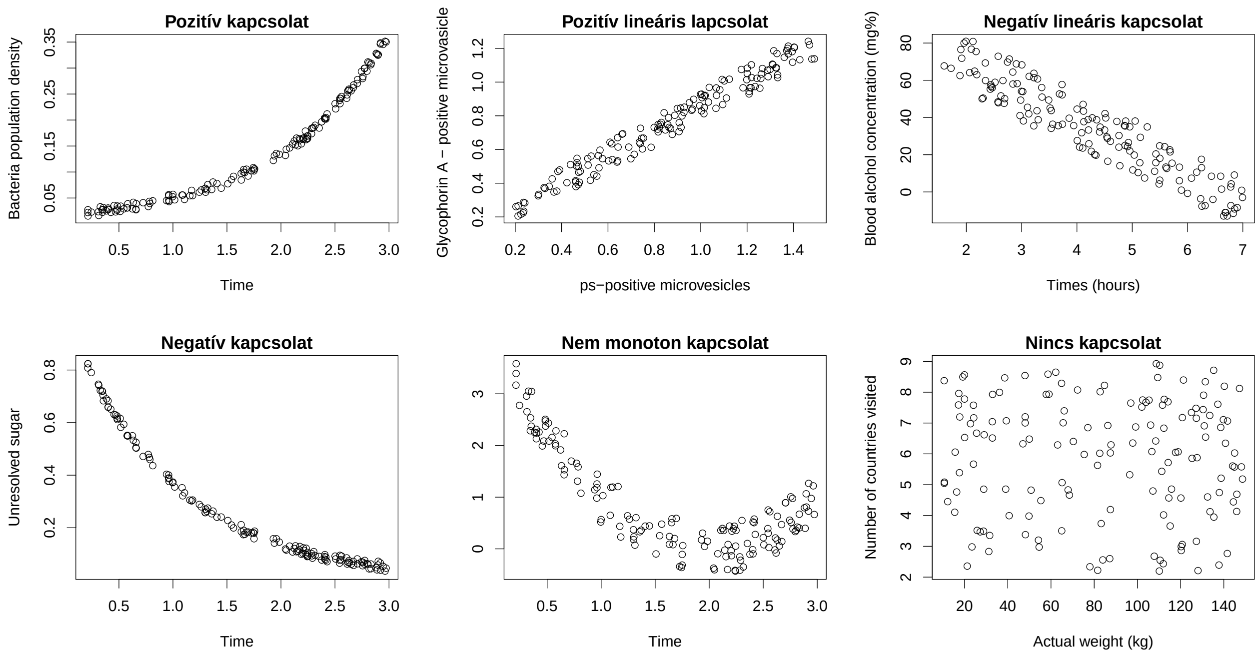
Regressziószámítás - Lineáris regresszió

Szűcs Mónika, Griechisch Erika, Rárosi Ferenc
SZTE ÁOK-TTIK Orvosi Fizikai és Orvosi Informatikai Intézet

Utoljára frissítve: 2020. június 4.

Élettudományokkal foglalkozó kutatások során gyakran mérnek több folytonos változót ugyanazokon a mintavételi egységeken *(pl.: testtömeg, vérnyomás, vércukorszint, testmagasság).* Ilyen esetekben felmerül a kérdés, hogy van-e valamilyen kapcsolat a két változó között. A regressziószámítás során egy függő és egy, vagy több független vagy magyarázó változó közötti kapcsolatot szeretnénk valamilyen függvénnyel leírni.

Lehetséges kapcsolatok két folytonos változó között



Jelen leckében csupán lineáris illetve linearizálható kapcsolatokat vizsgálunk kettő folytonos változó között.

1. Regressziós egyenes

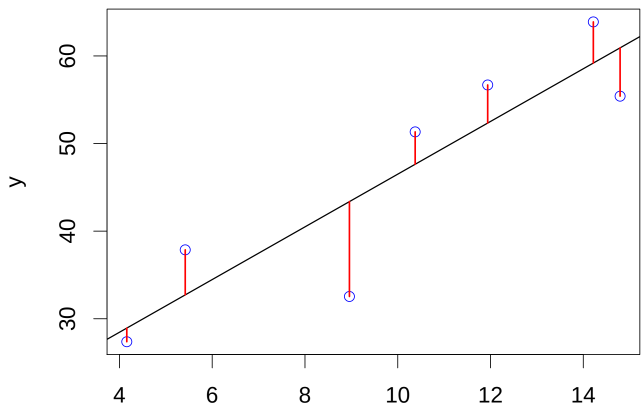
A regressziós egyenes a pontdiagramra legjobban illeszkedő egyenes, a függő (y) és a független (x) változó közötti lineáris kapcsolatot írja le.

$$y = a \cdot x + b$$

ahol a az illesztett egyenes meredeksége ; b az egyenes y-tengelymetszete.

A regressziós egyenes illesztése a legkisebb négyzetek módszerével történik. Meghatározzuk, hogy a mintapontok és az egyenes függőleges távolságainak (reziduumok) négyzetösszege mely a és b értékek esetén lesz minimális. A minimum meghatározása parciális deriválással történik, ezen módszer túlmutat jelen tananyag keretein, ezért nem részletezzük.

SZÉCHENYI 2020
Magyarország Kormánya
Európai Unió
Európai Szociális Alap
BEFEKTETÉS A JÖVŐBE



Az a és b értéke a legkisebb négyzetek módszere alapján:

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

A legjobban illeszkedő egyenes meredekségét (a) **regressziós együtthatónak** nevezzük.

Kapcsolat a korrelációs együttható (r) és a regressziós együttható (a) között:

$$r = a \cdot \frac{\mathrm{sd}_x}{\mathrm{sd}_y} \qquad a = r \cdot \frac{\mathrm{sd}_y}{\mathrm{sd}_x},$$

ahol sd_x a független, sd_y a függő változó megfigyelt értékeinek szórása

2. Hipotézisvizsgálat a regressziós együtthatóra *t* próbával

Célja annak vizsgálata van-e lineáris kapcsolat a függő (y) és független (x) változó között, pontosabban a regressziós együttható 0-e?

Hipotézisek:

- H_0: $a_{pop} = 0$ a regressziós egyenes meredeksége 0 a populációban, nincs lineáris kapcsolat a két változó között.
- H_1: $a_{pop} \neq 0$ a regressziós egyenes meredeksége nem 0 a populációban, van lineáris kapcsolat a két változó között.

Számolás és döntés:

I. Próbastatisztika

$$t = \frac{a}{\mathrm{SE}(a)},$$

ahol $\mathrm{SE}(a)$ regressziós együttható standard hibája.

Ha a nullhipotézis igaz, a próbastatisztika (t) Student féle t-eloszlást követ $n-2$ szabadságfokkal.

Döntési szabály:

- Ha $|t| < t_\alpha$, $t \in (-t_\alpha, t_\alpha)$: H_0-t elfogadjuk, a korrelációs együttható szignifikánsan eltér 0-tól α szinten, nincs elegendő okunk a lineáris kapcsolat feltételezéséhez.

- Ha $|t| > t_\alpha$, $t \notin (-t_\alpha, t_\alpha)$: H_0-t elvetjük, a korrelációs együttható szignifikánsan eltér 0-tól α szinten, elegendő okunk van a lineáris kapcsolat feltételezéséhez.
  ahol t_α a t-eloszlád α szignifikancia szinthez és $n-2$ szabadsági fokhoz tartozó kritikus értéke.

II. p-érték

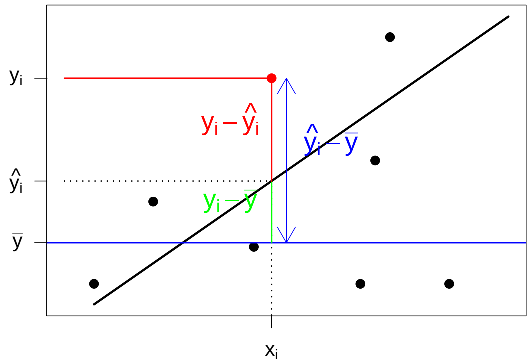
Döntési szabály:

- Ha $p > \alpha$, H_0-t elfogadjuk.
- Ha $p < \alpha$, H_0-t elvetjük.

3. Teljes variancia felbontása

$$SST = SSR + SSE$$

- SST $= \sum_{i=1}^{n}(y_i - \overline{y})^2$ a függő változó (y) teljes varianciája
- SSR $= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ y varianciájának a magyarázó változótól (x) való lineáris függéséből eredő variancia
- SSE $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ egyéb hatásokból adódó variancia, ahol $\hat{y}_i = ax_i + b$



4. A determinációs együttható

megmutatja, hogy a függő változó y varianciájának hány százaléka magyarázható az x-től való függéssel. A determinációs együttható értéke a korrelációs együttható négyzete százalékban kifejezve.

.

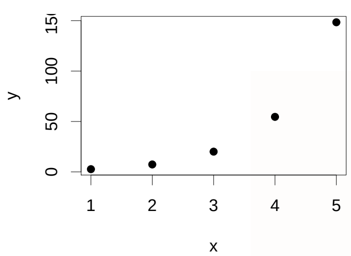
5. Nem lineáris kapcsolatok

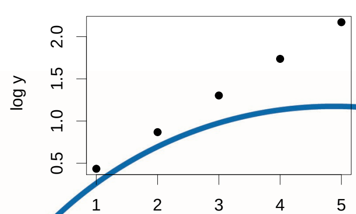
Abban az esetben, ha feltételezett kapcsolat nem lineáris (exponenciális, logaritmikus, hatvány), a változók transzformálásával a probléma visszavezethető lineáris regresszióra.

Kapcsolat	független változó	függő változó
lineáris	x	y
exponenciális	x	$\log y$
logaritmikus	$\log x$	y
hatvány	$\log x$	$\log y$

- Exponenciális kapcsolat

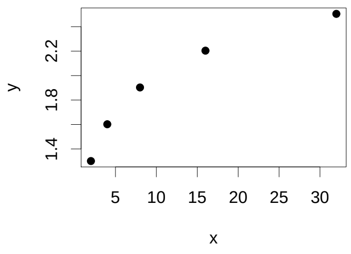
x	y	$\log y$
1	2.72	0.434
2	7.39	0.869
3	20.09	1.303
4	54.60	1.737
5	148.41	2.171

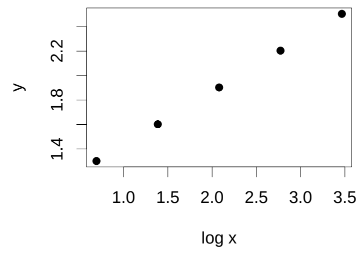




- Logaritmikus kapcsolat

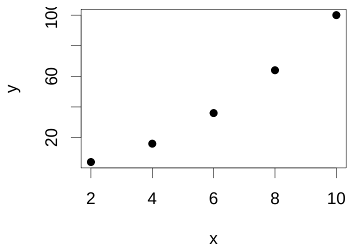
x	y	$\log x$
2	1.30	0.69
4	1.60	1.39
8	1.90	2.08
16	2.20	2.77
32	2.50	3.47

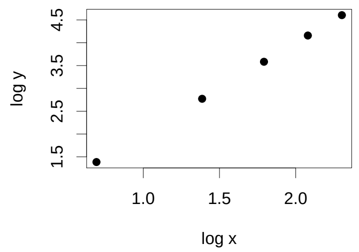




- Hatvány kapcsolat

x	y	$\log x$	$\log y$
2	4	0.69	1.39
4	16	1.39	2.77
6	36	1.79	3.58
8	64	2.08	4.16
10	100	2.30	4.61





Felhasznált irodalom

- Reiczigel Jenő, Harnos Andrea, Solymosi Norbert: Biostatisztika nem statisztikusoknak, Pars Kft. (2014)
- Reiczigel Jenő: Válogatott fejezetek a biostatisztikából, SZIE ÁOTK, (2005), https://docplayer.hu/279567-Reiczigel-jeno-valogatott-fejezetek-a-biostatisztikabol.html (hozzáférés: 2018.08 31.)

Jelen tananyag a Szegedi Tudományegyetemen készült
az Európai Unió támogatásával.
Projekt azonosító: EFOP-3.4.3-16-2016-00014