Valószínűségszámítás és Statisztika

10. előadás

2020. április 28.

χ-négyzet próba homogenitásvizsgálatra

- Homogenitásvizsgálat:

$H_0 : \xi_1, ..., \xi_n$ és $\eta_1, ..., \eta_m$ ugyanolyan eloszlásúak

- Hasonlóan járunk el, mint korábban

$$\bigcup_{i=1}^{r} C_i = \mathbf{R}$$

$$\nu_i = \left| \left\{ j : \xi_j \in C_i \right\} \right|, \mu_i = \left| \left\{ j : \eta_j \in C_i \right\} \right|, \; i = 1, 2, ..., r,$$

A tesztstatisztika:

$$\chi^2 = nm \sum_{i=1}^{r} \frac{\left(\nu_i / n - \mu_i / m \right)^2}{\nu_i + \mu_i} \xrightarrow[n,m \to \infty]{} \chi^2_{r-1}$$

Ki tanul jobban?

2009. január 5-ei vizsga

Jegy	Férfi	Nő	Összesen
1	47	4	51
2	11	1	12
3	11	2	13
4	9	2	11
5	8	2	10
Összesen	86	11	97
Átlag	2,1	2,7	2,1

$$C_1 = \{1;2\}, C_1 = \{3;4;5\}$$

$$\nu_i = \left|\left\{ j : \xi_j \in C_i \right\}\right|, \mu_i = \left|\left\{ j : \eta_j \in C_i \right\}\right|, \; i = 1,2,$$

$$\nu_1 = 58, \nu_2 = 28, \mu_1 = 5, \mu_2 = 6, n = 86, m = 11$$

A tesztstatisztika:

$$\chi^2 = 86 \cdot 11 \left(\frac{\left(\frac{58}{86} - \frac{5}{11}\right)^2}{58+5} + \frac{\left(\frac{28}{86} - \frac{6}{11}\right)^2}{28+6} \right) = 2.071$$

$$P(\chi_1^2 > 2.71) = 10\% \Rightarrow$$

Nem tudjuk elutasítani az egyforma képesség hipotézisét!

χ-négyzet próba függetlenségvizsgálatra

- H_0 hipotézis: az A_1, A_2 , ..., A_r és B_1, B_2 , ..., B_s teljes eseményrendszerekre teljesül a függetlenség.

$$\sum_{i,j} \frac{\left(\nu_{ij} - np_i q_j\right)^2}{np_i q_j}$$

- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az rs-1 szabadságfokú χ-négyzet eloszlás 1- α kvantilise, elutasítjuk a nullhipotézist.

Becsléses eset

- Általában, ha az illesztendő eloszlást nem ismerjük – csak a családját - becsüljük a paramétereit. Ekkor a próbastatisztika szabadságfoka annyival csökken, ahány paramétert becsültünk.
- Függetlenségvizsgálatnál általában nem ismerjük a teljes eseményrendszer tagjainak valószínűségét, így r-1+s-1 valószínűséget kell becsülnünk. A szabadságfok ekkor tehát

  rs-1-r-s+2=(r-1)(s-1).

ν_{ij}: $A_i B_j$ gyakorisága

$\nu_{i\bullet}$: A_i gyakorisága

$\nu_{\bullet j}$: B_j gyakorisága

A tesztstatisztika

$$n \sum_{i,j} \frac{\left(\nu_{ij} - \frac{\nu_{i\bullet}\nu_{\bullet j}}{n}\right)^2}{\nu_{i\bullet}\nu_{\bullet j}} \xrightarrow[n\to\infty]{} \chi^2_{(r-1)(s-1)}$$

$r = s = 2$esetben

$$n \frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1\bullet}\nu_{2\bullet}\nu_{\bullet 1}\nu_{\bullet 2}} \xrightarrow[n\to\infty]{} \chi^2_1$$

Szívbetegek diétája

- http://onlinestatbook.com/case_studies/diet.html
- The subjects, 605 survivors of a heart attack, were randomly assigned follow either (1) a diet close to the "prudent diet step 1" of the American Heart Association (control group) or (2) a Mediterranean-type diet consisting of more bread and cereals, more fresh fruit and vegetables, more grains, more fish, fewer delicatessen foods, less meat. An experimental canola-oil-based margarine was used instead of butter or cream. The oils recommended for salad and food preparation were canola and olive oils exclusively. Moderate red wine consumption was allowed.
- Over a four-year period, patients in the experimental condition were initially seen by the dietician, two months later, and then once a year. Compliance with the dietary intervention was checked by a dietary survey and analyses of plasma fatty acids. Patients in the control group were expected to follow the dietary advice given by their physician.

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

	Cancers	Deaths	Nonfatal illness	Healthy	Total
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)	303
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)	302
Total	22	38	33	512	605

- $\chi^2 = n \sum_{i,j} \frac{\left(\nu_{i,j} - \frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n}\right)^2}{\nu_{i,\cdot}\nu_{\cdot,j}} = 16{,}55$
- $P(\chi_3^2 > 16{,}55\ ) = 0{,}0009 \Longrightarrow$

elutasítjuk a hipotézist

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.
- Egyébként közelítés: például Nadarajah módszerével (hasonló a Parzen-Rosenblatt becsléshez).

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:

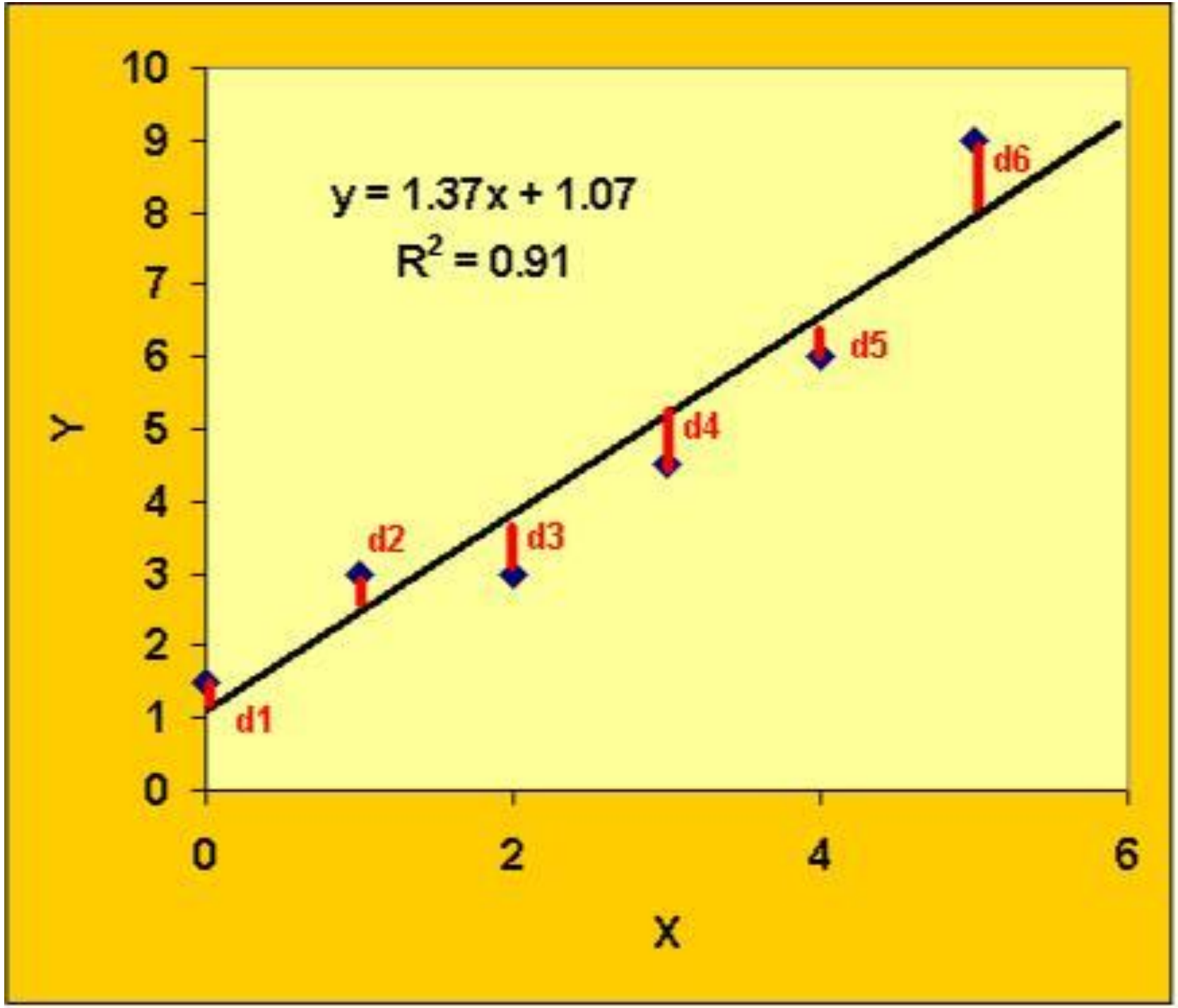
$$E(X \mid F=0) = \frac{\sum_{i=0}^{2} iP(X=i, F=0)}{P(F=0)} = \frac{\sum_{i=0}^{2} iP(F=0 \mid X=i)P(X=i)}{\sum_{i=0}^{2} P(F=0 \mid X=i)P(X=i)}$$

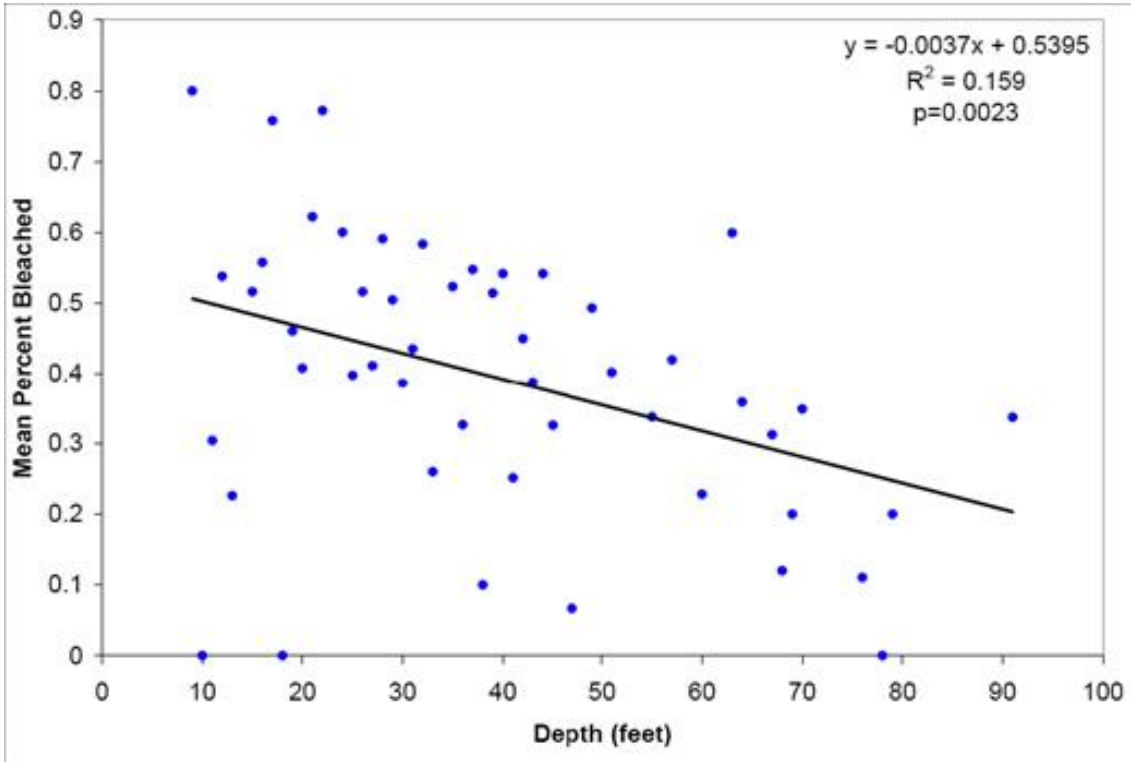
- Az eredmények: E(X|F=2)=2, E(X|F=1)=4/3, E(X|F=0)=2/3.

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,…,*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.


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 a becsülendő együtthatók

- $\Sigma(Yi - (aX_i + b))^2 \rightarrow \min$

- 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: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{a}X_i - \hat{b})^2}{n-2}$

Normális eloszlású eset

- $Y_i = aX_i + b + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$
- ε_i függetlenek
- Likelihood függvény:

$$L(\mathrm{y}, a, b, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - aX_i - b)^2}{2\sigma^2}\right)$$

$$= (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - aX_i - b)^2}{2\sigma^2}\right)$$

Hipotézisvizsgálat/1

H_0: a =0 tesztelése t-próbával:

$$t_{n-2} = \frac{\hat{a}\sqrt{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{a}X_i - \hat{b})^2}}$$

- Ebből konfidencia intervallum is kapható a-ra

Hipotézisvizsgálat/2

- $H_0: b = 0$ tesztelése t-próbával:

$$t_{n-2} = \frac{\hat{b}\sqrt{n(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^{n}(Y_i - \hat{a}X_i - b)^2}\sqrt{\sum_{i=1}^{n} {X_i}^2}}$$

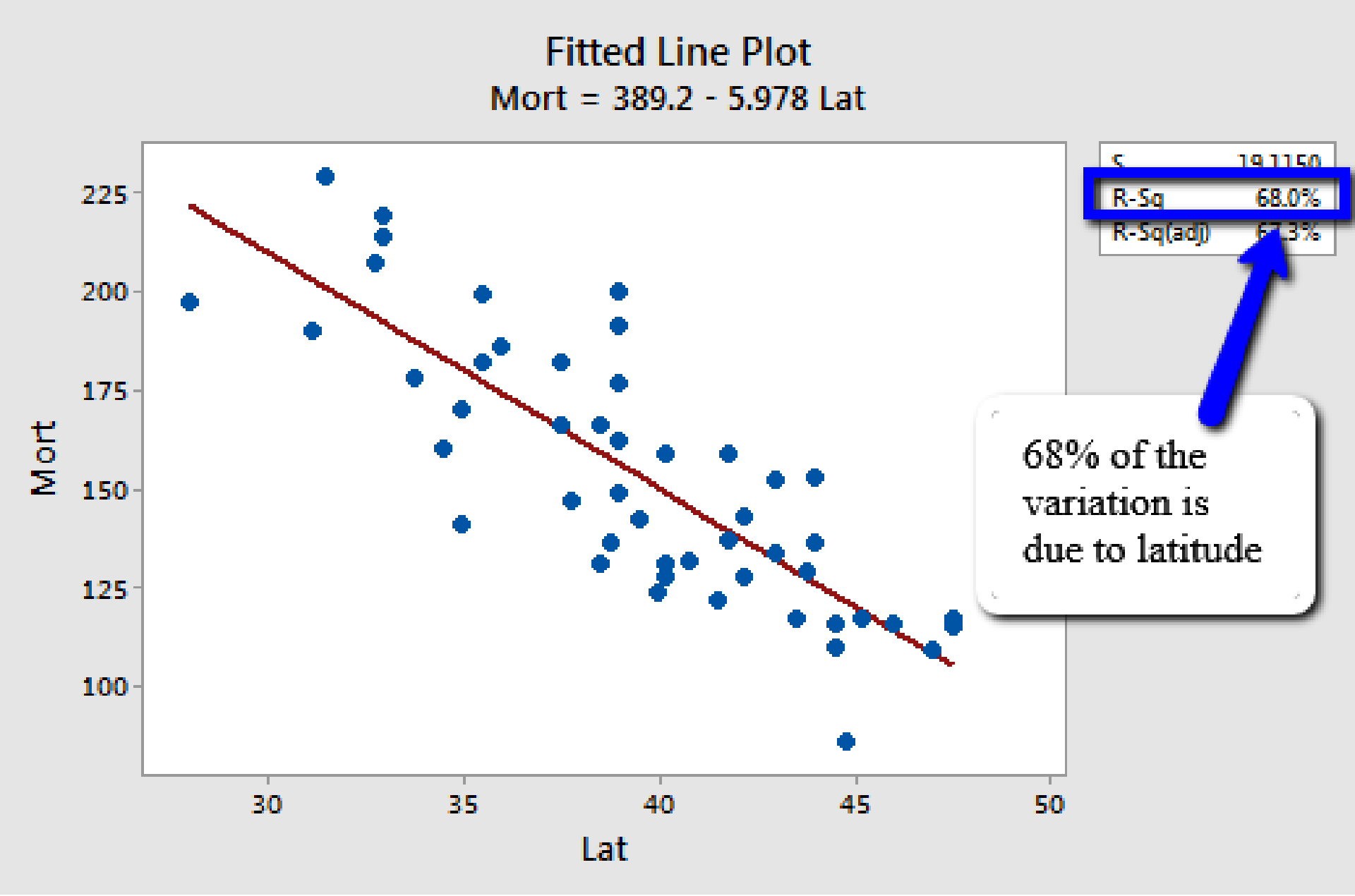
- Ebből konfidencia intervallum is kapható b-re

Szóródások

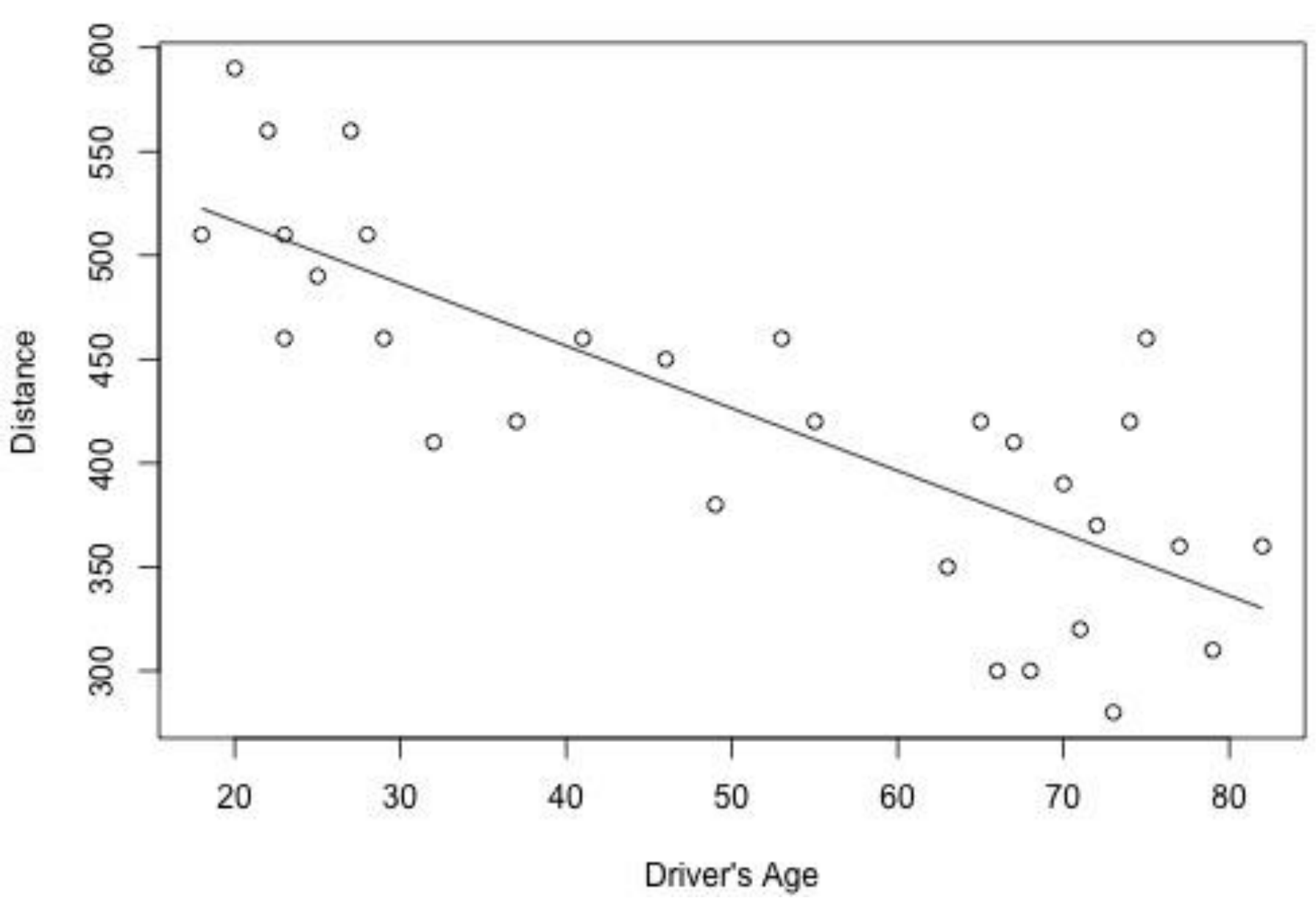
- Teljes ingadozás: $\sum_{i=1}^{n}(y_i-\bar{y})^2$
- Reziduális négyzetösszeg: $\sum_{i=1}^{n}(y_i-\hat{a}x_i-\hat{b})^2=\sum_{i=1}^{n}(y_i-\bar{y})^2-\frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- A megmagyarázott variabilitás részaránya:

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

éppen a tapasztalati korrelációs együttható négyzete

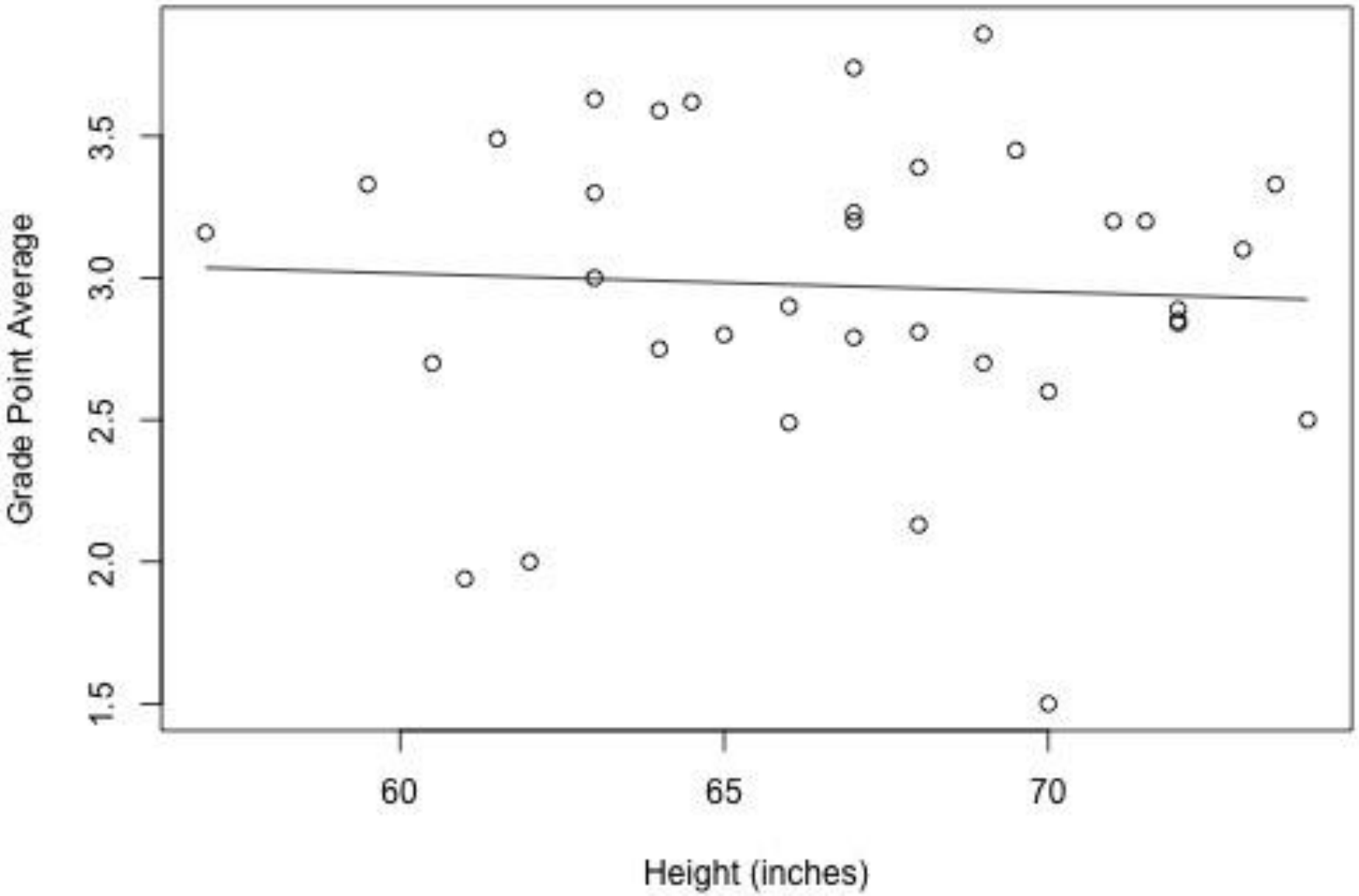


https://onlinecourses.science.psu.edu/stat501/node/255



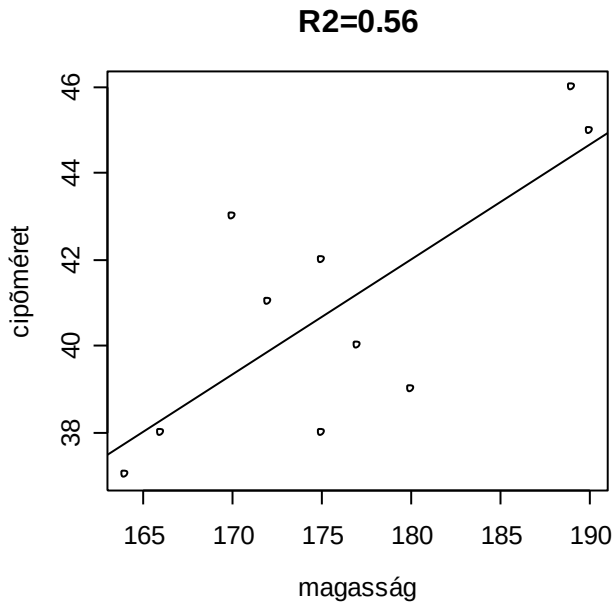
r^2=64,2% , r=-0,801

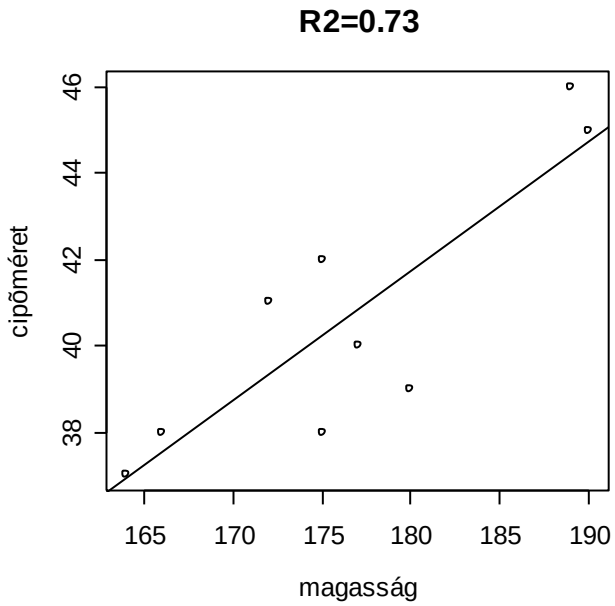
https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/01simple/signdist.jpeg



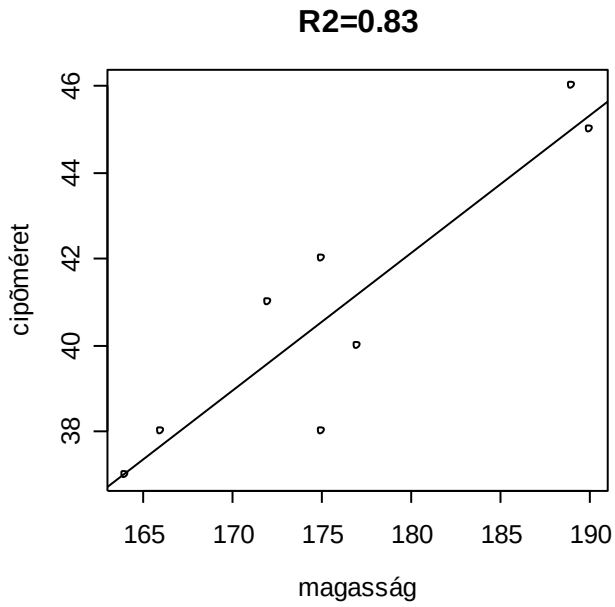
r^2=0,3% , r=-0,053

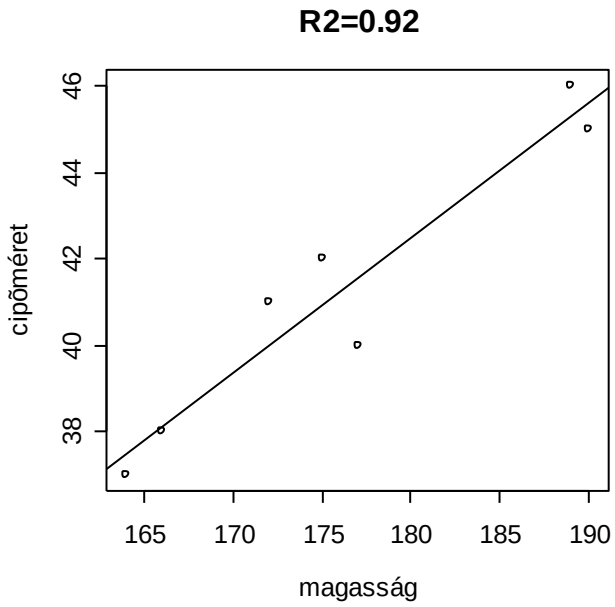
https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/01simple/heightgpa.jpeg

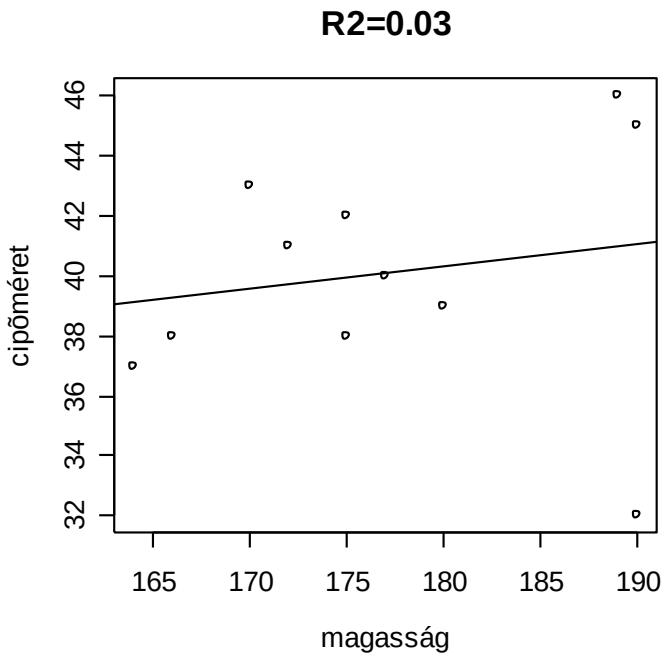


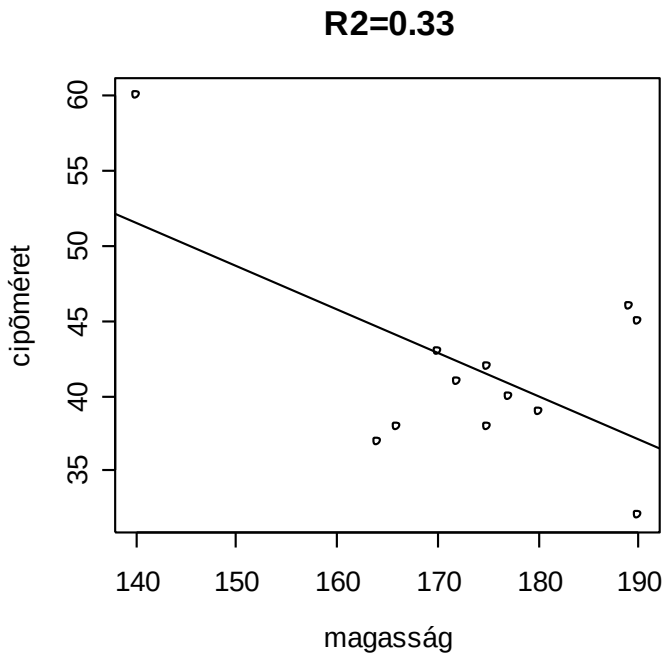


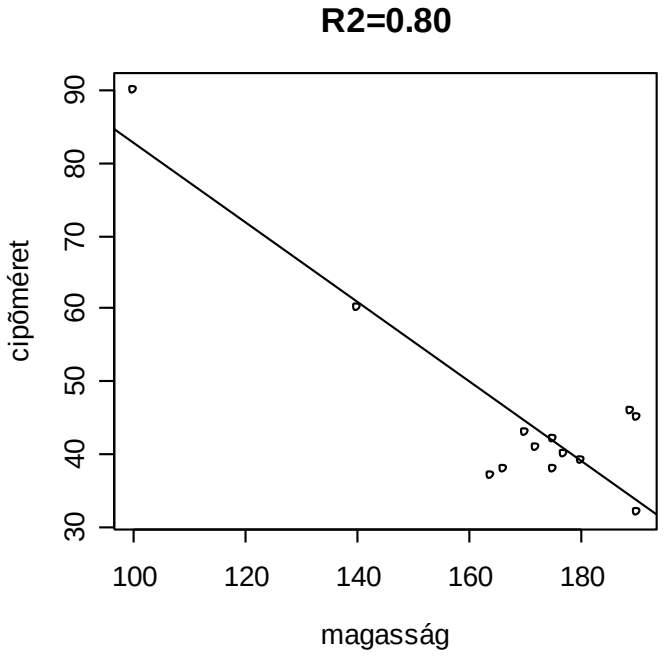
A regresszió vizsgálata

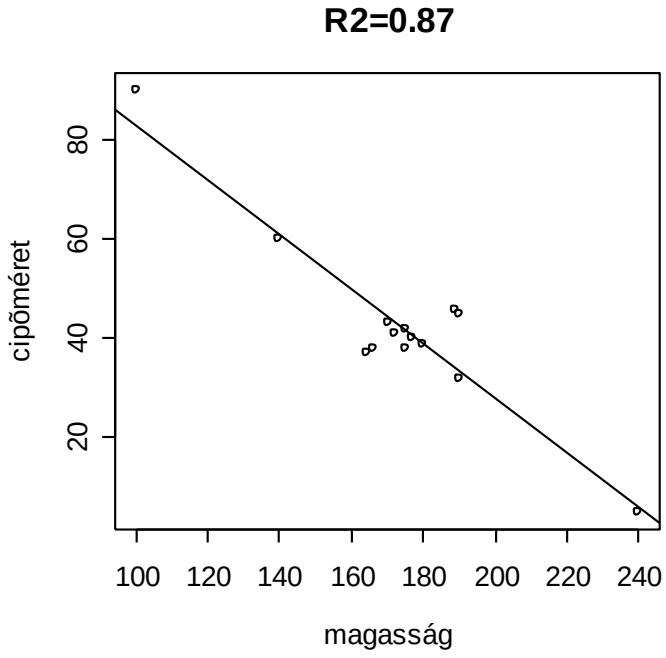


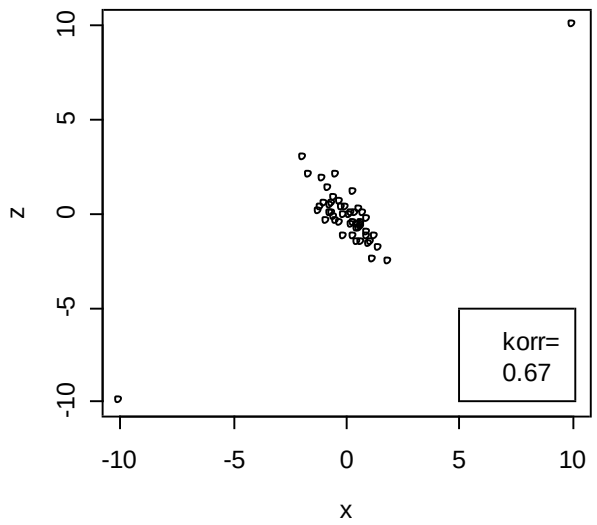


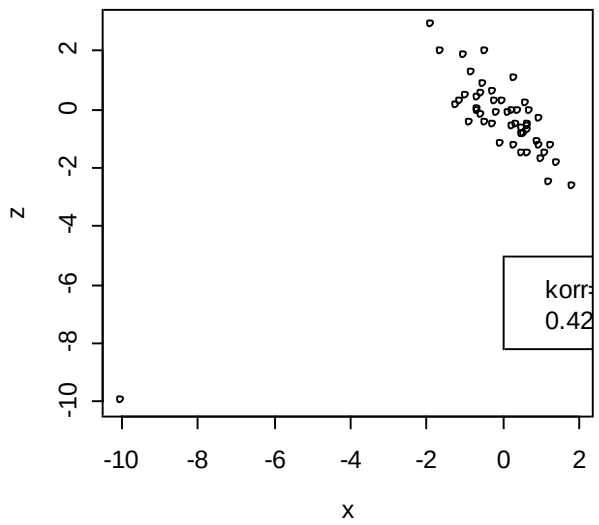

R2=0.03
cipõméret
magasság
32
34
36
38
40
42
44
46
165
170
175
180
185
190

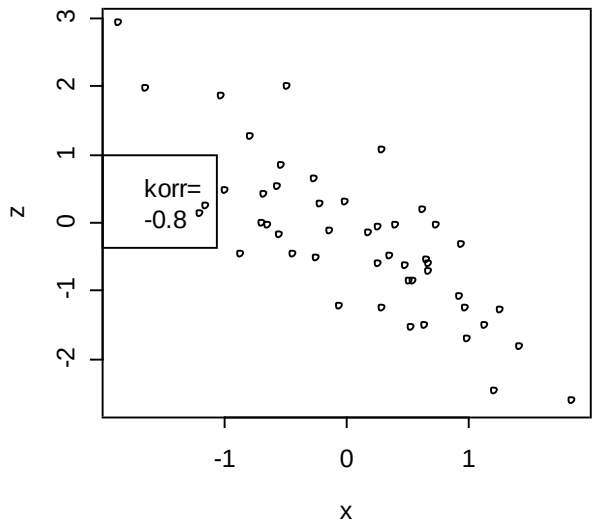

R2=0.33
cipõméret
magasság
35
40
45
50
55
60
140
150
160
170
180
190

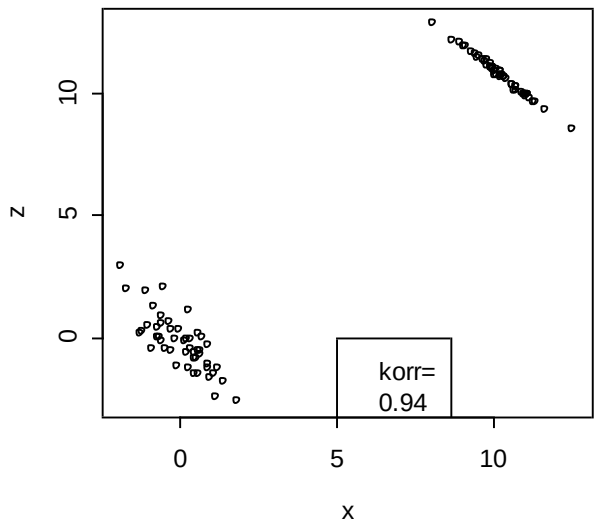

R2=0.80
cipõméret
magasság
30
40
50
60
70
80
90
100
120
140
160
180


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

Egyszerű lineáris modell általánosítása

- $Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \ldots + a_{(k-1)} X_{i,(k-1)} + b + \varepsilon_i$, $i = 1,\ldots,n$
- ($X_{i,j}$-k a magyarázó változók, ε_i hibák. $E(\varepsilon_i)=0$, általában itt is feltesszük, hogy normális eloszlásúak, függetlenek)
- a_j-k, b a becsülendő együtthatók.

Lineáris modell

Legyen

$$\mathbf{X} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,(k-1)} & 1 \\ \vdots & \cdots & \vdots & \vdots \\ X_{n,1} & \cdots & X_{n,(k-1)} & 1 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \\ b \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{k-1} \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Ekkor az előző egyenletrendszer átírható mátrixos alakba:

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Feladat: $\boldsymbol{\beta}$ becslése.

Legkisebb négyzetes becslés

$$\hat{\boldsymbol{\beta}}: \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{k}\hat{\beta}_j X_{i,j}\right)^2 = \min_{\mathbf{b}} \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{k} b_j X_{i,j}\right)^2$$

Tétel: Amennyiben $\mathbf{b}_0$ az $\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\mathbf{b}_0$ egyenlet megoldása, úgy legkisebb négyzetes becslés.

$rang(\mathbf{X}) = k$ esetén $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$.

Bizonyítás: $\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{k} b_j X_{i,j}\right)^2 = \left(\mathbf{Y}-\mathbf{X}\mathbf{b}\right)^T\left(\mathbf{Y}-\mathbf{X}\mathbf{b}\right) =$

$$\left(\mathbf{Y}-\mathbf{X}\mathbf{b}_0+\mathbf{X}\mathbf{b}_0-\mathbf{X}\mathbf{b}\right)^T\left(\mathbf{Y}-\mathbf{X}\mathbf{b}_0+\mathbf{X}\mathbf{b}_0-\mathbf{X}\mathbf{b}\right) =$$

$$\left(\mathbf{Y}-\mathbf{X}\mathbf{b}_0\right)^T\left(\mathbf{Y}-\mathbf{X}\mathbf{b}_0\right)+\left(\mathbf{b}_0-\mathbf{b}\right)^T\mathbf{X}^T\mathbf{X}\left(\mathbf{b}_0-\mathbf{b}\right)+$$

$$2\left(\mathbf{b}_0-\mathbf{b}\right)^T\left(\mathbf{X}^T\mathbf{Y}-\mathbf{X}^T\mathbf{X}\mathbf{b}_0\right) \geq \left(\mathbf{Y}-\mathbf{X}\mathbf{b}_0\right)^T\left(\mathbf{Y}-\mathbf{X}\mathbf{b}_0\right)$$

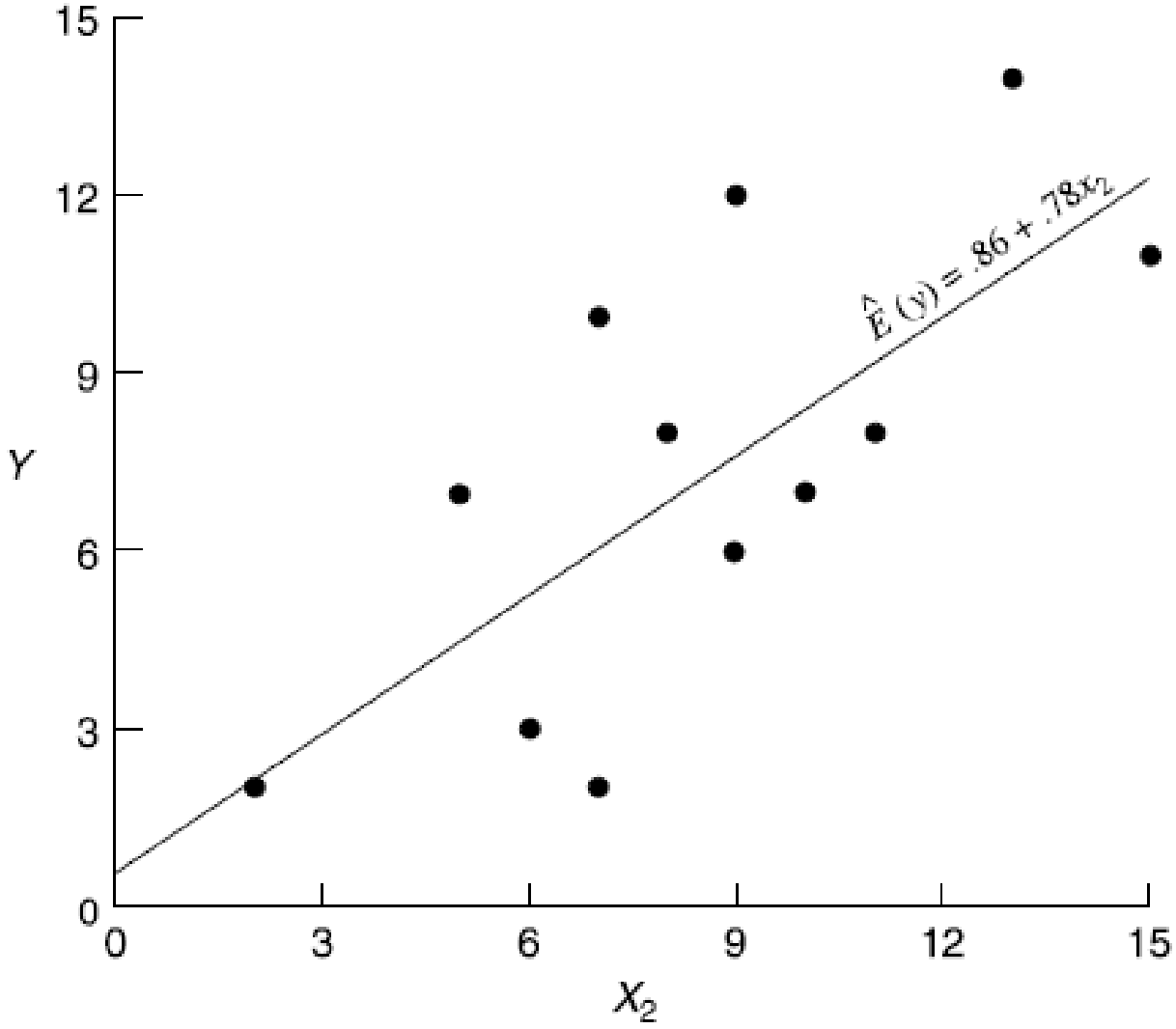
Observation Number	y	x_1	x_2
1	2	0	2
2	3	2	6
3	2	2	7
4	7	2	5
5	6	4	9
6	8	4	8
7	10	4	7
8	7	6	10
9	8	6	11
10	12	6	9
11	11	8	15
12	14	8	13

$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \\ 1 & 4 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 10 \\ 1 & 6 & 11 \\ 1 & 6 & 9 \\ 1 & 8 & 15 \\ 1 & 8 & 13 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} .97476 & .24290 & -.22871 \\ .24290 & .16207 & -.11120 \\ -.22871 & -.11120 & .08360 \end{pmatrix},$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 5.3754 \\ 3.0118 \\ -1.2855 \end{pmatrix}.$$

$$\hat{y} = 1.86 + 1.30x_1,$$
$$\hat{y} = .86 + .78x_2,$$
$$\hat{y} = 5.37 + 3.01x_1 - 1.29x_2.$$



Kisdolgozat gyakorlása

1. X_1, X_2, … független, N(3, 2^2) eloszlású valószínűségi változók.

   a) Mennyi $X_1 + 2X_2 + 1$ várható értéke?

A: 4	B: 10	C: 6	D: 8

   b) Milyen eloszlású $(X_1+X_2+ \ldots +X_{10})/10$?

A: N(3, 2^2)	B: N(3, 0,4)	C: a 3 konstans	D: N(0,3, 0,4)

   c) Mennyi $X_1 - X_2$ szórásnégyzete?

A: 0	B: 8	C: 16	D: egyéb

   d) Mennyi X_1 és $-X_2$ (mínusz X_2) korrelációja?

A: 20	B: 1	C: 0	D: -1

2. Egy kísérletsorozatnál megfigyeléseink a következők:

3, 2, 2, 1.

(a válaszoknál 2 tizedesjegyre kerekítettünk)

a) Mennyi a minta korrigált tapasztalati szórásnégyzete?

A: 0,67	**B: 6**	**C: 4,5**	**D: 0,5**

b) Feltételezzük, hogy megfigyeléseink azonos eloszlású Poisson eloszlásúak. Mennyi a Poisson paraméter momentum módszer szerinti becslése?

A: 1	**B: 0,25**	**C: 0**	**D: 2**