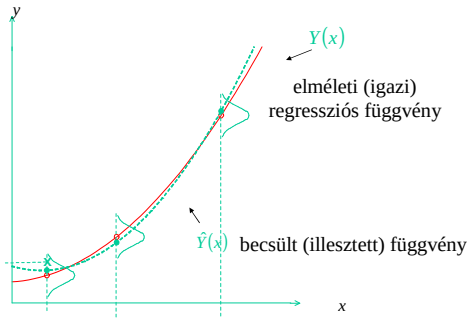
Regresszió1

REGRESSZIÓ-ANALÍZIS

Függvények illesztése

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).



Az illesztett függvényről alkotott elképzelés kétféle lehet:

- görbét (interpolációs formulát) kívánunk illeszteni, amely jól kezelhető és a szükséges pontossággal reprezentálja a mérési adatokat,
- a változók közötti kauzális összefüggést leíró modellt illesztünk, amelynek paraméterei fizikai értelemmel bírnak, így extrapolációra is használhatók.

Modell:

$Y(x) = \varphi(x; \alpha, \beta, \gamma, ...)$ igazi (elméleti) érték

$y = Y + \varepsilon$ y a függő változó mért értéke
ε a mérési hiba

$E(y|x) = Y$

$E(\varepsilon) = 0$ $Var(\varepsilon) = Var(y|x) = \sigma_y^2$

Feltételezések

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

Regresszió1

46. példa

No	run order	x	y
1	3	0	0.58
2	5	0.05	0.7
3	4	0.08	2.88
4	2	0.1	3.42
5	1	0.12	3.53
6	6	0.15	5.21

Kísérletileg vizsgálták az x független változó és az y függő változó közötti összefüggést.
x értéke pontosan beállítható,
y értéke az Y valódi érték körül ingadozik, ~$N(0,\sigma^2)$ $\sigma_y^2 = \text{konst}$

Egyenes illesztése, $\sigma_y^2 = \text{konst}$

A legkisebb négyzetek módszere szerinti becslési kritérium:

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

$$Y_i = \beta_0 + \beta x_i = \alpha + \beta\left(x_i - \bar{x}\right) \qquad \beta_0 = \alpha - \beta\bar{x} \quad \text{tengelymetszet}$$

$$\hat{Y}_i = b_0 + b x_i = a + b\left(x_i - \bar{x}\right) \qquad b_0 = a - b\bar{x}$$

$$\phi = \sum_i \left(y_i - b_0 - b x_i\right)^2 = \min.$$

Az ún. normálegyenletek:

$$\frac{\partial \phi}{\partial b_0} = -2\sum\left[y_i - b_0 - b x_i\right] = 0$$

$$\frac{\partial \phi}{\partial b} = -2\sum\left[y_i - b_0 - b x_i\right] x_i = 0$$

Mivel $\sum x_i \neq 0$

a b_0 és b becslések nem függetlenek

Átrendezve

$$\sum y_i = n b_0 + b\sum x_i$$

$$\sum y_i x_i = b_0 \sum x_i + b\sum x_i^2$$

Regression Summary for Dependent Variable: y (regr1)
R= .95061604 R²= .90367086 Adjusted R²= .87958858
F(1,4)=37.524 p<.00360 Std.Error of estimate: .62136

N=6	Beta	Std.Err. of Beta	B	Std.Err. of B	t(4)	p-level
Intercept			0.05196	0.504033	0.103084	0.922858
x	0.950616	0.155185	32.01651	5.226581	6.125708	0.003598

A modellt más alakban illesztve $Y_i = \alpha + \beta\left(x_i - \bar{x}\right)$

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

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

Átrendezve (normálegyenletek):

$$\sum y_i = na + b\sum\left(x_i - \bar{x}\right)$$

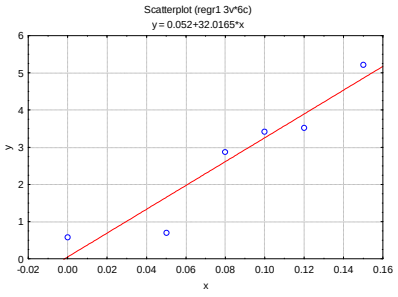
$$\sum y_i\left(x_i - \bar{x}\right) = a\sum\left(x_i - \bar{x}\right) + b\sum\left(x_i - \bar{x}\right)^2$$

Az a és b becslések itt függetlenek, mivel

$$\sum\left(x_i - \bar{x}\right) = 0$$

$$\bar{x} = \frac{\sum x_i}{n}$$

Graphs>Scatterplots



$$\sum y_i = na \qquad \sum y_i\left(x_i - \bar{x}\right) = b\sum\left(x_i - \bar{x}\right)^2$$

a és b függetlenül kaphatók meg a két normálegyenletből

$$a = \frac{\sum_i y_i}{n} \qquad b = \frac{\sum_i y_i\left(x_i - \bar{x}\right)}{\sum_i \left(x_i - \bar{x}\right)^2}$$

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

Regresszió1

$\sum (y_i - \hat{Y}_i)^2 = \chi^2 \sigma_y^2$ ha $Y = \alpha + \beta(x - \bar{x})$

(a valódi összefüggés egyenes)

$s_r^2 = \frac{\sum (y_i - \hat{Y}_i)^2}{n-2} = \frac{\chi^2 \sigma_y^2}{\nu} = s_y^2$ ha!

reziduális szórásnégyzet

$\hat{\sigma}_y^2 = s_y^2 = s_r^2$

Regression Summary for Dependent Variable: y (regr1)
R= .95061604 R²= .90367086 Adjusted R²= .87958858
F(1,4)=37.524 p<.00360 Std.Error of estimate: .62136

N=6	Beta	Std.Err. of Beta	B	Std.Err. of B	t(4)	p-level
Intercept			0.05196	0.504033	0.103084	0.922858
x	0.950616	0.155185	32.01651	5.226581	6.125708	0.003598

13

A négyzetösszegek felbontása

$y_i - Y_i = (y_i - \hat{Y}_i) + (\hat{Y}_i - Y_i)$

behelyettesítve

$Y = \alpha + \beta(x - \bar{x})$? és

$\hat{Y} = a + b(x - \bar{x})$!

$y_i - Y_i = (y_i - \hat{Y}_i) + (a_i - \alpha_i) + (b - \beta)(x_i - \bar{x})$

16

A becslések tulajdonságai

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

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

$E(a) \equiv E\left(\frac{\sum y_i}{n}\right) = \alpha$ torzítatlan

$Var(a) = \frac{\sum \sigma_y^2}{(n)^2} = \frac{\sigma_y^2}{n}$ konzisztens

$E(b) = \beta$ torzítatlan

$Var(b) = \frac{\sum (x_i - \bar{x})^2 \sigma_y^2}{\left(\sum (x_i - \bar{x})^2\right)^2} = \frac{\sigma_y^2}{\sum (x_i - \bar{x})^2}$ konzisztens

14

$\sum (y_i - Y_i)^2 = \sum (y_i - \hat{Y}_i)^2 + n(a_i - \alpha_i)^2 + (b - \beta)^2 \sum (x_i - \bar{x})^2$

ha $Y = \alpha + \beta(x - \bar{x})$

az eltérés forrása	eltérés-négyzetösszeg (SSQ)	df	SSQ várható értéke
a és α	$(a - \alpha)^2 n$	1	σ_y^2
b és β	$(b - \beta)^2 \sum (x_i - \bar{x})^2$	1	σ_y^2
a regressziós egyenes körül: y_i és $\hat{Y}_i$	$\sum (y_i - \hat{Y}_i)^2$	$n-2$	$(n-2)\sigma_y^2$
teljes	$\sum (y_i - Y_i)^2$	n	$n\sigma_y^2$

az egyetlen kiszámítható

17

$E(\hat{Y}) = E[a + b(x - \bar{x})] = \alpha + \beta(x - \bar{x}) = Y$

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

Hogyan lenne csökkenthető?

n, σ^2, x (allokáció)

15

$\sum (y_i - \hat{Y}_i)^2 = \chi^2 \sigma_y^2$ ha $Y = \alpha + \beta(x - \bar{x})$

(a valódi összefüggés egyenes)

$s_r^2 = \frac{\sum (y_i - \hat{Y}_i)^2}{n-2} = \frac{\chi^2 \sigma_y^2}{\nu} = s_y^2$ ha!

reziduális szórásnégyzet

$\hat{\sigma}_y^2 = s_y^2 = s_r^2$

18

Regresszió1

R-square: a modell magyarázta változás aránya

modell illeszkedési hiba (reziduum)

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

teljes

Statistic	Value
Multiple R	0.7302
Multiple R²	0.5332
Adjusted R²	0.5177
F(1,30)	34.2729
p	0.0000
Std.Err. of Estimate	273.0284

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

19

A becslések szórása

$$Var(a) = \frac{\sigma_y^2}{n} \qquad s_a = \frac{s_y}{\sqrt{n}}$$

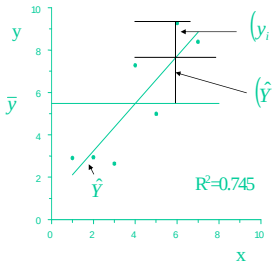
$$Var(b) = \frac{\sigma_y^2}{\sum (x_i - \bar{x})^2} \qquad s_b = \frac{s_y}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$\hat{\sigma}_y^2 = s_y^2 = s_r^2$$

22

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

teljes



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

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

20

$\hat{Y}_i = b_0 + bx_i$ b_0 és b **nem** függetlenek

$s_{b_0} = s_{\hat{Y}(x=0)} = \sqrt{s_a^2 + s_b^2 \bar{x}^2}$

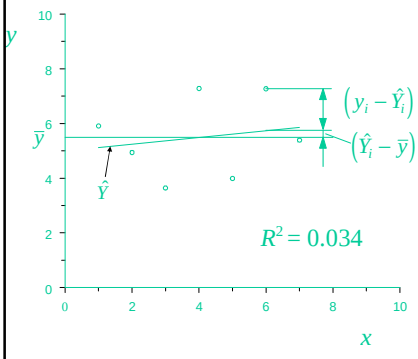
$t_0 = \frac{b_0 - 0}{s_{b_0}}$ nem független próba $t_0 = \frac{b-0}{s_b}$

(tengelymetszet) (iránytangens)

β_0 szignifikáns? β szignifikáns?

23

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



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

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

21

Konfidencia-sáv: az igazi egyenes egy pontjára

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

$$\hat{Y} - t_{\alpha/2} s_{\hat{Y}} < Y < \hat{Y} + t_{\alpha/2} s_{\hat{Y}}$$

Eshetnek kívül pontok?

24

Regresszió1

Jóslási sáv: egy újonnan mérendő *y** értékre

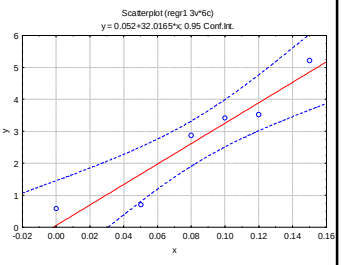
$$s_{y-\hat{Y}} = s_r\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}} = \sqrt{s_r^2+s_a^2+s_b^2(x-\bar{x})^2}$$

$$\hat{Y}-t_{\alpha/2}s_{y^*-\hat{Y}} < y^* < \hat{Y}+t_{\alpha/2}s_{y^*-\hat{Y}}$$

Konfidencia-sáv: az igazi egyenes egy pontjára

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

$$\hat{Y}-t_{\alpha/2}s_{\hat{Y}} < Y < \hat{Y}+t_{\alpha/2}s_{\hat{Y}}$$

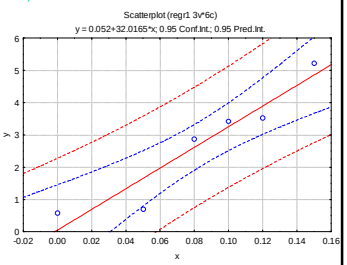


Eshetnek kívül pontok?

Jóslási sáv: egy újonnan mérendő *y** értékre

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

$$\hat{Y}-t_{\alpha/2}s_{y^*-\hat{Y}} < y^* < \hat{Y}+t_{\alpha/2}s_{y^*-\hat{Y}}$$



Eshetnek kívül pontok?

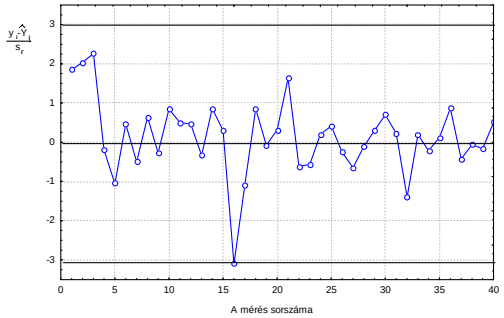
A regresszió feltételezéseinek ellenőrzése

- az illesztett függvény alkalmassága
- $Var(\varepsilon) = Var(y|x) = \sigma_y^2$ konstans
- a különböző *i* mérési pontokban elkövetett ε_i mérési hibák egymástól függetlenek
- *y* az *x* minden értékénél normális eloszlású, vagyis

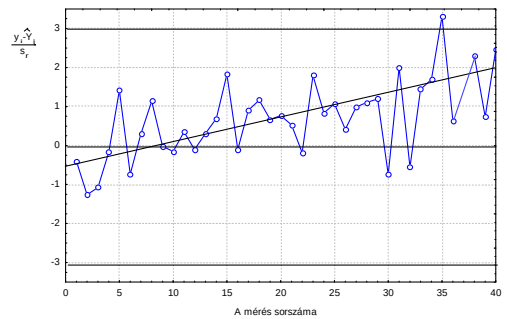
$$\varepsilon_i \sim N(0,\sigma^2)$$

a reziduumok ábrázolásával

A feltételek ellenőrzése a reziduumok ábrázolásával
Reziduumok a mérések sorszámának függvényében: extrém értékek

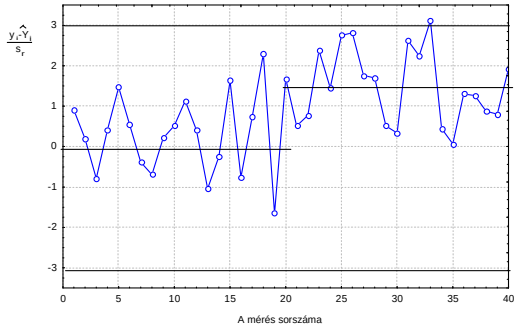


Reziduumok a mérések sorszámának függvényében: trend



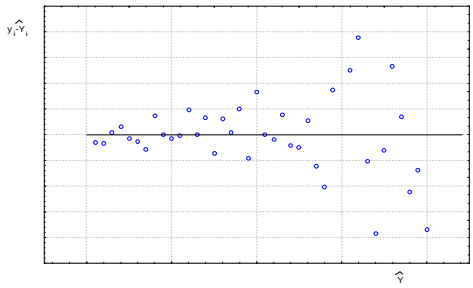
Regresszió1

Reziduumok a mérések sorszámának függvényében: ugrás



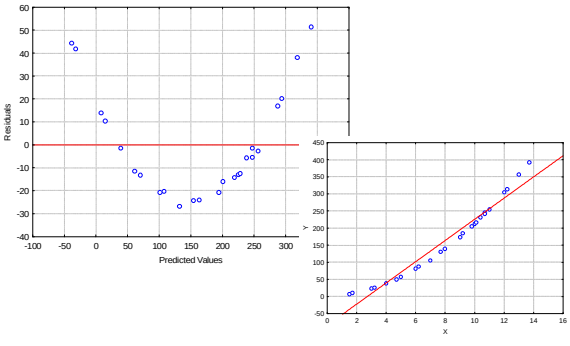
31

Reziduumok az *Y* függvényében: a szórás (variancia) változása



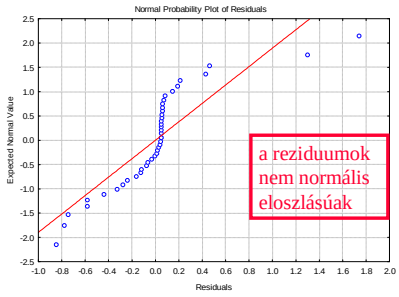
32

Reziduumok az *Y* függvényében:a függvény alkalmatlansága



33

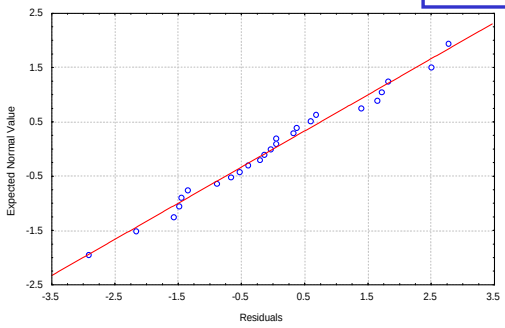
A reziduumok normális eloszlása



34

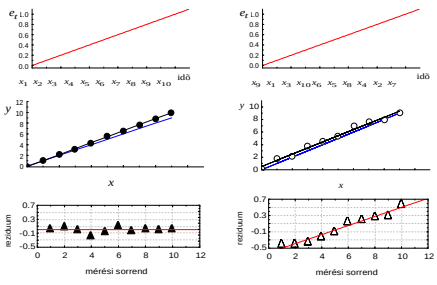
A reziduumok normális eloszlása

a reziduumok normális eloszlásúak



35

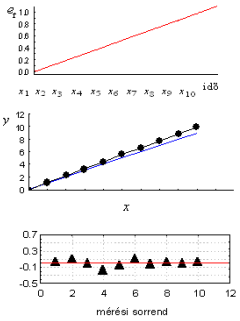
A mérések sorrendje



36

Regresszió1

Kézenfekvő, hogy *x* növekvő sorrendjében mérünk

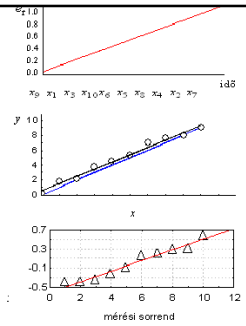


y az *x* hatása + a mérési ingadozás + az idő hatása (sorrend): az illesztett függvény az *x* és az idő hatásának összegét becsüli

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

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.

A növekvő *x* sorrendjében végzett mérés tehát olyan torzítást okoz, amit nincs módunk a reziduumok vizsgálatával észrevenni.

A véletlenszerű sorrendben végzett mérésnél is van torzítás (és a szórás is nagyobb lesz), de még egy jelenségről (az idő hatásáról) is értesülünk, amit az adatok korrekciójánál illetve a módszer fejlesztésénél felhasználhatunk.