# Lineáris regresszió-analízis

Statisztika1

# Bevezetés

- Számos olyan jelenség van, amelyet nem tudunk közvetlenül mérni
  - nehéz számszerűsíteni
  - jövőben bekövetkező dologra utal
- predikció, előrejelzés
- Nem feltétlenül jelent oksági viszonyt, csak annyit jelent, hogy egy változó értékei bejósolhatóak egy másik változó értékeinek ismeretében!!!

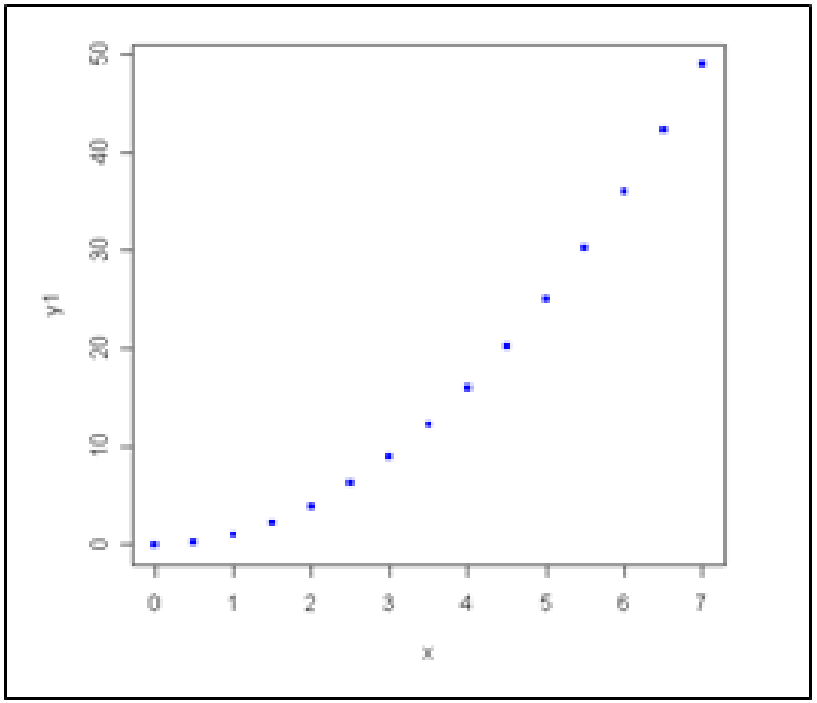
- „prediktor” (előjelző) vagy más néven független változó
- „célváltozó”, más néven függő változó
- Ha pontosan meghatároztuk a változók közötti kapcsolatot, akkor a prediktor változók értékeit használhatjuk arra, hogy a célváltozó értékeit megbecsüljük hasonló személyek esetében.

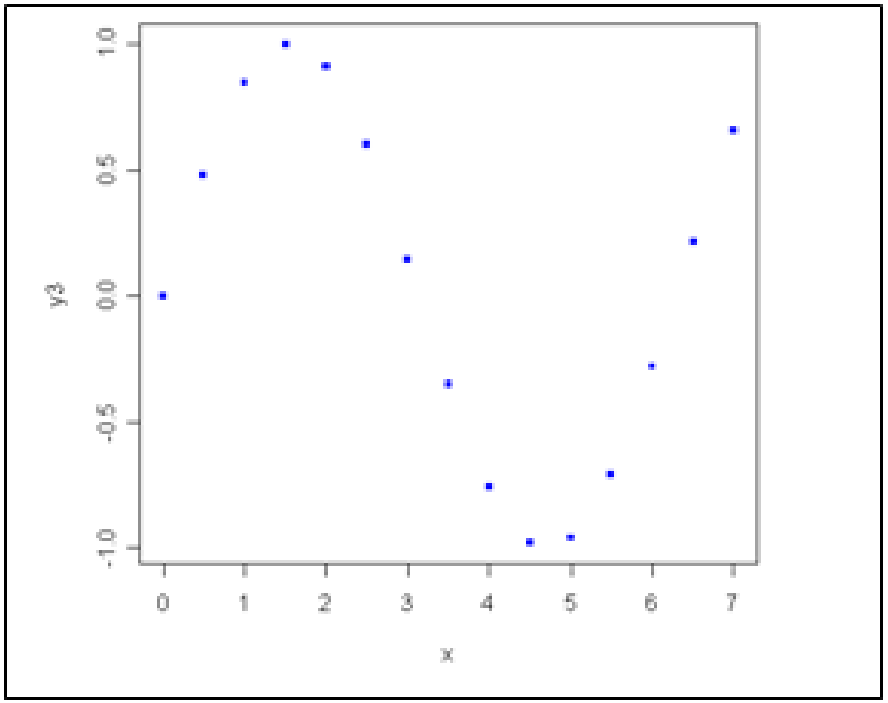
# Milyen mértékű elégedettséget von maga után a fizetésemelés?

| | fizetés (ezer) | elégedettség | | | fizetés (ezer) | elégedettség |
|---|---|---|---|---|---|---|
| | 44 | 50 | | | 44 | 30 |
| | 66 | 50 | | | 66 | 45 |
| | 89 | 50 | | | 89 | 60 |
| | 155 | 50 | | | 155 | 100 |
| | 130 | 50 | | | 130 | 85 |
| átlag: | 96,8 | 50 | | átlag: | 96,8 | 64 |

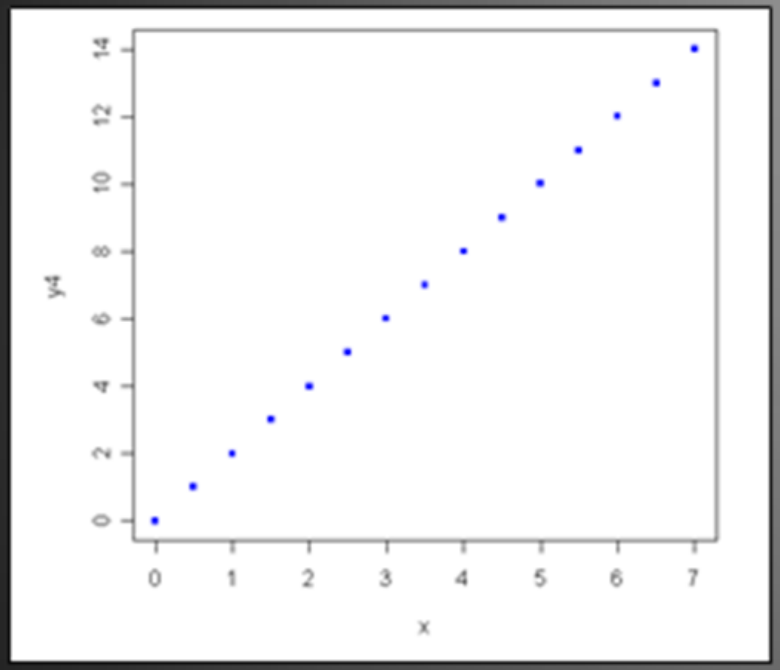
# Szisztematikus kapcsolat

- Ahhoz, hogy egy változó alapján becsülni tudjunk egy másik változót, valamilyen **szisztematikus kapcsolatnak** kell fennállnia a két változó között. (szabályos alakzat)
- Ebben az esetben az egyik változó (X) valamilyen függvénye a másik változónak (Y), és függvényszerű kapcsolatot találunk X és Y között.


y1
0
10
20
30
40
50
0
1
2
3
4
5
6
7
x


y3
-1.0
-0.5
0.0
0.5
1.0
0
1
2
3
4
5
6
7
x

# Lineáris összefüggés



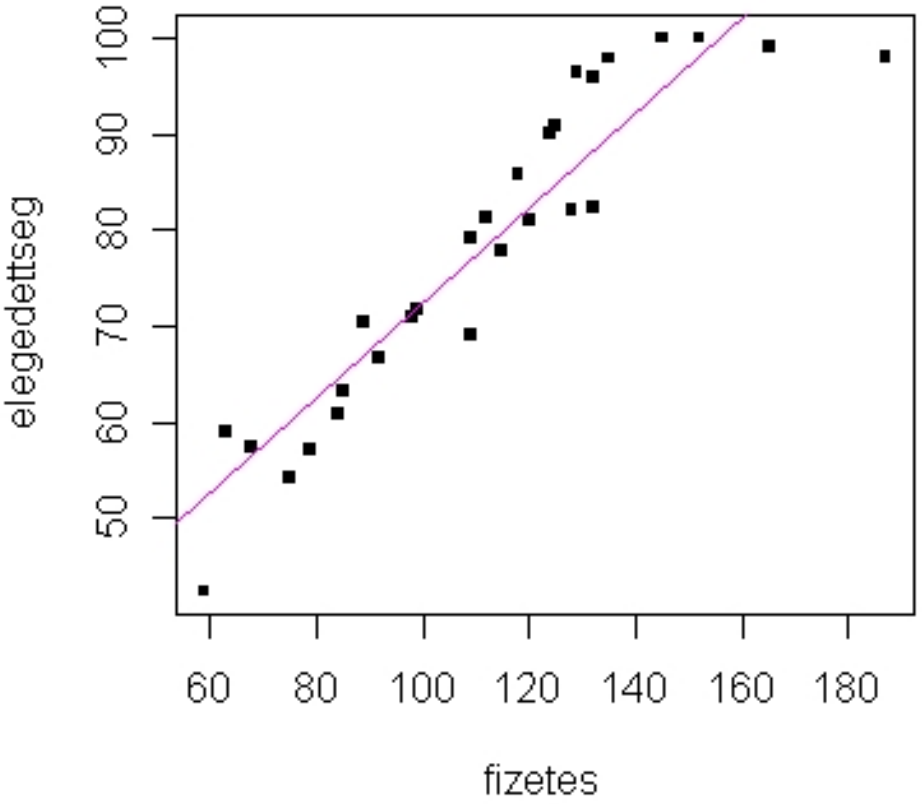

- A szisztematikus kapcsolatok egyik legegyszerűbb formája, amikor lineáris a kapcsolat két változó között.
- Milyen mértékű változás várható az Y változóban, ha X adott mértéknyit változik?

- Ennek az együttjárásnak a szorosságát egy mérőszám, a korrelációs együttható mutatja: r
- Függvényszerű összefüggés:

  elégedettség = 23 + 0,5 * fizetés

# Lineáris (egyenes) kapcsolat jellemzése

- Egyenes
- $Y = b_0 + b_1 X$

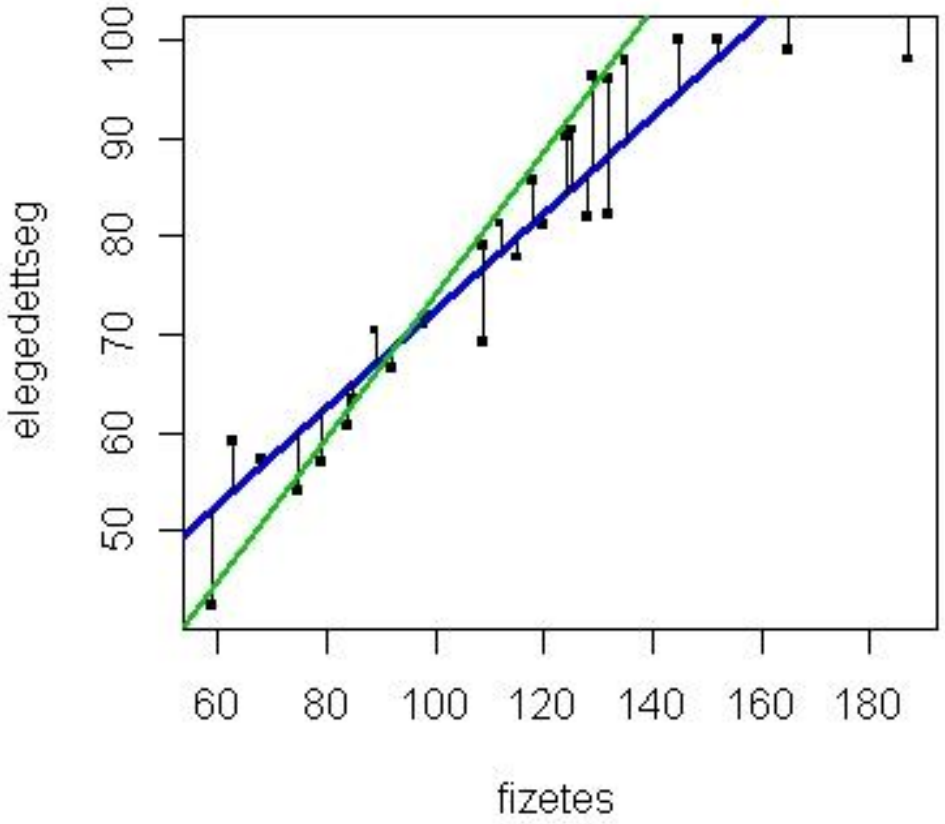


# A lineáris regressziós modell

- Egyenlet:

$$Y = b_0 + b_1 X$$

- A pontok az esetek többségében nem egy egyenesre esnek
- Számtalan olyan egyenes lehet, amelyet a ponthalmazra (a személyek adatinak a halmazára) illesztünk. Van legjobb?
- Legkisebb négyzetek elve: úgy kell $b_0$ és $b_1$-t meghatározni, hogy az együttes eloszlásukra illesztett egyenes körül lévő pontok variabilitását minimalizáljuk, azaz az egyenesnek az egyes pontoktól mért távolsága a lehető legkisebb legyen


elegedettseg
50
60
70
80
90
100
60
80
100
120
140
160
180
fizetes

- Legkisebb négyezetek elve

$$\sum (Y - \hat{Y})^2$$

- a kapott együtthatók nem a valódi, mért adatok együtthatói, azokat ugyanis nem ismerjük
- Az együtthatókból a regressziós egyenlet segítségével visszaszámolhatjuk az „eredeti” adatokat, vagyis megnézhetjük, hogy az ismert független változóhoz az egyenlet alapján milyen függő változóbeli értékek tartoznak.

- Mivel az adatok, a pontok, nem az elméleti egyenesre esnek, vagyis az Y valódi értékei eltérnek a $b_0$ és $b_1$ segítségével becsült értékektől, így ezzel a különbséggel számolni kell, ez a statisztikai értelemben vett hiba (ε).
- Ily módon a lineáris regresszió egyenletét a következőképpen egészítjük ki

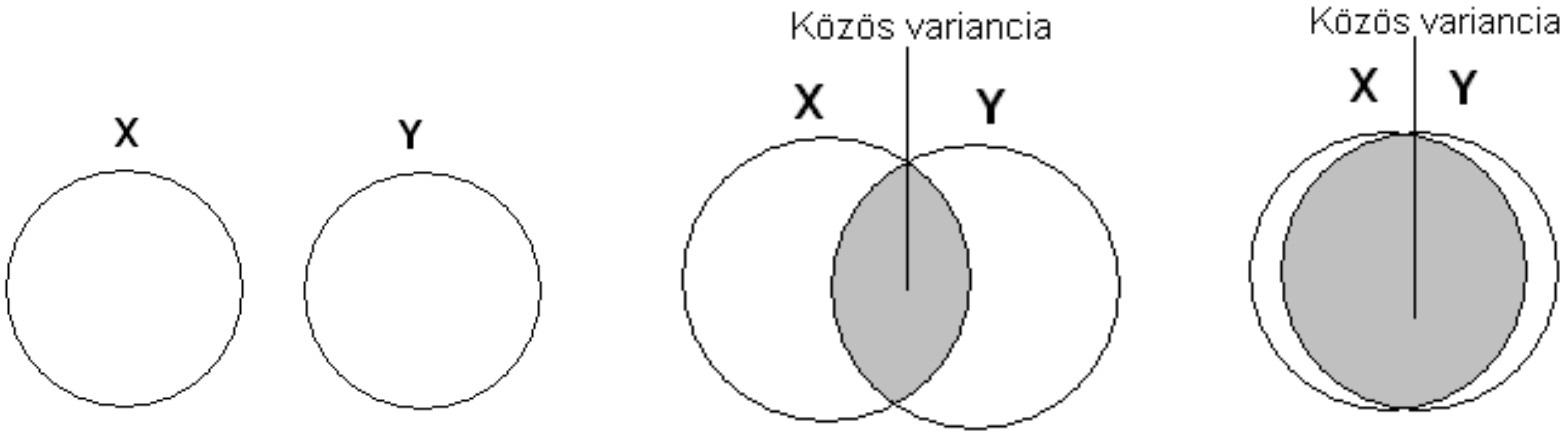
$$Y = b_0 + b_1 X + \varepsilon$$

# Mire használhatjuk a lineáris regresszió analízist?

- Annak meghatározására, hogy a független változók hatással vannak-e a függő változóra: Van-e összefüggés?
- Annak meghatározására, hogy a független változó milyen mértékben magyarázza a függő változó ingadozását: A kapcsolat erőssége.
- A kapcsolat formájának és struktúrájának meghatározása: matematikai egyenlőség.
- Predikció: függő változó értékeinek az előrejelzése.
  - Ha a független változó értékei köz nem szerepel egy érték, de a megfigyelt tartományban van (megfigyelt min. és max. érték között), akkor következtethetünk a függő változó értékére (interpoláció), ha a megfigyelt tartományon kívül van, akkor ezt nem tehetjük meg (extrapoláció).

# Négyzetes korrelációs együttható

- Kapcsolat szorossága: Pearson-féle korrelációs együttható
- Ha a korrelációs együtthatót a négyzetre emeljük ($r^2$), akkor megkapjuk, hogy Y varianciájának mekkora hányadát magyarázza X varianciája, vagy fordítva, X varianciájának mekkora hányadát magyarázza Y varianciája.
- Az arányt legtöbbször százalékos formában adjuk meg


X
Y
Közös variancia
X
Y
Közös variancia
X
Y

# A regressziós együtthatók vizsgálata

- A lineáris modellben a $b_1$ meredekségi paraméter azt mutatja, hogy X egységnyi változásához mekkora Y-beli változás tartozik.
- Ha $b_1=0$, akkor azt mondhatjuk, hogy X és Y lineárisan függetlenek egymástól.
- A becslés alapja a legkisebb négyzetek elve.

- A $b_1$ valószínűségi eloszlása normál eloszlás

$$b_1 \sim N(\beta_1, \sigma_{b_1})$$

- Az együtthatók vizsgálatakor a kiindulási hipotézis az, hogy $\beta_1 = 0$.
- Ez azt jelentené, hogy a lineáris egyenlet a következőképpen néz ki:

$$Y = b_0 + \varepsilon$$

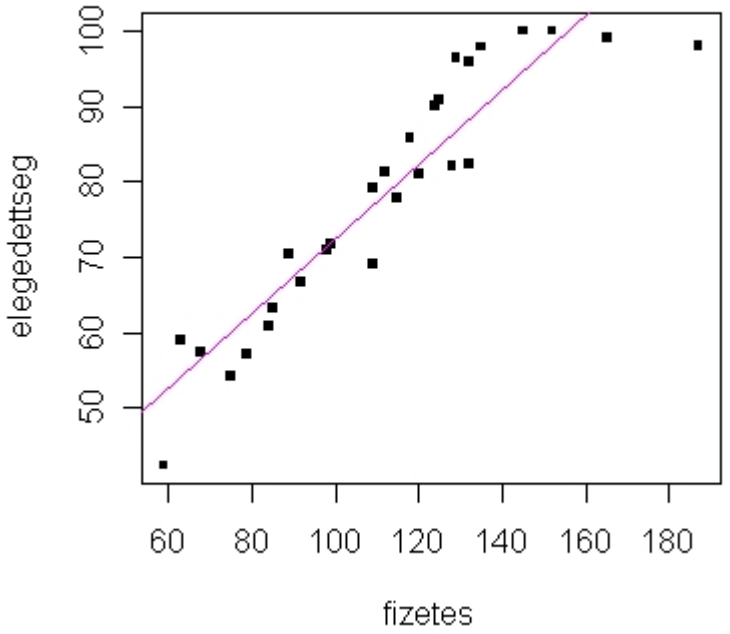
- A hipotézistesztelést elvégezhetjük annak ismeretében, hogy $\frac{b_1 - \beta_1}{\hat{\sigma}_{b_1}}$ t-eloszlást mutat

(n-2) szabadsági fokkal.

- Így a pontos tesztstatisztika a következőképpen néz ki :

$$t = \frac{b_1 - 0}{\hat{\sigma}_{b_1}} = \frac{b_1}{\hat{\sigma}_{b_1}}$$

- plot(elegedettseg~fizetes, data=d, cex=5, pch=".")
  model<-lm(elegedettseg~fizetes, data=d)
  abline(model, col=6)



- summary(lm(elégedettség~fizetés))

```
Call:
lm(formula = elégedettség ~ fizetés)

Residuals:
       1        2        3        4        5
-11.3812   0.3622  11.9575  -2.2211   1.2825

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.06500    8.87052   2.262   0.1087
fizetés      0.14803    0.02611   5.669   0.0109 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.647 on 3 degrees of freedom
Multiple R-squared:  0.9146,    Adjusted R-squared:  0.8861
F-statistic: 32.13 on 1 and 3 DF,  p-value: 0.01087
```