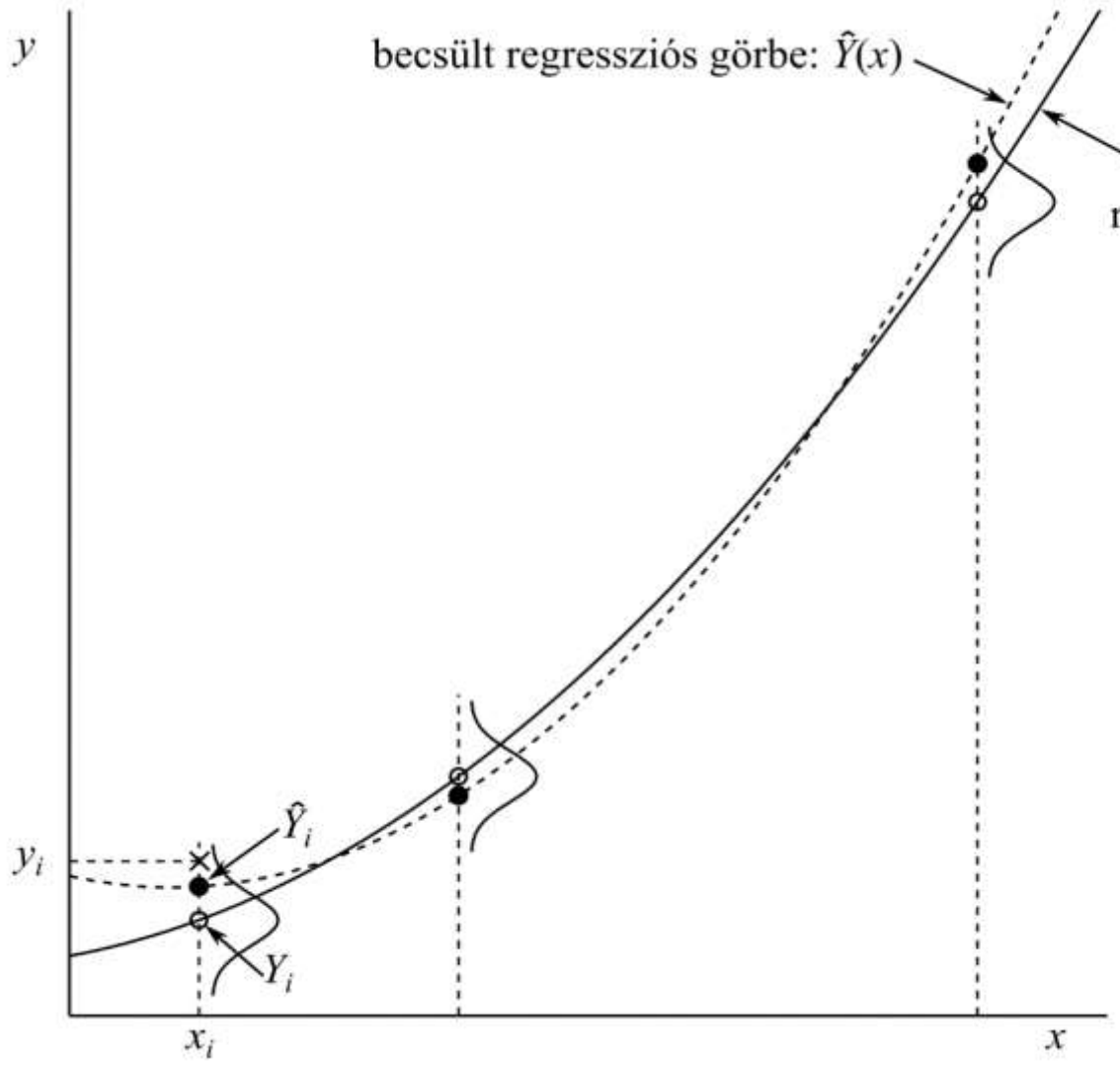
REGRESSZIÓANALÍZIS

Lineáris regresszió

Jelölések



y_i mért érték

$\hat{Y}_i$ becsült érték

Y_i elméleti érték

A regresszióanalízis feladatai

- a függvénykapcsolat ($Y(x)$ elméleti regressziós függvény) paramétereinek becslése,
- a függvény alkalmasságának vizsgálata,
- a paraméterekre vonatkozó hipotézisek vizsgálata (pl. átmegy-e az elméleti regressziós egyenes az origón, ill. meredeksége szignifikánsan különbözik-e zérustól),
- konfidencia-intervallum ill. konfidencia-sáv számítása a függvény paramétereire és az $Y(x)$ tapasztalati vagy empirikus regressziós görbére (a becsült függvényre).

Modell

Valamely (pl. fizikai) törvényszerűség értelmében az x független változó bizonyos értékénél a függő változó értéke $Y = \varphi(x)$.

A mérési pontatlanságok vagy a függvénykapcsolatban nem szereplő, de a jelenséget befolyásoló tényezők miatt Y helyett y értéket mérünk, melyre ha az ingadozások véletlenszerűek és kis hatásúak:

$$E(y \mid x) = Y, \quad \text{vagy}$$

$$y = Y + \varepsilon \qquad \text{és} \qquad E(\varepsilon) = 0 \qquad Var(\varepsilon) = \sigma^2$$

A regresszióanalízis feltételei

- $Y(x) = f(x;\alpha,\beta,\gamma,...)$ az ismert vagy feltételezett függvénykapcsolat alakja, ahol α, β, γ a függvény konstansai (paraméterei)
- y az x minden értékénél normális eloszlású, vagyis az ε_i mérési hibák $N(0,\sigma^2)$ normális eloszlásúak
- Var(y) = konstans, illetve y-nak vagy x-nek ismert függvénye;
- a különböző i mérési pontokban elkövetett mérési hibák egymástól függetlenek
- x (független változó) hibamentes

Egyváltozós lineáris regresszió ismétlés nélküli mérések esetén, $\sigma_{y_i}^2$ konstans

A függvénykapcsolat ($Y(x)$ elméleti regressziós függvény) alakja

Elméleti függvény: $Y_i = \alpha + \beta\left(x_i - \bar{x}\right)$ $\qquad \alpha' = \alpha - \beta\bar{x}$

Becsült függvény: $\hat{Y}_i = \hat{\alpha} + \hat{\beta}\left(x_i - \bar{x}\right)$ $\qquad \hat{\alpha}' = \hat{\alpha} - \hat{\beta}\bar{x}$

$\hat{Y}_i = a + b\left(x_i - \bar{x}\right)$ $\qquad a' = a - b\bar{x}$

Egyenes meredeksége:	β
Egyenes tengelymetszete:	α’

A függvénykapcsolat ($Y(x)$ elméleti regressziós függvény) paramétereinek becslése

A becslési kritérium a legkisebb négyzetek módszere elv alapján:

$$\phi = \sum_i \left(y_i - \hat{Y}_i \right)^2 = \min.$$

$$\phi = \sum \left[y_i - \hat{\alpha} - \hat{\beta}\left(x_i - \bar{x} \right) \right]^2 = \min.$$

A normálegyenletek:

$$\frac{\partial \phi}{\partial \hat{\alpha}} = -2\sum\left[y_i - \hat{\alpha} - \hat{\beta}\left(x_i - \bar{x}\right)\right] = 0$$

$$\frac{\partial \phi}{\partial \hat{\beta}} = -2\sum\left[y_i - \hat{\alpha} - \hat{\beta}\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) = 0$$

Átrendezve:

$$\sum y_i = n\hat{\alpha} + \hat{\beta}\sum\left(x_i - \bar{x}\right) \longrightarrow \hat{\alpha} \equiv a = \frac{\sum y_i}{n} = \bar{y}$$

$$\sum y_i\left(x_i - \bar{x}\right) = \hat{\alpha}\sum\left(x_i - \bar{x}\right) + \hat{\beta}\sum\left(x_i - \bar{x}\right)^2 \longrightarrow \hat{\beta} \equiv b = \frac{\sum y_i\left(x_i - \bar{x}\right)}{\sum\left(x_i - \bar{x}\right)^2}$$

Az $\hat{\alpha}$ és $\hat{\beta}$ becslések egymástól függetlenek!

$$\hat{\alpha}' \equiv a' = \hat{\alpha} - \hat{\beta}\bar{x}$$

1. példa

Kísérletileg vizsgálták a koncentráció (*x*, független változó) és az abszorbancia (*y*, függő változó) közötti összefüggést.
x értéke pontosan beállítható,
y értéke az *Y* valódi érték körül ingadozik.

sorrend	x	y
9	0,102	4,141
1	0,200	6,715
5	0,311	9,202
2	0,398	11,526
7	0,508	14,775
4	0,603	17,642
8	0,715	20,554
3	0,804	23,363
6	0,920	26,000

1. példára: az egyenes paramétereinek becslése

sorrend	x	y	xi-xátlag	(xi-xátlag)^2	yi(xi-xátlag)
9	0,102	4,141	-0,405	0,164	-1,676
1	0,200	6,715	-0,307	0,094	-2,060
5	0,311	9,202	-0,196	0,038	-1,802
2	0,398	11,526	-0,109	0,012	-1,254
7	0,508	14,775	0,001	0,000	0,018
4	0,603	17,642	0,096	0,009	1,698
8	0,715	20,554	0,208	0,043	4,280
3	0,804	23,363	0,297	0,088	6,944
6	0,920	26,000	0,413	0,171	10,744
szumma	4,561	133,918		0,620	16,892
átlag	0,5068	14,880			

$$\hat{\alpha} \equiv a = \frac{\sum y_i}{n} = \bar{y}$$

$$\hat{\beta} \equiv b = \frac{\sum y_i \left(x_i - \bar{x}\right)}{\sum \left(x_i - \bar{x}\right)^2}$$

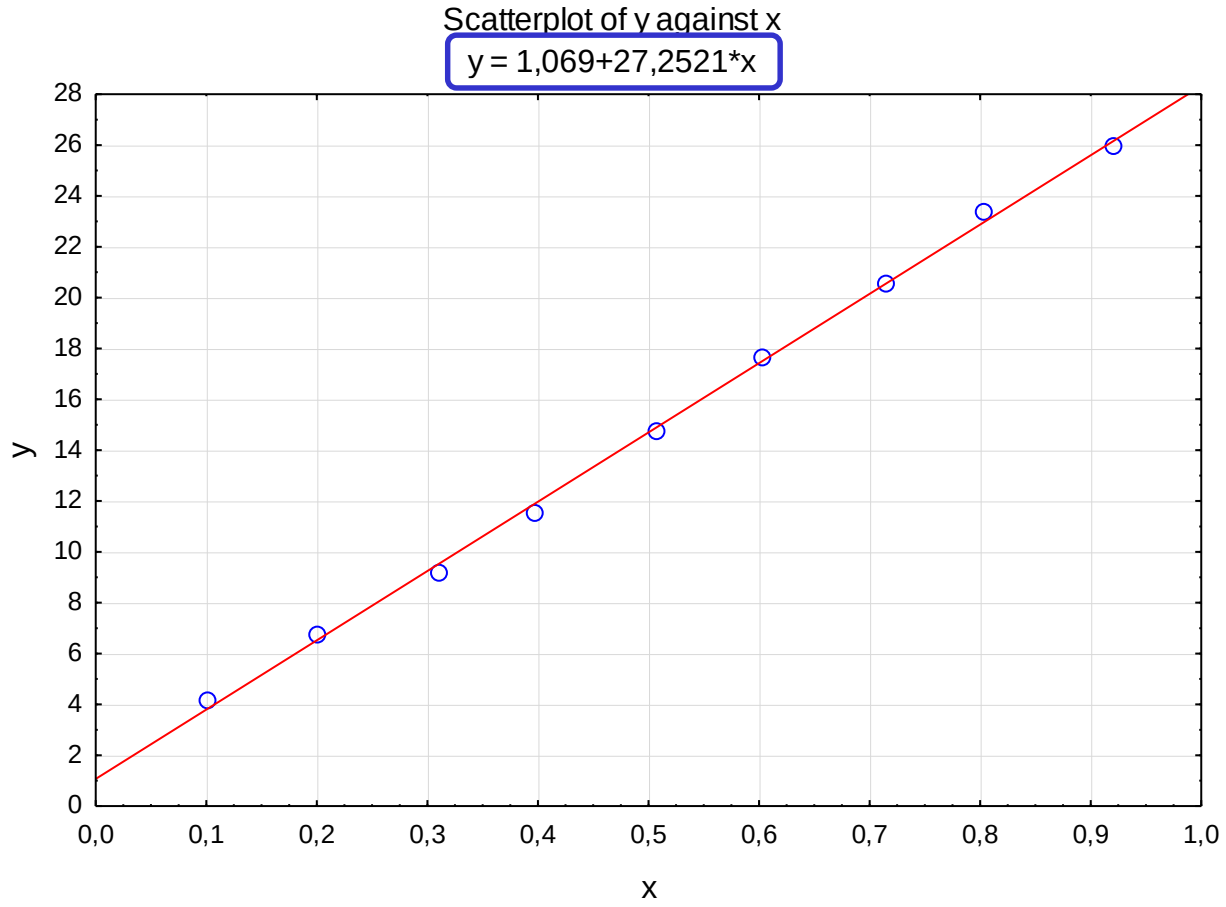
$$\hat{\alpha}' \equiv a' = a - b\bar{x}$$

$$a = \frac{133{,}918}{9} = 14{,}880$$

$$b = \frac{16{,}892}{0{,}620} = 27{,}252$$

$$a' = 14{,}880 - \\ -27{,}252 * 0{,}5068 = 1{,}069$$

Graphs > Scatterplots



$$Y_i = \alpha + \beta(x_i - \bar{x})$$
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$$

A becslések tulajdonságai

$$\hat{\alpha} = \frac{\sum y_i}{n} = \bar{y}$$
$$\hat{\beta} = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$E(\hat{\alpha}) \equiv E\left(\frac{\sum y_i}{n}\right) = \frac{1}{n}\sum E(y_i) = \frac{1}{n}\sum E(\alpha + \beta(x_i - \bar{x}) + \varepsilon_i) =$$

$$= \frac{1}{n}\sum(\alpha + \beta(x_i - \bar{x}) + E(\varepsilon_i)) = \frac{1}{n}(n\alpha + \beta\sum(x_i - \bar{x})) = \alpha$$ torzítatlan

$$Var(\hat{\alpha}) = Var\left(\frac{\sum y_i}{n}\right) = \frac{1}{n^2} Var(\sum y_i) = \frac{1}{n^2} * n * \sigma_y^2 = \frac{\sigma_y^2}{n}$$ konzisztens

Hasonlóan levezethető:

$E(\hat{\beta}) = \beta$ torzítatlan

$$Var(\hat{\beta}) = \frac{\sigma_y^2}{\sum (x_i - \bar{x})^2}$$ konzisztens

Reziduális szórásnégyzet

hiba = a mért érték és az elméleti érték különbsége

reziduum = a mért érték és a becsült érték különbsége

$$\hat{\sigma}_y^2 = s_r^2 = \frac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{n-2}$$

1. példára:

- reziduális szórásnégyzet számítása

- az egyenes paramétereinek bizonytalanságának számítása

				reziduum	reziduum^2
sorrend	x	y	y_becsült	(y-y_becsült)	(y-y_becsült)^2
9	0,102	4,141	3,849	0,292	0,085
1	0,200	6,715	6,519	0,196	0,038
5	0,311	9,202	9,544	-0,342	0,117
2	0,398	11,526	11,915	-0,389	0,152
7	0,508	14,775	14,913	-0,138	0,019
4	0,603	17,642	17,502	0,140	0,020
8	0,715	20,554	20,554	0,000	0,000
3	0,804	23,363	22,980	0,383	0,147
6	0,920	26,000	26,141	-0,141	0,020
				szumma:	SS_{error} = 0,598
				SS/(n-2)=	MS_{error} = 0,0854

$$s_r^2 = \frac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{n-2} = \frac{0{,}598}{9-2} = 0{,}0854$$

$$s_a^2 = \frac{s_r^2}{n} = \frac{0{,}0854}{9} = 0{,}0095$$

$$s_b^2 = \frac{s_r^2}{\sum_i (x_i - \bar{x})^2} = \frac{0{,}0854}{0{,}620} = 0{,}1377$$

Konfidencia-intervallum az elméleti regressziós egyenes meredekségére

Általánosan:

$$P\left(\hat{\beta} - t_{\alpha/2} s_{\hat{\beta}} < \beta \leq \hat{\beta} + t_{\alpha/2} s_{\hat{\beta}}\right) = 1 - \alpha$$

1. példa adataival:

$$P\left(27{,}252 - 2{,}365 \cdot 0{,}371 < \beta \leq 27{,}252 + 2{,}365 \cdot 0{,}371\right) = 0{,}95$$

$$P\left(26{,}374 < \beta \leq 28{,}130\right) = 0{,}95$$

Konfidencia-intervallum az elméleti regressziós egyenes tengelymetszetére 1.

Tengelymetszet: az x=0 pontban vett függvényérték

$$\boxed{a' = \widehat{\alpha'} = \hat{Y}(x = 0)}$$

Általánosan:

$$P\left(\hat{\alpha}' - t_{\alpha/2} s_{\hat{\alpha}'} < \alpha' \leq \hat{\alpha}' + t_{\alpha/2} s_{\hat{\alpha}'}\right) = 1 - \alpha$$

Az elméleti regressziós egyenes egy pontjának bizonytalansága

$$E\left(\hat{Y}_x\right) = E\left[\hat{\alpha} + \hat{\beta}\left(x - \bar{x}\right)\right] = \alpha + \beta\left(x - \bar{x}\right) = Y_x$$

$$Var\left(\hat{Y}_x\right) = Var\left(\hat{\alpha}\right) + \left(x - \bar{x}\right)^2 Var\left(\hat{\beta}\right) = \sigma_y^2\left[\frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2}\right]$$

Kísérletből becsülve:

$$s^2_{\hat{Y}_X} = s^2_{\hat{\alpha}} + \left(x - \bar{x}\right)^2 s^2_{\hat{\beta}} = s_r^2\left[\frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2}\right]$$

$$a' = \widehat{\alpha'} = \hat{Y}(x = 0)$$

Konfidencia-intervallum az elméleti regressziós egyenes tengelymetszetére 2.

1. példa adataival:

$$s^2_{\hat{Y}_{x=0}} = s_r^2 \left[\frac{1}{n} + \frac{\left(0 - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} \right] = 0{,}0854 \left[\frac{1}{9} + \frac{\left(0 - 0{,}5068\right)^2}{0{,}620} \right] = 0{,}0449$$

$$P\left(1{,}069 - 2{,}365 \cdot 0{,}212 < \hat{Y}(x = 0) \leq 1{,}069 + 2{,}365 \cdot 0{,}212 \right) = 0{,}95$$

$$P\left(0{,}568 < \hat{Y}(x = 0) \leq 1{,}570 \right) = 0{,}95$$

Hipotézisvizsgálat az elméleti regressziós egyenes meredekségére

Vizsgálja meg 5%-os szignifikanciaszinten, hogy az elméleti egyenes meredeksége 27-e!

$H_0 : \beta = 27$ $\qquad$ $H_0 : \beta \neq 27$

$$t_0 = \frac{\hat{\beta} - E\left[\hat{\beta}\right] \| H_0}{s_{\hat{\beta}}} = \frac{27{,}252 - 27}{0{,}371} = 0{,}679$$

$$t_0 = t_{0{,}025} = 2{,}365$$

Döntés?

Hipotézisvizsgálat az elméleti regressziós egyenes tengelymetszetére

Vizsgálja meg 5%-os szignifikanciaszinten, hogy az elméleti egyenes tengelymetszete nulla-e!

$H_0 : \alpha' = 0$ $\qquad$ $H_0 : \alpha' \neq 0$

$$t_0 = \frac{\hat{\alpha}' - E\left[\hat{\alpha}'\right]H_0}{s_{\hat{\alpha}'}} = \frac{1{,}069 - 0}{0{,}212} = 5{,}04$$

$$t_0 = t_{0{,}025} = 2{,}365$$

Döntés?

A lineáris regresszió eredménytáblázata a STATISTICA programban – 1.

Statistics > Multiple Regression

95%-os konfidencia-intervallumok

Parameter Estimates (Linreg_orai_pelda in Workbook2)
Sigma-restricted parameterization

Effect	y Param.	y Std.Err	y t	y p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept	1,06901	0,211862	5,04577	0,001487	0,56803	1,56998
x	27,25213	0,371234	73,40950	0,000000	26,37430	28,12996

$\hat{\alpha}'$ $\hat{\beta}$ $s_{\hat{\beta}}$ $s_{\hat{\alpha}'}$

próbastatisztikák (t_0) értékei

$$H_0 : \alpha' = 0 \longrightarrow t_0 = \frac{\hat{\alpha}' - 0}{s_{\hat{\alpha}'}}$$

$$H_0 : \beta = 0 \longrightarrow t_0 = \frac{\hat{\beta} - 0}{s_{\hat{\beta}}}$$

Konfidencia-intervallum
az elméleti regressziós egyenes egy pontjának (várható) értékére

Milyen intervallumban található 95%-os valószínűséggel az elméleti függvényérték az x=0,3 helyen?

Azaz adjunk 95%-os konfidencia-intervallumot az analitikai jel (y) várható értékére az x=0,3 helyen!

$$s^2_{\hat{Y}_{x=0,3}} = s_r^2\left[\frac{1}{n} + \frac{(0,3-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right] = 0,0854\left[\frac{1}{9} + \frac{(0,3-0,5068)^2}{0,620}\right] = 0,0154$$

$$P\left(9,245 - 2,365\cdot 0,124 < \hat{Y}(x=0,3) \le 9,245 + 2,365\cdot 0,124\right) = 0,95$$

$$P\left(8,952 < \hat{Y}(x=0,3) \le 9,538\right) = 0,95$$

Konfidencia-sáv

Kiszámítjuk különböző x értékeknél az elméleti függvényérték konfidencia-intervallumát, s az így kapott pontokat összekötjük.

Ez az a sáv, amin belül az **elméleti egyenes** 1-α valószínűséggel megtalálható.

$$P\left(\hat{Y} - t_{\alpha/2} s_{\hat{Y}} < Y \leq \hat{Y} + t_{\alpha/2} s_{\hat{Y}}\right) = 1 - \alpha$$

$$s_{\hat{Y}} = s_y \sqrt{\frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2}}$$

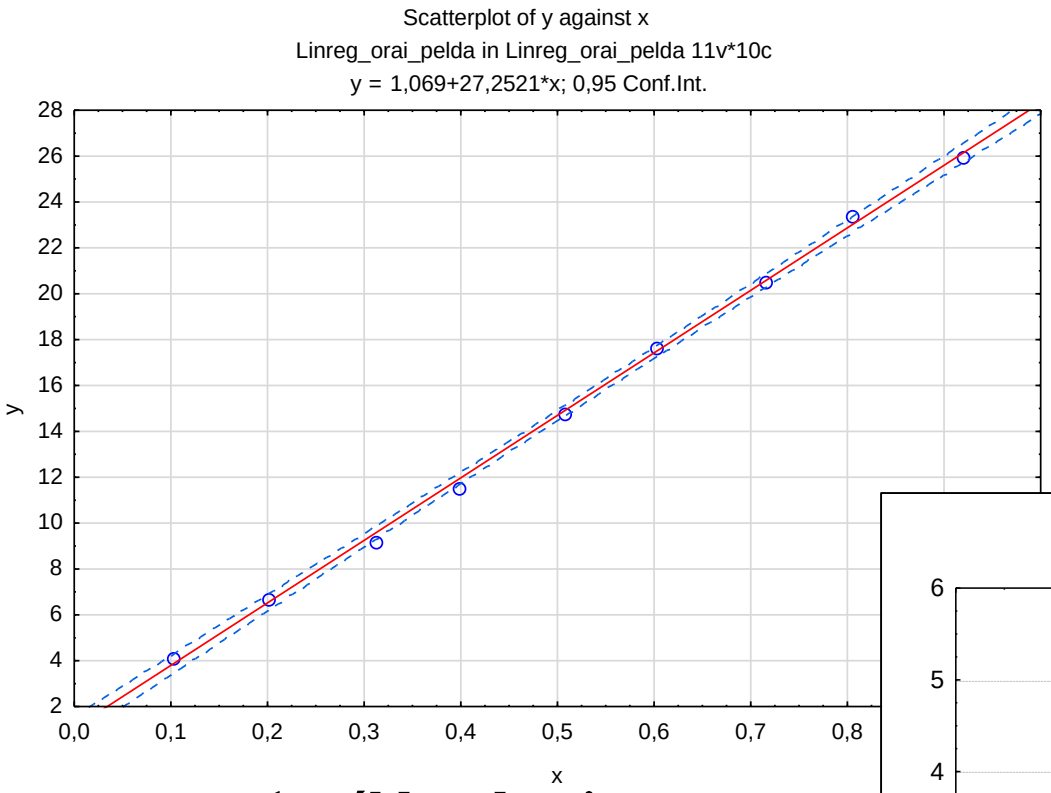
Mitől függ a konfidencia-sáv szélessége?

$$P\left(\hat{Y} - t_{\alpha/2} s_{\hat{Y}} < Y \leq \hat{Y} + t_{\alpha/2} s_{\hat{Y}}\right) = 1 - \alpha$$

$$Var\left(\hat{Y}_x\right) = \sigma_y^2 \left[\frac{1}{n} + \frac{\left(x - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2}\right]$$

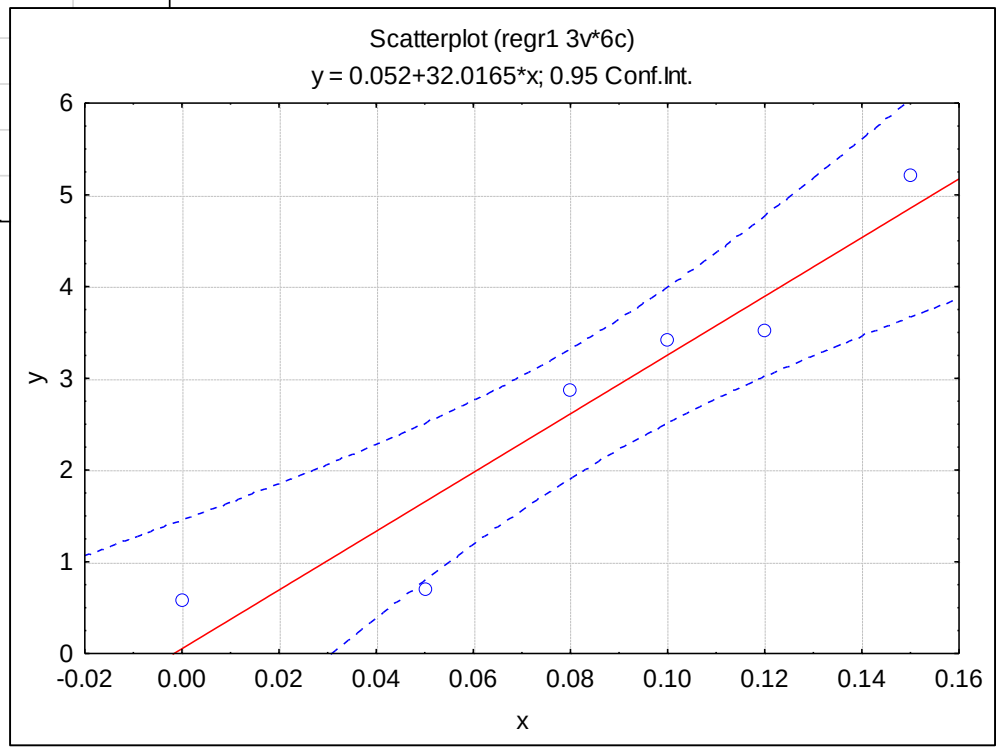
- mérések számától (n)
- mérési hibától (σ_y^2): milyen precízen mérünk
- x_i értékek elhelyezkedése
- milyen x értéknél kérdezzük

Konfidencia-sáv



1. példa adataira

egy másik adatsorra



Jóslási intervallum, jóslási sáv

A jóslási intervallum megmutatja hogy hol helyezkedik el egy újonnan mérendő *y** érték adott valószínűséggel.

$$P\left(\hat{Y} - t_{\alpha/2} s_{y^*-\hat{Y}} < y^* \le \hat{Y} + t_{\alpha/2} s_{y^*-\hat{Y}}\right) = 1-\alpha$$

$$s_{y-\hat{Y}} = s_r \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

Számoljuk ki az 1. példára a jóslási sáv szélességét az x=0,3 helyen!

$$s^2_{\hat{Y}_{x=0,3}} = s_r^2\left[1+\frac{1}{n}+\frac{\left(0,3-\bar{x}\right)^2}{\sum_i\left(x_i-\bar{x}\right)^2}\right] = 0,0854\left[1+\frac{1}{9}+\frac{\left(0,3-0,5068\right)^2}{0,620}\right] = 0,1$$

$$P\left(9,245-2,365\cdot 0,3175 < y^*_{x=0,3} \le 9,245+2,365\cdot 0,3175\right) = 0,95$$

$$P\left(8,494 < y^*_{x=0,3} \le 9,996\right) = 0,95$$

Gyakorló feladatok az 1. példához

1. Alátámasztják-e a mérési adatok 5%-os szignifikancia-szinten, hogy a 0,3-es koncentrációnál az abszorbancia (várható értéke) legfeljebb 9,5?

2. A kalibrációs egyenes felvétele után a 0,3-es koncentrációnál egy új mérést végeztek, s az abszorbanciára 8,8 adódott. Lehetséges ez? A vizsgálatot 5%-os szignifikanciaszinten végezze!

Determinációs együttható (R^2)

Megadja, hogy az ingadozás hányadrészét magyarázza a választott modell.

modell

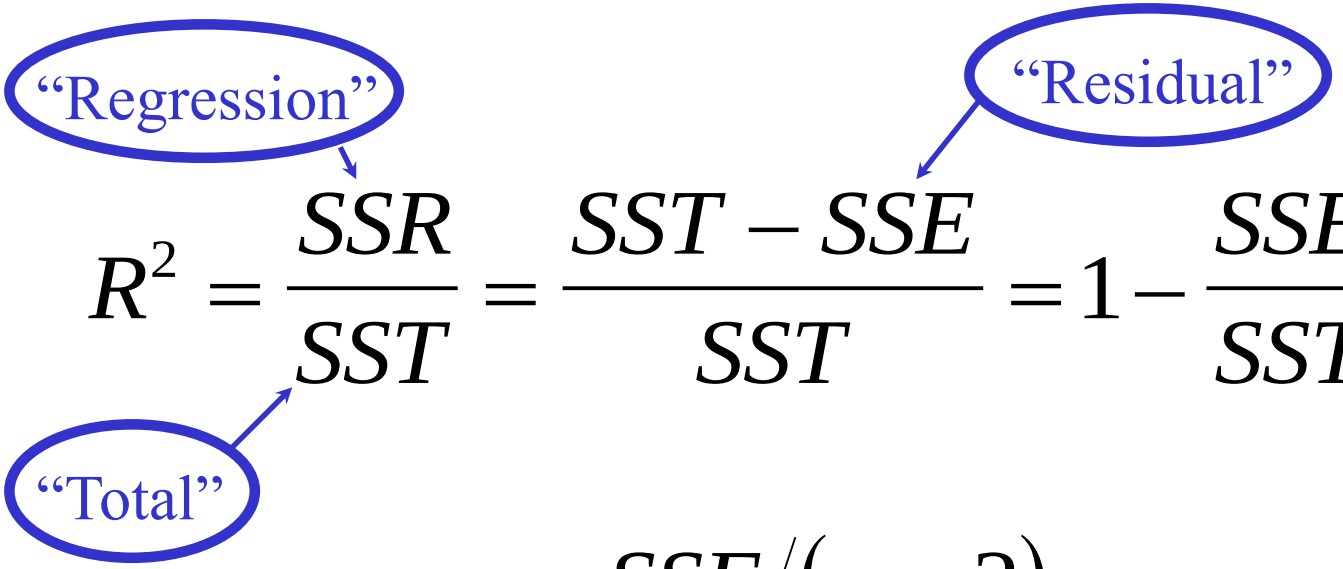
illeszkedési hiba (reziduum)

$$R^2 = \frac{\sum_i \left(\hat{Y}_i - \bar{y}\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2} = 1 - \frac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2}$$

teljes

$$R^2_{adj} = 1 - \frac{\dfrac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{n-p}}{\dfrac{\sum_i \left(y_i - \bar{y}\right)^2}{n-1}} = 1 - \frac{s_r^2}{s_T^2}$$

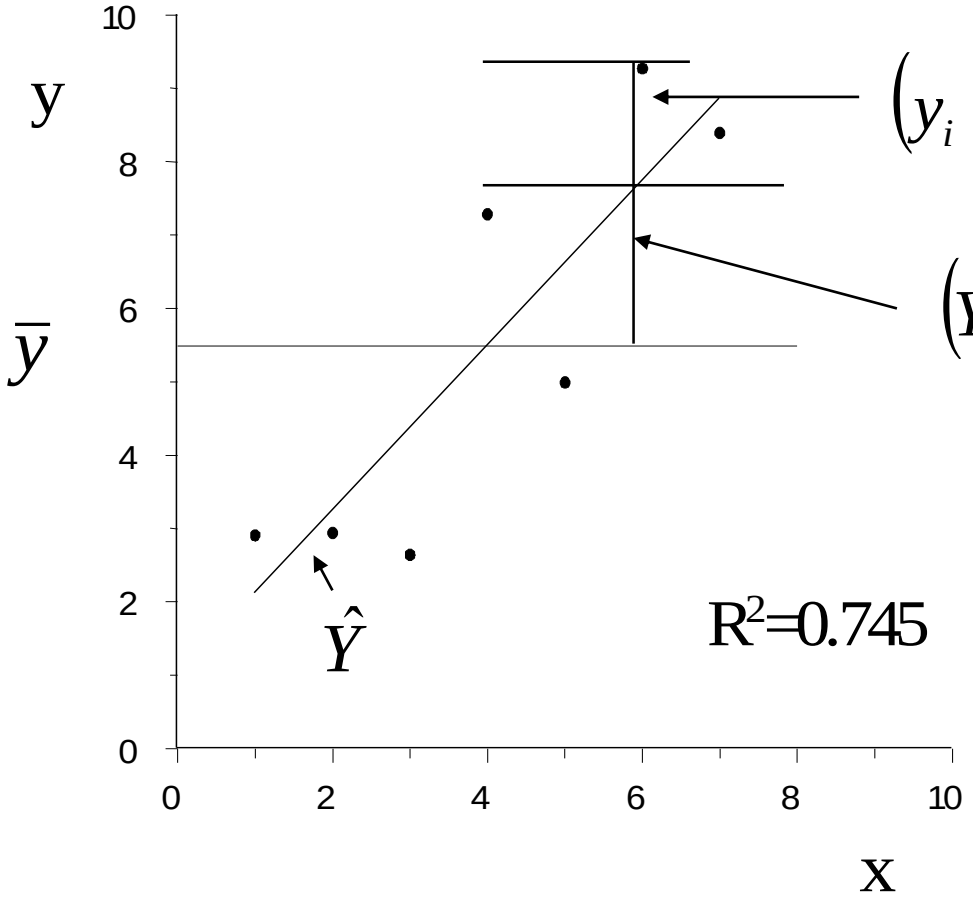
Determinációs együttható



$$R^2\_adj = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i \left(y_i - \hat{Y}_i\right)^2 + \sum_i \left(\hat{Y}_i - \bar{y}\right)^2$$

teljes

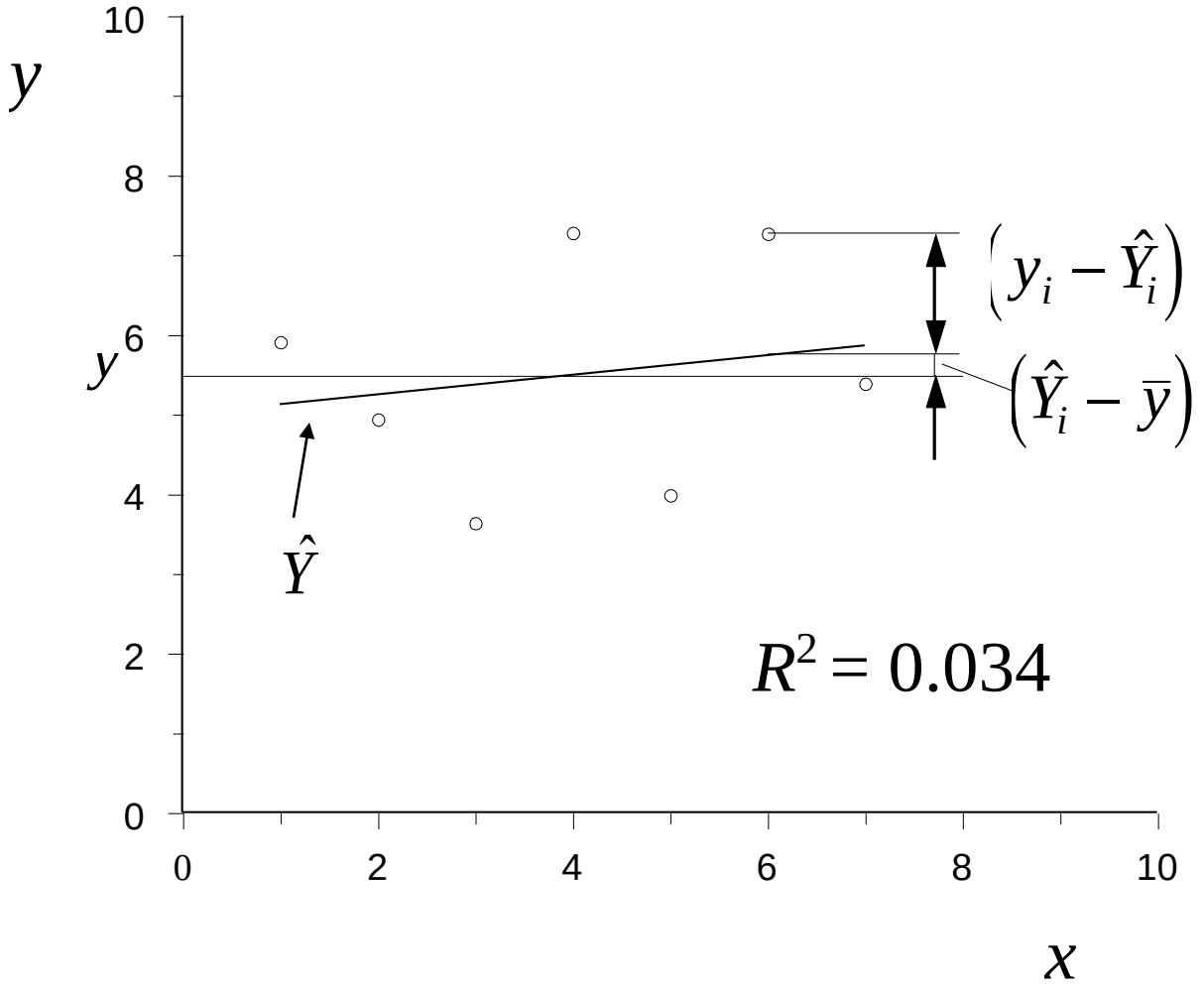


$\left(y_i - \hat{Y}\right)$ a teljes változás nem magyarázott része

$\left(\hat{Y} - \bar{y}\right)$ az egyenes magyarázta rész

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{y})^2$$

teljes



a teljes változás nem magyarázott része

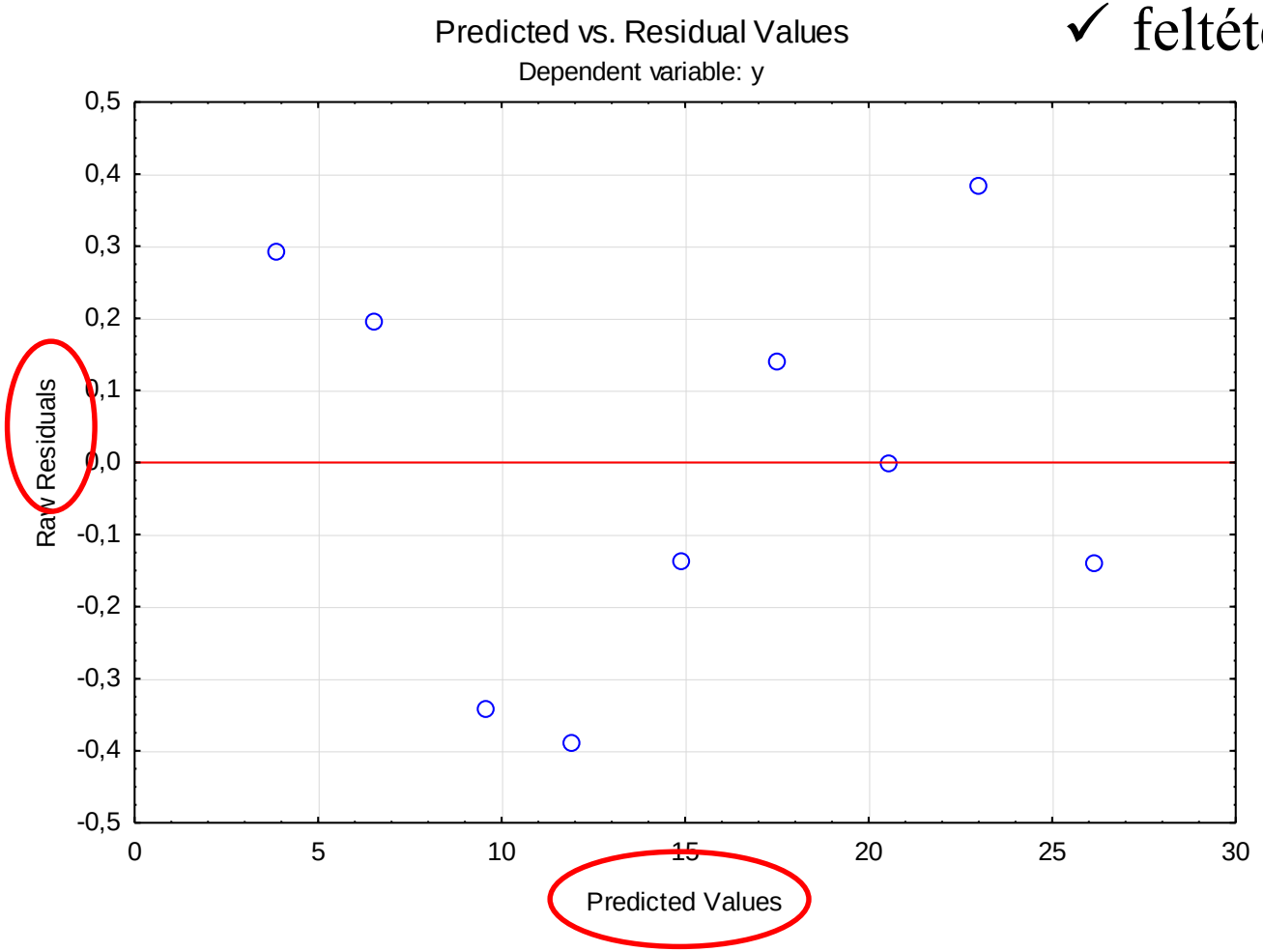
az egyenes magyarázta rész

A regresszió feltételeinek ellenőrzése: a reziduumok vizsgálata, reziduumábrák

1. példa adataival készült ábra

1. Var(*y*) = Var(ε) konstans ellenőrzése

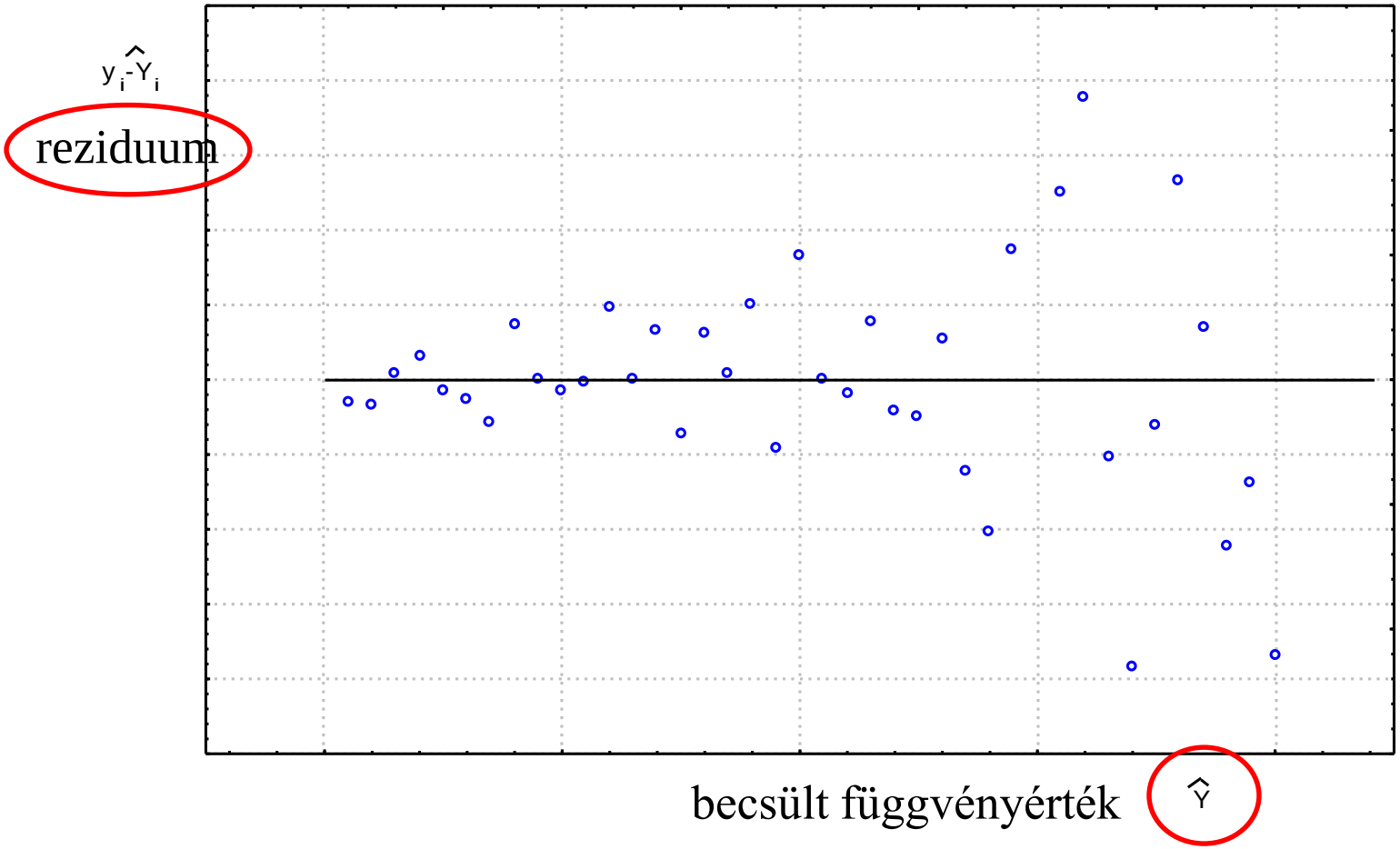
✓ feltétel teljesül



1. Var(*y*) = Var(ε) konstans ellenőrzése

másik adatsorra készült ábra

feltétel NEM teljesül

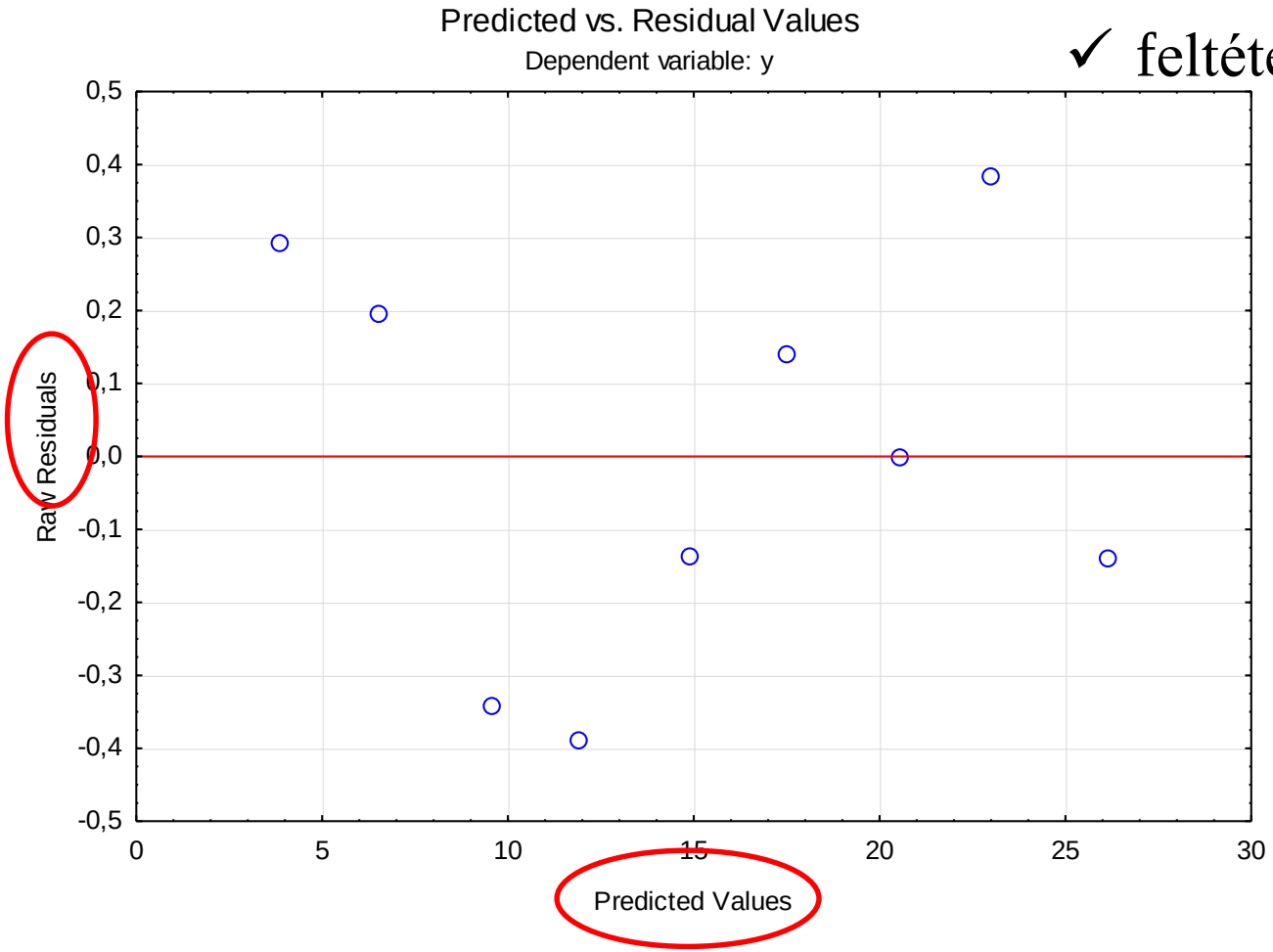


1. példa adataival készült ábra

2. A lineáris függvény adekvát-e

(az illesztett függvény megfelelően írja-e le a változást)

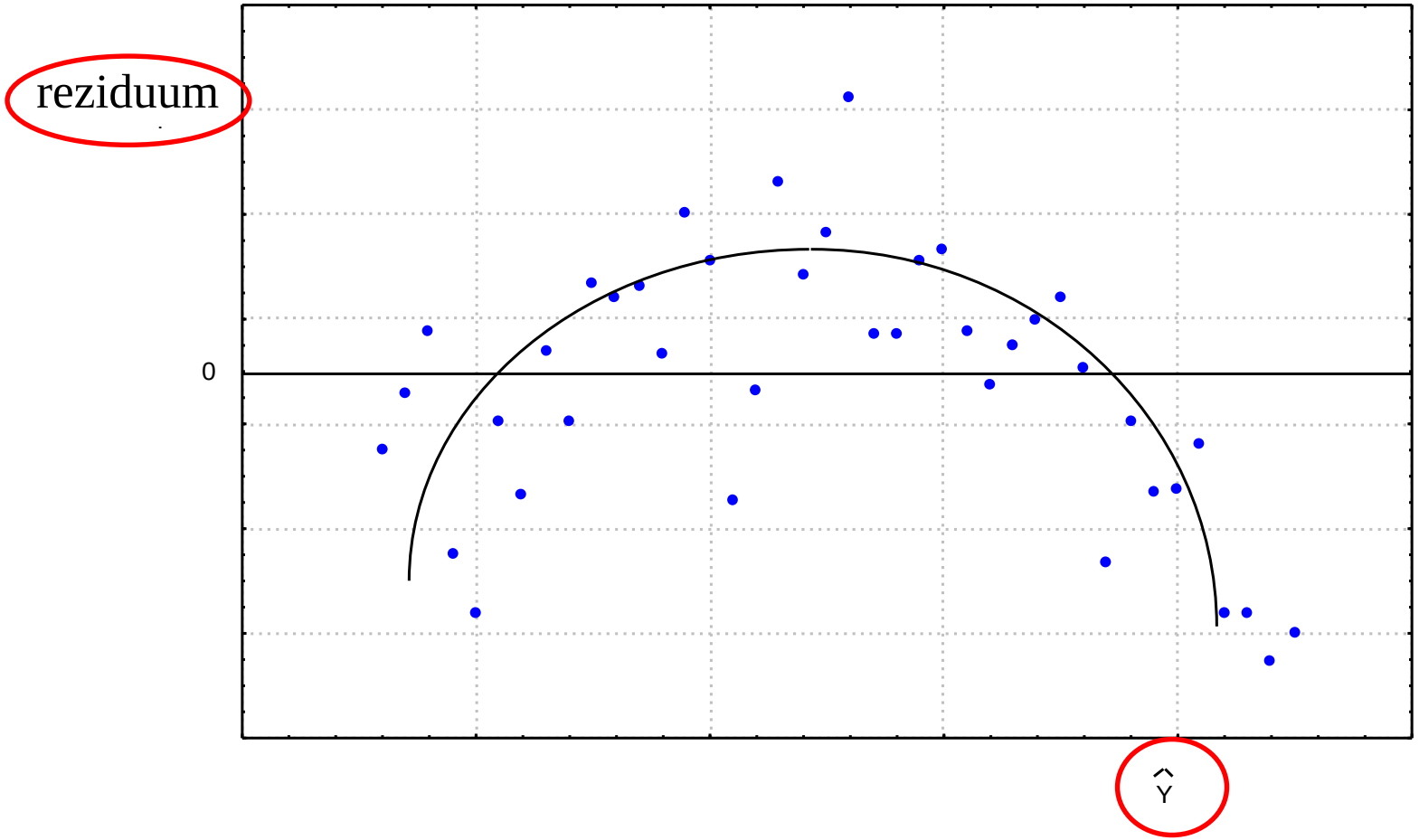
✓ feltétel teljesül



másik adatsorra készült ábra

2. A lineáris függvény adekvát-e
(az illesztett modell megfelelő-e)

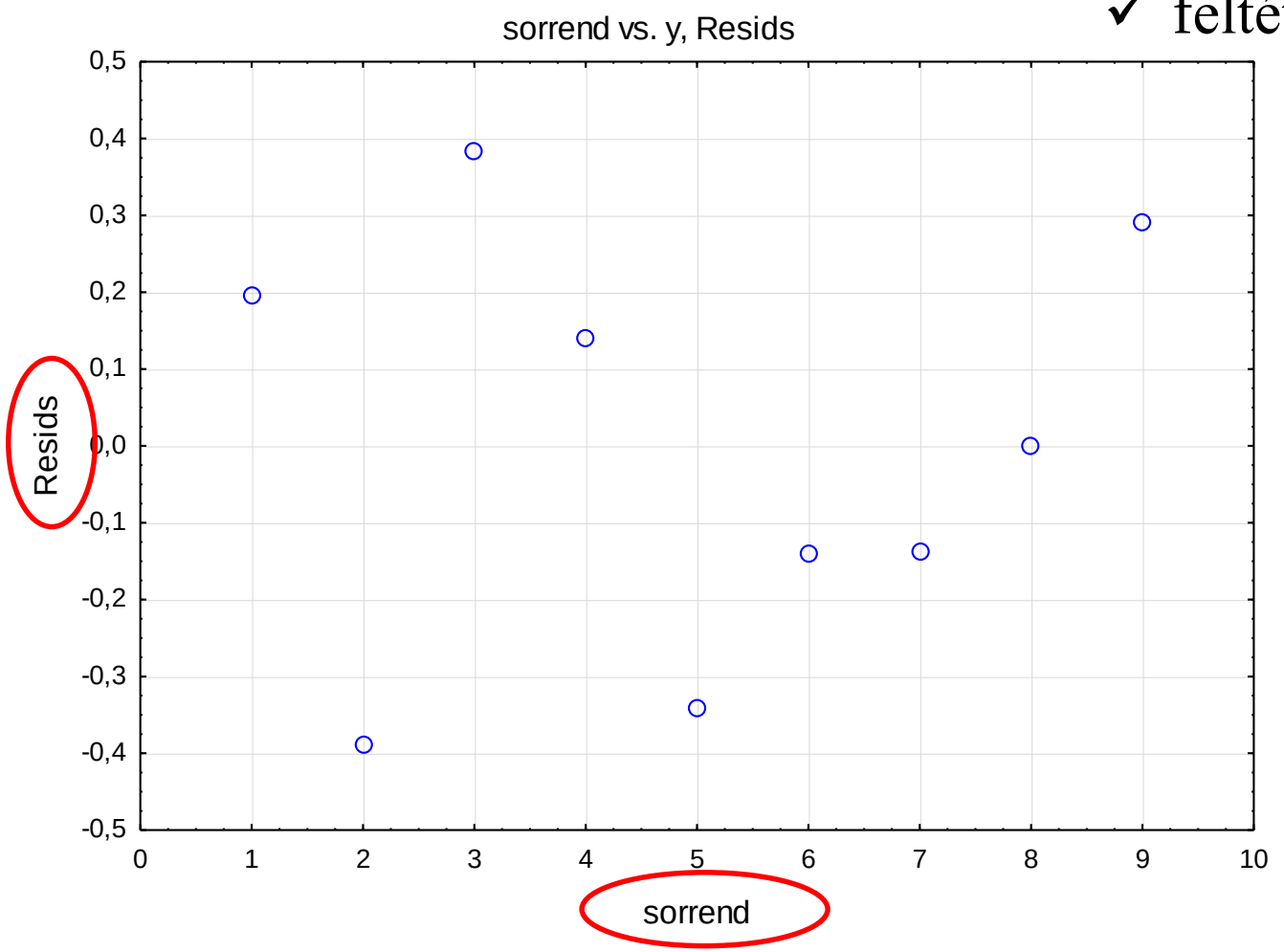
feltétel NEM teljesül



3. mérési hibák függetlenek-e egymástól

1. példa adataival készült ábra

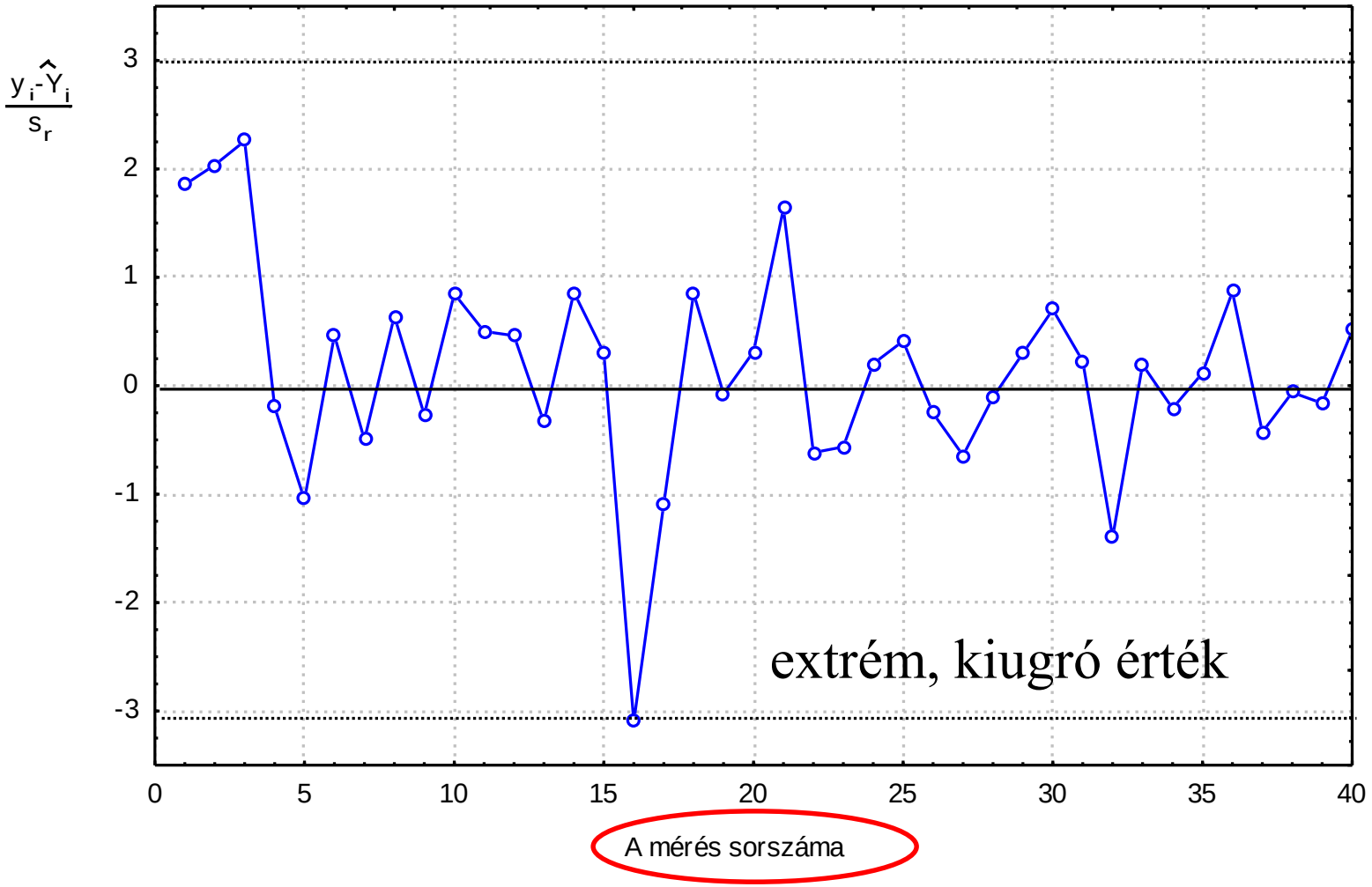
✓ feltétel teljesül



3. mérési hibák függetlenek-e egymástól

másik adatsorra készült ábra

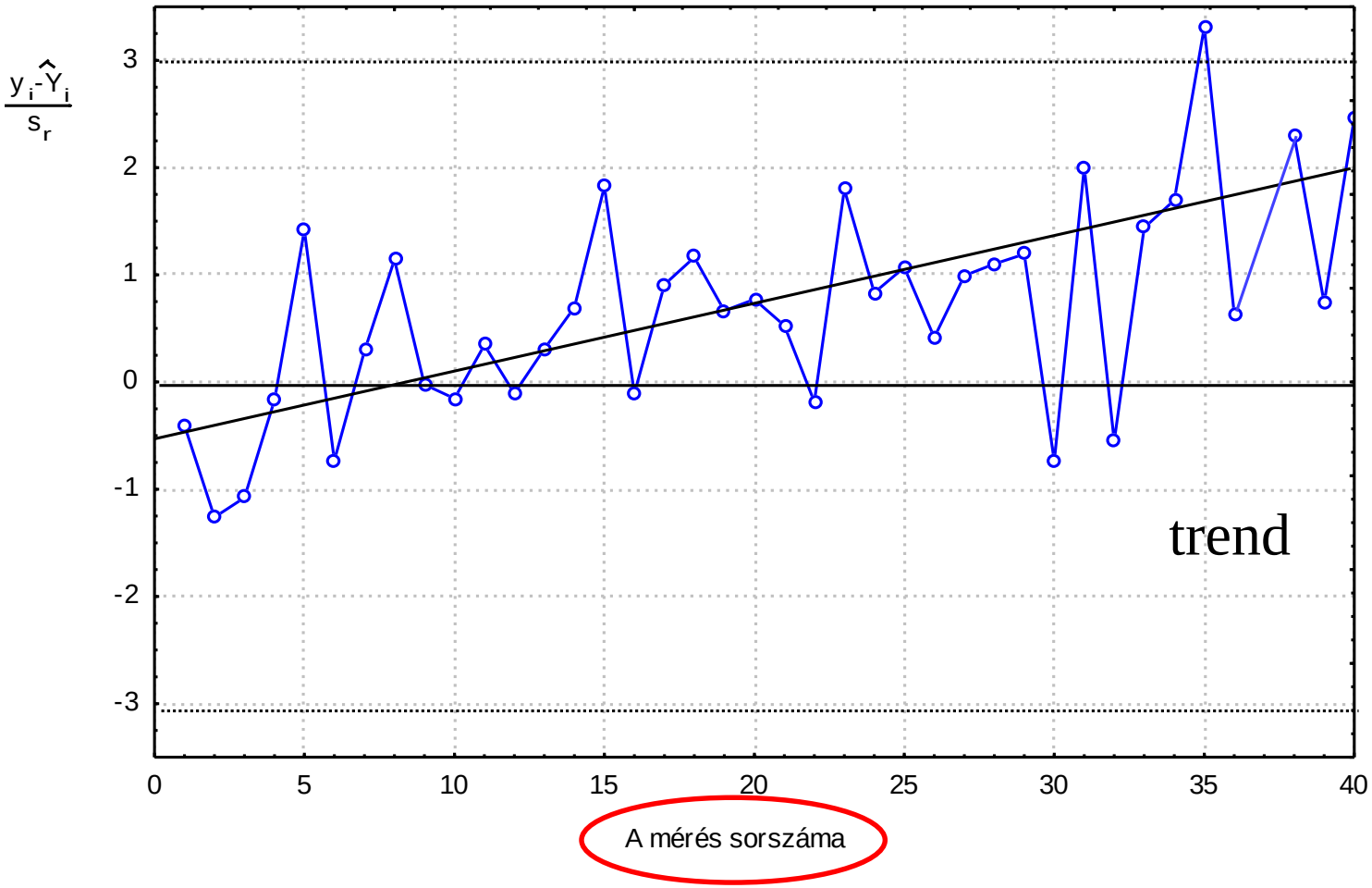
feltétel NEM teljesül



3. mérési hibák függetlenek-e egymástól

másik adatsorra készült ábra

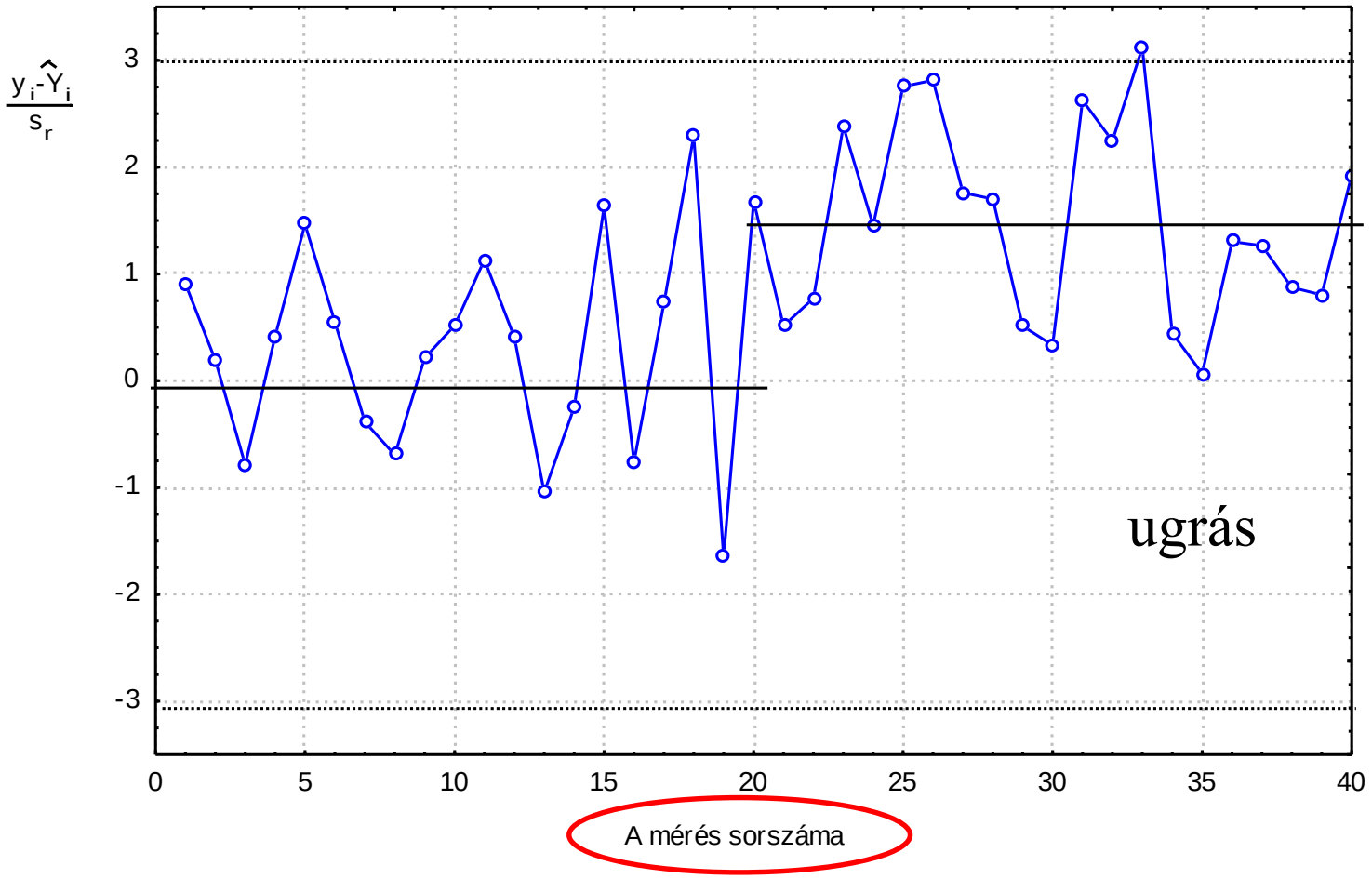
feltétel NEM teljesül



3. mérési hibák függetlenek-e egymástól

másik adatsorra készült ábra

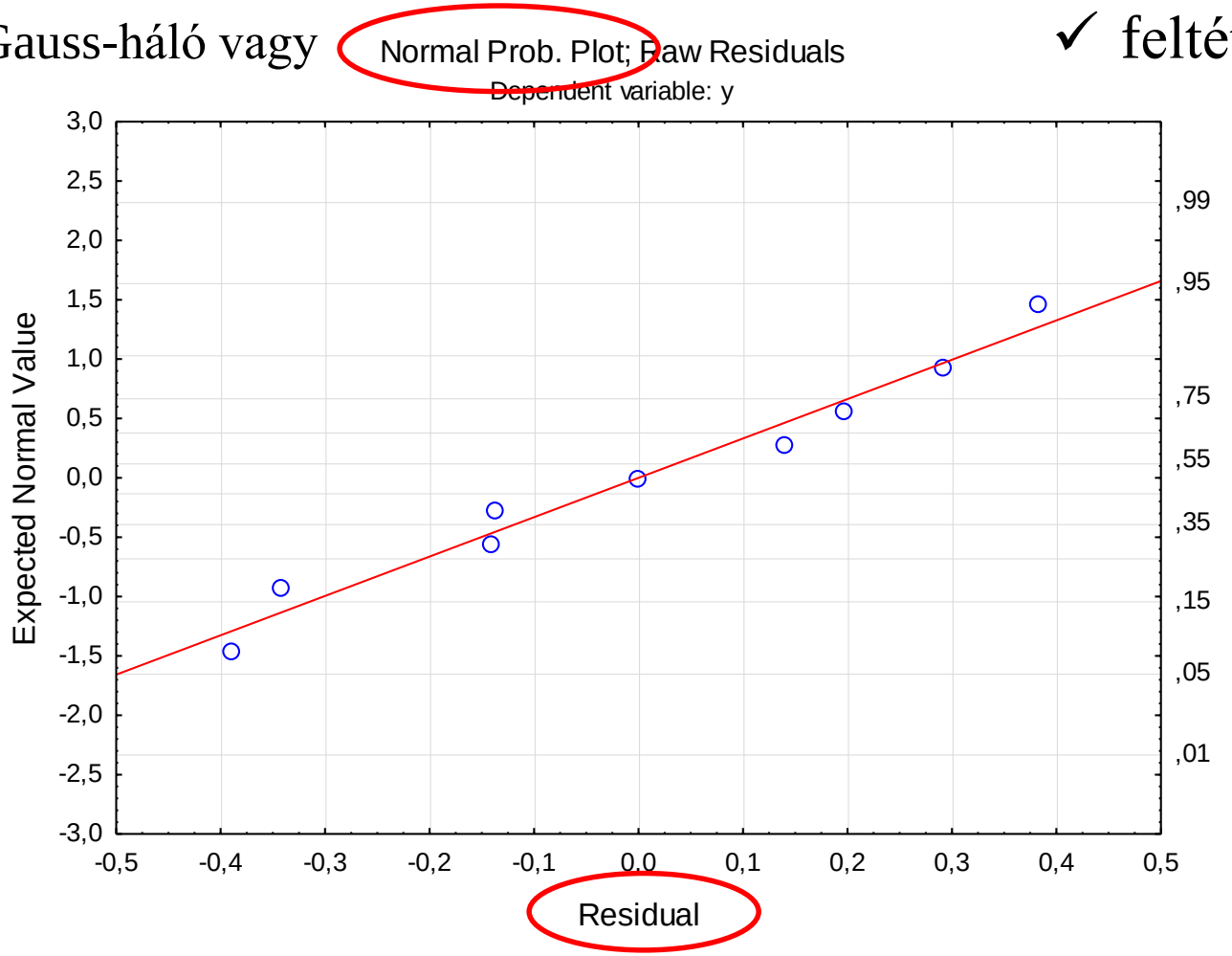
feltétel NEM teljesül



1. példa adataival készült ábra

4. mérési hibák normális eloszlásúak-e

Gauss-háló vagy

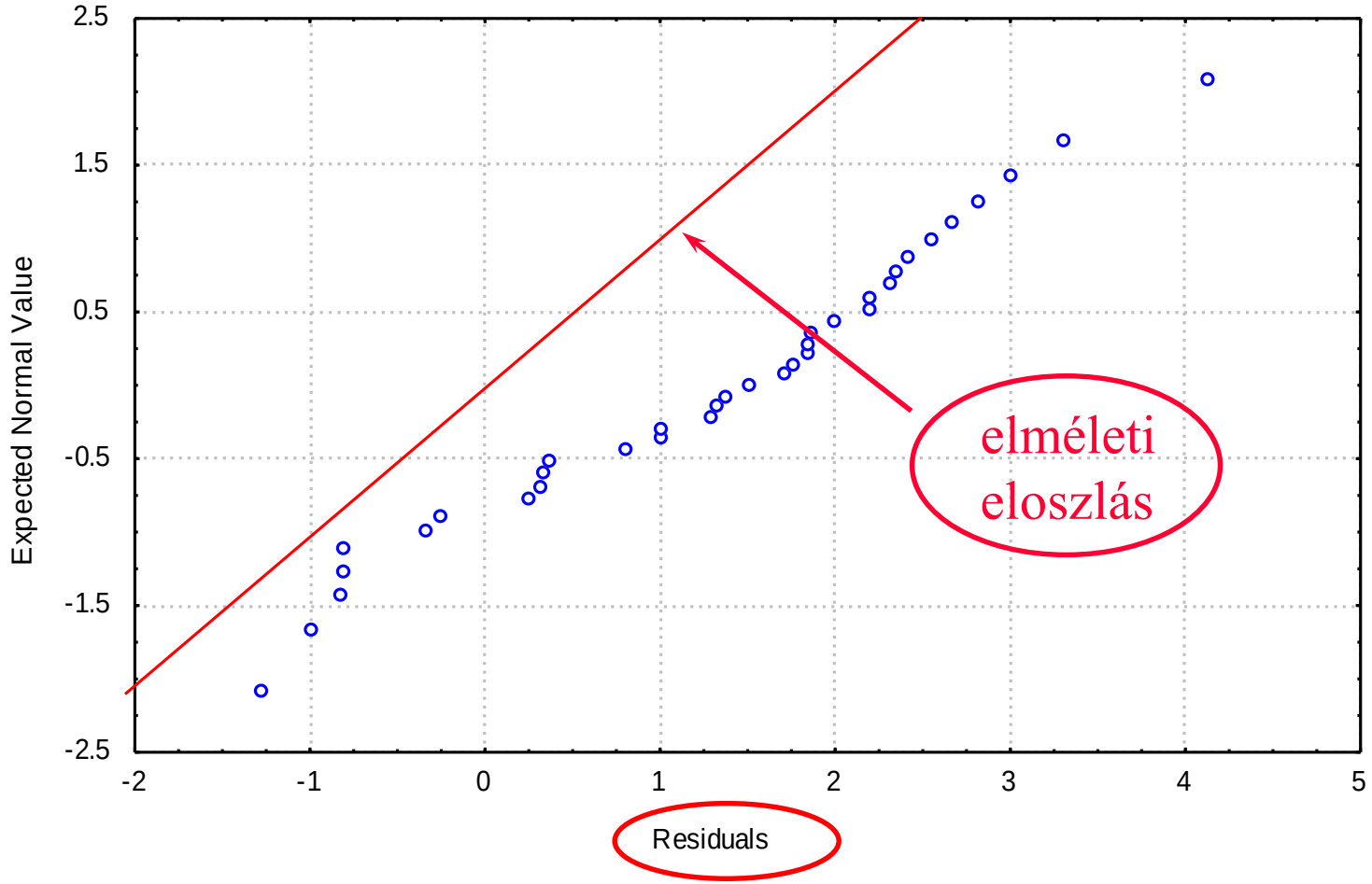


✓ feltétel teljesül

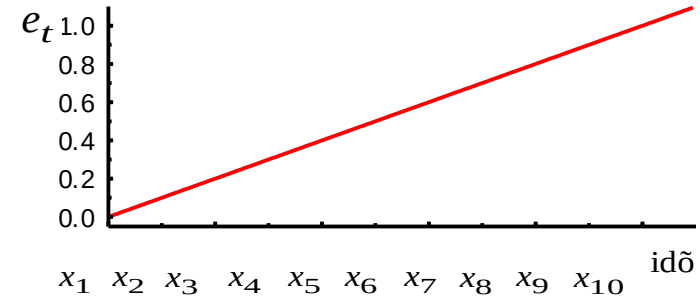
4. mérési hibák normális eloszlásúak-e

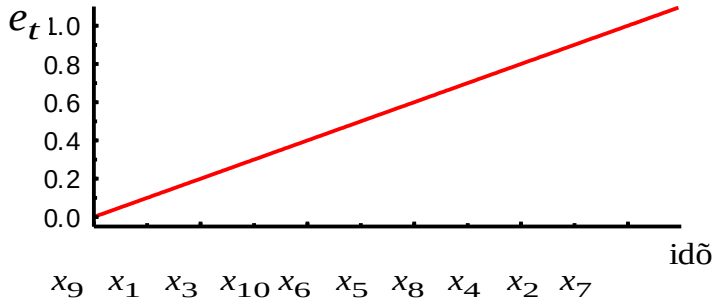
másik adatsorra készült ábra

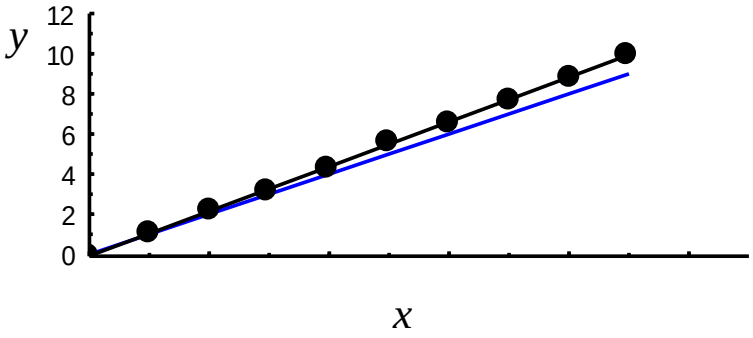
feltétel NEM teljesül

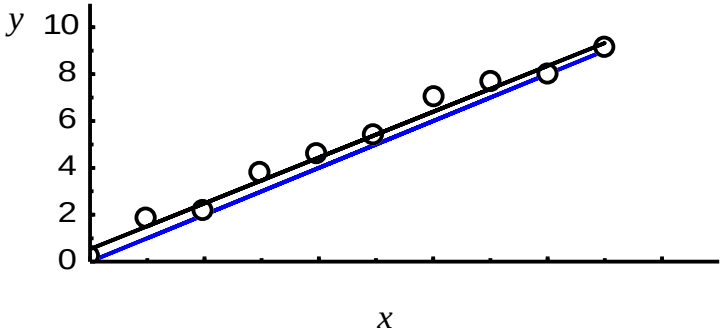


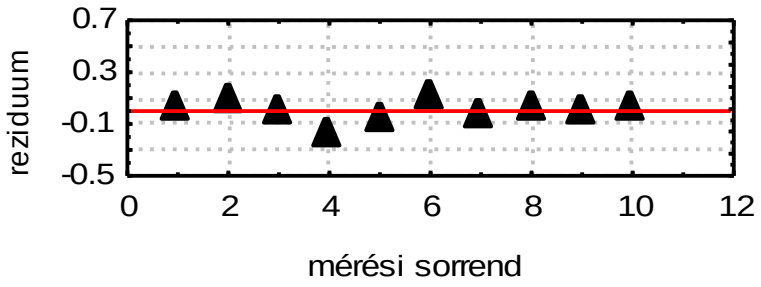
A mérések sorrendjének hatása

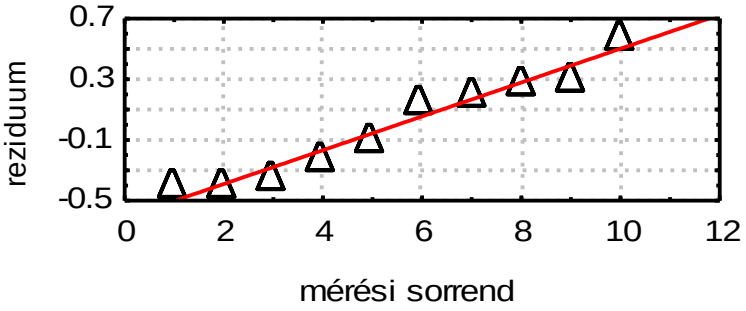








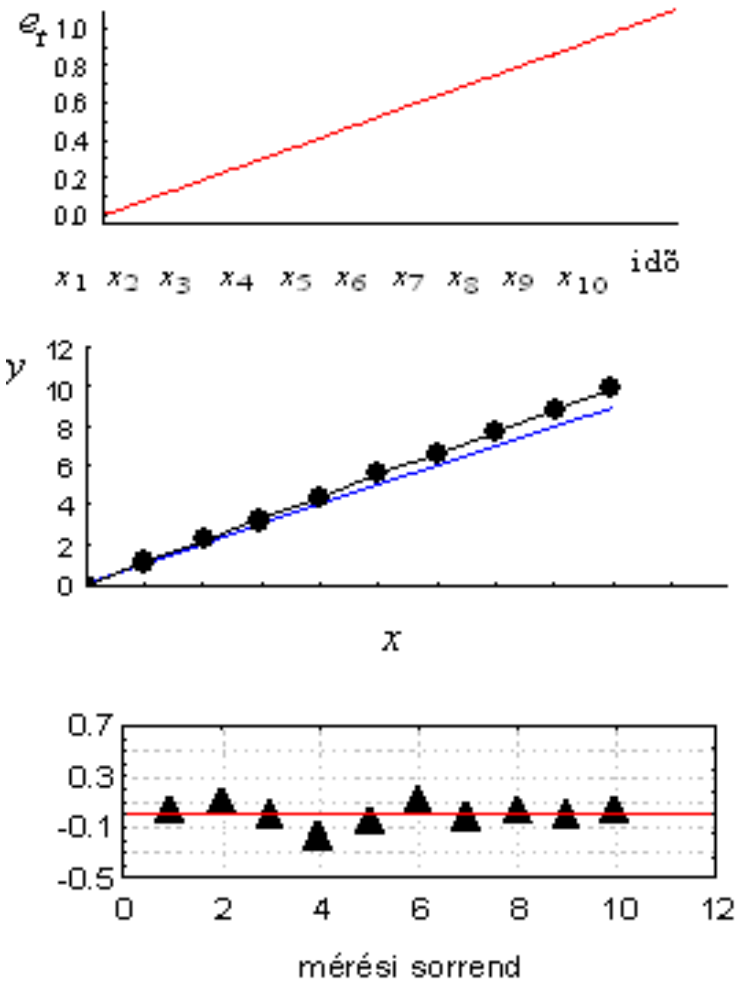




Kézenfekvő lenne, hogy <u>*x* növekvő sorrendjében mérünk</u>

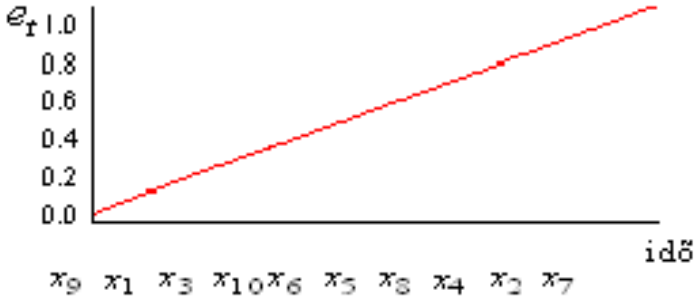
ekkor *y* az *x* hatása + a mérési ingadozás + az idő hatása (sorrend): az illesztett függvény az *x* és az idő hatásának összegét becsüli

Ha a reziduumokat a mérések sorszáma függvényében ábrázoljuk, nem látunk rendszerességet, mert a két hatás összegét leíró egyenes körül véletlenszerű az ingadozás.

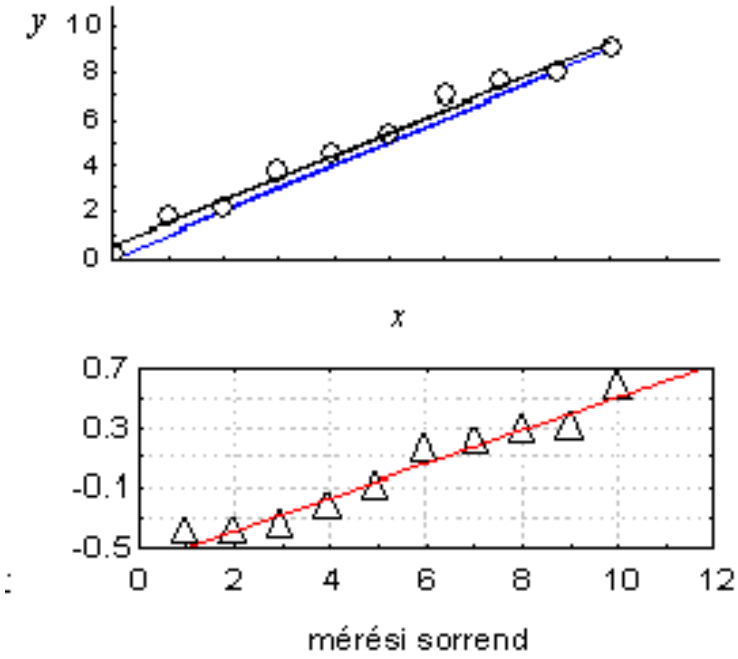


Vagyis nem szerzünk tudomást arról, hogy az *y* *x*-től való függésébe egy zavaró tényező (az idő) hatását is belemértük és beleszámoltuk: hamis az összefüggés.

Ha véletlenszerű sorrendben mérünk, az egyre nagyobb *x* értékekhez nem tartozik egyre hosszabb eltelt idő.



Itt is a két hatás összegét mérjük, de a két hatás nem mutat egy irányba, az idő járuléka nem nő monoton módon az *x* értékével. Az időbeliség egyrészt eltorzítja az összefüggést (nagyjából párhuzamosan fölfelé tolja el az egyenest), másrészt nagyobb szóródást okoz.



Ha a mérések sorszáma függvényében ábrázoljuk a reziduumokat, azok rendszerességet mutatnak, mert az *y*-nak az időtől való függéséről az illesztett egyenes nem ad számot, azt az egyenestől való eltérésként észleljük.