6. Variancia analízis

Több minta szórásnégyzetének (varianciájának) összehasonlításán alapul a statisztika egyik nagy fejezete, a variancia analízis. A vizsgálatok célja ennek alkalmazásakor ugyanaz, mint a két mintára kiterjedõ statisztikai próbáké volt: sokaságok egyezésének vagy eltérésének valószínûsítése. Meg kell jegyezni, hogy amikor a mintaszórások eltérés-valószínûségét F próbával határozzuk meg, a mintaelemek legyenek függetlenek és normális eloszlásúak.

A módszer lényegét számos variancia analitikus eljárás közül a legegyszerûbbön, az *"egytényezõs" variancia elemzésen* mutatjuk be.

Több sokasággal foglalkozunk, amelyekrõl feltesszük, hogy $N(\mu_i,\sigma^2)$ eloszlásúak, ahol μ_i az *i*-edik sokaság várható értéke, σ^2 pedig a sokaságok *megegyezõ* varianciája. Kérdésünk az, a minták elkerülhetetlen eltérése véletlen-e, avagy érvényesült valami olyan hatás, aminek alapján a sokaságok nem tekinthetõk megegyezõnek.

Szabatosan a

$$H_0 : \mu_I = \mu \qquad i = 1,2, \ldots m \qquad (6.1)$$

nullahipotézis elfogadásáról vagy elvetésérõl van szó. Vegyük észre, hogy különös módon most *szórások* összehasonlításával *középértékek* eltérésérõl ítélkezünk.

Az eljárást példán mutatjuk be. Tegyük fel, abban kívánunk dönteni, hogy három, L1, L2 és L3 laboratórium egyenlõ megbízhatóan dolgozik-e, avagy a laboratóriumokból érkezõ eredményeket fenntartással kell fogadni. A vizsgálathoz a három laboratórium 1,5 tömeg% ként tartalmazó gázolajat kap, amelyet ugyanazzal a (megegyezõ szórású) szabványos módszerrel kell megvizsgálnia. L1 labor $n_1 = 3$ párhuzamos mérést végez, L2 labor $n_2 = 5$-öt, L3 $n_3 = 4$-et. A beküldött eredményeket a 6.1 táblázatban bemutatott elrendezésû táblázatba foglaljuk. Itt x_{ij} jelenti a *j*-edik laboratórium *i*-edik mérési értékét. (A számszerû értékek a 6.3 táblázatban találhatók).

6.1 táblázat. A variancia analízis alapadatai

	j = 1	j = 2	j = 3	Sorösszegek /átlagok
i = 1	x_{11}	x_{12}	x_{13}	
i = 2	x_{21}	x_{22}	x_{23}	
i = 3	x_{31}	x_{32}	x_{33}	
i = 4		x_{42}	x_{43}	
i = 5		x_{52}		
Oszlopösszeg	$\sum_{i=1}^{n1} x_{i1}$	$\sum_{i=1}^{n2} x_{i2}$	$\sum_{i=1}^{n3} x_{i3}$	$\sum_{j=1}^{m}\sum_{i=1}^{nj} x_{ij}$
Elemszám	n_1	n_2	n_3	$n = \sum_{j=1}^{3} n_j$
Szab.fok	n_1 -1	n_2 -1	n_3 -1	n -3
Átlag	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x} = \frac{1}{n}\sum_{j=1}^{3}\sum_{i=1}^{nj} x_{ij}$
Eltérésnégyzet-összeg	$\sum_{i=1}^{3}(x_{i1} - \bar{x}_1)^2$	$\sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2$	$\sum_{i=1}^{4}(x_{3j} - \bar{x}_3)^2$	$\sum_{j=1}^{m}\sum_{i=1}^{nj}(x_{ij} - \bar{x}_j)^2 = \mathrm{SS}_{\mathrm{intra}}$
Variancia	$\frac{\sum_{i=1}^{3}(x_{i1} - \bar{x}_1)^2}{n_1 - 1}$	$\frac{\sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2}{n_2 - 1}$	$\frac{\sum_{i=1}^{4}(x_{3j} - \bar{x}_3)^2}{n_3 - 1}$	$\frac{1}{n-3}\sum_{j=1}^{m}\sum_{i=1}^{nj}(x_{ij} - \bar{x}_j)^2 = \mathrm{MS}_{\mathrm{intra}}$
Átlag csoportok között	$\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{nj} x_{ij}$			
Eltérésnégyzet-összeg csoportok között	$SS_{\mathrm{int}\,er} = \sum_{j=1}^{m} n_j\left(\bar{x}_j - \bar{x}\right)^2$			
Variancia csoportok között	$\mathrm{MS}_{\mathrm{int}\,er} = \frac{1}{m-1}\sum_{j=1}^{m} n_j\left(\bar{x}_j - \bar{x}\right)^2$			

A táblázatban látható két variancia, az MS_{intra} és MS_{inter} érték közül az elsõ a kénmeghatározó módszer szórását, véletlen hibáját becsli. A második a laborok középértékeinek eltérését tükrözi azok közös középértékétõl. Belátható, hogy ha a középértékek egymástól jobban eltérnek, mint amennyit a módszer szórása megenged, akkor a laboratóriumok között szignifikáns eltérés van. A döntés az MS_{intra} és MS_{inter} varianciák F próbáján alapul. Ha a kapott F nagyobb, mint a kritikus $F(\alpha,\nu_1,\nu_2)$ érték, akkor a (6.1) nullahipotézist elvetjük.

A variancia analízisnek ezeket a lépéseit a 6.2 táblázat mutatja.

6.2 táblázat. A variancia analízis erdményei

	SS	Szab.fok	MS	$\hat{F}$	*p*
Csoportokon belül	SS_{intra}	ν_{intra}	MS_{intra}		
Csoportok között	SS_{inter}	ν_{inter}	MS_{inter}	--	--
Összesen	SS_{total}	ν_{total}	--	--	--

A táblázat legalsó sorában az

$$SS_{intra} + SS_{inter} = SS_{total} = \sum_{j=1}^{m}\sum_{i=1}^{nj}\left(x_{ij} - \bar{x}\right)^2 \qquad (6.2)$$

egyenlõségnek kell (matematikai okokból) teljesülnie. Ez hasznos ellenõrzési lehetõség. Ugyanez áll a szabadsági fokokra is. A táblázatban szereplõ *p* érték azt adja meg, hogy a kapott $\hat{F}$ hányadosnál *nagyobb* értékek mely valószínûséggel fordulnak elõ. A statisztikus döntést nyilván ennek alapján is meg lehet hozni. A variancia analízis algoritmusai azonban legtöbbször kérik az α tévedési valószínûséget és megadják F kritikus értékét.

A bevezetésben bemutatott példa számszerû eredményeit a 6.3 táblázat mutatja be.

6.3 táblázat. Egytényezõs variancia analízis

	L1	L2	L3	sorösszeg	Sorátlag
x1.	1,5	1,6	1,3		
x2.	1,55	1,72	1,3		
x3.	1,47	1,4	1,4		
x4.		1,48	1,45		
x5.		1,55			
x_{ij} összegek	4,52	7,75	5,45	17,72	
Mérésszámok	3	5	4	12	
Szab.fokok	2	4	3	9	
Átlagok	1,506666667	1,55	1,3625		1,476666
Eltérésnégyzet-összeg	0,0032666	0,0588	0,016875	0,0789416	
Varianciák	0,0016333	0,0147	0,005625		0,008771
VARIANCIA ANALÍZIS					
Tényezõk	SS	df	MS	$\hat{F}$	p-érték
Csoportok között	0,081725	2	0,0408625	4,658661459	0,040851902
Csoporton belül	0,078941667	9	0,008771296		
Összesen	0,160666667	11			

A kritikus F érték 5% tévedést megengedve, egyoldalas kérdésfeltevésnél 4.256 lenne. Ennél a kapott F érték nagyobb, így a nullahipotézist, miszerint a laboratóriumok egyformán dolgoznak elvetjük. *p* értékbõl látjuk, hogy a döntés nem módfelett biztos, hiszen, ha "igazságosabbak" akarunk lenni, és csak 3% tévedést vállalnánk, a laboratoriumokat már nem tartanók különbözõnek.

7. Összefüggések vizsgálata

A matematikai statisztika eddig tárgyalt fejezetei többnyire *egy* valószínûségi változóval foglalkoztak, ha pedig többel, akkor is feltételezték azok egymástól való függetlenségét. Nyilvánvaló ugyanakkor, hogy vizsgált rendszereket leíró, azokra ható változók közötti összefüggések elsõrendû fontosságúak.

Az összefüggések többféle szempontból tárgyalhatók, pl. abból, hogy okságiak-e, avagy nem azok, hogy van-e róluk elõzetes ismeretünk avagy csak tapasztalati leírásunk stb. A továbbiakban aszerint tegyünk különbséget, hogy két (vagy több) *valószínûségi változó* összefüggésével kell foglalkozni, avagy nem valószínûségi *(determinisztikus) változók* hatnak egy valószínûségi változóra. Ez utóbbi változó legtöbbször azért tekintendõ valószínûségi változónak, mert pontos, valódi értékét rátelepedett hiba terheli:

$$Y = Y_{valódi} + \varepsilon \qquad (7.1)$$

Ebben az esetben Y eloszlása megegyezik ε eloszlásával, annyi különbséggel, hogy ha ε várható értéke 0, akkor Y-é $Y_{valódi}$.

7.1 Valószínûségi változó függése determinisztikus változó(k)tól.

Bizonyos, hogy a gyakorlatban változót hibátlanul nem lehet mérni vagy beállítani, elvben tehát determinisztikus változó nincs. Az azonban mégis elfogadható, hogy egyes független változók két-három nagyságrenddel pontosabbak, mint a függõ, így nem okoz nagy hibát, ha azokat determinisztikusnak tekintjük.

7.1.1 A legkisebb négyzetek elve

Akár valószínûségi változó függ determinisztikus változó(k)tól, akár determinisztikus, szükségünk van egy matematikai összefüggésre (modellre), amelyik a függést leírja. Jelöljük a modellt F-fel. Ilyen modell lehet egy origón áthaladó vagy általános helyzetû egyenes, egy exponenciális függvény stb. A modellnek állandói (konstansai, paraméterei) vannak (meredekség, tengelymetszet) és független változói. Legyen az elöbbiek jele *a*1, *a*2,… utóbbiaké *x*1, *x*2…, de egyszerûség kedvéért tekintsünk most egyetlen *x*-t. A független változókat "prediktoroknak" vagy "regresszoroknak" is nevezik. A számított (jósolt, predikált) érték jel legyen *y*.

A modell tehát teljesen általánosan így fest:

$Y = F(x1,x2,\ldots a1,a2\ldots)$

Ha a paraméterek ismertek, beállított független változóknál y kiszámítható. A feladatot azonban általában meg kell elõzze a paraméterek meghatározása (*becslése*), egyfajta "kalibráció". Ismert *x* értékeknél párhuzamos kisérletekben meghatározzunk *y* mért értékeket, és *a* paramétereket tekintjük ismeretleneknek. Ha a mérések pontosak lennének, bármelyikbõl

ki lehetne számítani az ismeretlen *a*1, *a*2, …paramétereket. A kapott *y* értékeket azonban ismeretlen hibával mérjük:

$$\begin{aligned} &F(x_1,a1,a2,...) = y_1 + \varepsilon_1 \\ &F(x_2,a1,a2,\ldots) = y_2 + \varepsilon_2 \\ &\ldots \\ &F(x_n,a1,a2,\ldots) = y_n + \varepsilon_n \end{aligned} \qquad (7.2)$$

ezért a (7.2) egyenletekbõl mérésrõl mérésre más *a* paraméterek adódnának. Megállapodás szerint azokat az *a*1, *a*2, … értékeket fogadjuk el optimálisnak, amelyeknél a mért és számított értékek különbségnégyzeteinek összege minimális:

$$Q = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \min \qquad (7.3)$$

$\hat{y}$ az F(*x,a,b*) modellel számított értéket jelöli.

Ez a követelmény a legkisebb négyzetek elve. Gyakorlati alkalmazására a következõ fejezet ad példát.

Az ε_i hibákról nemcsak azt szokták feltételezni, hogy várható értékük 0, hanem azt is, szórásuk megegyezik. Ez az un. *homoszkedasztikus* eset. Ha ugyanis a mérési hibák *x* változó mentén változnak (*heteroszkedasztikus* eset), a fellépõ nagy eltérések (azok négyzetei) aránytalanul eltorzítják a minimum helyét, ezzel a paraméterek értékét. Ilyen esetben az eltéréseket *súlyozni* szokás, amivel a minimum követelmény így alakul:

$$\sum_{i=1}^{n} w_i \left(y_i - F(x_i, a1, a2,...) \right)^2 = \min \qquad (7.4)$$

A súly általában az adott *x* változóértéknél érvényes variancia reciproka:

$$w_i = \frac{1}{s_i^2} \qquad (7.5)$$

A súlyozott legkisebb négyzetek módszerére más esetekben, pl. az *y* változóra alkalmazott transzformáció miatt is szükség lehet.

7.1.2 Egyenes paramétereinek becslése (lineáris regresszió)

a) Az egyenes állandói

A lineáris regressziónál az egyenes ismert egyenletének érvényességét tételezzük fel:

$$y = F(x, a, b) = a + bx \qquad (7.6)$$

A paraméterek becslésére *n* darab x_i, y_i értékpárt használunk fel. A becslés gondolatmenetének megfelelõen minimálni kell a mért és számított y értékek eltérése négyzetének összegét :

$$Q = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2 = \min \qquad (7.7)$$

($\hat{y}$ az F(*x,a,b*) modellel számított értéket jelöli. A további összefüggésekben az egyszerûség kedvéért a szummázás jelénél az i indexet elhagyjuk, sõt, ahol nem zavaró, az index a változók mellõl is hiányzik).

Az *a* és *b* paraméterek függvényében *Q* négyzetösszeg nyilvánvalóan ott lesz minimális, ahol *Q*-nak *a* és *b* szerinti parciális deriváltjai 0 értékûek lesznek. Fenn kell tehát állnia, hogy

$$\frac{\partial Q}{\partial a} = -2 \sum \left(y_i - a - bx_i \right) = 0 \qquad (7.8)$$

$$\frac{\partial Q}{\partial b} = -2 \sum \left(y_i - a - bx_i \right) x_i = 0 \,. \qquad (7.9)$$

A kapott egyenleteket egyszerûsítve, az összegezéseket tagonként végrehajtva és azokat rendezve az

$$a\ n \quad + \quad b \sum x = \sum y \qquad (7.10)$$

$$a \sum x + b \sum x^2 = \sum xy \qquad (7.11)$$

lineáris egyenletrendszer adódik, amelybõl megoldás után a meredekségre a

$$\hat{b} = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \left(\sum x \right)^2} \qquad (7.12)$$

összefüggés, a tengelymetszetre pedig (7.9) egyenlet *n*-nel való osztása után az

$$\hat{a} = \bar{y} - b\bar{x} \qquad (7.13)$$

képlet adódik.

(7.12) képlet könnyen számítható tényezõket tartalmaz. Mind számlálója, mind nevezõje aritmetikai mûveletekkel átalakítható úgy is, hogy a képlet jobban megjegyezhetõ és többet mondó alakú legyen:

$$\hat{b} = \frac{\sum \left(x_i - \bar{x} \right) \left(y_i - \bar{y} \right)}{\sum \left(x_i - \bar{x} \right)^2} \qquad (7.14)$$

A paraméterbecslés ismertetett elve és a (7.10), (7.11).egyenletek *többváltozós* lineáris összefüggések paramétereinek becslésére is általánosíthatók. Az

$$y = a_0 + a_1 x1 + a_2 x2 + \ldots + a_m xm \qquad (7.15)$$

lineáris modell paramétereinek becslései *n* darab mérésbõl az *m* + 1 ismeretlenes

$$\begin{aligned}
&a_0 n + a_1 \sum x1_i + a_2 \sum x2_i + \ldots + a_m \sum xm_i = \sum y_i \\
&a_0 \sum x1_i + a_1 \sum x1_i^2 + a_2 \sum x1_i x2_i + \ldots + a_m \sum x1_i xm_i = \sum x1_i y_i \\
&a_0 \sum x2_i + a_1 \sum x1_i x2_i + a_2 \sum x2_i^2 + \ldots + a_m \sum x2_i xm_i = \sum x2_i y_i \\
&\ldots \\
&a_0 \sum xm_i + a_1 \sum x1_i xm_i + a_2 \sum x2_i xm_i + \ldots + a_m \sum xm_i^2 = \sum xm_i y_i
\end{aligned} \quad (7.16)$$

lineáris egyenletrendszer megoldásával lehet megkapni.

b) Szórásbecslések

Tételezzük fel, hogy az *y* mennyiség valóban lineáris függvénye *x* független változónak. Ebben az esetben csak a mérési hiba az oka annak, hogy a mért pontok nem esnek pontosan a becsült egyenesre. Ebbõl következik, hogy ebben az esetben az

$$s = \sqrt{\frac{\sum \left(y - \hat{y} \right)^2}{n-2}} \quad (7.17)$$

mennyiség a *mérési hiba becslése*. (A szabadsági fok azért *n* – 2, mert az egyenes két állandója két megkötést jelent az *y* mért értékek között). Ez a becslés egyébként

$$\begin{aligned}
s &= \sqrt{\frac{1}{n-2}\left(\sum \left(y - \bar{y} \right)^2 - \frac{\left[\sum \left(x - \bar{x} \right)\left(y - \bar{y} \right) \right]^2}{\sum \left(x - \bar{x} \right)^2} \right)} = \\
&= \sqrt{\frac{1}{n-2}\left(\sum \left(y - \bar{y} \right)^2 - b \sum \left(x - \bar{x} \right)\left(y - \bar{y} \right) \right)}
\end{aligned} \quad (7.18)$$

módon is számítható.

Felvetõdik ezekután a *paraméterek* és a kapott paraméterekkel *számitott* $\hat{y}$ értékek szórásának becslése. Ezek rendre a következõk:

$$s_a = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum \left(x - \bar{x} \right)^2}} \quad (7.19)$$

$$s_b = s \sqrt{\frac{1}{\sum \left(x - \bar{x} \right)^2}} \quad (7.20)$$

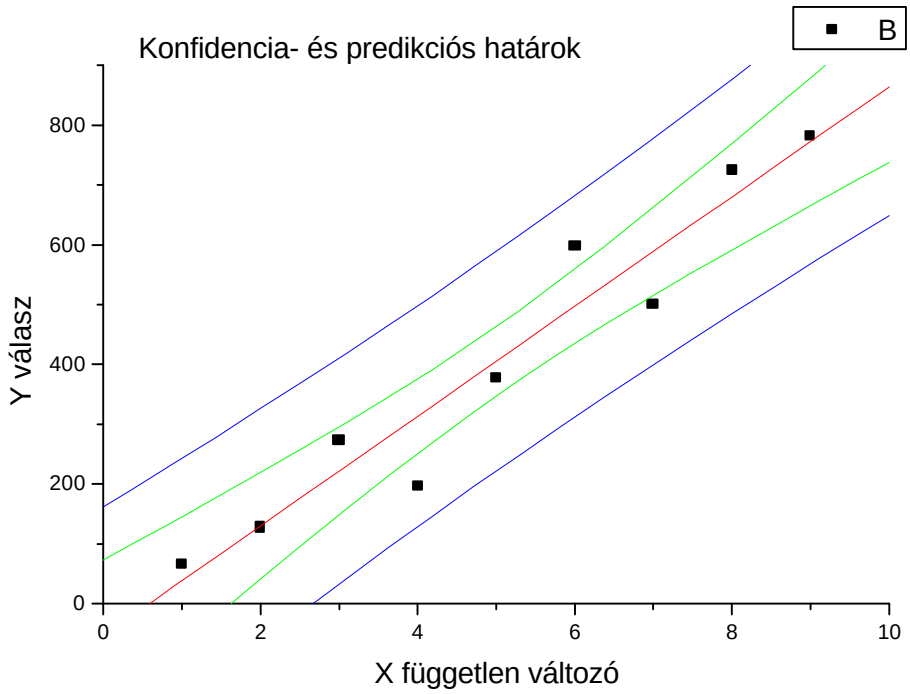
Normális eloszlású *y* értékek esetén adott valószínûségû konfidencia tartományt is megadhatunk a paraméterekhez.

$$\begin{aligned}
&\hat{a} \pm t_{n-2,\alpha} s_a \\
&\hat{b} \pm t_{n-2,\alpha} s_b
\end{aligned} \quad (7.21$$

A becsült paraméterekkel bármely x* változóhoz kiszámítható egy $\hat{y}$ érték várható értékének szórása:

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{\sum\left(x - \bar{x}\right)^2}} \qquad (7.22)$$

Ez az *x**-tól függõ mennyiség a regressziós egyenes felett és alatt megadja a konfidencia "övet", tájékoztat arról, milyen határok között mozog a regressziós egyenes 1 – α valószínûséggel.(7.1 ábra). Mint látható, a konfidencia tartomány az egyenes "súlypontjában" (az $x^* = \bar{x}$ pontban) legkeskenyebb és az egyenes két széle felé nõ.



7.1 ábra. Regressziós egyenes és a megbízhatósági övek. Fekete négyzetek: mért értékek. Egyenes: regresszós egyenes. Belsõ öv: konfidencia határok, külsó öv: predikciós határok.

Bármely *x** változónál *jövõben mért y* várható helyének bizonytalansága nagyobb. A *jóslási (predikciós)* szórás:

$$s_{y\cdot} = s\sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{\sum\left(x - \bar{x}\right)^2}} \qquad (7.23),$$

amit a 7.1 ábrán a külsõ öv mutat meg.

c) Az illesztés jósága

A 7.2 ábrát tanulmányozva látható, hogy hogy az $y_i - \hat{y}_i$ távolság nem az egyetlen, amelyiket definiálni lehet. Beszélhetünk az $y_i - \bar{y}$ távolságról, és az $\hat{y}_i - \bar{y}$ távolságokról is. Belátható, hogy az $\hat{y}_i - \bar{y}$ távolságok

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad (7.24)$$

négyzetösszege ill. az azokból számított *"modell okozta szórás"* arról vall, miért térnek el az *y* mért értékek átlaguktól amiatt, hogy azok *x* függvényei. Az is érthetõ, hogy ha ez a szórás összemérhetõ a kisérleti szórást becslõ, $y_i - \hat{y}_i$ különbségek

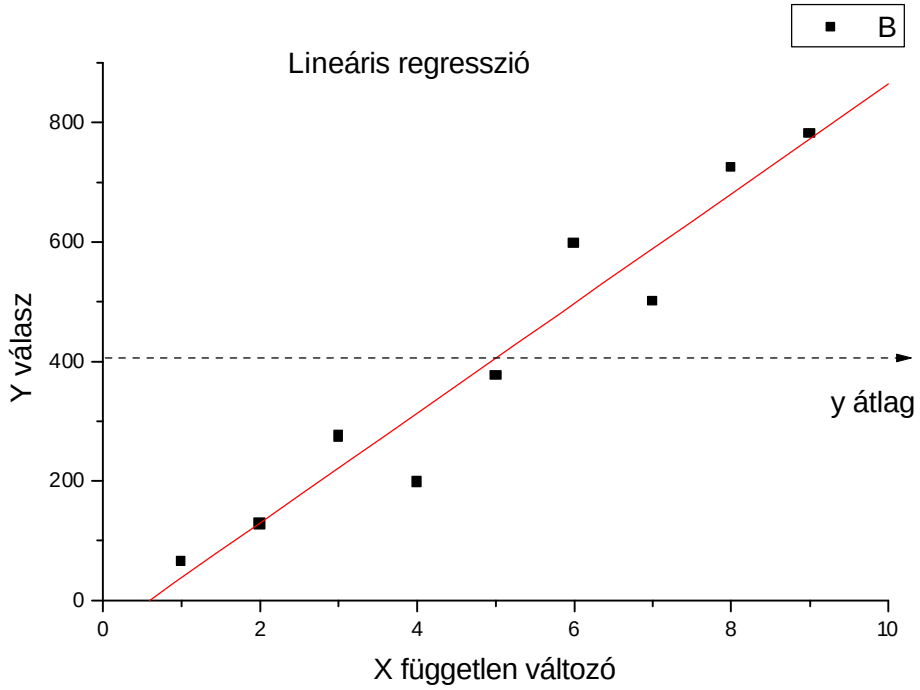
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (7.25)$$

négyzetösszegébõl számított *"reziduális szórással"*, akkor kétséges a függés léte. Ezért számos esetben a regressziós számítást variancia analízis (l. 6. pont) követi, ahol ennek a két szórásnak négyzetét (varianciáját) F próbával hasonlítják össze. Minél nagyobb F, annál biztosabb a függés. Megjegyezhetõ, hogy a *modell okozta* eltérésnégyzetösszeg és a *reziduális* eltérésnégyzetösszeg kiadja az $y_i - \bar{y}$ távolságok

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (7.26)$$

totális négyzetösszegét, mert

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \qquad (7.27)$$



7.2 ábra. A regresszió megítéléséhez

A regresszió jóságát szokás a modell okozta eltérésnégyzetösszeg (7.24) és a teljes (totális) eltérésnégyzetösszeg (7.26) hányadosával is jellemezni.

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7.28}$$

Ez a mennyiség azt adja meg, hogy az *y* értékek *x*-menti változásának hányadrésze tulajdonítható a lineárisnak tekintett függésnek. Az r^2 hányados az r korrelációs együttható (l. 7.2 pont) négyzete. Nem túl érzékeny mutató. Ha a mért pontok valamennyien pontosan az egyenesen vannak, értéke 1, de szemmelláthatóan szóró és nem is lineárisan függõ mért értékeknél is viszonylag magas (0.9 feletti) lehet.

7.1.3 Nemlineáris paraméterbecslés

Nemlineáris (pl. hatványfüggvénnyel leírható, reciprokos, exponenciális) össszefüggések paramétereinek becslésére a legkisebb négyzetek módszere ugyancsak alkalmazható. Az optimális paramétereket megadó

$$\sum_{i=1}^{n} (y_i - F(x_i, a1, a2, ...))^2 = \min \tag{7.29}$$

kritériumban az F függvény *nemlineáris* , így a deriválás után (ha az lehetséges) kapott (7.8), (7.9) összefüggésekre emlékeztetõ egyenletek nem lineárisak és megoldásukhoz a numerikus matematika erre alkalmas módszerei használhatók.

Egyes esetekben nem kell ehhez a nehéz eljáráshoz folyamodni. Az

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \qquad (7.30)$$

függvényben például helyettesíthetünk. Legyen $x1 = x, \quad x2 = x^2, \quad x3 = x^3$. Ezzel a (7.30) egyváltozós összefüggés az

$$y = a_0 + a_1 x1 + a_2 x2 + a_3 x3 \qquad (7.31)$$

háromváltozós, ám lineáris fügvénnyé alakult, amelyikbõl a lineáris regresszió szabályai szerint a keresett paraméterek meghatározhatók. (l. a 7.16 egyenletrendszert). Hasonlóan lehet eljárni pl. az $y = a_o + a_1 \ln x$ modell esetén is.

Függvények gyakran úgy linearizálhatók, hogy a transzformáció az *y* függõ változót is érinti. Az

$$y = ae^{-bx} \qquad (7.32)$$

összefüggés mindkét oldalát logaritmizálva az

$$\ln y = \ln a - bx \qquad (7.33)$$

lineáris függvényhez jutunk, amelynek paraméterei lineáris regresszióval becsülhetõk.

Az

$$\ln k = c - a / T \qquad (7.34)$$

összefüggés $y = \ln k$ és $x = 1/T$ helyettesítéssel lineárissá alakítható.

A mért értékeket érintõ átalakításoknál azonban figyelni kell arra a következményre, hogy az eredetileg (alkalmasint) egyenlõ nagyságú hibák a transzformációk után eltérõkké, sõt esetleg aszimmetrikusakká válnak, igy a becsült paraméterek torzítottak lehetnek és hibáikról gondos munka után lehet nyilatkozni. A *súlyozott legkisebb négyzetek módszerének* alkalmazása mindenképpen indokolt.

7.2 Valószínûségi változók összefüggése

Két valószínûségi változó ugyancsak összefügghet. Az összefüggés abban nyilvánul meg, hogy az egyik változó növekedése vagy csökkenése együttjár a másik változó csökkenésével vagy növekedésével. Ezt az összefüggést a kovariancia méri:

$$C(X,Y) = E[(X - E(X)) (Y - E(Y))] \qquad (7.35)$$

Ez a mennyiség pozitív, ha az X és Y valószínûségi változók együtt mozognak, negatív, ha ellentétesen. Szokás a kovarianciát a két változó szórásával osztva a –1 és +1 határok közé szorítani. A kapott mennyiség a korrelációs együttható:

$$\rho = \frac{C(X,Y)}{\sigma_X \sigma_Y} \tag{7.36}$$

Minél inkább megközelíti ρ a +1 vagy –1 értéket, annál szorosabb a két változó közötti összefüggés. Ha el is éri, a két változó egymás lineáris függvénye. Ha a korrelációs együttható 0, akkor a két változó *korrelálatlan*. Ha két változó független, akkor korrelálatlan is. Fordítva a megállapítás nem érvényes. Attól, hogy ρ 0-értékû, a két változó között még lehet függvénykapcsolat. Kivételt ez alól a normális eloszlású változók képviselnek.

A korrelációs együtthatót mintából az

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{\sum xy - \sum x \sum y / n}{\sqrt{\left(\sum x^2 - \left(\sum x\right)^2 / n\right)\left(\sum y^2 - \left(\sum y\right)^2 / n\right)}} \tag{7.37}$$

statisztika becsli.

Ha a korrelációs együttható szignifikáns 0 értékét akarjuk megvizsgálni, a

$$\hat{t} = |r| \sqrt{\frac{n-2}{1-r^2}} \tag{7.38}$$

értéket kell kiszámítani. Ha ez nagyobb, mint $t_{n-2,\,\alpha}$, akkor a $H_0 : r = 0$ nullahipotézist a tévedési valószínûséggel elvetjük.

Korrelált valószínûségi változók lineáris összefüggését lehet regresszióval vizsgálni. A regressziós egyenes azonban más helyzetû, ha Y-t X függvényében, avagy X-et Y függvényében vizsgáljuk.

Általában fennáll, hogy a korreláció nem jelent szükségképpen oksági kapcsolatot. A függõ és független változó fogalma ebben a környezetben gyakran értelmezhetetlen.