak nevezzük.
- A rugalmasság előjelének is van jelentősége, az a derivált előjelét mutatja meg – vagyis azt, hogy $f(x)$ az x_0 helyen monoton növekvő vagy csökkenő.
- Általában öt esetet szoktunk elkülöníteni:

  - Ha $\varepsilon = 0$, akkor a függvény *tökéletesen rugalmatlan*. Az $f(x) = c$ konstansfüggvény például minden pontjában ilyen.

  - Ha $0 < |\varepsilon| < 1$, akkor *rugalmatlan*.

  - Ha $|\varepsilon| = 1$, akkor a függvény *egységnyi rugalmasságú*.

  - Ha $|\varepsilon| > 1$, akkor *rugalmas*.

  - Ha a derivált nem létezik, ezért ε nem értelmezhető, akkor *tökéletesen rugalmas* függvényről beszélhetünk.

- Ez szorosan kapcsolódik a regresszióhoz.

$$SST = SSR + SSE$$

teljes négyzetösszeg:

$$SST = \sum (y - \bar{y})^2 = \sum d_y^2$$

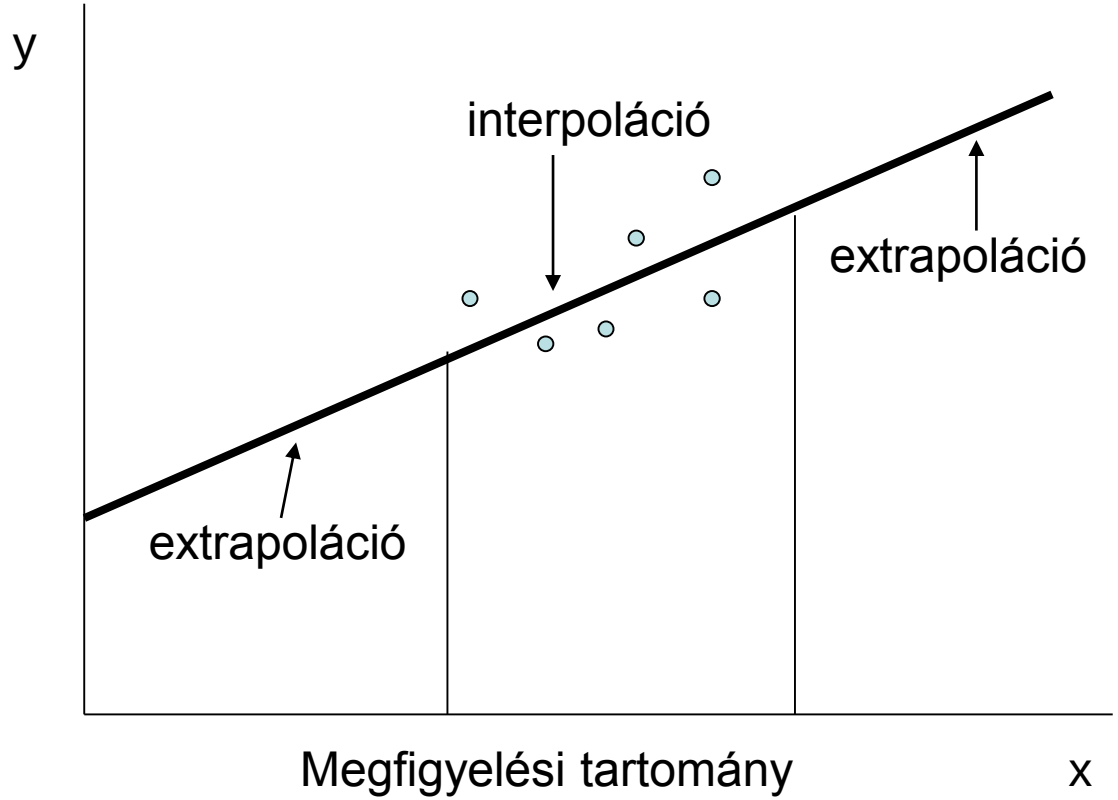
regressziós vagy magyarázott négyzetösszeg:

$$SSR = \sum (\hat{y} - \bar{y})^2 = \sum d_{\hat{y}}^2$$

maradék vagy hiba négyzetösszeg:

$$SSE = \sum (\hat{y} - y)^2 = \sum e^2$$

extrapoláció ⟺ interpoláció



Ha a magyarázó változók száma (k) több (k>1), akkor sok(több)változós lineáris modellről beszélünk:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \varepsilon$$

Feltételezzük, hogy valamennyi változóra *n* számú megfigyelésünk van, amelyeket célszerűen vektorokba, illetve mátrixba rendezhetünk:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1\, x_{11}\; x_{21} \cdots x_{k1} \\ 1\, x_{12}\; x_{22} \cdots x_{k2} \\ \vdots\;\vdots\;\;\vdots\;\;\ddots\;\vdots \\ 1\, x_{1n}\; x_{2n} \cdots x_{kn} \end{bmatrix}, \quad B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}, \quad \text{és} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

A modell feltételek vizsgálata

- A **multikollinearitás** úgy is megfogalmazható, hogy a magyarázó változók között korreláció van.
- Multikollineáris esetben mind a becslés, mind a paraméterek értelmezése megnehezedik, hiszen a magyarázó változók hatásait nem lehet egyértelműen szétválasztani.
- Minden változó hatása minden más változóban is megjelenik, a becslések bizonytalanná válnak.

- **Multikollinearitás**
- Mintabeli tulajdonság – mintán kívül nem alkalmazható.
- **Ellenőrzése:**
  – Többszörös determinációs együtthatóval,
  – F-próbával (F>Fkrit),
  – VIF-mutató,
  - Tolerancia mutató.

- Ez a mutató azt mutatja, hogy a *j*-edik változó becsült együtthatójának tényleges varianciája hányszorosa annak, ami a multikollinearitás teljes hiányának esete lenne.
- Ezért ezt a mutatószámot a *j*-edik változóhoz tartozó variancia inflációs tényezőnek (Variance Inflator Factor) VIF_j mutatónak nevezzük:

$$VIF_j = \frac{1}{1 - R_j^2}$$

- A *VIFj -mutató reciprokát tolerancia**mutatónak** nevezzük.*

$$Tolerancia = \frac{1}{VIF_j}$$

- ***Értéke:*** $0 \leq \text{Tolerancia} \leq 1$.
- Minél nagyobb a multikollinearitás mértéke annál közelebb van a mutató értéke a nullához.

Modell építési lehetőségek

- Enter (forced entry)
- Hierachikus (blockwise)
- Stepwise módszerek (forward, backward, stepwise)
- Kevert módszerek

Többszörös regresszió használatához a feltételek

- **Linearitás a függő változóval:** ha ez nincs, akkor alábecsüljük az y-t, pontatlan a modell.
- **Mintaszám:** kis elemszám növeli a β-hibát. Ökölszabály betartása többváltozós vizsgálatoknál.
- **Nincsenek több dimenziós extrém (outlier) értékek:** az együtthatók torzítását eredményezik.
- **Nincs multikollinearitás:** az összefüggő változók nem értelmezhetők. Ki kell hagyni a kevésbé fontos változót.
- **Minden változónak van varianciája:** nincs konstans változónk.
- **Nincs kovariáns** (külső befolyásoló változó).
- **Független hibák:** a belső korrelációk a CI-t, a szignifikancia értékeket torzítják.
- **Hiba normális eloszlása:** a normalitás sérülése, más feltételek sérüléseként keletkezik.
- **Változók típusai:** dummy-változó is engedett (pl. nem).
- Fennáll a homoszkedaszticitás (varianciák homogenitása vagy szóráshomogenitás): a heteroszkedaszticitás rontja a konfidencia-intervallumokat, torzítja a szignifikancia értékeket.

Shrinkage módszerek

- **Ridge regresszió**
- **Lasso**
- **Legkisebb szög regresszió (2004)**