Valószínűségszámítás és Statisztika

12. előadás

2023. május 25.

χ-négyzet próba

- H_0 hipotézis: az $A_1, A_2, \ldots, A_r$ teljes eseményrendszerre teljesül $P(A_1) = p_1, P(A_2) = p_2, \ldots, P(A_r) = p_r$
- A tesztstatisztika:

$$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}$$

ami aszimptotikusan $r - 1$ szabadságfokú χ-négyzet eloszlású, ha igaz a nullhipotézis.

- Kritikus tartomány: ha a statisztika értéke nagyobb, mint az $r - 1$ szabadságfokú χ-négyzet eloszlás $1 - \alpha$ kvantilise, elutasítjuk a nullhipotézist.

χ-négyzet próba illeszkedésvizsgálatra

- Illeszkedésvizsgálat:

$H_0 : \xi_1, ..., \xi_n \ \ F$ eloszlásfüggvényűek

- Visszavezetjük az előző esetre

$$A_i = \{\xi \in C_i\}, i = 1, 2, ..., r, \bigcup_i C_i = \mathbf{R}$$

Diszkrét esetben gyakran: $A_i = \{\xi = x_i\}, i = 1, 2, ..., r$

Példa

- Mi lehet egy vezető által okozott károk számának eloszlása?
- Poisson eloszlású-e?

Kár-szám	0	1	2	3	4	5	6	7	>7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006

Becsléses χ-négyzet próba

- H_0 hipotézis: az A_1, A_2 , ..., A_r teljes eseményrendszerre teljesül:

$$P(A_i) = p_i(\vartheta_1, ..., \vartheta_s), i = 1, 2, ..., r$$

$\vartheta_1, ..., \vartheta_s$ ismeretlen paraméterek.

A tesztstatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} \xrightarrow[n\to\infty]{} \chi^2_{r-s-1},$$

ahol

$$\hat{p}_i = p_i(\hat{\vartheta}_1, ..., \hat{\vartheta}_s).$$

Példa (folyt.)

Kár-szám	0	1	2	3	4	5	6	7	>7	Össze-sen
Veze-tők száma	129524	16267	1966	211	31	5	1	1	0	148006
np_i *Poisson*	128 433	18 218	1 292	61	2,2	0,06	0,001	3E-05	5E-07	
Np_i *Neg. bin.*	129 541	16 237	1 962	234	28	3,3	0,39	0,05	0,006	

$n = 148006,\ r = 5$

$A_i = \{\xi = i\}, i = 0,1,2,3$

$A_4 = \{\xi \geq 4\}$

Poisson eset:

$\hat{\lambda}$=0.709

$$\sum_{i=0}^{4} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} \sim \chi^2_{5-1-1}$$

$$\sum_{i=0}^{4} \frac{\left(\nu_i - n\hat{p}_i\right)^2}{n\hat{p}_i} > 200$$

$$P\left(\chi_3^2 > 17.7\right) = 0.05\% \Rightarrow$$

Elutasítjuk Poisson eloszlás hipotézisét!

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:

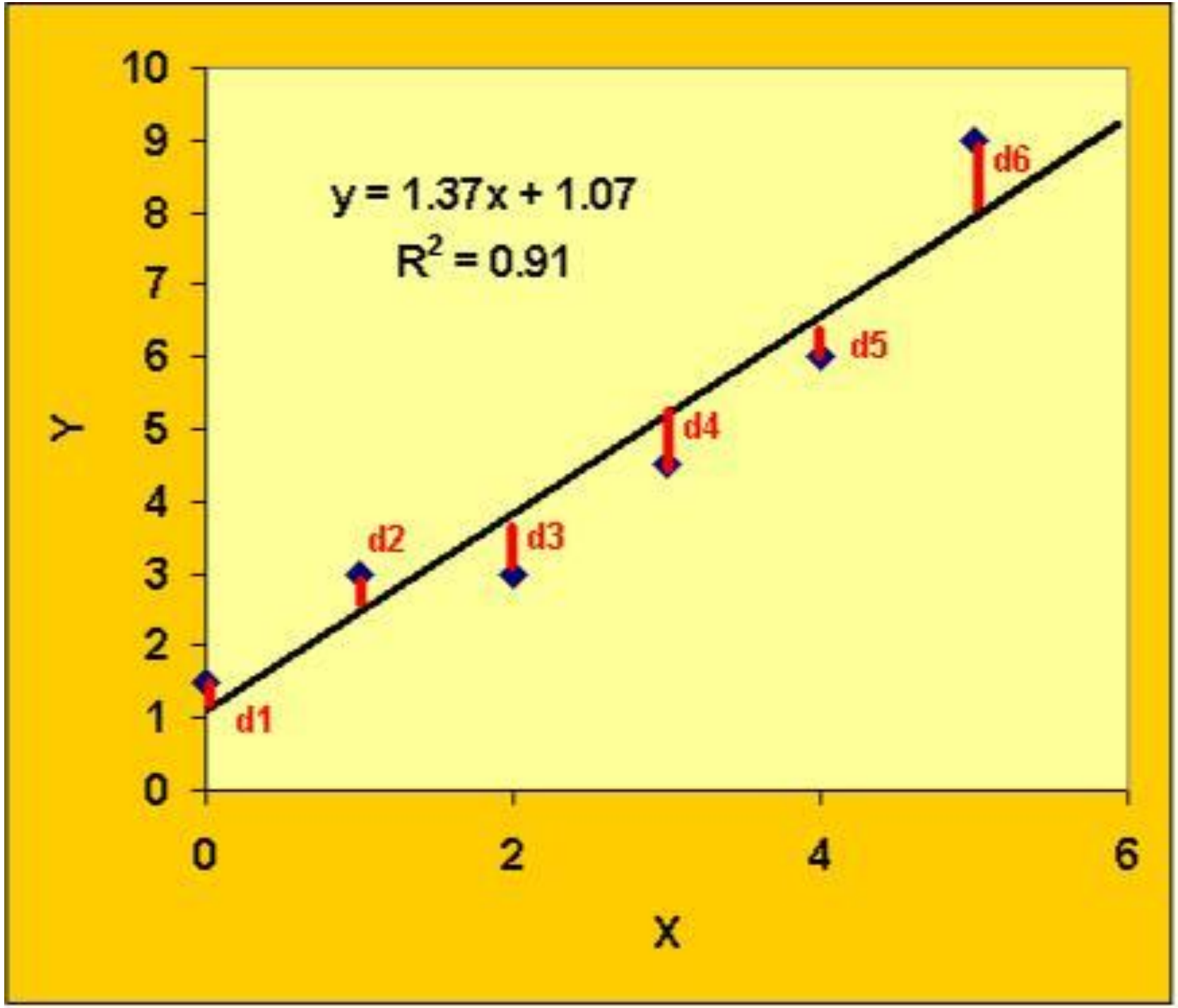
$$E(X \mid F=0) = \frac{\sum_{i=0}^{2} iP(X=i, F=0)}{P(F=0)} = \frac{\sum_{i=0}^{2} iP(F=0 \mid X=i)P(X=i)}{\sum_{i=0}^{2} P(F=0 \mid X=i)P(X=i)}$$

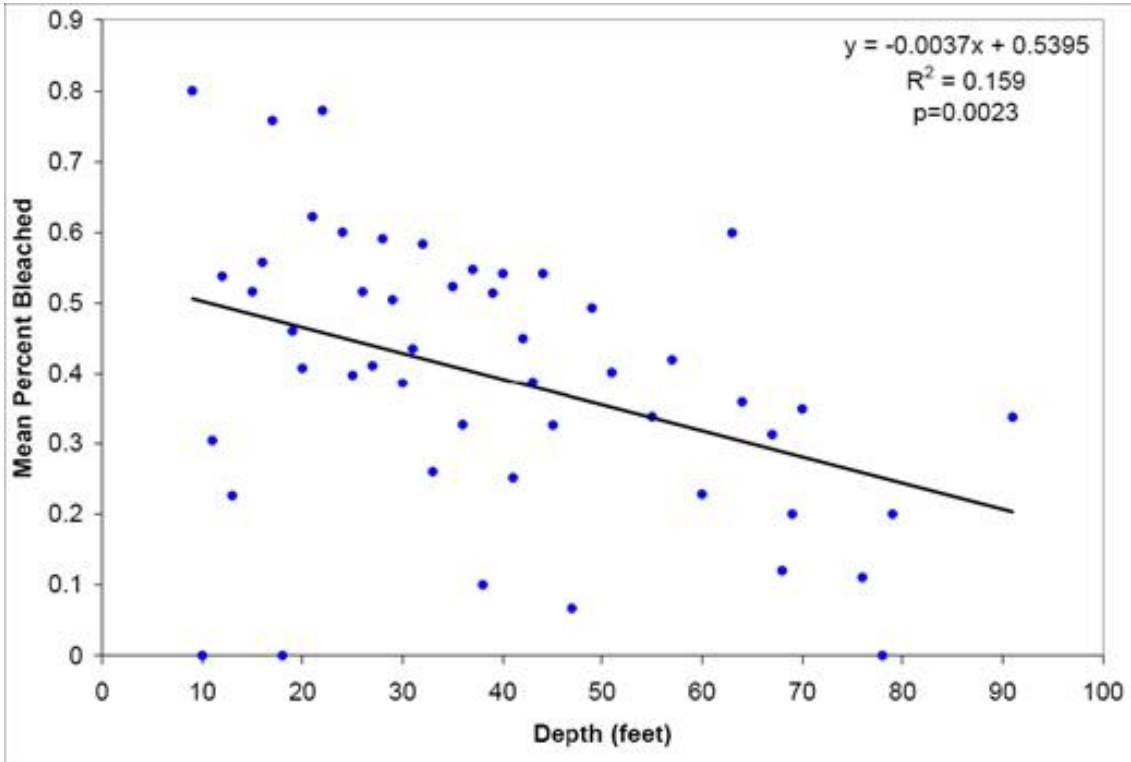
- Az eredmények: E(X|F=2)=2, E(X|F=1)=4/3, E(X|F=0)=2/3.

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:

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

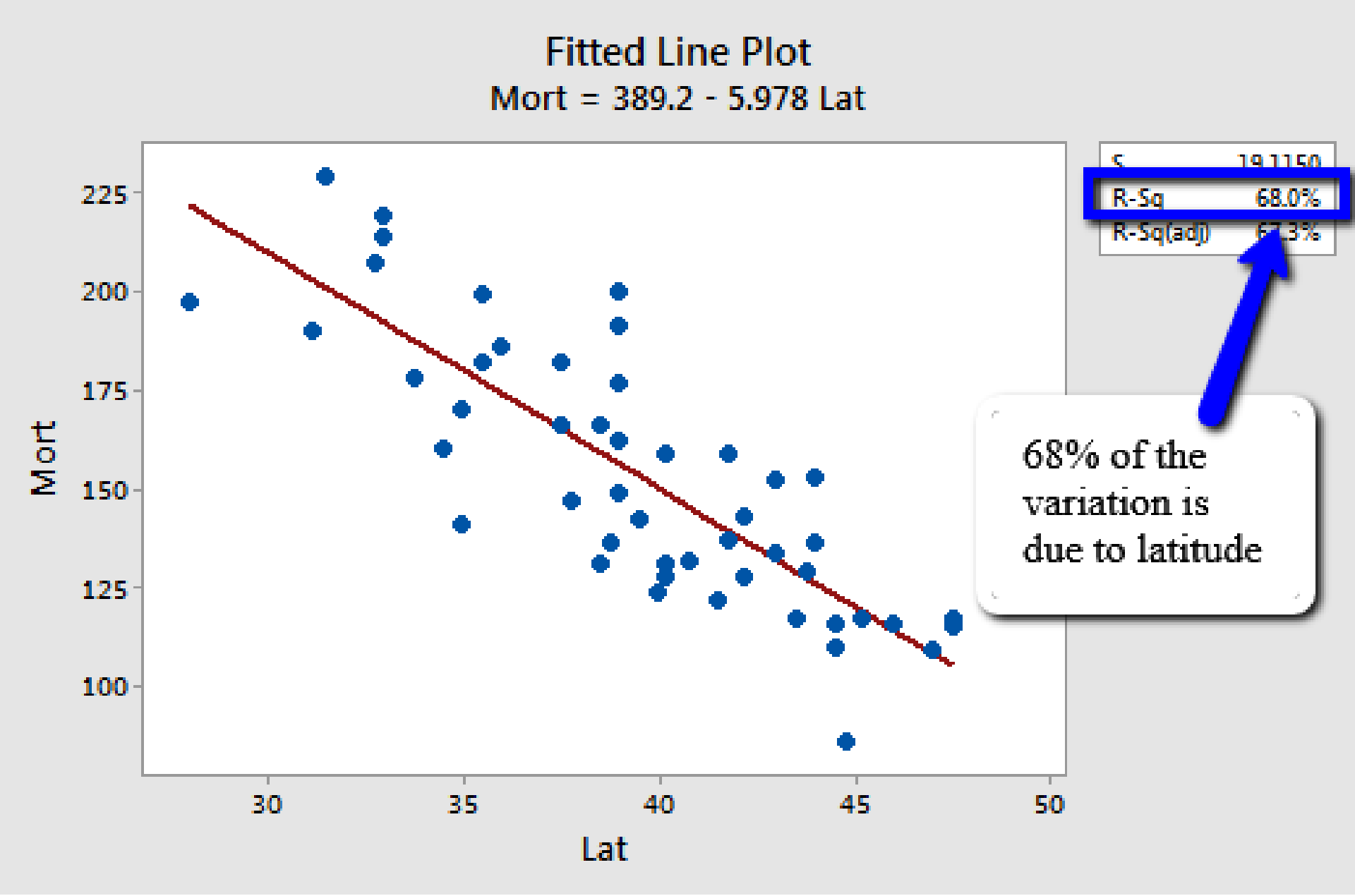
- Ebből konfidencia intervallum is kapható b-re

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