Valószínűségszámítás és Statisztika

12. előadás

2017. december 11.

Y közelítése X függvényével

- Gyakori eset, hogy nem ismerjük a számunkra érdekes mennyiség (Y) pontos értékét (pl. holnapi részvény-árfolyam, vízállás, időjárás). Van viszont információnk hozzá kapcsolódó mennyiségről (X, mai értékek).
- Feladat: olyan f_0 megtalálása, amelyre $f_0(X)$ a lehető legjobb közelítése Y-nak.
- Matematikailag: f_0 a megoldása a $\min_f E(Y - f(X))^2$ szélsőérték-problémának (legkisebb négyzetes becslés).
- Ha az együttes eloszlás ismert (nem teljesen reális, de a megfigyelések alapján közelíthető), akkor megoldható a feladat.

Valószínűségszámításból tanultak

$E\left(Y-a\right)^2$ minimumhelye: EY

$E\left(Y-f\left(X\right)\right)^2$ minimumhelye: $f_0(x)=E\left(Y\middle|X=x\right)$

lineáris függvények esetében:

$E\left(Y-aX-b\right)^2$ minimumhelye:

$$a=\frac{\operatorname{cov}(X,Y)}{D^2X}=\frac{\operatorname{corr}(X,Y)DY}{DX}$$

$$b=EY-aEX$$

Példa

- Annyi érmével dobtunk újra, amennyi fejet kaptunk 2 érmével dobva. Csak azt tudjuk, hogy hány fejet kaptunk a második dobásnál. Közelítsük ennek segítségével az első dobás eredményét.
- Például $F=0$ esetre:

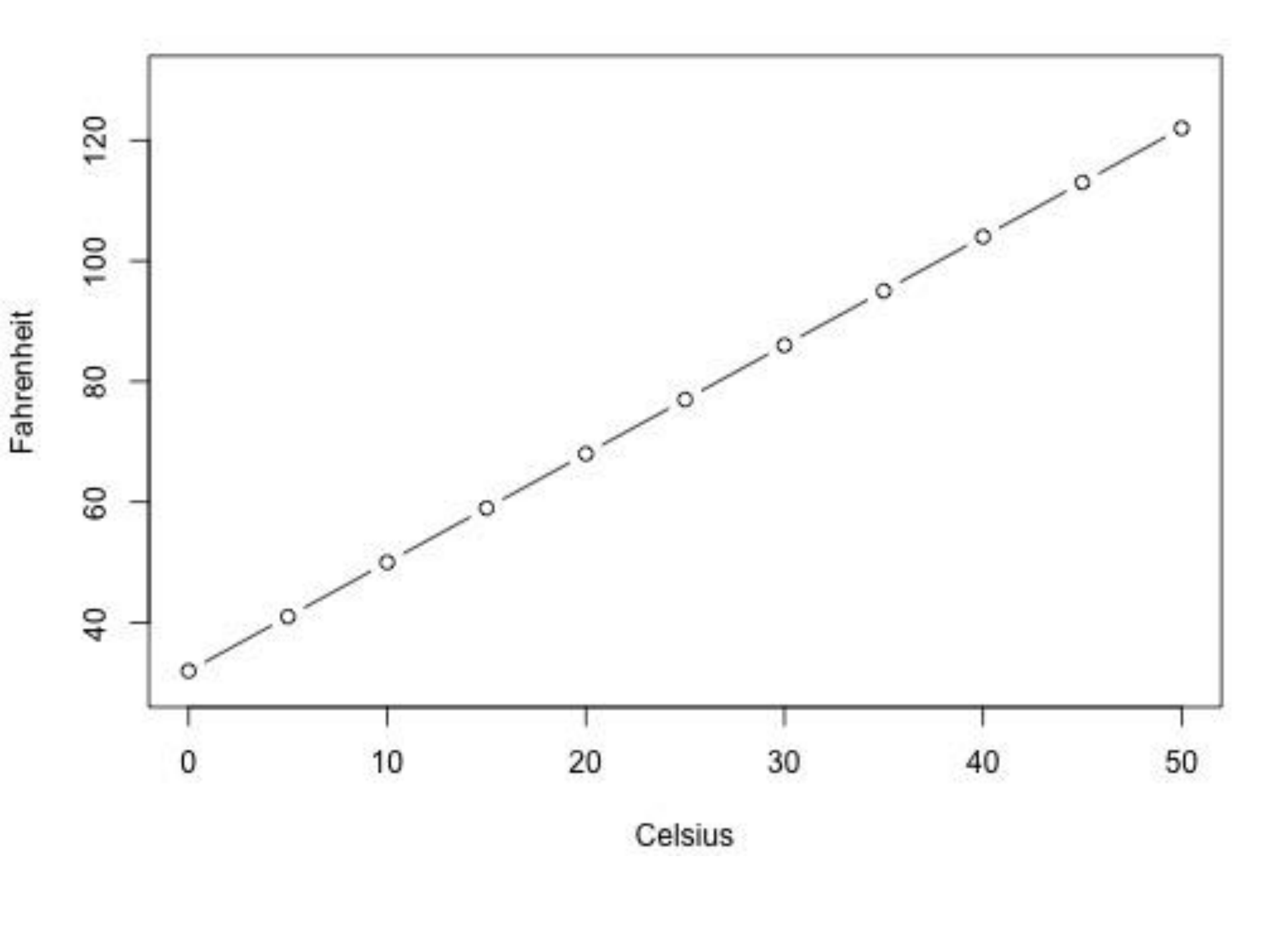
$$E(X \mid F=0) = \frac{\sum_{i=0}^{2} iP(X=i, F=0)}{P(F=0)} = \frac{\sum_{i=0}^{2} iP(F=0 \mid X=i)P(X=i)}{\sum_{i=0}^{2} P(F=0 \mid X=i)P(X=i)}$$

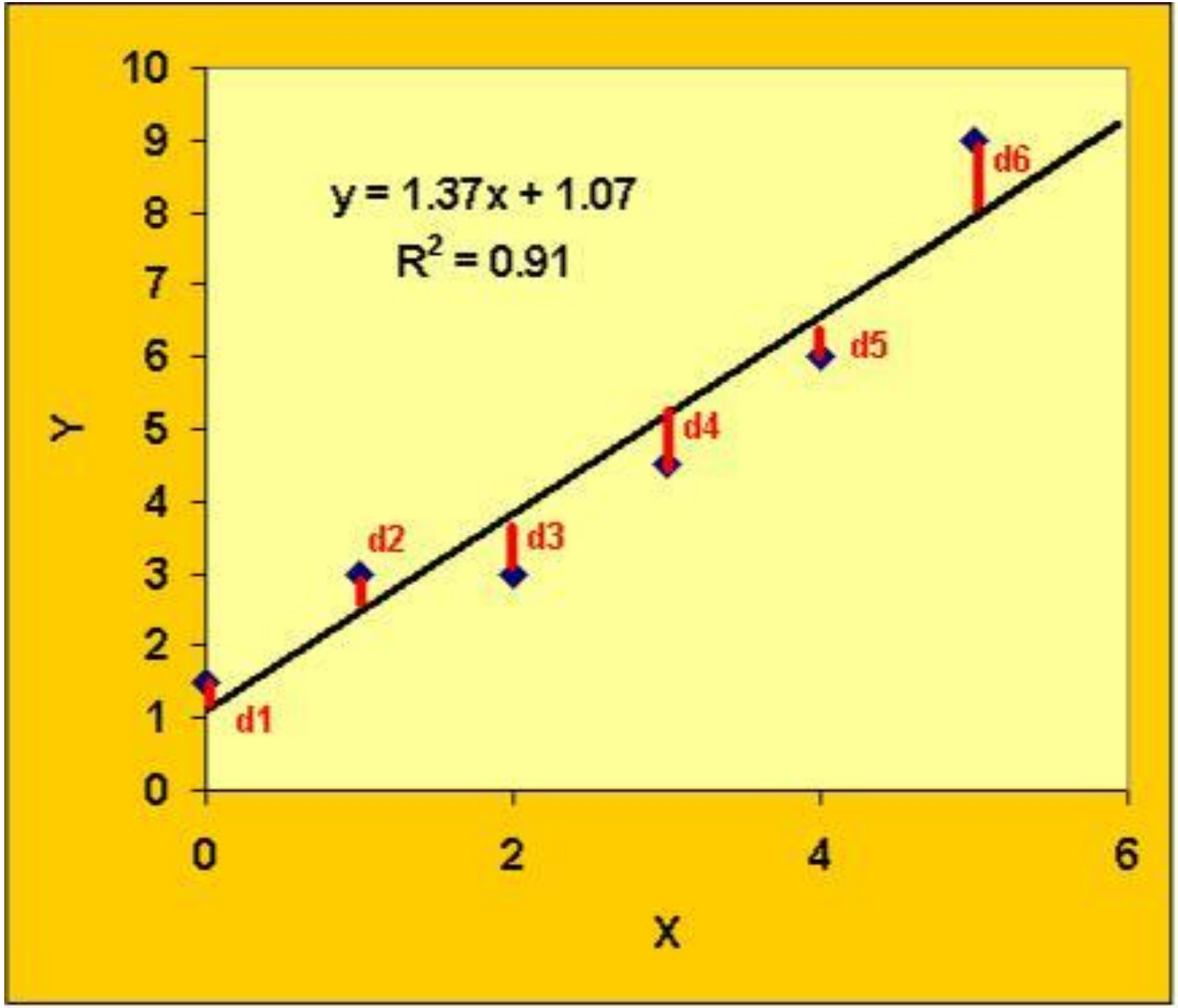
- Az eredmények: $E(X|F=2)=2$, $E(X|F=1)=4/3$, $E(X|F=0)=2/3$.

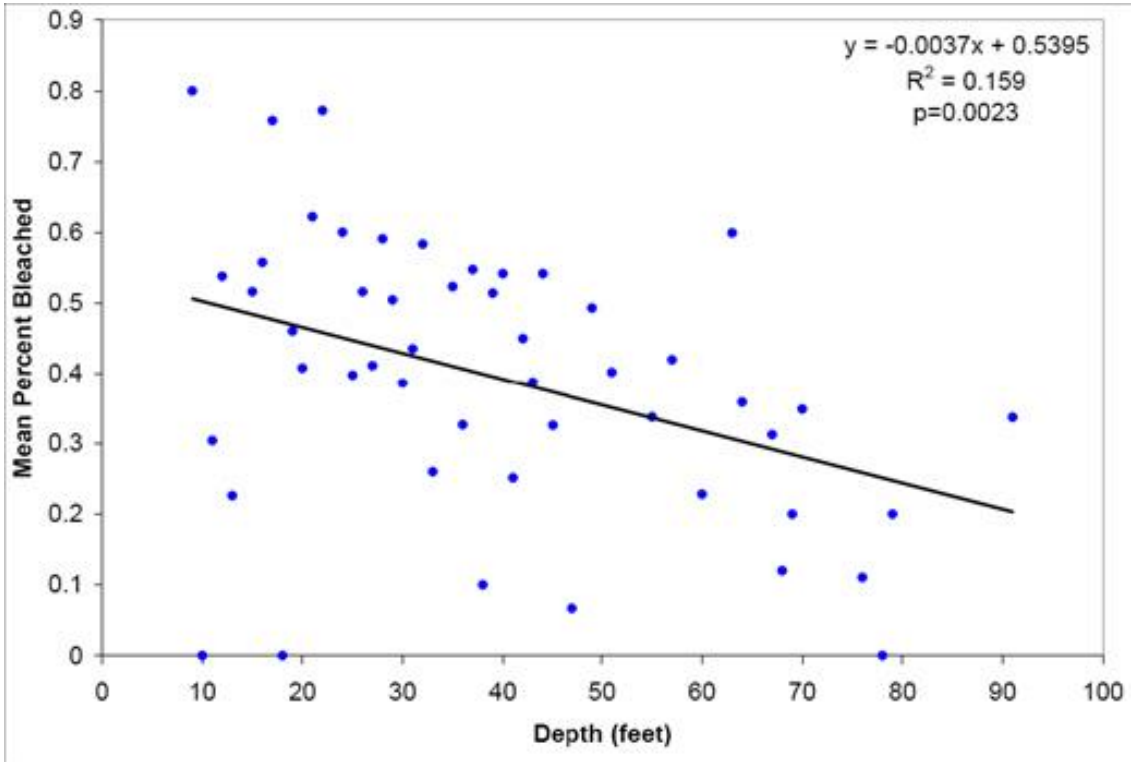
Az $aX+b$ egyenes tulajdonságai

- Ez a legkisebb négyzetes eltérést adó a lineáris függvények között (a fenti megoldás valóban minimum)
- Elnevezés: regressziós egyenes
- Átmegy az $(E(X),E(Y))$ ponton
- Példa: Kockával dobunk, majd ha k az eredmény, az $1,\ldots,k$ cédulák közül húzunk egyet. Nem tudjuk a húzás eredményét, csak a kockadobását. Hogyan tippeljünk a húzott számra (a legkisebb négyzetes eltérést adó becslést keressük)?
  $E(h|K=k)=(k+1)/2$

  az univerzálisan legjobb közelítés, tehát a legjobb lineáris közelítés is.


Fahrenheit
40
60
80
100
120
0
10
20
30
40
50
Celsius


y = 1.37x + 1.07
R2 = 0.91
d1
d2
d3
d4
d5
d6
Y
X
0
1
2
3
4
5
6
7
8
9
10
0
2
4
6


y = -0.0037x + 0.5395
R2 = 0.159
p=0.0023
Mean Percent Bleached
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
Depth (feet)

Lineáris modell

- $Y_i = aX_i + b + \varepsilon_i$ (X_i a magyarázó változó értéke, ε_i független, azonos eloszlású hiba. $E(\varepsilon_i)=0$, $D(\varepsilon_i)=\sigma$, általában feltesszük, hogy normális eloszlású. *a, b, σ* a becsülendő együtthatók)
- $\Sigma(Y_i-(aX_i+b))^2 \rightarrow \min$
- Megoldás:

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

A becslések szórása

$$D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, D(\hat{b}) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Az x* pontban előrejelzett érték $\hat{a}x^* + \hat{b}$

és ennek szórása $\sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

A szórásbecslésnél σ helyett annak becsült értékét használjuk: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2}{n-2}$

Hipotézisvizsgálat/1

H_0:a=0 tesztelése t-próbával:

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

- Ebből konfidencia intervallum is kapható a-ra

Hipotézisvizsgálat/2

- H_0:b=0 tesztelése t-próbával:

$$t_{n-2} = \frac{\hat{b}\sqrt{n(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \hat{a}x_i - b)^2}\sqrt{\sum_{i=1}^{n} x_i^{\,2}}}$$

- Ebből konfidencia intervallum is kapható b-re

Szóródások

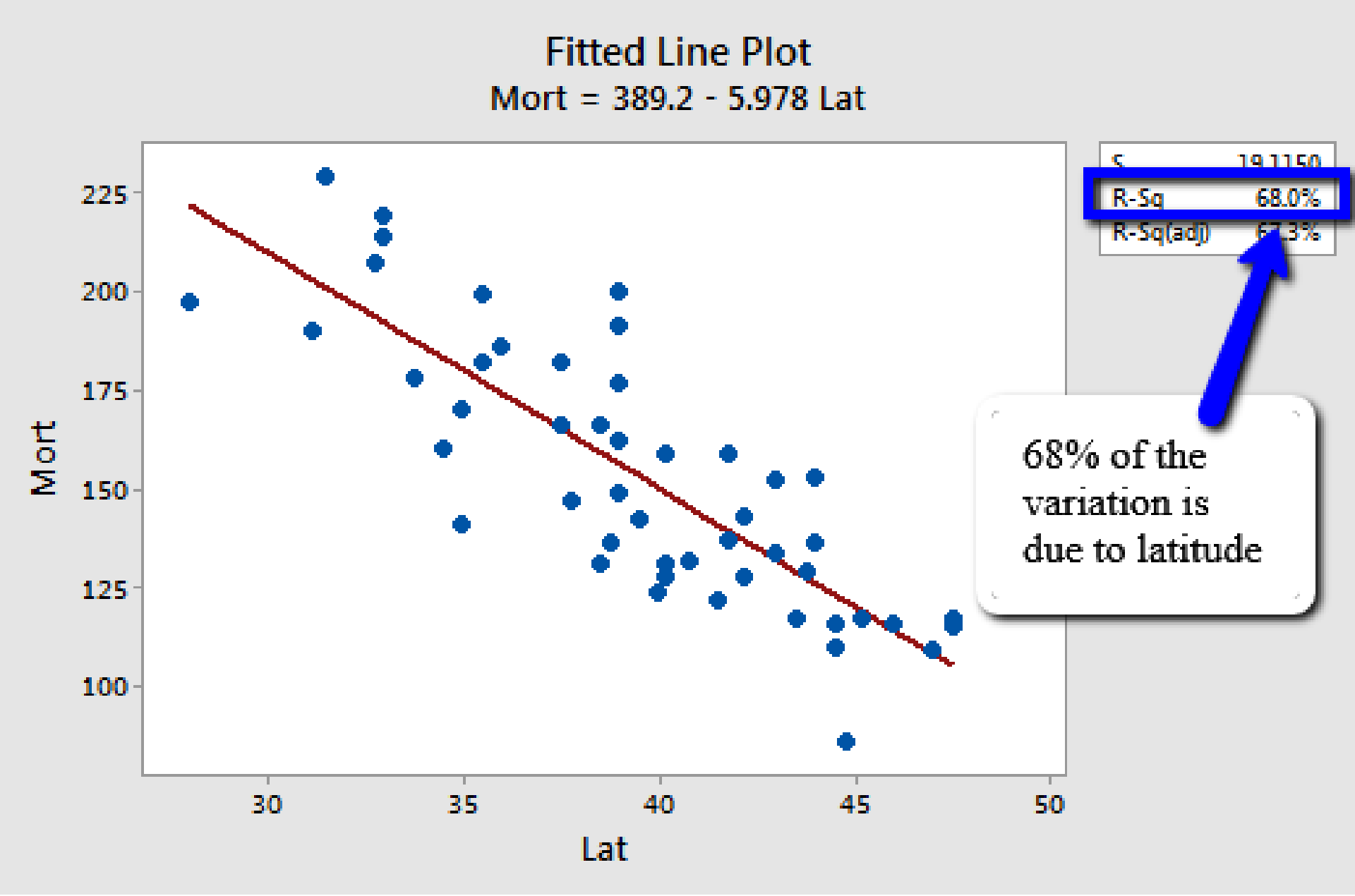
- Teljes ingadozás: $\sum_{i=1}^{n}(y_i-\bar{y})^2$
- Reziduális négyzetösszeg:

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

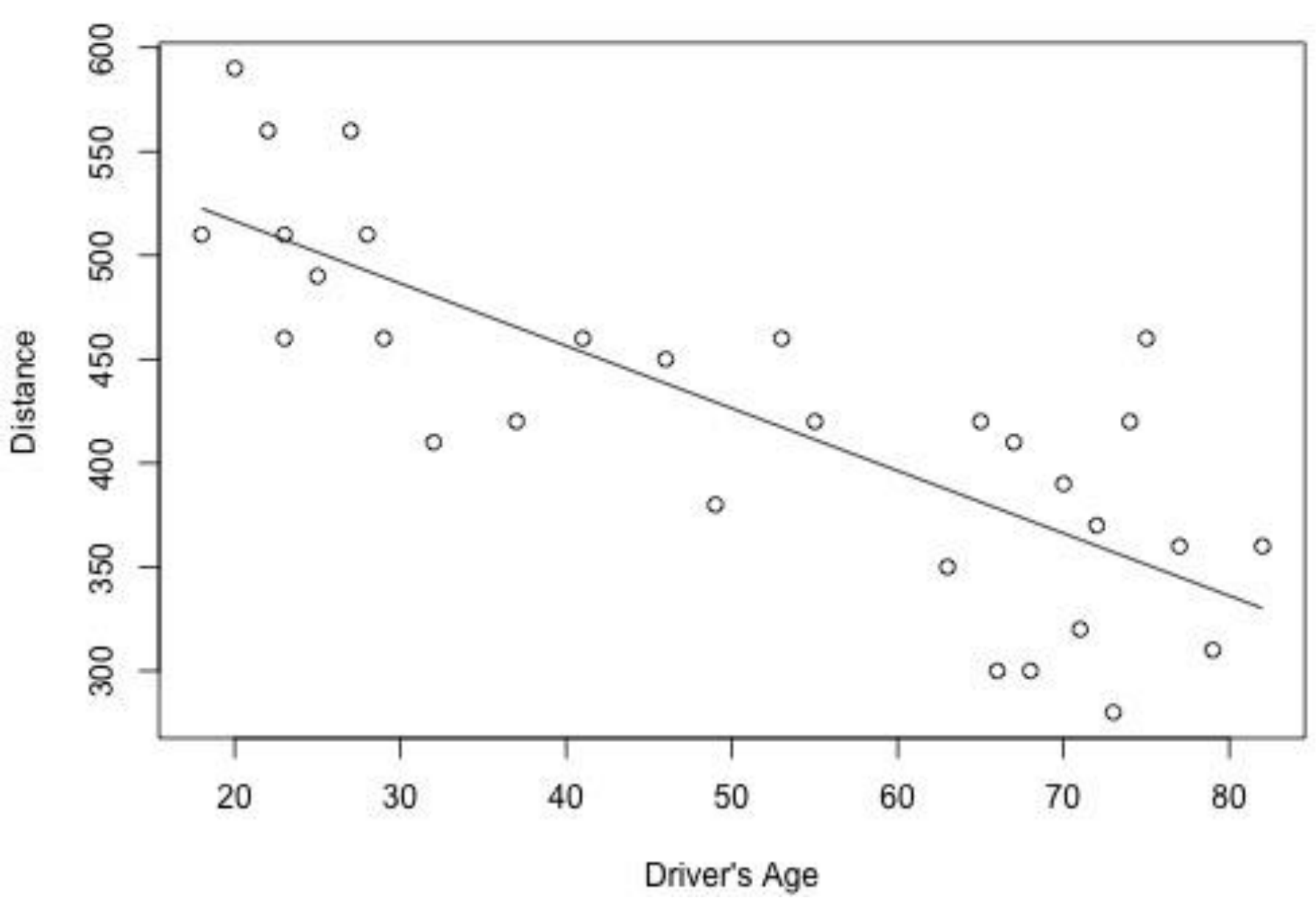
- A megmagyarázott variabilitás részaránya:

$$R^2=\frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

éppen a tapasztalati korrelációs együttható négyzete

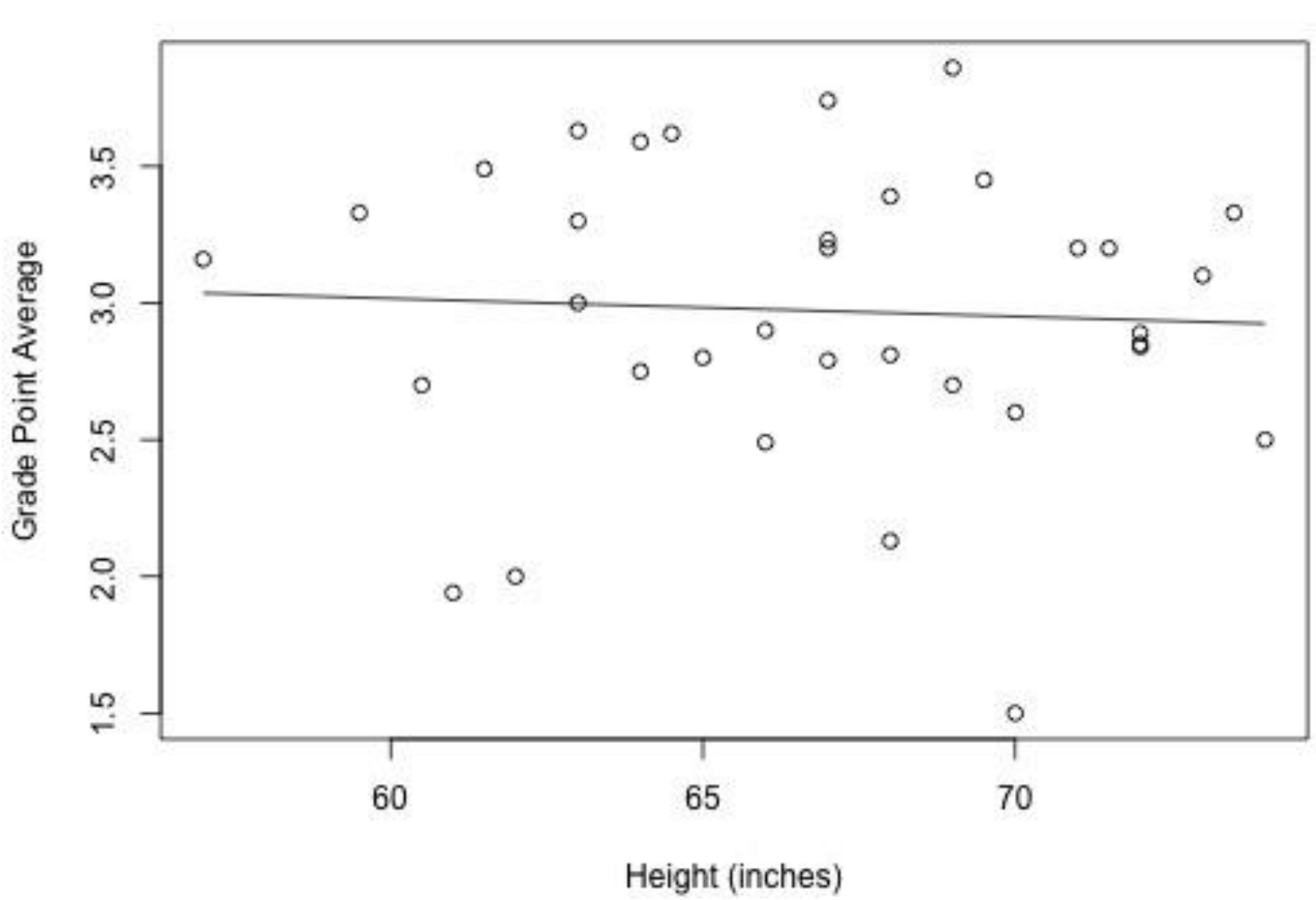


https://onlinecourses.science.psu.edu/stat501/node/255



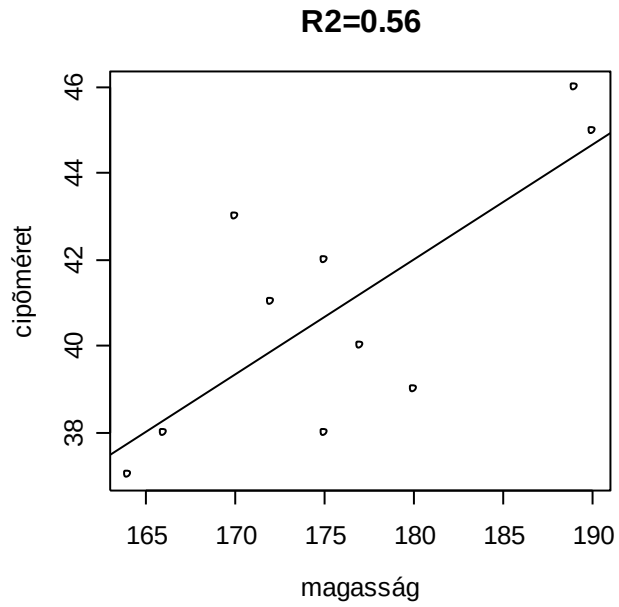
r^2=64,2% , r=-0,801

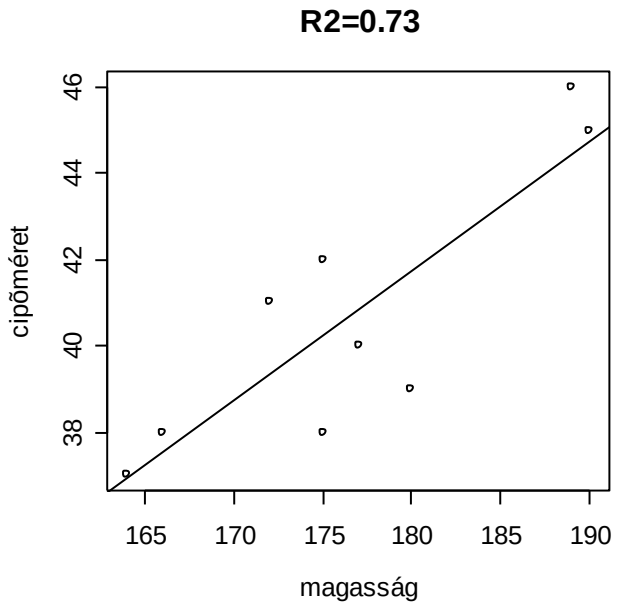
https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/01simple/signdist.jpeg



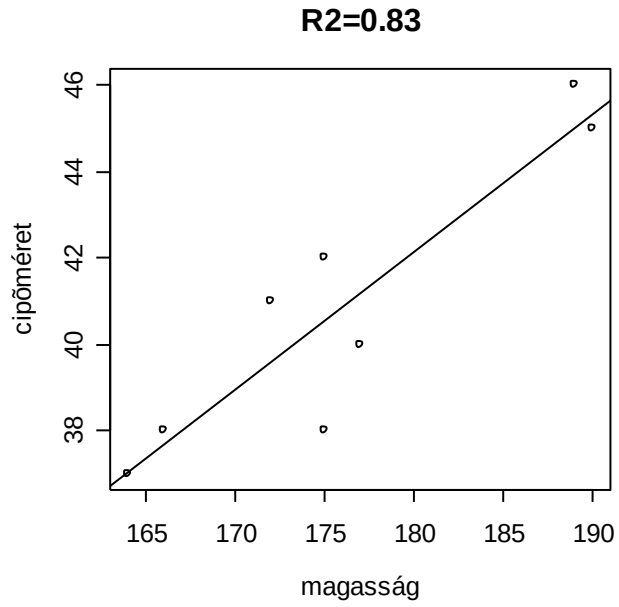
r^2=0,3% , r=-0,053

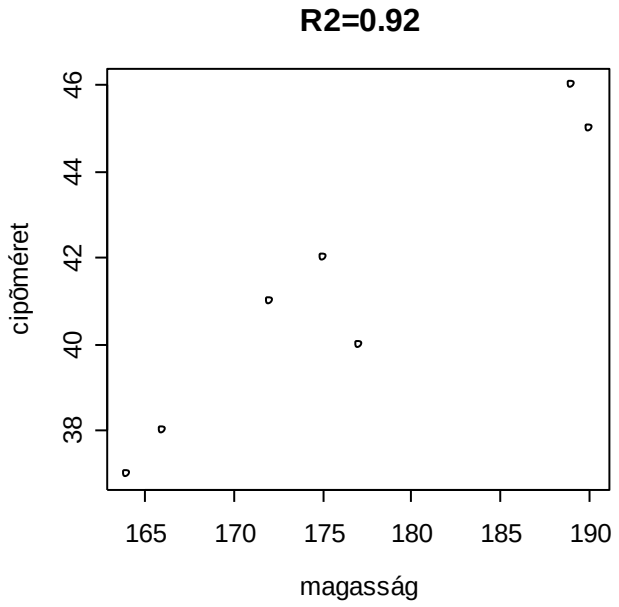
https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/01simple/heightgpa.jpeg

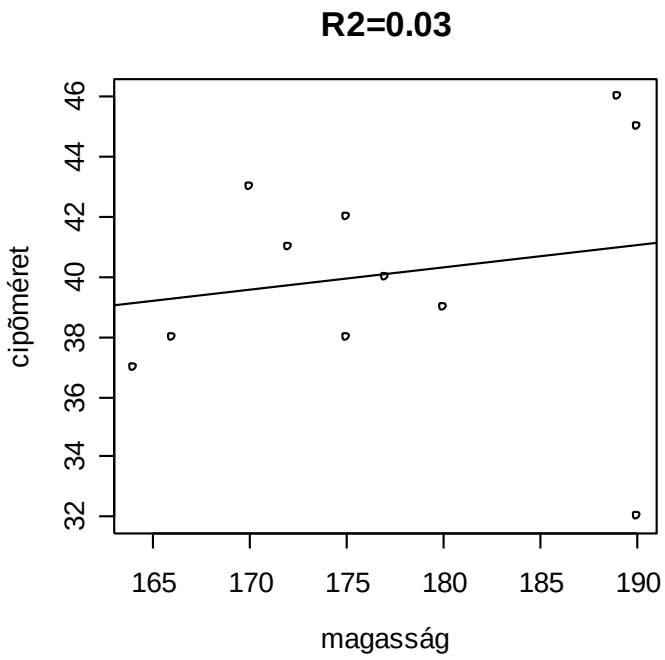


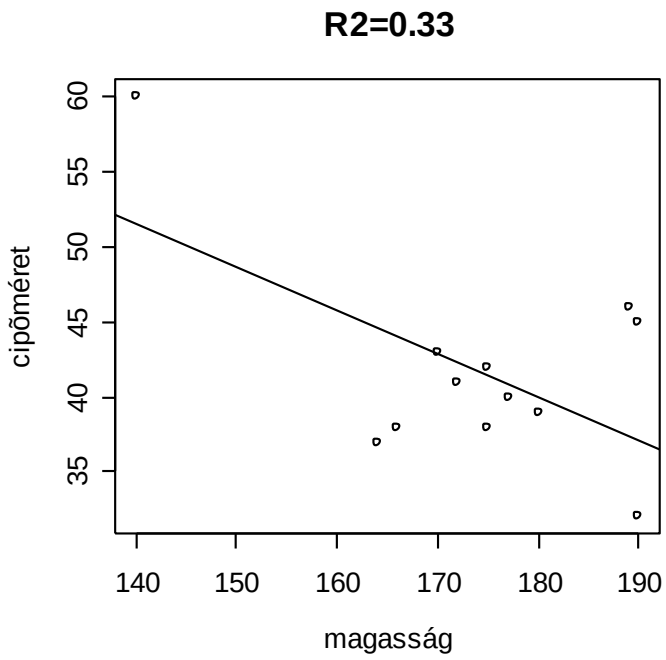


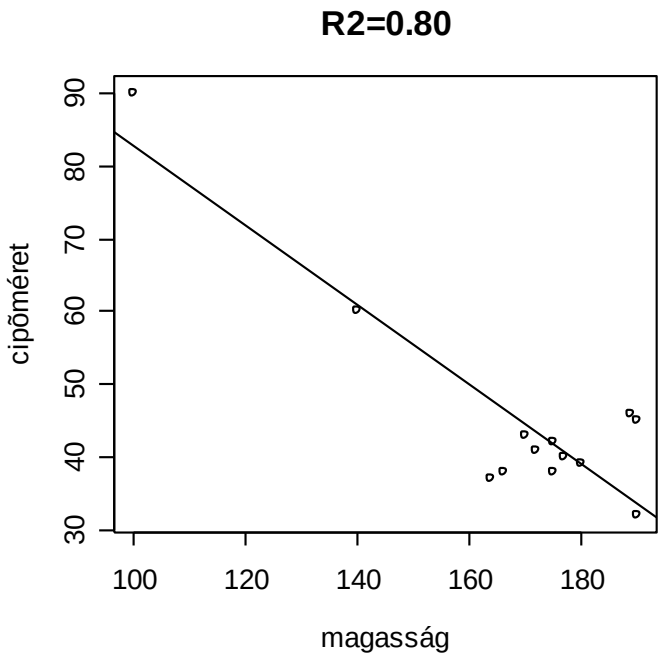
A regresszió vizsgálata

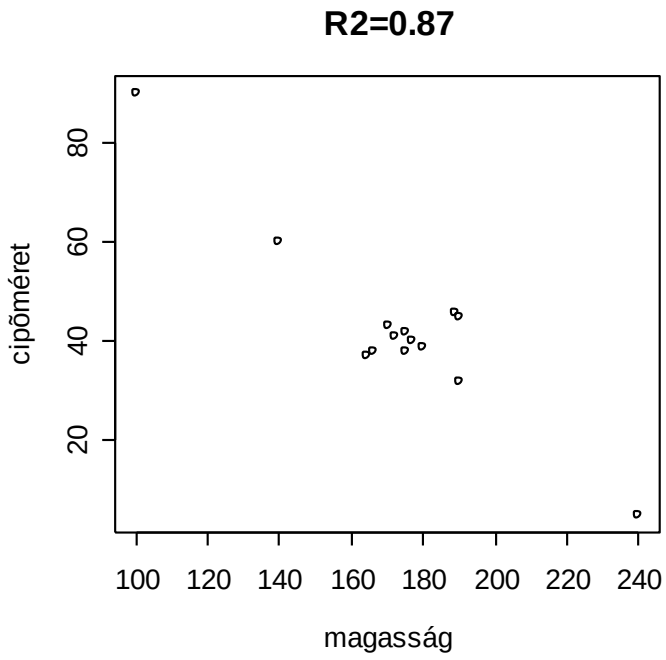


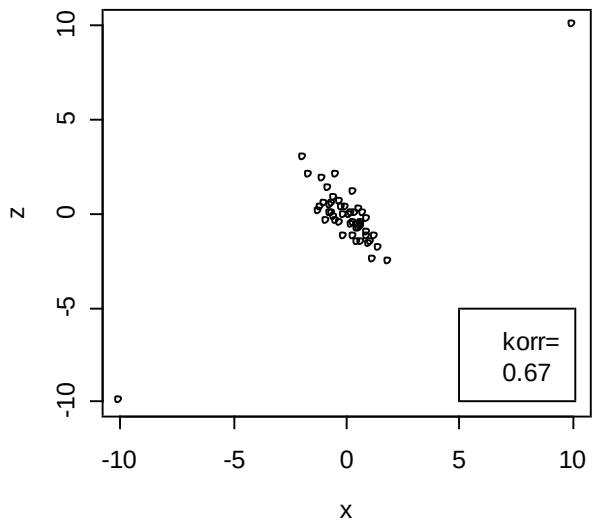


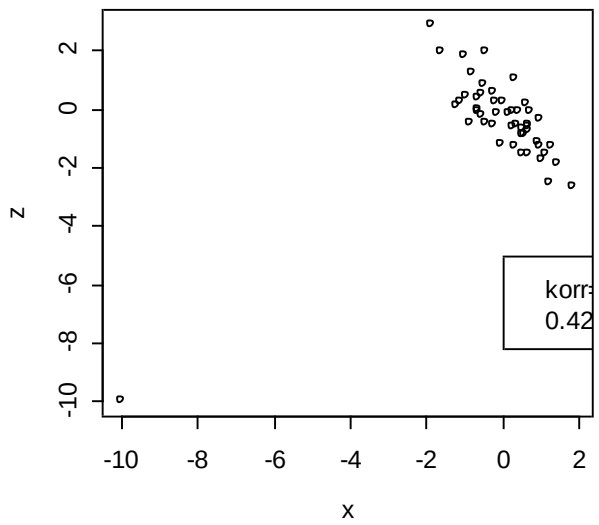

R2=0.03
cipõméret
magasság
32
34
36
38
40
42
44
46
165
170
175
180
185
190

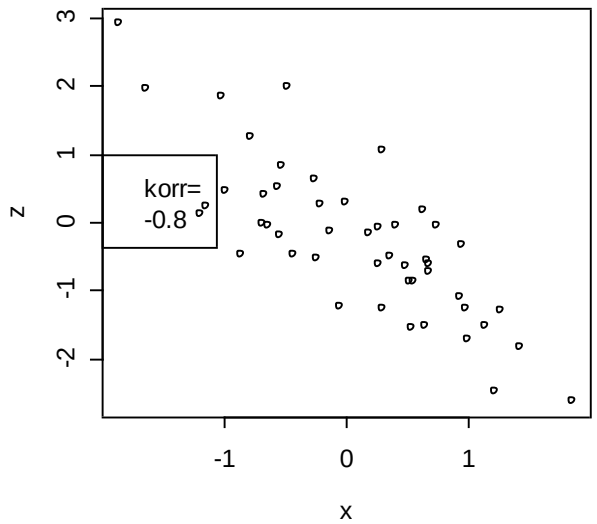

R2=0.33
cipõméret
magasság
35
40
45
50
55
60
140
150
160
170
180
190

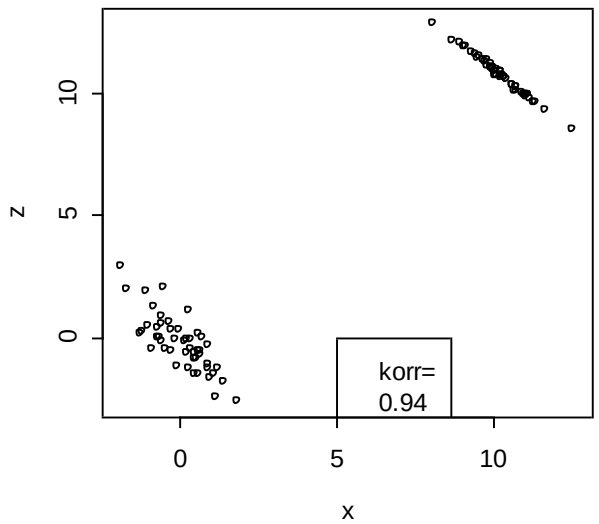

R2=0.80
cipõméret
magasság
30
40
50
60
70
80
90
100
120
140
160
180


R2=0.87
cipõméret
magasság
20
40
60
80
100
120
140
160
180
200
220
240









A tapasztalati korreláció vizsgálata

Igen érzékeny a kiugró értékekre