Regresszióanalízis

Lineáris regresszió

Modell:

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

Y helyett y értéket mérünk, $E(y \mid x) = Y$, vagy

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

Amennyiben nincsen ismert és igazolt fizikai összefüggés, nem lehetünk elõre meggyõzõdve az illesztett függvény alkalmasságáról.

A regresszióanalízis során feltételezzük, hogy

- *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 különbözõ *i* mérési pontokban elkövetett mérési hibák egymástól függetlenek;
- $Y(x) = f(x, \alpha,\beta,\gamma, ...)$ az ismert vagy feltételezett függvénykapcsolat alakja, ahol α, β, γ a függvény konstansai (paraméterei).

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

A becslési kritérium:

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

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

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

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

A normálegyenletek:

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

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

Átrendezve:

$$\sum y_i = nb_0 + b\sum x_i$$

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

Ha $\sum x_i \neq 0$

a b_0 és b becslések egymástól nem függetlenek

A normálegyenletek az $Y_i = \alpha + \beta\left(x_i - \bar{x}\right)$ modell illesztésekor

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

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

Átrendezve:

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

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

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

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

Az a és b becslések egymástól függetlenek, mert

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

tehát az *a* és *b* becsült paraméterek egymástól függetlenül kaphatók meg a két normálegyenletbõl:

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

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

A becslések tulajdonságai:

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

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

$$E(b) = \beta$$

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

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

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

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

$$s_a = \frac{s_r}{\sqrt{n}} \qquad s_b = \frac{s_r}{\sqrt{\sum\left(x_i - \bar{x}\right)^2}}$$

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

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

A konfidenciatartományok a *t*-eloszlás alapján számíthatók.

1. példa

Kísérletileg vizsgálták az *x* független változó és az *y* függő változó közötti összefüggést. Az *x* független változó értéke pontosan beállítható, az *y* függő változó értéke azonban a *Y* valódi érték körül ingadozik. A mérési adatok a következő táblázatban láthatók, az *y* értéke szerint növekvő sorrendbe rendezve. A tényleges mérési sorrendet a táblázat második oszlopa tartalmazza. Feltételezve, hogy *y* normális eloszlású, valamint azt hogy az *y* és *x* közötti függvénykapcsolat lineáris, adjunk becslést az egyenes paramétereire!

No	mérési sorrend	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

Excel eredmények

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.95061604
R Square	0.90367086
Adjusted R Square	0.87958858
Standard Error	0.62135527
Observations	6

R^2

s_r reziduális szórás

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14.48747052	14.48747052	37.5243	0.003597945
Residual	4	1.544329481	0.38608237		
Total	5	16.0318			

s_r^2

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.05195755	0.504033217	0.103083577	0.922858	-1.347465911	1.451381
x	32.0165094	5.22658099	6.125708087	0.003598	17.50516417	46.527855

b_0

b

Determinációs együttható:

"Regression" "Residual"

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

"Total"

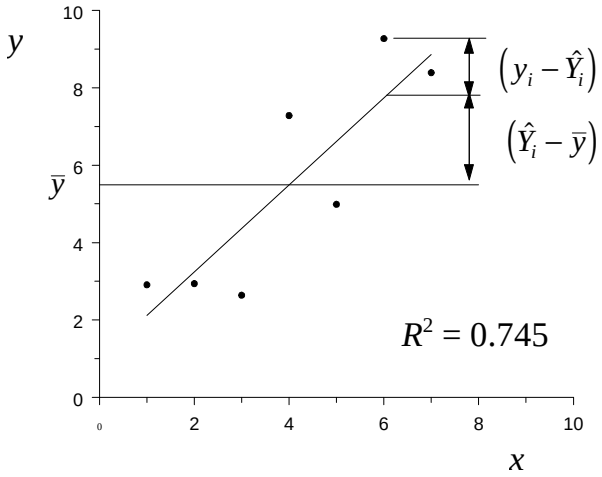
$$R^2\_adj = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$$

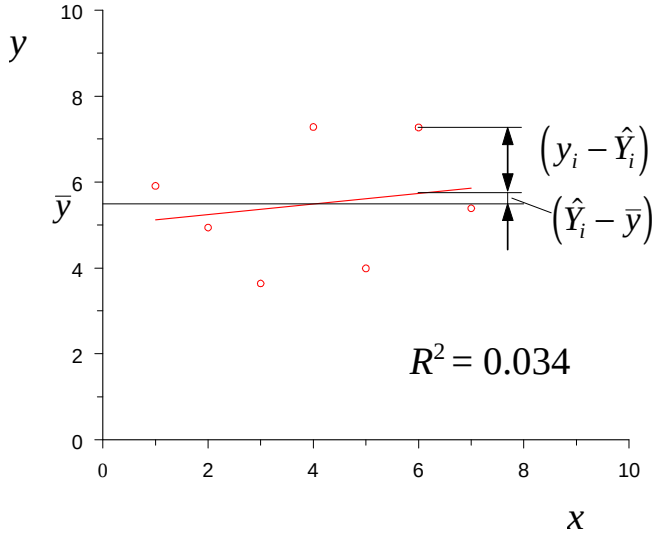
$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{y})^2$$

SST = SSE + SSR

d.f.: n-1 = n-2 + 1

R^2 = SSR/SST





ANOVA

	df	*SS*
Regression	1	14.48747052
Residual	4	1.544329481
Total	5	16.0318

SSR ***SSE*** ***SST***

***n* - 2**

$$s_r^2 = \frac{SSE}{n-2}$$

RESIDUAL OUTPUT

$\hat{Y}$ $y - \hat{Y}$ $\left(y - \hat{Y}\right)/s_r$

Observation	*Predicted y*	*Residuals*	*Standard Residuals*
1	0.05195755	0.528042453	0.849823731
2	1.65278302	-0.952783019	-1.53339493
3	2.6132783	0.266721698	0.429257965
4	3.25360849	0.166391509	0.26778804
5	3.89393868	-0.363938679	-0.585717539
6	4.85443396	0.355566038	0.572242734

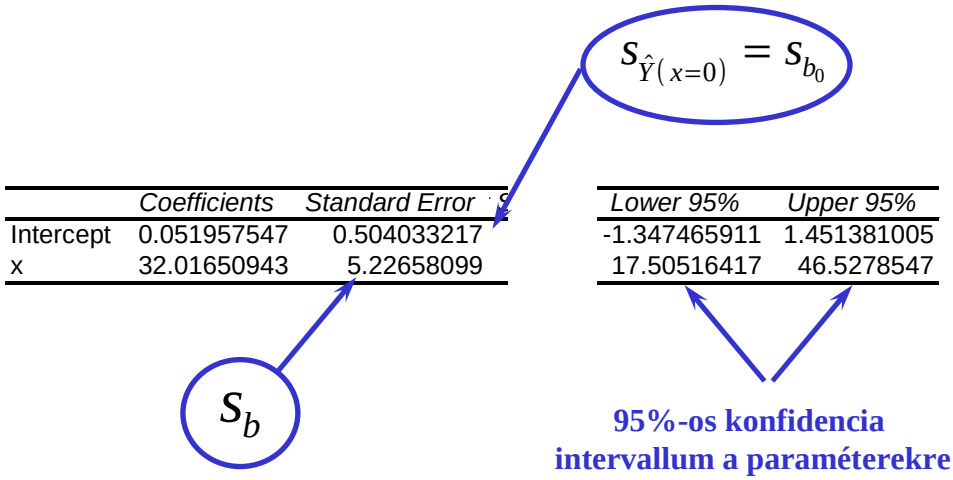
$$\sum_{i=1}^{n} (\text{Residual})^2 = \text{SSE}$$

$$s_a = \frac{s_r}{\sqrt{n}} \qquad s_b = \frac{s_r}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}}$$

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

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

A konfidenciatartományok a *t*-eloszlás alapján számíthatók.



Konfidencia sáv az *Y*(*x*) valódi értékre

$$\hat{Y}_{fölsõ} = \hat{Y} + t_{0.05/2}(4)s_{\hat{Y}}$$

$$\hat{Y}_{alsó} = \hat{Y} - t_{0.05/2}(4)s_{\hat{Y}}$$

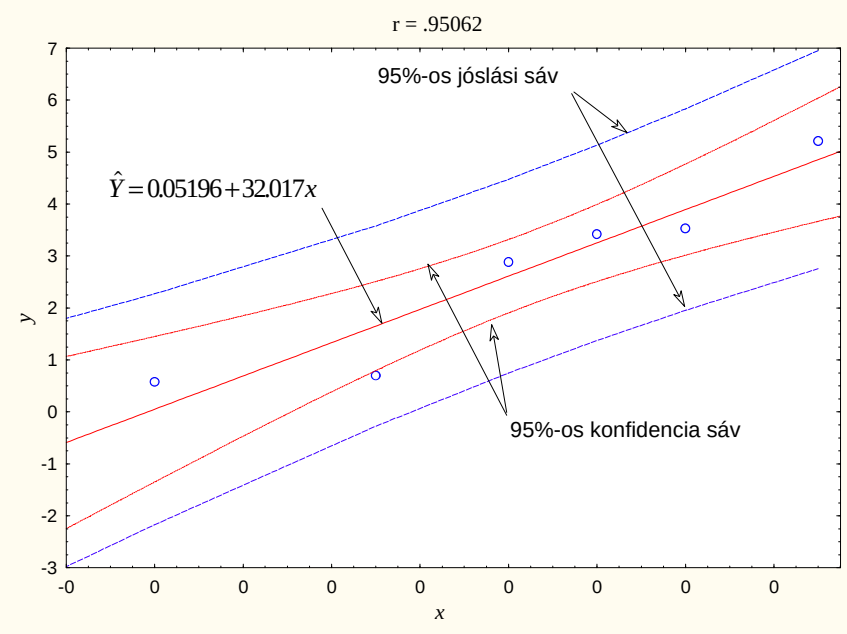
x	Yhat	s_Yhat	Yhat_alsó	Yhat_fölsõ
0	0.05	0.50	-1.35	1.45
0.05	1.65	0.31	0.80	2.51
0.08	2.61	0.25	1.91	3.32
0.1	3.25	0.27	2.51	4.00
0.12	3.89	0.32	3.01	4.78
0.15	4.85	0.43	3.66	6.05

Jóslási intervallum

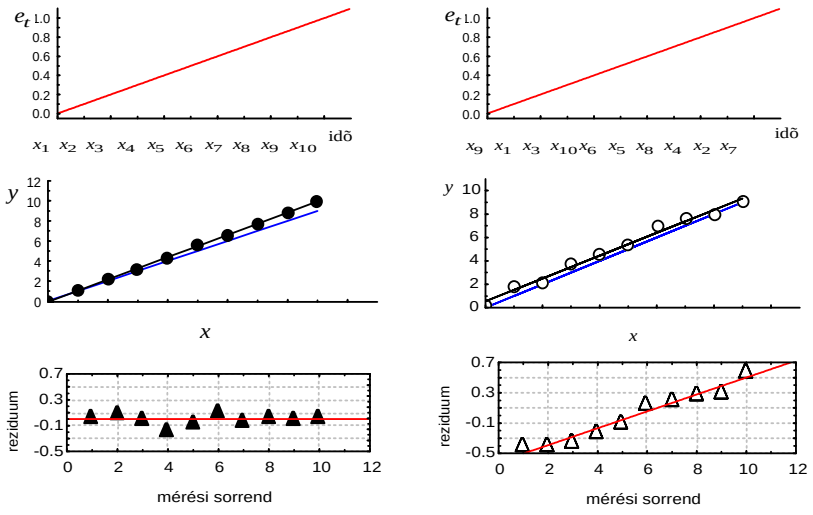
$$s_{y-\hat{Y}} = s_r\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}} = \sqrt{s_r^2 + s_a^2 + s_b^2(x-\bar{x})^2}$$

intervallum: $\hat{Y}(x) \pm t_{\alpha/2}s_{y-\hat{Y}}$

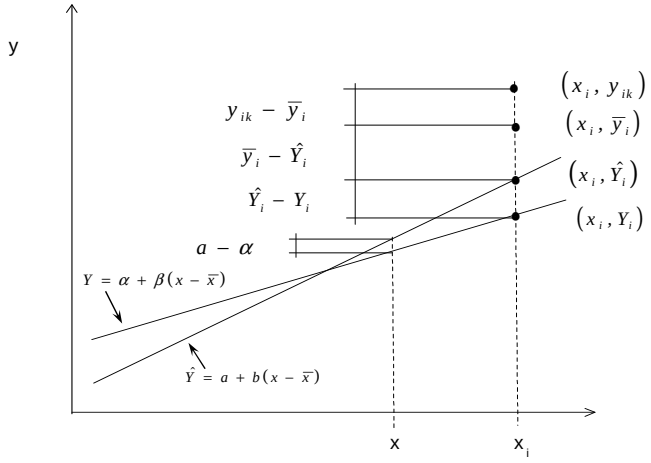
(1- α) a valószínűsége annak, hogy *x* adott értékénél egy későbbi mérés eredménye a számított intervallumba esik.



A mérések sorrendje



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



$$SST = SSE + SSR$$

$$SST = SSrepl + SSres + SSR$$

Ismétlésekbõl számított négyzetösszeg

Reziduális négyzetösszeg

A szabadsági fokok száma:

$$\sum_{i=1}^{n} p_i - 1 \quad = \quad \sum_{i=1}^{n} (p_i - 1) \quad + \quad n-2 \quad + \quad 1$$

$$s_e^2 = \frac{SSrepl}{\sum_i (p_i - 1)} \qquad s_r^2 = \frac{SSres}{n-2}$$

Az s_e^2 csoportokon belüli *error szórásnégyzet* a variancia torzítatlan becslése, függetlenül az *Y* függvény alakjától.

Az s_r^2 *reziduális szórásnégyzet* csak akkor becslése σ_y^2-nak, ha a tapasztalati regressziós függvény "megfelelõ alakú", vagyis az elméleti regressziós függvény lineáris. Esetünkben tehát akkor, ha $Y = \alpha + \beta(x - \bar{x})$.

A hipotézis vizsgálatára az *F*-próbát használjuk:

$$F = \frac{s_r^2}{s_e^2} = \frac{\chi_r^2 \sigma^2 / \nu_r}{\chi_e^2 \sigma^2 / \nu_e}$$

Ha az s_r^2 / s_e^2 arány nem halad meg egy F_α kritikus értéket, mondhatjuk, hogy a mérési adatok nem mondanak ellent annak a nullhipotézisnek, amely szerint az elméleti és tapasztalati regressziós görbe matematikailag azonos alakú.

Ha elfogadjuk a nullhipotézist, egyben azt állítjuk, hogy s_e^2 és s_r^2 egyaránt σ^2 torzítatlan becslései. A kettõ együtt több információt nyújt, mint bármelyik külön-külön, mivel az így egyesített szórásnégyzet nagyobb szabadsági fokú (tehát kisebb varianciájú) becslése σ^2 -nak, mint akár s_e^2, akár s_r^2 .
Célszerű tehát a két becslést egyesíteni.

$$\hat{\sigma}^2 = s^2 = \frac{s_e^2 \nu_e + s_r^2 \nu_r}{\nu_e + \nu_r} = \frac{\sum_i \sum_k \left(y_{ik} - \bar{y}_i\right)^2 + \sum_i p_i \left(\bar{y}_i - \hat{Y}_i\right)^2}{\left(\sum p_i - n\right) + (n-2)}$$

2. példa

Kalibrációs eljárás során a táblázatban közölt adatokat mérték, *x* a koncentráció, *y* a mért jel.
Illesszünk egyenest a mérési adatokra.

x_i	i	y_{ik} ha k					p_i
		1	2	3	4	5	
20	1	2.0046	2.1167	2.0059	2.1028	2.1053	5
14	2	1.5404	1.4737	1.5205	1.5372	1.4512	5
10	3	1.0043	1.0059	1.1068	1.0036	-	4
5	4	0.5756	0.6248	0.5701	0.6275	-	4
1.25	5	0.1952	0.2362	0.1954	0.2437	0.2455	5
						$\sum p_i =$	23

x	y
5	0.6248
20	2.1053
1.25	0.1954
14	1.5404
20	2.0059
1.25	0.1952
20	2.1167
5	0.5756
1.25	0.2362
5	0.6275
14	1.4512
1.25	0.2455
20	2.0046
5	0.5701
10	1.0059
10	1.0036
10	1.1068
1.25	0.2437
10	1.0043
20	2.1028
14	1.5205
14	1.5372
14	1.4737

Az adatok a mérési sorrendjében kerülnek be az input file-ba, tehát a programok számára általában ugyanaz az x - y adatok szerkezete, mint ismétlés nélküli mérések esetén.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.997696
R Square	0.995398
Adjusted R Square	0.995179
Standard Error	0.04772
Observations	23

$$s^2 = \frac{\mathrm{SSrepl} + \mathrm{SSres}}{\sum_i p_i - 2}$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10.34309	10.34309	4542.0869	4.98E-26
Residual	21	0.0478205	0.002277		
Total	22	10.39091			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.09427	0.0180397	5.225696	3.524E-05	0.056754	0.131786
x	0.098729	0.0014649	67.39501	4.985E-26	0.095682	0.101775

$$\bar{x} = \frac{5 \cdot 20 + 5 \cdot 14 + 4 \cdot 10 + 4 \cdot 5 + 5 \cdot 1.25}{23} = 10.2717$$

$$a = \frac{\sum p_i \bar{y}_i}{\sum p_i} = \frac{25.4933}{23} = 1.10839$$

$$b = \frac{\sum p_i \bar{y}_i \left(x_i - \bar{x} \right)}{\sum p_i \left(x_i - \bar{x} \right)^2} = \frac{104.7655}{1061.1141} = 0.09873$$

$$\hat{Y} = 1.10839 + 0.09873(x - 10.2717) = 0.0943 + 0.09873x$$

$$s_e^2 = \frac{\sum_i \sum_k \left(y_{ik} - \bar{y}_i \right)^2}{\sum_i p_i - n} = \frac{0.032587}{23 - 5} = 1.810 \cdot 10^{-3}$$

Annak ellenõrzésére, hogy az alkalmazott lineáris modell megfelelõ-e, F-próbát végzünk. Az Excel táblázat segítségével számítsuk ki a reziduális szórásnégyzetet, majd végezzük el a próbát!

$$\left(\sum_i p_i - 2 \right) s^2 = s_e^2 \left[\sum \left(p_i - 1 \right) \right] + (n - 2) s_r^2$$

$$21 \cdot 0.002277 = 18 \cdot 1.810 \cdot 10^{-3} + 3 s_r^2$$

$$s_r^2 = 5.070 \cdot 10^{-3}$$

$$F_0 = \frac{5.070 \cdot 10^{-3}}{1.810 \cdot 10^{-3}} = 2.804$$

Az *F*-eloszlás kritikus értéke 95 % -os egyoldali szinten ($\alpha = 0.05$), ha a számláló szabadsági foka 3, a nevezõé 18: $F_{0.05}(3, 18) = 3.16$.

Azt mondhatjuk, hogy a számított egyenes (a tapasztalati regressziós görbe) a mérési pontokat megfelelõen leírja.

$$s_b^2 = \frac{s_y^2}{\sum p_i \left(x_i - \bar{x}\right)^2} = \frac{2.277 \cdot 10^{-3}}{1061.11} = 2.146 \cdot 10^{-6}$$

$$s_b = 1.46 \cdot 10^{-3}$$

$$s_a^2 = \frac{s_y^2}{\sum p_i} = \frac{2.277 \cdot 10^{-3}}{23} = 9.901 \cdot 10^{-5}$$

$$s_a = 9.95 \cdot 10^{-3}$$

$$s_{\hat{Y}}^2 = s_a^2 + s_b^2 (x - \bar{x})^2 =$$
$$= 9.901 \cdot 10^{-5} + 2.146 \cdot 10^{-6} \cdot (x - 10.2717)^2$$

$$s_{\hat{Y}(x=0)}^2 = 9.901 \cdot 10^{-5} + 2.146 \cdot 10^{-6} \cdot 10.2717^2 = 3.254 \cdot 10^{-4}$$

$$s_{b_0} = s_{\hat{Y}(x=0)} = 0.01804$$

Egyváltozós lineáris regresszió ismételt mérések esetén, $\sigma_{y_i}^2$ nem konstans

A becslési kritérium: $$\sum_i \sum_k \left(\frac{y_{ik} - \hat{Y}_i}{\sigma_{y_i}} \right)^2 = min.$$

A négyzetösszeg felbontható:

$$\sum_i \sum_k \left(\frac{y_{ik} - \bar{y}_i}{\sigma_{y_i}} \right)^2 + \sum_i p_i \left(\frac{\bar{y}_i - \hat{Y}_i}{\sigma_{y_i}} \right)^2 = min.$$

$$\frac{\sigma_{y_i}^2}{p_i} = \sigma_{\bar{y}_i}^2$$

A variancia nem konstans, hanem *x*-nek ismert függvénye:

$$Var\left[y|x_i\right] = \sigma_{y_i}^2 = \sigma^2 h^2\left(x_i\right)$$

ahol σ^2 x -tõl független konstans.

A minimalizálandó függvény:

$$\sum_i p_i \frac{\left(\bar{y}_i - \hat{Y}_i\right)^2}{\sigma^2 h^2\left(x_i\right)} = \frac{1}{\sigma^2}\sum_i w_i p_i \left(\bar{y}_i - \hat{Y}_i\right)^2$$

$$\sum_i w_i p_i \left(\bar{y}_i - \hat{Y}_i\right)^2 = \sum_i w_i p_i \left[\bar{y}_i - a - b\left(x_i - \bar{x}\right)\right]^2 = min$$

ahol w_i az ún. súly: $w_i = \frac{\sigma^2}{\sigma_{y_i}^2} = \frac{1}{h^2\left(x_i\right)}$

Ha $\bar{x} = \frac{\sum w_i p_i x_i}{\sum w_i p_i}$

az *a* és *b* becsült paraméterek egymástól függetlenül kaphatók meg a két normálegyenletbõl:

$$a = \frac{\sum_i w_i p_i \bar{y}_i}{\sum_i w_i p_i} \qquad b = \frac{\sum_i w_i p_i \bar{y}_i \left(x_i - \bar{x}\right)}{\sum_i w_i p_i \left(x_i - \bar{x}\right)^2}$$

Kalibrációs egyenes:
a regressziós egyenlet megoldása a független változóra

Az egyenes egyenlete: $\hat{Y} = a + b(x - \bar{x})$

Most *y* a független, de sztochasztikus változó (ötször mérve 5 különbözõ abszorbanciát kapunk), *x* a függõ változó, amelynek becslése

$$\hat{x} = \hat{x}(y) = \bar{x} + \frac{y - a}{b}$$

várható értéke (és valódi értéke) *X*. (Az $\hat{x}$ becslés valószínûségi változó, mivel *y*, *a* és *b* valószínûségi változók.)

$\hat{x}$ konfidencia-intervalluma:

segédváltozó $z \equiv y - a - b(X - \bar{x})$

$$t = \frac{z - E(z)}{s_z} \qquad \left(\nu = \sum p_i - 2\right)$$

$$E(z) = Y - \alpha - \beta(X - \bar{x}) = 0$$

$$Var(z) = Var(y) + Var(a) + (X - \bar{x})^2 Var(b)$$

Ha *y* *n* mérés átlagértéke, értelemszerûen írandó *y* helyébe, és

$$Var(\bar{y}) = \frac{Var(y)}{n}$$

$$Var(z) = \sigma^2 \left[\frac{1}{wn} + \frac{1}{\sum w_i p_i} + \frac{(X - \bar{x})^2}{\sum w_i p_i (x_i - \bar{x})^2} \right]$$

Az s_z^2 becslést úgy kapjuk, hogy *Var(z)* elõbbi kifejezésében a *w* súlyok helyett beírjuk a $h^2(x)$ függvény reciprokának becslését, σ^2 becsléséül pedig az s^2-statisztikát használhatjuk.

$$P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha \quad ; \qquad t = \frac{z - 0}{s_z} = \frac{y - a - b(X - \bar{x})}{s_z}$$

Az *X*-re másodfokú kifejezés átrendezése után a konfidenciaintervallum

$$\bar{x} + \frac{y - a}{b_{\alpha/2}} - t_{\alpha/2} \frac{s}{b_{\alpha/2}} \sqrt{\left(\frac{1}{wn} + \frac{1}{\sum w_i p_i} \right) \frac{b_{\alpha/2}}{b} + \frac{(\hat{x} - \bar{x})^2}{\sum w_i p_i (x_i - \bar{x})^2}} < X <$$

$$< \bar{x} + \frac{y - a}{b_{\alpha/2}} + t_{\alpha/2} \frac{s}{b_{\alpha/2}} \sqrt{\left(\frac{1}{wn} + \frac{1}{\sum w_i p_i} \right) \frac{b_{\alpha/2}}{b} + \frac{(\hat{x} - \bar{x})^2}{\sum w_i p_i (x_i - \bar{x})^2}}$$

ahol

$$b_{\alpha/2} = b - \frac{t_{\alpha/2}^2 s_b^2}{b}$$

Az *X*-re másodfokú kifejezés átrendezése után a konfidenciaintervallum

$$P\left(\bar{x}+\frac{y-a}{b_{\alpha/2}}-\Delta<X<\bar{x}+\frac{y-a}{b_{\alpha/2}}+\Delta\right)=1-\alpha$$

ahol

$$\Delta=\frac{t_{\alpha/2}}{b_{\alpha/2}}\sqrt{\left(h^2(\hat{x})s^2+\frac{s^2}{\sum_i w_i p_i}\right)\frac{b_{\alpha/2}}{b}+(\hat{x}-\bar{x})^2\frac{s^2}{\sum_i w_i p_i\left(x_i-\bar{x}\right)^2}}$$

és $\quad b_{\alpha/2}=b\left(1-\frac{t_{\alpha/2}^2 s_b^2}{b^2}\right)$

s_a^2 -val és s_b^2-vel kifejezve

$$\Delta=\frac{t_{\alpha/2}}{b_{\alpha/2}}\sqrt{\left(h^2(\hat{x})s^2+s_a^2\right)\frac{b_{\alpha/2}}{b}+s_b^2(\hat{x}-\bar{x})^2}$$

Ha $b >> s_b$, $\quad b_{\alpha/2}\cong b$, így az elõzõ kifejezés egyszerûsödik

$$P(\hat{x}-\Delta<X\le\hat{x}+\Delta)=1-\alpha$$

ahol

$$\Delta=\frac{t_{\alpha/2}}{b}\sqrt{h^2(\hat{x})\frac{s^2}{n}+s_a^2+s_b^2(\hat{x}-\bar{x})^2}$$

Az összefüggések s_{b_0}, s_b, $\overline{x}$ felhasználásával, ha $b >> s_b$:

$$\hat{x} = \frac{y - b_0}{b}; \qquad P(\hat{x} - \Delta < X < \hat{x} - \Delta) = 1 - \alpha$$

ahol

$$\Delta = \frac{t_{\alpha/2}}{b}\sqrt{h^2(\hat{x})\frac{s_y^2}{n} + s_{b_0}^2 + s_b^2\left(\hat{x}^2 - 2\hat{x}\overline{x}\right)}$$

3. példa

A 2. példában kapott regressziós egyenest kalibrációs összefüggésként használjuk. Az ismeretlen koncentrációjú oldattal végzett 5 mérés átlagértéke 1.25. Adjunk becslést és 95 %-os konfidencia-intervallumot az oldat koncentrációjára (X-re).

$$\overline{y} = 1.25 \qquad n = 5$$

$$\hat{x} = \frac{\overline{y} - b_0}{b} = \frac{1.25 - 0.09427}{0.09873} = 11.706$$

$$s_b^2 = 2.146 \cdot 10^{-6}; \qquad t_{0.05/2}(21) = 2.080; \qquad \frac{t_{\alpha/2}^2 s_b^2}{b} = 9.52 \cdot 10^{-4}$$

$$b_{\alpha/2} = b - \frac{t_{\alpha/2}^2 s_b^2}{b} = 0.09864 \qquad \frac{b_{\alpha/2}}{b} \approx 1 \qquad h^2(x) = 1$$

s_{b_0}, s_b, $\overline{x}$ felhasználásával:

$$\Delta = \frac{2.080}{0.09873}\sqrt{\frac{2.277 \cdot 10^{-3}}{5} + (0.01804)^2 + \left(1.465 \cdot 10^{-3}\right)\left(11.7061^2 - 2 \cdot 11.7061 \cdot 10.2717\right)}$$

$$\Delta = 1.028$$

A konfidencia-intervallum:

$$P(11.706 - 1.028 < X < 11.706 + 1.028) = 0.95$$

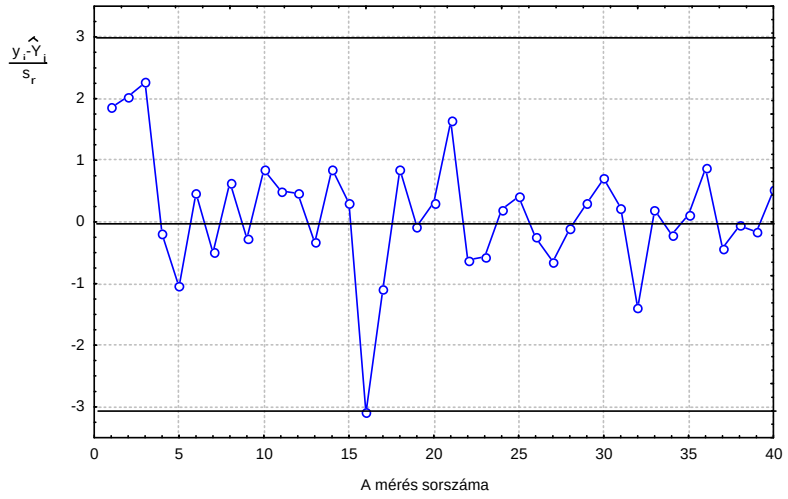
$$P(10.7 < X < 11.7) = 0.95$$

A regresszió feltételeinek ellenõrzése; a reziduumok vizsgálata

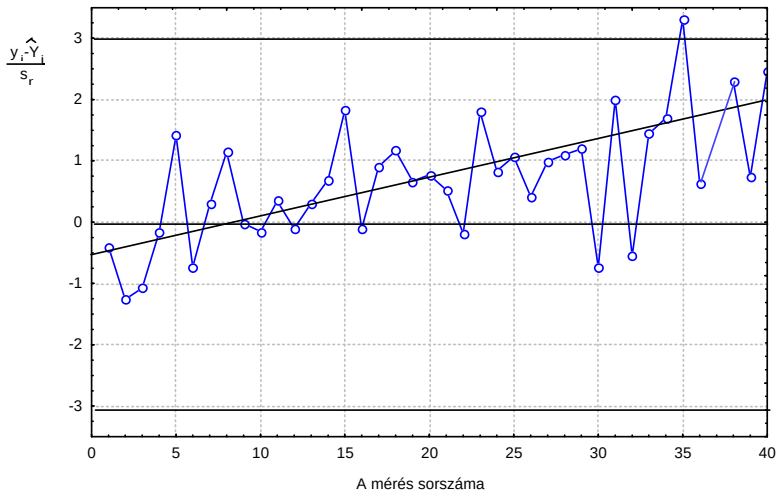
A regresszióanalízis során feltételeztük, hogy

- *y* az *x* minden értékénél normális eloszlású, vagyis az ε mérési hibák $N(0,\sigma^2)$ normális eloszlásúak;
- Var(*y*) = Var(*y* | *x*) = konstans, illetve *y*-nak vagy *x*-nek ismert függvénye;
- a különbözõ *i* mérési pontokban elkövetett mérési hibák egymástól függetlenek;
- E(*y* | *x*) = *Y*(*x*) = *f*(*x*, α,β,γ, ...) az ismert vagy feltételezett függvénykapcsolat alakja, ahol α, β, γ a függvény konstansai (paraméterei).

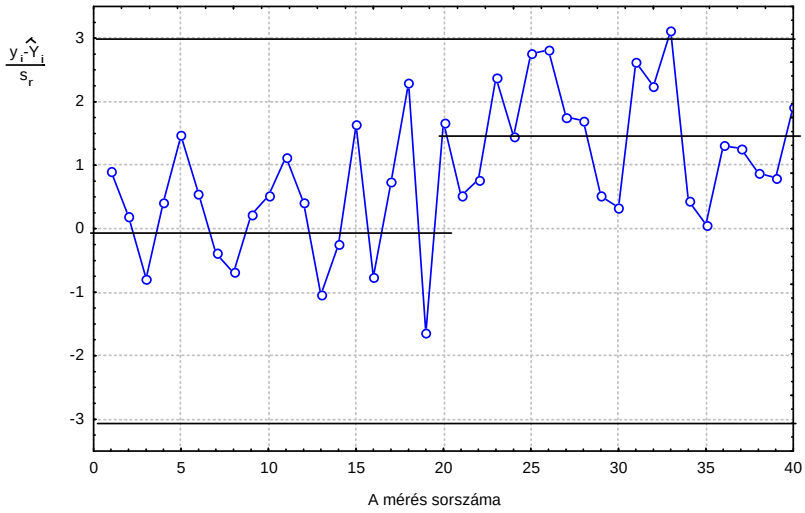
1. **Reziduumok a mérések sorszámának függvényében: extrém értékek**



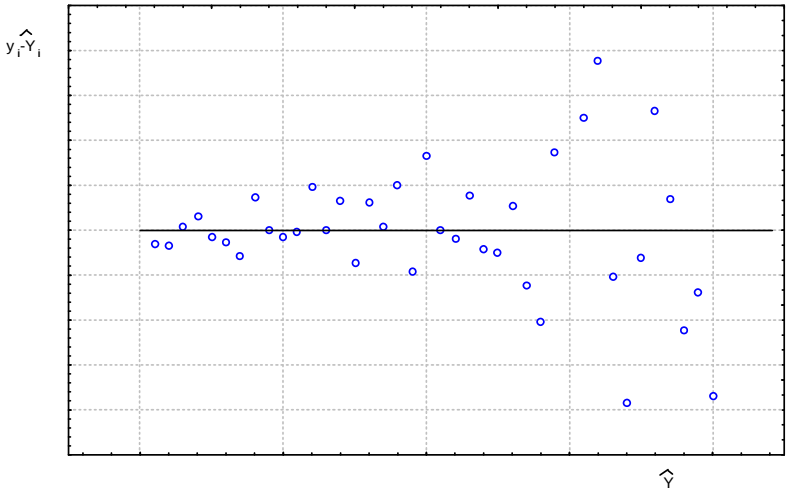
2. **Reziduumok a mérések sorszámának függvényében: trend**



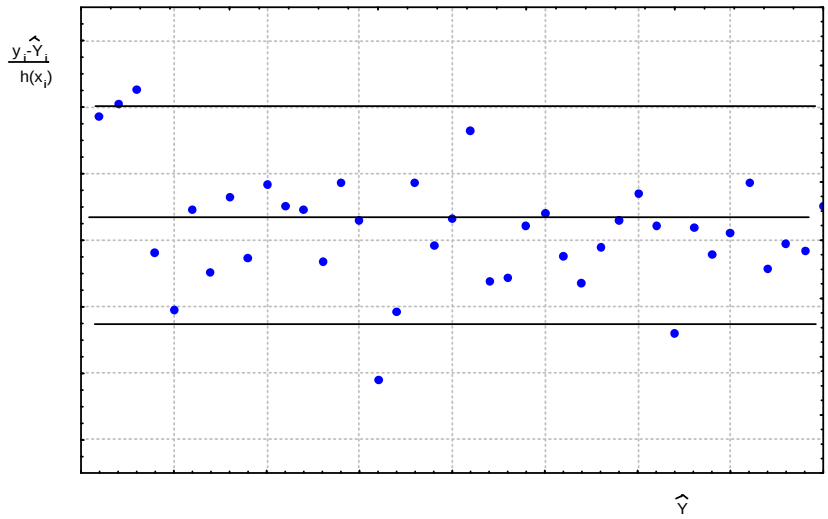
3. Ugrás (Szintváltozás a reziduumok vizsgálatánál)



4. A szórás (variancia, mérési pontosság) változása



A $h^2(x)$ függvény megfelelõen írja le változását:



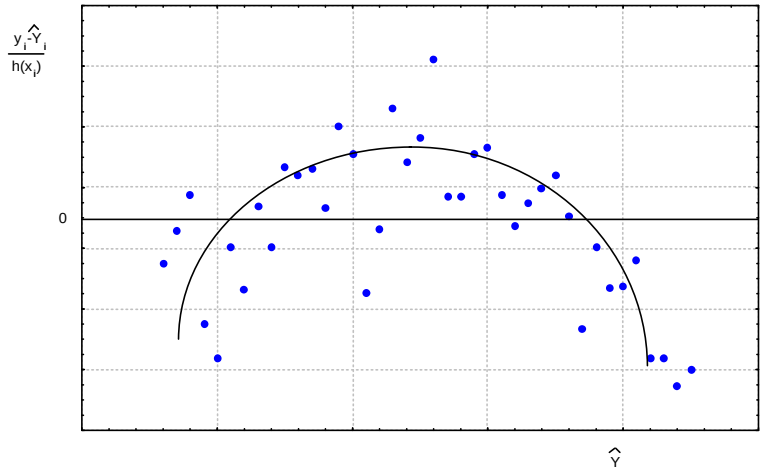
5. Normalitás

Az $\frac{y_i - \hat{Y}_i}{h(x)}$ közelítõleg zérus várható értékû normális eloszlású kell legyen az 1…4. feltételezések szerint.

A normalitást statisztikai próbával vizsgálhatjuk (χ^2 -próba, Kolmogorov – Szmirnov próba).

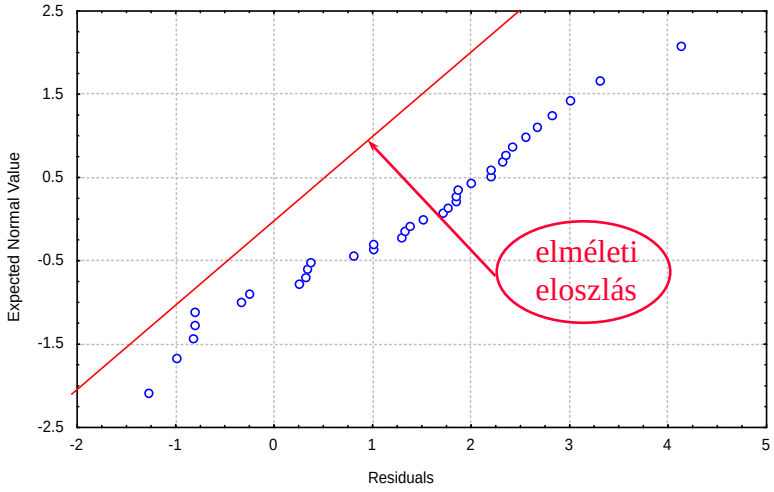
A normalitást úgy is vizsgálhatjuk, hogy ún. valószínûségi papíron (Gauss hálón) ábrázoljuk $\frac{y_i - \hat{Y}_i}{h(x)}$ értékét

A reziduumok eloszlása nem normális, az illesztett modell nem megfelelõ:



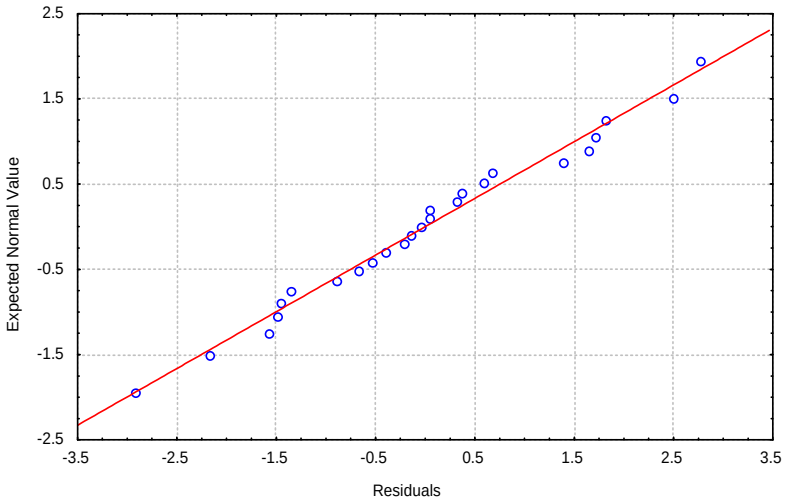
A reziduum értékek ábrázolása Gauss-hálón.

a reziduumok nem normális eloszlásúak



a reziduumok normális eloszlásúak

A reziduum értékek ábrázolása Gauss-hálón.



Kétváltozós lineáris regresszió

Az elméleti regressziós függvény:

$$Y = \alpha + \beta_1\left(x_1 - \bar{x}_1\right) + \beta_2\left(x_2 - \bar{x}_2\right)$$

A becslési kritérium:

$$\phi = \sum_i \left(y_i - \hat{Y}_i\right)^2 = \sum_i \left[y_i - a - b_1\left(x_{1i} - \bar{x}_1\right) - b_2\left(x_{2i} - \bar{x}_2\right)\right]^2 = \min.$$

A becsülendõ paraméterek szerint deriválva, és a deriváltakat nullával egyenlõvé téve kapjuk a normálegyenleteket:

$$na + b_1\sum\left(x_{1i} - \bar{x}_1\right) + b_2\sum\left(x_{2i} - \bar{x}_2\right) = \sum y_i$$

$$a\sum\left(x_{1i} - \bar{x}_1\right) + b_1\sum\left(x_{1i} - \bar{x}_1\right)^2 + b_2\sum\left(x_{1i} - \bar{x}_1\right)\left(x_{2i} - \bar{x}_2\right) = \sum y_i\left(x_{1i} - \bar{x}_1\right)$$

$$a\sum\left(x_{2i} - \bar{x}_2\right) + b_1\sum\left(x_{1i} - \bar{x}_1\right)\left(x_{2i} - \bar{x}_2\right) + b_2\sum\left(x_{2i} - \bar{x}_2\right)^2 = \sum y_i\left(x_{2i} - \bar{x}_2\right)$$

A becsült paraméterek akkor függetlenek egymástól, ha

$$\sum_i\left(x_{1i} - \bar{x}_1\right) = 0 \;;\quad \sum_i\left(x_{2i} - \bar{x}_2\right) = 0 \;;$$

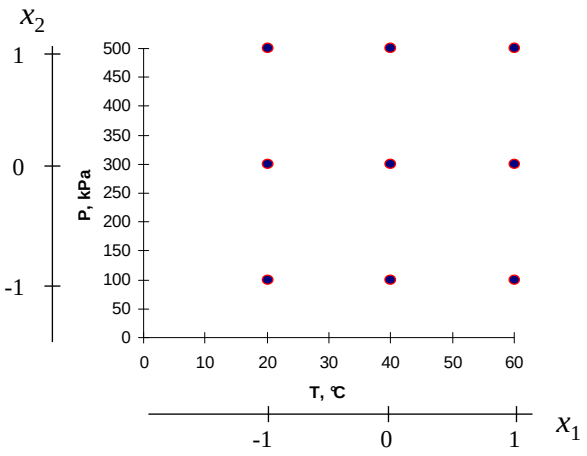
és

$$\sum_i\left(x_{1i} - \bar{x}_1\right)\left(x_{2i} - \bar{x}_2\right) = 0$$

ortogonális kísérleti terv

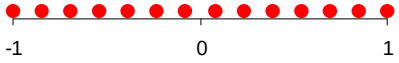
Szempontok a független változók értékeinek megválasztásához

Egymástól független becsült paraméterek (ortogonalitás)



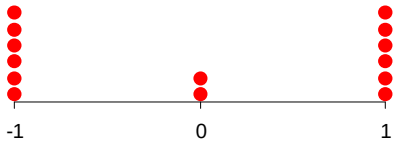
A paraméter minél pontosabb becslése

a)



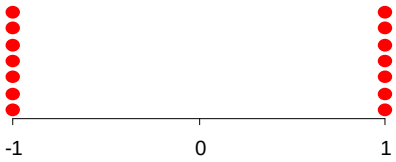
$$\frac{s_b}{\sigma} = 0.43$$

b)



$$\frac{s_b}{\sigma} = 0.29$$

c)



$$\frac{s_b}{\sigma} = 0.27$$

Többváltozós lineáris regresszió

Legyen r a független változók száma. A kísérletsorozat eredményeit a következő táblázatos formában szokásos írni:

$$\begin{array}{cccccccc} x_{11} & x_{21} & \cdots & x_{j1} & \cdots & x_{r1} & & y_1 \\ x_{12} & x_{22} & \cdots & x_{j2} & \cdots & x_{r2} & & y_2 \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{1i} & x_{2i} & \cdots & x_{ji} & \cdots & x_{ri} & & y_r \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{jn} & \cdots & x_{rn} & & y_n \end{array}$$

A modell

$$Y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_r x_{ri}$$

ahol x_{0i} az általános írásmód érdekében bevezetett fiktív változó.

Az x_{0i} elemek értéke 1.

A tapasztalati regressziós egyenes

$$\hat{Y}_i = b_0 x_{0i} + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_r x_{ri}$$

A kétváltozós regressziónál mondottakhoz hasonlóan a b_j becslések egymástól nem függetlenek.

Az egyes változók szignifikanciájának vizsgálata

Eldöntendõ, hogy $q < r$ változó figyelembevétele r változóhoz képest nem rontja-e a közelítést.

A q ill. r számú változóra a mért pontok és a becsült sík közötti eltérések négyzetösszege, ha minden i pontban csak egy y mérés van:

$$S_q = \sum_i \left(y_i - \sum_{j=0}^{q} b_{jq} x_{ji} \right)^2 \qquad S_r = \sum_i \left(y_i - \sum_{j=0}^{r} b_{jr} x_{ji} \right)^2$$

$\hat{Y}(q)$ $\hat{Y}(r)$

Tegyük fel, hogy r változó biztosan elég (hibátlan a regressziós egyenlet alakja), ekkor az

$$\left[y_i - \hat{Y}_i(r)\right]$$

eltérések normális eloszlásúak, σ_y^2 (konstansnak feltételezett) varianciával; az eltérések S_r négyzetösszegének szabadsági foka n-(r+1)

Ha q változó is elég (H_0 nullhipotézis), az $\left[y_i - \hat{Y}_i(q)\right]$ eltérések is normális eloszlásúak, σ_y^2 varianciával; az eltérések S_q négyzetösszegének szabadsági foka n-(q+1)

Ha a nullhipotézis igaz, az

$$F_0 = \frac{s_q^2}{s_r^2} = \frac{S_q / (n-q-1)}{S_r / (n-r-1)}$$

hányados F-eloszlású $n - q - 1$ és $n - r - 1$ szabadsági fokkal.

F-próba

S_q és S_r különbsége szintén normális eloszlású eltérések négyzetösszege, szabadsági foka $r - q$:

$$F_0 = \frac{s^2_{r-q}}{s^2_r} = \frac{\left(S_q - S_r\right)/\left(r-q\right)}{S_r/\left(n-r-1\right)}$$

F-próba

Bármelyik módszerrel elvégezhetõ az *F*-próba, a második érzékenyebb (**általános regressziós próba**).

Ha az arány a kritikus *F* értéket meghaladja, el kell vetnünk a nullhipotézist, amely szerint $r - q$ változó hatása nem szignifikáns.
Természetesen $r - q = 1$ is lehet, ekkor azt vizsgáljuk, hogy adott egyetlen változó hatásának (lineáris) figyelembevétele javítja-e a közelítést.

Minthogy a becslések egymástól nem függetlenek, az elõbbi vizsgálat *t*-próbával nem végezhetõ el.

Ha a normális eloszlás feltételezése nem jogos, az itt leírt vizsgálati módszer hamis eredményeket ad!

Regresszió más, a független változóban nemlineáris, de a paraméterekben lineáris függvényekkel

$$Y = \beta_0 + \beta_1 z + \beta_2 exp\left(-\frac{z^2}{2}\right) + \beta_3 log(z)$$

Vezessük be a következõ jelöléseket:

$x_1 = z$ $\quad x_2 = exp\left(-\frac{z^2}{2}\right)$ $\quad x_3 = log(z)$

Ezekkel $Y = \sum_j \beta_j x_j$

A becslési probléma és az eredmények statisztikai elemzése teljesen azonos a többváltozós lineáris regresszionál leírtakkal.

Polinom illesztése

Legyenek olyan mérési adataink, amelyeknél az *y* függõ változó nem lineáris, hanem polinommal leírható függvénye a *z* független változónak.

Mivel a *z* független változó értéke pontosan beállítható és nem terheli mérési hiba, tetszõleges hatványa is pontosan ismert, tehát determinisztikus független változóként kezelhetõ.

Bevezetve az $x_1 = z$, $x_2 = z^2$, ..., $x_k = z^k$ jelöléseket, a feladat a többváltozós lineáris regresszióra vezethetõ vissza.

$$\hat{Y} = b_0 + b_1 z + b_2 z^2 + ... + b_k z^k = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

Mivel x_j értékek nem függetlenek egymástól, a becsült b_j együtthatók erõsen korreláltak lesznek.