II. Lineáris regresszió

4. A regresszióanalízis alapjai; egyváltozós lineáris regresszió

4.1. A regresszióanalízis alapjai

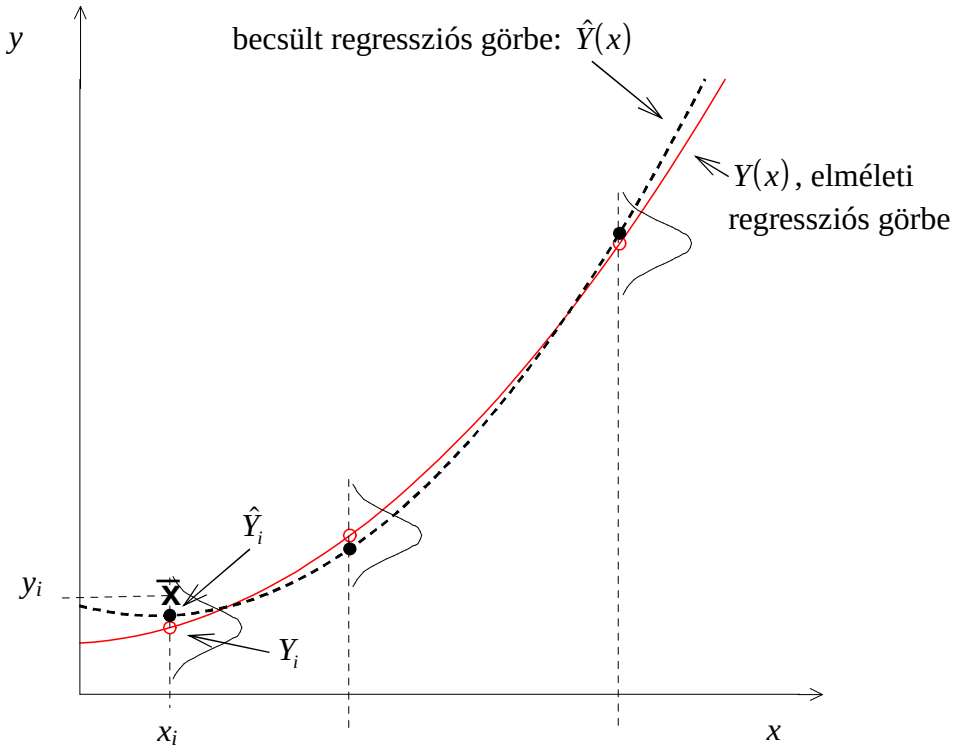
A regresszióanalízis feladatai a következők:

a) a függvénykapcsolat (az $Y(x)$ elméleti regressziós függvény) paramétereinek becslése,

b) ha lehetséges, a függvény alkalmasságára vonatkozó hipotézis vizsgálata,

c) a paraméterekre vonatkozó hipotézisek vizsgálata (pl. az elméleti regressziós egyenes átmegy-e az origón, ill. meredeksége szignifikánsan különbözik-e zérustól),

d) konfidenciaintervallum, ill. konfidenciasáv számítása a függvény paramétereire és az $\hat{Y}(x)$ tapasztalati vagy empirikus regressziós görbére (a becsült függvényre).



4-1. ábra. Egyedi mért érték, elméleti regressziós függvény és becsült regressziós görbe kapcsolata

Az, hogy a függvény konstansait paramétereknek nevezzük, összhangban van az első fejezetben említett paraméterfogalommal, lévén ezek is a sokaságnak a tulajdonságai,

a becsült paraméterek pedig a mérési adatokat (a mintát) leíró függvény (a tapasztalati regressziós görbe) konstansai, és így a minta jellemzői.

Az illesztett függvényről alkotott elképzelés kétféle lehet:

a) görbét (interpolációs formulát) kívánunk illeszteni, amely jól kezelhető, és a szükséges pontossággal reprezentálja a mérési adatokat,

b) a változók közötti oksági összefüggést leíró modellt illesztünk, amelynek paraméterei fizikai értelemmel bírnak, így extrapolációra is használhatók.

A regresszió technikája a két esetben azonos, de a becsült paraméterekre vonatkozó statisztikai vizsgálatoknak elsősorban fizikai tartalom szempontjából megalapozott modellillesztés esetén van értelmük.

Ebben a fejezetben az *x* független változókat determinisztikus (nem valószínűségi) változóknak tekintjük, azaz a független változók értékeit a kísérletező választhatja meg, és pontosan ismeri azokat.

Vizsgáljuk a következő modellt: valamely (pl. fizikai) törvényszerűség értelmében az *x* független változó bizonyos értékénél a függő változó értéke $Y = \varphi(x)$.

A mérés során mérési pontatlanságok vagy az egyéb, a függvénykapcsolatban nem szereplő, de a jelenséget befolyásoló hatások (pl. nem elég pontosan állandóan tartott hőmérséklet, ajtócsapkodás, légáram stb.) miatt *Y* helyett valamely *y* értéket mérünk, amelyre, ha az ingadozások véletlenszerűek (azonos valószínűséggel pozitív vagy negatív irányúak), és igen kis hatásúak, igaz, hogy $E(y|x) = Y$, vagy $y = Y + \varepsilon$, ahol ε a hiba és $E(\varepsilon) = 0$. Általában feltehető, hogy *y* eloszlása *Y* körül normális eloszlás, varianciája:

$$Var(y|x) = \sigma_y^2 = Var(\varepsilon). \qquad (4.1)$$

Amennyiben nincsen ismert és igazolt fizikai összefüggés a változók között (pontosabban a változók várható értékei között), hanem éppen ilyet keresünk, vagy csak a mérési eredményeket leíró, esetleg minden kauzális meggondolás nélkül alkotott függvénykapcsolatot keresünk (approximáció), a meggondolások ugyanúgy érvényesek, azzal a különbséggel, hogy nem lehetünk előre meggyőződve a függvény alkalmasságáról.

A regresszióanalízis során feltételezzük, hogy

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

Ha a feltételek nem teljesülnek, akkor is elvégezhetjük függvények (modellek vagy görbék) illesztését, de az ismertetendő statisztikai vizsgálatok nem használhatók, ill. a kapott becslések tulajdonságai mások lesznek.

Nevezetesen, ha az ε mérési hibák nem normális eloszlásúak, a maximum-likelihood-becslésmódszer alkalmazásakor a másik, de ismert eloszlás sűrűségfüggvényét kell helyettesítenünk. Ha az eloszlás nem ismert, a maximum-likelihood-módszer nem is használható, de más becslési módszerek (pl. a legkisebb négyzetek) igen.

Amennyiben a hiba varianciája nem ismert, és nem tudjuk, hogy állandó-e, azt a becslésnél nem használhatjuk, így a kapott becslések kevesebb információt fognak tartalmazni, nagyobb bizonytalanságúak lesznek.

Vannak olyan becslési módszerek is, amelyek akkor alkalmazhatók, ha a mérési hibák a különböző *i* pontokban egymástól nem függetlenek, de összefüggésük ismert (Bard, 1974).

Fontos megjegyezni, hogy az ismertetendő becslési kritériumok és hipotézisvizsgálati módszerek nagymértékben épülnek a fölsorolt feltételezésekre. A fejezet további részében vizsgálatunkat szűkítsük le az egyváltozós lineáris függvénykapcsolat becslésére! A többváltozós lineáris függvények illesztésével a következő fejezetben foglalkozunk.

4.2. Lineáris regresszió, ismétlés nélküli mérések, σ_y^2 konstans

Vizsgáljuk a következő mérési adatokat:

$$x_1, y_1; x_2, y_2; \ldots x_i, y_i; \ldots x_n, y_n.$$

Az összefüggés alakja:

$$Y(x) = E\left(y\middle|x\right) = \alpha + \beta(x - \bar{x}); \tag{4.2}$$

$$y = \alpha + \beta(x - \bar{x}) + \varepsilon, \tag{4.3}$$

ahol $\bar{x} = \dfrac{\sum x_i}{n}$. Az $\bar{x}$-ot tartalmazó transzformált alak előnyösebb az

$$Y(x) = \alpha' + \beta x \tag{4.4}$$

alaknál, elsősorban azért, mert α becslése így független β-étól (l. később); $\alpha' = \alpha - \beta\bar{x}$, a tengelymetszet.

4.2.1. Becslés a legkisebb négyzetek módszerével

A legkisebb négyzetek módszere szerinti becslési kritérium:

$$\sum\left[y_i - \hat{Y}(x_i)\right]^2 = \min. \tag{4.5}$$

A továbbiakban az összegzés mindig *i* szerint (a mérési pontok sorszáma szerint) végzendő; az egyszerűség kedvéért $\sum_i$ helyett csak a $\sum$ jelölést használjuk.

Esetünkben:

$$\phi = \sum\left[y_i - \hat{\alpha} - \hat{\beta}\left(x_i - \bar{x}\right)\right]^2 = \min. \tag{4.6}$$

Ennek a függvénynek kell az $\hat{\alpha}$ és $\hat{\beta}$ szerinti minimumát megkeresni:

$$\frac{\partial\phi}{\partial\hat{\alpha}} = -2\sum\left[y_i - \hat{\alpha} - \hat{\beta}\left(x_i - \bar{x}\right)\right] = 0\,, \tag{4.7}$$

$$\frac{\partial\phi}{\partial\hat{\beta}} = -2\sum\left[y_i - \hat{\alpha} - \hat{\beta}\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) = 0\,. \tag{4.8}$$

Ezek az ún. normálegyenletek. Átrendezve:

$$\sum y_i = \hat{\alpha}n + \hat{\beta}\sum\left(x_i - \bar{x}\right),$$

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

de mivel $\sum\left(x_i - \bar{x}\right) = \sum x_i - n\bar{x} = 0$, az első egyenletből

$$\hat{\alpha} \equiv a = \frac{\sum y_i}{n} = \bar{y}\,, \tag{4.9}$$

a másodikból

$$\beta \equiv b = \frac{\sum y_i\left(x_i - \bar{x}\right)}{\sum\left(x_i - \bar{x}\right)^2}\,. \tag{4.10}$$

A két egyenlet egymástól függetlenül oldható meg, ezt az $Y = \alpha + \beta\left(x_i - \bar{x}\right)$ írásmódnak köszönhetjük, és éppen ebben áll az $\hat{\alpha}$ és $\hat{\beta}$ becslések függetlensége. Az $\hat{\alpha} = a$ becsült paraméter nem a tengelymetszet, hanem $\bar{y}$.

4.2.2. Maximum-likelihood-becslés

A feltételes sűrűségfüggvény normális eloszlás esetén:

$$f\left(y|x\right) = \frac{1}{\sqrt{2\pi}\sigma_y}\exp\left\{-\frac{1}{2}\left[\frac{y_i - Y\left(x_i;\alpha,\beta\right)}{\sigma_y}\right]^2\right\}. \tag{4.11}$$

A likelihood-függvény az egyes *y* értékekhez tartozó sűrűségfüggvények szorzata, ha a különböző *i* mérési pontokban elkövetett hibák egymástól függetlenek:

$$L = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_y}\exp\left\{-\frac{1}{2}\left[\frac{y_i - Y\left(x_i\right)}{\sigma_y}\right]^2\right\} =$$

$$= (2\pi)^{-n/2}\,\sigma_y^{-n}\,\exp\left\{-\frac{1}{2}\sum_i\left[\frac{y_i - Y(x_i)}{\sigma_y}\right]^2\right\}. \qquad (4.12)$$

A likelihood-függvény logaritmusa:

$$L' \equiv \ln L = -n/2\ln(2\pi) - n\ln\sigma_y - \frac{1}{2}\sum_i\left[\frac{y_i - Y(x_i)}{\sigma_y}\right]^2. \qquad (4.13)$$

A logaritmusfüggvény monoton növekvő függvény lévén, mindegy, hogy *L* vagy *L'* maximumát keressük. Az *L* és *L'* függvényeknek az *Y*(*x*) regressziós görbe konstansai, vagyis a becsült paraméterek (*a*, *b*,...) szerinti maximuma ott van, ahol a Σ minimuma, vagyis *L* maximalizálása helyett a

$$\sum\left[\frac{y_i - \hat{Y}(x_i)}{\sigma_y}\right]^2 = \min. \qquad (4.14)$$

szélsőérték-feladatot kell megoldani.

Mivel σ_y^2 konstans, ez a következő alakot ölti:

$$\sum\left[y_i - \hat{Y}(x_i)\right]^2 = \sum\left[y_i - a - b(x_i - \bar{x})\right]^2 = \min. \qquad (4.15)$$

A kapott kritérium normális eloszlás és konstans variancia esetén azonos a legkisebb négyzetek (4.5) becslési kritériumával, így *a* és *b* kifejezése is azonos a (4.9) és (4.10) képletekkel.

σ_y^2 maximum-likelihood-becslése

A (4.13) képlet szerinti likelihood-függvény szélsőértékhelye:

$$\frac{\partial L'}{\partial\hat{\sigma}_y} = -\frac{n}{\hat{\sigma}_y} + \frac{1}{\hat{\sigma}_y^3}\sum\left[y_i - Y(x_i)\right]^2 = 0. \qquad (4.16)$$

$\sigma_y^2 \neq 0$, ezért szorozhatjuk vele az egyenlet mindkét oldalát:

$$-n + \frac{1}{\hat{\sigma}_y^2}\sum\left[y_i - Y(x_i)\right]^2 = 0,$$

$$\hat{\sigma}_y^2 = \frac{\sum\left[y_i - Y(x_i)\right]^2}{n}, \qquad (4.17)$$

amelyről bebizonyítható, hogy torzítatlan becslés, de nem lehet kiszámítani, mivel az $Y(x)$ függvény konstansai (esetünkben α és β) nem ismertek.

Ha α és β helyébe *a* és *b* értékét helyettesítjük [vagyis $Y(x)$ helyett annak becslését, $\hat{Y}(x)$ -ot használjuk], a becslés torzított lesz, de ez a torzítás korrigálható.

Az $\hat{\alpha}$ és $\hat{\beta}$ becslések tulajdonságai

Az $\hat{\alpha}$ becslés várható értéke

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

és mivel $(x_i - \bar{x})$ nem valószínűségi változó:

$$E(a) = \frac{1}{n}\sum\left\{\alpha + \beta(x_i - \bar{x}) + E(\varepsilon_i)\right\} = \frac{1}{n}n\alpha + \frac{1}{n}\beta\sum(x_i - \bar{x}) = \alpha, \qquad (4.18)$$

figyelembe véve, hogy $\sum(x_i - \bar{x}) = 0$.

Hasonlóan belátható, hogy $E(b) = \beta$. Vagyis *a* és *b* torzítatlan becslése α-nak, ill. β-nak, amennyiben az elméleti regressziós függvény a feltételezett $Y = \alpha + \beta(x - \bar{x})$ alakú, ugyanis ezt felhasználtuk az előző levezetésekben, amikor *y* helyében $Y + \varepsilon$ került.

Megjegyzendő, hogy *a* és *b* normális eloszlású, mivel *y* normális eloszlású valószínűségi változók lineáris függvénye, és normális eloszlású változók bármely lineáris kombinációja szintén normális eloszlású (ún. addíciós tétel). Hasonlóan $\hat{Y}_i$ is normális eloszlású.

A becslések varianciája:

$$Var(a) = Var\left(\frac{\sum y_i}{n}\right) = \frac{1}{n^2}\sum Var(y_i) = \frac{1}{n^2}n\sigma_y^2 = \frac{\sigma_y^2}{n}; \qquad (4.19)$$

$$Var(b) = \frac{\sigma_y^2}{\sum(x_i - \bar{x})^2}. \qquad (4.20)$$

Látható, hogy *n* növekedésével *a* és *b* varianciája zérushoz tart, vagyis a becslés konzisztens is. Az is látható, hogy $Var(b)$ nagymértékben függ $\sum(x_i - \bar{x})^2$-től, vagyis az x_i mérési pontok megválasztásától, tehát a kísérletek megtervezésétől. Minél nagyobb (4.20) nevezője, azaz a választott x_i értékek minél inkább a vizsgált intervallum két szélén helyezkednek el, annál kisebb $Var(b)$, vagyis *b* becslése szempontjából ez a kísérleti terv optimális. Ebben az esetben azonban nem tudjuk ellenőrizni a választott lineáris függvény helyességét. A mérési pontok ilyen elrendezését akkor alkalmazzuk, ha a lineáris függvénykapcsolat fennállását előzőleg igazoltuk.

A becsült regressziós egyenes varianciája:

$$Var\left(\hat{Y}|x\right) \equiv \sigma_{\hat{Y}}^2 = E\left[\left(\hat{Y} - Y\right)^2\right] = Var\left[a + b(x - \bar{x})\right] =$$

$$= Var(a) + (x - \bar{x})^2 Var(b) = \sigma_y^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \right]. \qquad (4.21)$$

Vegyük észre, hogy a becslés varianciája az $(x - \bar{x}) = 0$-nak megfelelő helyen a legkisebb (azaz ha $x = \bar{x}$), valamint hogy a mérési pontok számának növekedésével csökken, és függ a mérési pontok elhelyezésétől (az $x_j - \bar{x}$ különbségektől), vagyis a kísérleti tervtől is.

Az eltérések négyzetösszegének vizsgálata

A mért *y* érték és az *Y* valódi érték közötti eltérés a következő módon írható fel:

$$y_i - Y_i = (y_i - \hat{Y}_i) + (\hat{Y}_i - Y_i). \qquad (4.22)$$

Feladatunk a lineáris függvény alkalmasságának vizsgálata, vagyis arra vagyunk kíváncsiak, hogy adekvát-e a függvény. Nullhipotézisünk az, hogy a változók között fennáll a (4.2) lineáris összefüggés, ellenhipotézisünk, hogy nem:

H_0: $Y = \alpha + \beta(x - \bar{x})$; $\qquad$ H_1: $Y \neq \alpha + \beta(x - \bar{x})$.

Írjuk a (4.22) kifejezésbe a nullhipotézisnek megfelelő lineáris függvényt:

$$y_i - Y_i = (y_i - \hat{Y}_i) + (a_i - \alpha_i) + (b - \beta)(x_i - \bar{x}). \qquad (4.23)$$

A (4.23) kifejezést négyzetre emelve és *i* szerint szummázva, a következőt kapjuk:

$$\sum (y_i - Y_i)^2 = \sum (y_i - \hat{Y}_i)^2 + n(a_i - \alpha_i)^2 + (b - \beta)^2 \sum (x_i - \bar{x})^2 . \qquad (4.24)$$

A vegyes szorzatokat tartalmazó szummák mindegyike zérus.

4-1. táblázat

Az eltérés	Négyzetösszeg	Szabadsági fok	A négyzetösszeg 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
Az empirikus regresszsziós görbe körül $\bar{y}_i$ és $\hat{Y}_i$	$\sum (\bar{y}_i - \hat{Y}_i)^2$	$n - 2$	$(n - 2)\sigma_y^2$
Teljes	$\sum_i (y_i - Y_i)^2$	n	$n\sigma_y^2$

Az elméleti regressziós vonal körüli $\left(y_i - Y_i\right)$ eltérések a regresszióanalízis bevezetőjében ismertetett feltételezés értelmében normális eloszlásúak, σ_y^2 varianciával, így négyzetösszegük $\chi^2\sigma_y^2$ eloszlású, szabadsági foka *n*.

A (4.24) összefüggés jobb oldalán szereplő három négyzetösszeg közül az elsőnek $n-2$ [mivel $\hat{Y}_i$ és y_i között fennáll a (4.9) és (4.10) összefüggés], a másik kettőnek egy-egy a szabadsági foka. A négyzetösszegek szabadsági fokainak összege *n*, megegyezik a $\sum_i \left(y_i - Y_i\right)^2$ négyzetösszeg szabadsági fokával. Ha a nullhipotézis igaz [ekkor igaz a (4.23) szerinti algebrai felbontás], akkor a Fisher–Cochran-féle felbontási tétel mindkét feltétele teljesül, és a (4.24) jobb oldalán levő három négyzetösszeg mindegyike egymástól független, $\chi^2\sigma_y^2$ eloszlású. A H_0 nullhipotézis ellenőrzése χ^2-próbával lehetséges, ugyanis H_0 fennállása esetén a (4.24) jobb oldalán szereplő kifejezés mindegyikének várható értéke:

$$E\left(\chi^2\sigma_y^2\right) = \sigma_y^2 E\left(\chi^2\right) = \sigma_y^2 \nu. \qquad (4.25)$$

Az α és β értékének ismerete híján a három közül csupán a $\sum\left(y_i - \hat{Y}_i\right)^2$ négyzetösszeg, s így a következőképpen definiált szórásnégyzet (ún. reziduális szórásnégyzet) számítható ki:

$$s_r^2 = \frac{\sum\left(y_i - \hat{Y}_i\right)^2}{n-2}.$$

Ha a H_0 nullhipotézis igaz, a (4.25) értelmében s_r^2 torzítatlan becslése σ_y^2 -nak. Ekkor az $s_r^2(n-2)/\sigma_y^2$ kifejezés χ^2 eloszlású, szabadsági foka $n-2$. A nullhipotézis ellenőrzésére a χ^2-próbát akkor tudjuk elvégezni, ha előzetes kísérleti tapasztalatok alapján ismerjük σ_y^2 értékét. Ha a számított $s_r^2(n-2)/\sigma_y^2$ érték az α szignifikanciaszinthez tartozó elutasítási tartományba esik, a nullhipotézist elutasítjuk, vagyis α szignifikanciaszinten elutasítjuk a lineáris modellt. Sajnos a gyakorlati esetek döntő többségében σ_y^2 értéke nem áll rendelkezésre, tehát ismételt mérések nélkül nincs mód annak vizsgálatára, hogy a feltételezett egyenes adekvát-e.

Az adatok feldolgozásakor az egyenes illesztése után nemcsak azt kérdezhetjük, hogy az egyenes megfelelő-e az adatok leírására, hanem azt is, hogy szükség van-e erre az egyenesre, vagyis a vízszintes egyenes nem épp olyan jó-e. A vízszintes egyenes az $\hat{Y} = a = \bar{y}$ modellt jelenti, vagyis a $\beta = 0$ nullhipotézis vizsgálandó.

Amennyiben az egyenes adekvátnak bizonyult, vagy legalábbis azt joggal feltételezzük, a (4.24) jobb oldalának minden tagja $\chi^2\sigma_y^2$ eloszlású, így az utolsó is, 1 szabadsági fokkal. A H_0: $\beta = 0$ nullhipotézis igazsága esetén ez a $b^2\sum\left(x_i - \bar{x}\right)^2$ négyzetösszegre is igaz.

Az alkalmas próbastatisztika:

$$F_0 = \frac{\dfrac{b^2 \sum \left(x_i - \bar{x}\right)^2}{1}}{\dfrac{\sum \left(y_i - \hat{Y}_i\right)^2}{(n-2)}} = \frac{b^2 \sum \left(x_i - \bar{x}\right)^2}{s_r^2}, \tag{4.26}$$

ahol a nevezőben az $\hat{Y} = a + b(x - \bar{x})$ modellhez számított reziduális szórásnégyzet szerepel.

A szakkönyvekben a (4.26) kifejezés többféle azonos alakjával találkozunk. Belátható, hogy

$$b^2 \sum \left(x_i - \bar{x}\right)^2 = \sum_i \left(\hat{Y}_i - \bar{y}\right)^2 .$$

E képlet alapján szokás a szóban forgó négyzetösszeget a "regresszió négyzetösszegének" nevezni, mert azt fejezi ki, hogy a regressziós egyenes az *y* adatok átlagától mennyire tér el. Ugyanehhez az F_0 próbastatisztikához vezet az ún. általános regressziós próba is:

$$F_0 = \frac{\dfrac{S_R^{H_0} - S_R}{\Delta\nu}}{\dfrac{S_R}{\nu_R}}, \tag{4.27}$$

ahol S_R a teljes modellhez (itt $\hat{Y} = a + b(x - \bar{x})$) számított reziduális négyzetösszeg; $S_R^{H_0}$ a nullhipotézis szerinti redukált modellhez (itt $\hat{Y} = a = \bar{y}$) számított reziduális négyzetösszeg; $\Delta\nu$ a két négyzetösszeg szabadsági fokszámainak különbsége, itt $\Delta\nu = (n-2) - (n-1) = 1$; ν_R a teljes modell reziduális négyzetösszegének szabadsági fokszáma.

Esetünkre:

$$S_R^{H_0} = \sum_i \left(y_i - \bar{y}\right)^2 ; \quad \nu = n - 1; \tag{4.28a}$$

$$S_R = \sum_i \left[y_i - a - b\left(x_i - \bar{x}\right)\right]^2 ; \quad \nu = n - 2 . \tag{4.28b}$$

$S_R^{H_0}$ a következőképpen bontható föl algebrailag:

$$\sum_i \left(y_i - \bar{y}\right)^2 = \sum_i \left(y_i - \hat{Y}_i\right)^2 + \sum_i \left(\hat{Y}_i - \bar{y}\right)^2 . \tag{4.29}$$

Az első, az egyenlet bal oldalán lévő négyzetösszeget nevezhetjük "teljes" négyzetösszegnek, a második a reziduális, a harmadik a "regresszió" négyzetösszege. Vagyis a (4.27) általános regressziós próba számlálójában szereplő különbség éppen a "regresszió" négyzetösszege.

Az F_0 próbastatisztika aktuális értéke alapján dönthetünk arról, hogy a regresszió szignifikáns-e, vagyis a kérdéses függvény (itt az egyenes) jobban illeszkedik-e az adatokra, mint $\bar{y}$, a vízszintes egyenes.

Természetesen ez a próba csak akkor ad helyes eredményt, ha a teljesebb modell (itt az $\hat{Y} = a + b(x - \bar{x})$ egyenes) csakugyan megfelelő az adatok leírására. Ugyancsak természetesen, ha a mérési hibák nem normális eloszlásúak, vagy nem függetlenek egymástól, a (4.24)-ben szereplő négyzetösszegek nem lesznek $\chi^2\sigma_y^2$ eloszlásúak.

A (4.29) egyenlet bal oldalán szereplő "teljes" négyzetösszeget, valamint a felbontásával kapott reziduális és "regressziós" négyzetösszegeket felhasználhatjuk arra is, hogy a regressziós függvény illeszkedésének jóságát mérjük. Erre szolgál a determinációs együttható:

$$R^2 = \frac{\sum_i \left(\hat{Y}_i - \bar{y}\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2} = 1 - \frac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2} \tag{4.30}$$

Azt mondhatjuk, hogy R^2 az y_i mérési adatok $\bar{y}$ átlagtól való eltérésének az a része, amely az $\hat{Y}$ regressziós függvénnyel magyarázható.

Érdekes megjegyezni, hogy ez az R^2 determinációs együttható algebrai átrendezésekkel belátható módon az x és y adatok közötti r korrelációs együttható négyzete, amelyet az eredeti értelmében szigorúan véve csak akkor volna jogos használni, ha x és y egyaránt valószínűségi változók lennének (ún. korrelációs modell).

R^2 értéke erősen függ az x független változó értékeitől, ahogy az a számláló $b^2\sum\left(x_i - \bar{x}\right)^2$ írásmódjából rögtön látszik. Az R^2 értéke ugyancsak erősen függ a modell paramétereinek p számától. Szélső esetben, ha a modellnek ugyanannyi ($p = n$) paramétere van, mint mérési pont, például $n-1$-edfokú polinomot használunk, az $y_i - \hat{Y} \equiv 0$, vagyis $R^2 \equiv 1$ lesz.

Ezért szokás egy igazított R^2-et (R^2_{adj}) használni, amely nem a négyzetösszegeket, hanem a megfelelő szórásnégyzeteket tartalmazza:

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p} = 1 - \frac{\dfrac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{n-p}}{\dfrac{\sum_i \left(y_i - \bar{y}\right)^2}{n-1}} = 1 - \frac{s_r^2}{s_T^2}, \tag{4.31}$$

ahol s_T^2 az ún. teljes eltérés szórásnégyzete.

A becsült regressziós egyenes varianciája

Az $\hat{Y}$ becsült regressziós egyenes egy pontjának szórásnégyzete:

$$s_{\hat{Y}}^2 = \frac{\chi^2 \sigma_{\hat{Y}}^2}{\nu} = \frac{\chi^2 \sigma_y^2}{\nu}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]. \tag{4.32}$$

Felhasználva, hogy $s_y^2 = \frac{\chi^2 \sigma_y^2}{\nu}$,

$$\hat{\sigma}_{\hat{Y}}^2 \equiv s_{\hat{Y}}^2 = s_y^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]. \tag{4.33}$$

Ugyanígy kapjuk $Var(a)$ és $Var(b)$ becsléséül az s_a^2 és s_b^2 kifejezését:

$$\hat{\sigma}_a^2 \equiv s_a^2 = \frac{s_y^2}{n}, \tag{4.34}$$

$$\hat{\sigma}_b^2 \equiv s_b^2 = \frac{s_y^2}{\sum_i (x_i - \bar{x})^2}. \tag{4.35}$$

Ha a lineáris függvény adekvát, az s_r^2 reziduális szórásnégyzet $\chi^2 \sigma_y^2 / \nu$ eloszlású, így s_y^2-ként használhatjuk.

Az $\hat{Y}$ becsült regressziós egyenes egy pontjának szórásnégyzete (4.34) és (4.35) felhasználásával:

$$s_{\hat{Y}}^2 = s_a^2 + s_b^2(x-\bar{x})^2 = s_{a'}^2 + s_b^2\left(x^2 - 2x\bar{x}\right). \tag{4.36}$$

ahol $s_{a'}^2$ az a' tengelymetszet szórásnégyzete, $s_{a'}^2 \equiv s_a^2 + s_b^2 \bar{x}^2$. Az $s_{a'}^2$-vel felírt kifejezés azért igen hasznos, mert a számítógépi programok általában $s_{a'}^2$ értékét adják meg.

A becsült paraméterek konfidenciaintervalluma

A *t*-eloszlást definiáló összefüggés értelmében a következő statisztika *t*-eloszlású:

$$t = \frac{a-\alpha}{s_a} \quad \text{és} \quad \nu = n-2, \tag{4.37a}$$

ahol

$$s_a = \frac{s_y}{\sqrt{n}}. \tag{4.37b}$$

Hasonlóan

$$t = \frac{b - \beta}{s_b};\ \nu = n - 2, \tag{4.38a}$$

ahol
$$s_b = \frac{s_y}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}}, \tag{4.38b}$$

és

$$t = \frac{\hat{Y} - Y}{s_{\hat{Y}}};\ \nu = n - 2, \tag{4.39}$$

ahol $s_{\hat{Y}}$ a (4.33) kifejezés négyzetgyöke.

Az előbbi statisztikák alapján adott valószínűségi szintekhez α-ra és β-ra konfidenciaintervallumot, az Y függvényre pedig konfidenciasávot számíthatunk.

4-1. példa

Egy standard anyagból különböző koncentrációjú oldatokat készítettek, majd minden oldatból azonos mennyiséget (1 μl) gázkromatográfba fecskendezve, kísérletileg vizsgálták a kapott csúcs alatti terület (y) és a koncentráció (x) közötti összefüggést. A mérési adatok a 4-2. táblázatban láthatók, a koncentráció értéke szerint növekvő sorrendbe rendezve. A tényleges mérési sorrendet a táblázat második oszlopa tartalmazza.

Feltételezve, hogy a terület és a koncentráció közötti függvénykapcsolat lineáris, adjunk becslést az egyenes paramétereire (kalibrációs egyenes)!

4-2. táblázat

i	A mérés sorrendje	Mérési adatok		Számított adatok		
		x_i (mg/ml)	y_i (terület)	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2 \cdot 10^2$	$y_i\left(x_i - \bar{x}\right)$
1	3	0	0	–0.083333	0.69444	0
2	5	0.05	1681894	–0.033333	0.11111	–56063
3	4	0.08	2614987	–0.003333	0.00111	–8717
4	2	0.10	3297753	0.016667	0.02778	54963
5	1	0.12	3983787	0.036667	0.13444	146072
6	6	0.15	4978455	0.066667	0.44444	331897
Σ		0.50	16556876		1.41333	468152

$$\bar{x} = 0.083333; \qquad \bar{y} = 2759479;$$

$$a = \bar{y} = 2759479;$$

$$b = \frac{\sum y_i\left(x_i - \bar{x}\right)}{\sum\left(x_i - \bar{x}\right)^2} = \frac{468152}{1.41333 \cdot 10^{-2}} = 3.3124 \cdot 10^7;$$

$$\hat{Y} = 2759479 + 3.3124 \cdot 10^7 \cdot (x - 0.083333) = -843 + 3.3124 \cdot 10^7 \cdot x.$$

A becsült $\hat{Y}_i$ értékek és a reziduális szórásnégyzet számítása:

4-3. táblázat

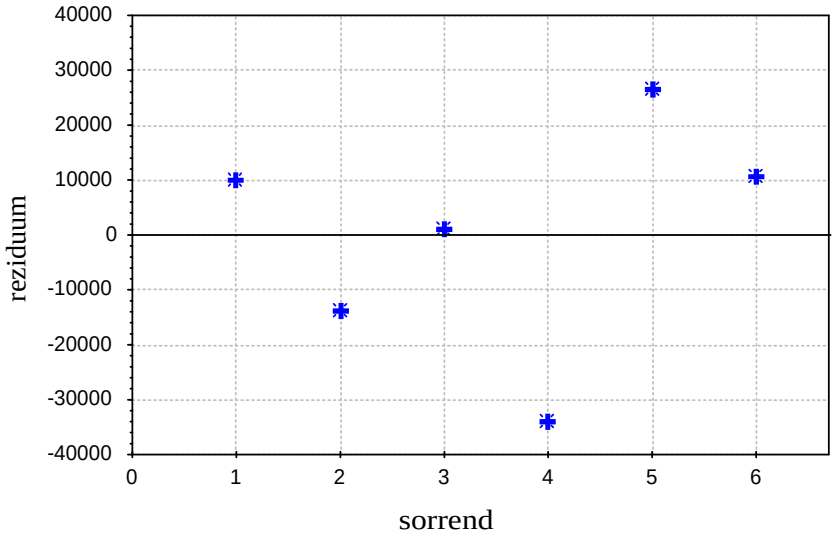
i	sor-rend	x_i	y_i	$\hat{Y}_i$	$y_i - \hat{Y}_i$	$\left(y_i - \hat{Y}_i\right)^2 \cdot 10^{-8}$	$\left(y_i - \hat{Y}_i\right)/s_r$
1	3	0	0	–843	843	0.007	0.0354
2	5	0.05	1681894	1655357	26537	7.042	1.1147
3	4	0.08	2614987	2649077	–34090	11.621	–1.4319
4	2	0.10	3297753	3311557	–13804	1.906	–0.5798
5	1	0.12	3983787	3974037	9750	0.951	0.4095
6	6	0.15	4978455	4967757	10698	1.144	0.4494
Σ						22.671	

$$s_r^2 = \frac{22.671 \cdot 10^8}{6 - 2} = 5.6678 \cdot 10^8; \qquad s_r = 2.3807 \cdot 10^4.$$

A regresszió feltételeinek ellenőrzése a reziduumok grafikus vizsgálatával

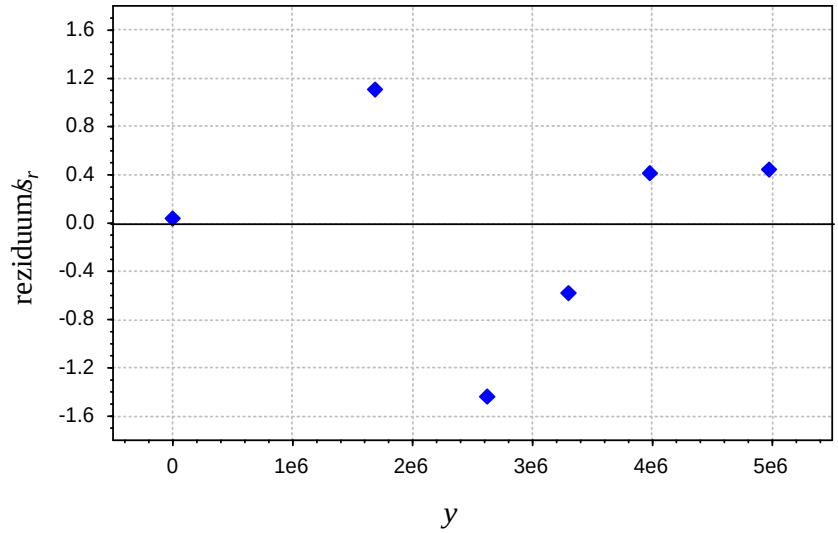
Ismételt mérések hiányában nincs lehetőség σ_y^2 modelltől független becslésére, ezért a regresszió feltételeit csak a reziduumok grafikus vizsgálatával tudjuk ellenőrizni.

- Az $y_i - \hat{Y}_i$ reziduumokat a mérések sorrendjében ábrázolva ellenőrizhetjük, hogy a mérések egymásutánjában a mérési hibáknak nincs-e egyirányú menete (az ε_i és ε_{i-1} hibák függetlenek-e).



4-2. ábra. A reziduumok a mérések sorrendjének függvényében

A 4-2. ábrán az adatoknak nincs egyirányú menete, tehát az ε_i hibák nem korreláltak.

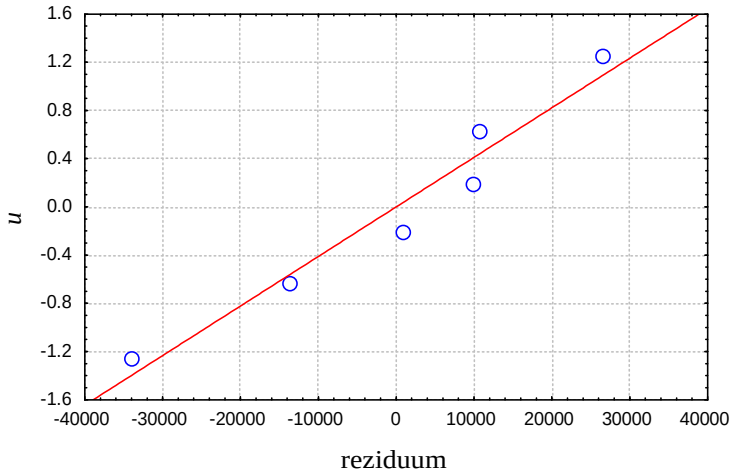


4-3. ábra. A standardizált reziduumok az y_i mérési adatok függvényében

- Az $\frac{\left(y_i - \hat{Y}_i\right)}{s_r}$ ún. standardizált reziduum értékeket a mért y_i függvényében ábrázoljuk (4-3. ábra). Mivel a pontok zérus körül véletlenszerűen ingadoznak, azaz nem tapasztalunk trendet, és a reziduumok eltolódást sem mutatnak, az illesztett lineáris függvény adekvát. Mivel az adatok azonos szélességű sávban ingadoznak, elfogadhatjuk azt a feltételezést is, hogy σ_y^2 konstans.

- Az ún. Gauss-háló (valószínűségi papír) vízszintes tengelyén a reziduumokat, függőleges tengelyén pedig az elméleti eloszlásfüggvényből visszaszámolt való-

színűségi változó (normal score) értékeket ábrázoljuk (4-4. ábra). Mivel az így előállított ábrán a pontok egyenes mentén helyezkednek el, és nem találunk kiugró pontot vagy szisztematikus eltérést, elfogadhatjuk azt a feltételezést, hogy az ε_i hibák normális eloszlást követnek.



4-4. ábra.
A reziduumok ábrázolása Gauss-hálón

Szignifikáns-e a regresszió (nem zérus-e a meredekség)?

Szükség van-e erre az egyenesre (vagyis a vízszintes egyenes nem épp olyan jó-e)? A vízszintes egyenes az $\hat{Y} = a = \bar{y}$ modellt jelenti, vagyis a $\beta = 0$ nullhipotézis vizsgálandó. A nullhipotézist a (4.26) általános regressziós próbával ellenőrizzük.

Regressziós négyzetösszeg:

$$\sum\left(\hat{Y}_i - \bar{y}\right)^2 = b^2 \sum\left(x_i - \bar{x}\right)^2 = \left(3.3124 \cdot 10^7\right)^2 \cdot 1.41333 \cdot 10^{-2} = 1.5507 \cdot 10^{13}.$$

Reziduális négyzetösszeg:

$$\sum\left(y_i - \hat{Y}_i\right)^2 = 2.2671 \cdot 10^9, \quad s_r^2 = 5.6678 \cdot 10^8.$$

$$F_0 = \frac{b^2 \sum\left(x_i - \bar{x}\right)^2}{\frac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{n-2}} = \frac{b^2 \sum\left(x_i - \bar{x}\right)^2}{s_r^2} = \frac{1.5507 \cdot 10^{13}}{5.6678 \cdot 10^8} = 2.736 \cdot 10^4.$$

A Függelék IV. táblázatából $F_{0.05}(1,4) = 7.71$, a próbastatisztika aktuális értéke ezt meghaladja, tehát a nullhipotézist elutasítjuk, azaz szükség van az egyenesre, a regresszió szignifikáns (a vízszintes egyenes nem megfelelő). Statisztikai programok használatakor általában nincs szükségünk a IV. táblázatra, mivel a programok meg-

adják a próbastatisztika számított értékéhez tartozó elsőfajú hiba valószínűségét, amelyet általában *p*-vel jelölnek. Ez az érték jelenleg rendkívül kicsi ($p << 0.05$).

A 4-4. táblázatban a statisztikai programok eredményközlésének megfelelő elrendezésben (ún. ANOVA táblázat) megadjuk a (4.29) összefüggés bal és jobb oldalán szereplő négyzetösszegeket (S), szabadsági fokaikat (ν), valamint a szórásnégyzetek értékét. A táblázat utolsó előtti oszlopában az F_0 próbastatisztika értékét, utolsó oszlopában pedig az elsőfajú hiba valószínűségét találjuk (a nullhipotézis érvényessége esetén ekkora a valószínűsége annak, hogy F_0 értéke 27360, vagy annál nagyobb legyen).

4-4. táblázat

négyzetösszeg	S	ν	s^2	F_0	p
$\sum_i \left(\hat{Y}_i - \bar{y}\right)^2$	$1.5507 \cdot 10^{13}$	1	$1.5507 \cdot 10^{13}$	$2.736 \cdot 10^{4}$	$8 \cdot 10^{-9}$
$\sum_i \left(y_i - \hat{Y}_i\right)^2$	$0.0002 \cdot 10^{13}$	4	$5.6678 \cdot 10^{8}$		
$\sum_i \left(y_i - \bar{y}\right)^2$	$1.5509 \cdot 10^{13}$	5			

Mivel *b* normális eloszlás szerint ingadozik β körül, *t*-próbával is vizsgálható, hogy az egyenes meredeksége szignifikánsan különbözik-e zérustól. A próbastatisztika:

$$t_0 = \frac{b}{s_b} = \frac{b\sqrt{\sum \left(x_i - \bar{x}\right)^2}}{s_r} = \frac{3.3124 \cdot 10^7 \sqrt{1.41333 \cdot 10^{-2}}}{2.3807 \cdot 10^4} = \frac{3.9379 \cdot 10^6}{2.3807 \cdot 10^4} = 165.41$$

Ez éppen az előbbi F_0 négyzetgyöke. Általánosan is igaz, hogy egy ν szabadsági fokú *t*-eloszlású valószínűségi változó négyzete $F(1, \nu)$ eloszlású, amint erről a *t* és *F* táblázat összevetésével is meggyőződhetünk.

Számítsuk ki a determinációs együtthatót!

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{Y}_i\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2} = 1 - \frac{2.269 \cdot 10^9}{1.5509 \cdot 10^{13}} = 0.99985$$

A teljes eltérés-négyzetösszeg 99.985%-át „magyarázza” az illesztett egyenes.

Áthalad-e az igazi egyenes az origón?

A nullhipotézis: H_0: $Y(x = 0) = \alpha' = 0$.

A próbastatisztika kiszámításához szükségünk van az a' tengelymetszet $s_{a'}$ szórására, azaz $s_{\hat{Y}}$ értékére az $x = 0$ helyen.

$$s_a^2 = \frac{s_r^2}{n} = \frac{5.6678 \cdot 10^8}{6} = 9.446 \cdot 10^7, \quad s_b^2 = \frac{s_r^2}{\sum_j \left(x_j - \bar{x}\right)^2} = \frac{5.6678 \cdot 10^8}{1.41333 \cdot 10^{-2}} = 4.010 \cdot 10^{10},$$

$$s_{a'}^2 \equiv s_a^2 + s_b^2 \bar{x}^2 = 9.447 \cdot 10^7 + 4.010 \cdot 10^{10} \cdot 0.08333^2 = 3.7295 \cdot 10^8,$$

$$t_0 = \frac{a'}{s_{a'}} = \frac{-843}{\sqrt{3.7295 \cdot 10^8}} = -0.043, \quad t_{0.05/2}(4) = 2.776, \quad -2.776 < t_0 < 2.776.$$

Mivel a próbastatisztika értéke a 0.05 szignifikanciaszinthez tartozó elfogadási tartományba esik, nem utasítjuk el azt a nullhipotézist, hogy az egyenes átmegy az origón.

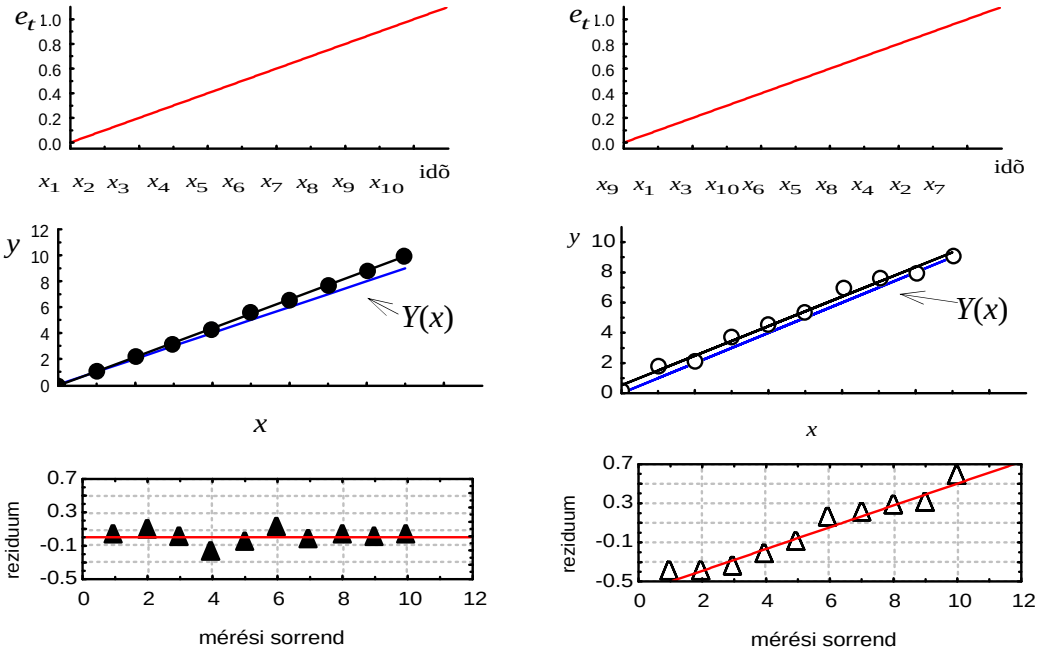
4.2.3. A mérések sorrendje

Képzeljük el a következő helyzetet: egy anyag bemért koncentrációja függvényében mérjük az abszorbanciáját (analitikai jel), és az összefüggést lineáris függvénnyel írjuk le. A méréseket nem tudjuk nagyon rövid idő alatt elvégezni, és az eltelt idő alatt az anyag nedvességet szív magába, ami az abszorbanciáját is megváltoztatja.

A 4-5. ábra fölső képei szemléltetik az abszorbanciának a koncentrációtól és az eltelt idő alatt felvett nedvességtől való függését külön-külön, ha nincs mérési bizonytalanság. A valóságos mérésnél a koncentráció és a nedvesség hatása együtt jelentkezik, ráadásul a mérést mérési hiba is terheli. A kísérletezőn múlik, hogy az egyes méréseket milyen sorrendben végzi el.

- Az egyik (nagyon kézenfekvő) lehetőség, hogy a koncentráció növekvő sorrendjében mér, ezt mutatja a bal oldali ábra (a bal oldali fölső ábrán az időtengelyre bejelölt *x* adatok nagyság szerinti sorrendben vannak). Ekkor a két hatás összege körül ingadozó pontokat kap az *y* tengelyen *x* függvényében, vagyis az illesztett függvény a két hatás összegét becsüli (a két függvény összege lesz). Ez látható a bal oldali középső képen, ahol $Y(x)$ az abszorbancia valódi koncentrációfüggvénye.

  Ha a kísérletező gondos, és az egyenes illesztése után a reziduumokat is ábrázolja a mérések sorszáma (bal oldali legalsó ábra) ill. az *x* változó függvényében, nem lát rendszerességet (zérus körül véletlenszerű ingadozást tapasztal), mert a két hatás összegét leíró egyenes körül véletlenszerű az ingadozás. Tehát a kísérletezőnek minden oka megvan, hogy azt higgye, az illeszkedés megfelelő, és nem szerez tudomást róla, hogy az abszorbanciának a koncentrációtól való függésébe egy zavaró tényező (az időben fölvett nedvesség) hatását is belemérte és beleszámolta. Amikor használja az összefüggést, tipikusan kalibrációs egyenesként, akkor az ismeretlen koncentrációjú anyag abszorbanciájából számolja a koncentrációt, de ez hamis lesz, mert a nedvességtartalom eltéríti.



4-5. ábra. A mérések végrehajtásának sorrendje

- A másik lehetőség, hogy nem a növekvő koncentráció sorrendjében mér, hanem véletlenszerű sorrendben, ebbe beleértve, hogy ha egy koncentrációnál több ismételt mérést végez, azok nem egymás után következnek. Ez látható a jobb oldali fölső ábrán, ahol az időtengelyre bejelölt *x* adatok nem nagyság szerinti sorrendben vannak. Ekkor az egyre nagyobb koncentrációértékekhez nem tartozik egyre nagyobb nedvességtartalom, vagyis bár itt is a két hatás összegét méri, de a két hatás nem mutat egy irányba, a nedvességtartalom (az idő) járuléka nem nő monoton módon a koncentrációval. Ilyenkor az időbeliség egyrészt eltorzítja az összefüggést (nagyjából párhuzamosan fölfelé tolja el az egyenest), másrészt nagyobb szóródást okoz.

  Ha a kísérletező ábrázolja a reziduumokat *x* függvényében (az ábrán ez nem látható), az előbbi esetben tapasztaltnál valamivel nagyobb ingadozást lát, de az most is véletlenszerű. Ha azonban a mérések sorszáma függvényében ábrázolja a reziduumokat, azok rendszerességet mutatnak, mert az abszorbanciának az időtől (a nedvességtartalomtól) való függéséről az illesztett egyenes nem ad számot, azt az egyenestől való eltérésként észleljük.

A növekvő koncentráció 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.