Szegedi Orvostudományi Egyetem Számitástechnikai Központ

Linearizáló transzformáció segitségével kapott paraméterbecslések jóságának vizsgálata szimulációval

Eller József, Győri István

Bevezetés. Vizsgált módszerek leirása.

Az orvosi és biológiai kutatásokban gyakran előforduló feladat egy biológiai rendszer müködését leiró

$$y = f(\underline{x}, \underline{\alpha}) \qquad (1)$$

$$\underline{x} = (x_1,\ldots,x_k) \quad \underline{\alpha} = (\alpha_1,\ldots,\alpha_p)$$

egyenlet α paramétereinek becslése, az $\underline{x}$ független változó különböző $\underline{x}_i$ értékei mellett megfigyelt

$$\tilde{y}_i = f(\underline{x}_i, \underline{\alpha}) + \varepsilon_i, \quad i=1,\ldots,n \qquad (2)$$

mérések alapján, ahol feltesszük, hogy a mintavételnél elkövetett ε_i additiv mérési hibák függetlenek és azonos, $N(0, \sigma^2)$ eloszlásúak. Ekkor ismeretes [5], hogy az $\underline{\alpha}$ paramétervektor maximum-likelihood becslése a legkisebb négyzetek módszerével nyerhető, az

$$S = \sum_{i=1}^{n} /\tilde{y}_i - f(\underline{x}_i, \underline{\alpha})/^2 \qquad (3)$$

hibanégyzetösszeg $\underline{\alpha}$ szerinti minimalizálásával, és az igy nyert $\hat{\underline{\alpha}}$ becslés rendelkezik a maximum-likelihood becslések összes kedvező tulajdonságával /konzisztens, aszimptotikusan normális, ill. hatásos/. Ha az f regresszió függvény lineáris $\underline{\alpha}$-ban, akkor - csupán a hibák korrelálatlanságának, zérus várható értékének és azonos szórásának feltevése mellett - a lineáris regresszió Gauss-Markov-féle elmélete szerint $\hat{\underline{\alpha}}$ torzitatlan, és az összes lineáris becslés között minimális szórásnégyzetü becslés [5]. Ha f nem lineáris $\underline{\alpha}$-ban, akkor az S négyzetösszeg minimalizálása csak valamilyen iterációs módszerrel végezhető el, melyhez megfelelő kezdetiérték megadása szükséges. Négyzetösszegek minimalizálására igen alkalmasak a Gauss-Newton-módszer különböző variánsai /lásd [2]/.

Az általunk vizsgált esetben az f regressziós függvény ugyan nemlineáris az $\underline{\alpha}$ paraméterekben, de a függő változó $z(\underline{x},y)$, a független változók $\underline{u}(\underline{x})$ és a paraméterek $\underline{\beta}(\underline{\alpha})$ transzformációjával az (1) egyenlet $z = u_1\beta_1+\ldots+u_p\beta_p$ lineáris alakra hozható, tehát egy

$$z(\underline{x},f(\underline{x},\underline{\alpha})) = \sum_{j=1}^{p} u_j(\underline{x})\beta_j(\underline{\alpha}) \qquad (4)$$

alakú azonosság áll fenn. A regressziós modell linearizáló traszformációját a következő egyszerü példán fog-

juk szemléltetni:

$$\left.\begin{aligned} y &= \sqrt{\alpha_1 + \alpha_2 x}, \\ y^2 &= \alpha_1 + \alpha_2 x; \end{aligned}\right\} \qquad (5)$$

itt

$$z(x,y) = y^2, \quad \underline{\beta}(\underline{\alpha}) = \underline{\alpha}, \quad \underline{u}(x) = (1,x) .$$

Az eredeti nemlineáris regressziós függvény illesztése - az iterációs eljárásokkal kapcsolatos kezdetiérték-ill. konvergenciaproblémák miatt - még számítógéppel is komoly numerikus nehézségekkel járhat*, illetve gyakran nem áll megfelelő program rendelkezésre. Ezért linearizálható regressziós függvény esetén általánosan elterjedt módszer az eredeti (1) függvénynek az $(\underline{x}_i, \tilde{y}_i)$ adatokhoz való illesztése helyett a (4) lineáris függvénykapcsolat illesztése a transzformált $(\underline{u}_i, \tilde{z}_i)$ adatokhoz, ahol

$$\underline{u}_i = \underline{u}(\underline{x}_i) \quad \text{és} \quad \tilde{z}_i = z(\underline{x}_i, \tilde{y}_i), \qquad i=1,\ldots,n. \qquad (6)$$

---

* Transzformációval linearizálható függvények illesztésére speciális módszer adható meg [3], amely nem igényel kezdetiértéket és némileg kevesebb számolást kiván iterációs lépésenként a Gauss-Newton-módszernél, viszont konvergenciatulajdonságai megegyeznek azéval.

Tehát a linearizációs módszer esetén az eredeti S hibanégyzetösszeg helyett a transzformált modell hibáinak

$$S_L = \sum_{i=1}^{n} (\tilde{z}_i - u_1\beta_1 - \ldots - u_p\beta_p)^2 \qquad (7)$$

négyzetösszegét minimalizáljuk; ezt az eljárást a továbbiakban L-modszernek fogjuk nevezni. A (7) kifejezés minimumaként kapott $\underline{\beta}_L$ paramétert visszatranszformálva egy $\underline{\alpha}_L = \underline{\alpha}(\underline{\beta}_L)$ becslés adódik $\underline{\alpha}$-ra.

Azonban a lineáris regresszió alkalmazása a transzformált modellre statisztikailag nem megalapozott. A transzformált modell hibáira nem teljesül többé a zérus várható érték és azonos szórás feltétele, a normalitás még kevésbé, mivel a nemlineáris transzformáció eltorzitja a hibaeloszlást. Ez legegyszerübben az emlitett (5) példán mutatható be; itt a modell

$$\left.\begin{aligned} \tilde{y}_i &= \sqrt{\alpha_1 + \alpha_2 x_i} + \varepsilon_i \,, \\ \tilde{y}_i^2 &= \alpha_1 + \alpha_2 x_i + 2\sqrt{\alpha_1 + \alpha_2 x_i}\,\varepsilon_i + \varepsilon_i^2 \,; \end{aligned}\right\} \qquad (8)$$

a transzformáció utáni hibák:

$$\eta_i = 2\sqrt{\alpha_1 + \alpha_2 x_i}\,\varepsilon_i + \varepsilon_i^2 \,, \qquad (9)$$

tehát

$$M(\eta_i) = M(\varepsilon_i^2) = \sigma^2 > 0 \,,$$

és

$$D^2(\eta_i) = 4(\alpha_1 + \alpha_2 x_i)\,\sigma^2 + 3\sigma^4 \quad \text{függ } x_i\text{-től},$$

igy általában torzitott becslést kapunk az L-módszerrel.

Ha az ε_i mérési hibák kicsik /$\varepsilon_i \approx 0$/, akkor a transzformált modell η_i hibáira

$$M(\eta_i) \approx 0 \quad \text{és} \quad D^2(\eta_i) \approx \sigma^2 \left[\frac{\partial z(\underline{x}_i, \tilde{y}_i)}{\partial y}\right]^2 ; \qquad (10($$

tehát a

$$w_i = \left[\frac{\partial z(\underline{x}_i, y_i}{\partial y}\right]^{-2} \qquad (11)$$

súlyok alkalmazásával közelitőleg kielégithetők a lineáris regresszió feltételei. Ebben az esetben az

$$S_Q = \sum_{i=1}^{n} w_i(\tilde{z}_i - \beta_1 u_1 - \ldots - \beta_p u_p)^2 \qquad (12)$$

súlyozott négyzetösszeg minimumaként adódó $\underline{\beta}_Q$ megoldást visszatranszformálva egy $\underline{\alpha}_Q = \underline{\alpha}(\underline{\beta}_Q)$ becsléshez jutunk. Ezt a módszert /a Michaelis-Menten-féle egyenletre/ Wilkinson [10] alkalmazta először. Kubicek és mtsai [6] az eredeti S hibanégyzetösszeg elsőrendü közelitésével jutottak ugyanehhez a módszerhez, és kvázilinearizált regressziónak ill. Q-módszernek nevezték.

Mind a (7) L-módszer, mind a (12) Q-módszer esetében világos az, hogy zérus ε_i hibák esetén ugyanazt a

megoldást adják, mint a nemlineáris regresszió, mégpedig a paraméterek pontos értékeit /determinisztikus eset/. Folytonossági okok miatt kis σ hibaszórásra is nagy valószinüséggel keveset térnek el a pontos paraméterektől, várhatóan a Q-módszer jobb eredményt ad az L-módszernél. Sajnos általában nem tudjuk eldönteni, hogy elég kicsi-e a hiba, és - ellentétben a fizikával és kémiával - a biológiában a mérési hibák viszonylag nagyok szoktak lenni (1). Ezért tartjuk fontos kérdésnek, hogy nagy hibák esetén mennyivel rosszabb eredményt ad a linearizációs módszer a legkisebb négyzetes becslésnél, ill. mennyit javit az előbbin a (11) súlyozás alkalmazása.

Ezt a problémát két egyszerübb, de a biológiában fontos szerepet játszó modell esetében kivánjuk vizsgálni. Az első a különféle növekedési és bomlási folyamatokat leiró exponenciális /egy-kompartment/ modell /lásd (7)/:

$$\tilde{y} = a \exp(bx) + \varepsilon, \quad a > 0, \quad x > 0 . \tag{13}$$

Mivel ε normális eloszlású, $\tilde{y}$ elvileg negativ is lehet, ezért itt a linearizálás csak az $\varepsilon > -a \exp(bx)$ feltétel mellett végezhető el. A linearizált modell:

$$\left.\begin{aligned} \log \tilde{y} &= \log a + bx + \eta , \\ \text{ahol} \qquad \eta &= \log\left(1 + \frac{\varepsilon}{a \exp(bx)}\right) \end{aligned}\right\} \tag{14}$$

a transzformált hiba. A linearizált modell hibájának $M(\eta \mid \varepsilon > - a \exp(bx))$ feltételes várható értékéről néhány alkalmas integrálbecsléssel belátható, hogy negatív lesz minden, a gyakorlatban szóbajövő pozitiv σ hibaszórás esetén. /Pontosabban $0 < \sigma < a \exp(bx)/\sqrt{3}$ esetén, ami ekvivalens azzal, hogy $1 > P(\varepsilon > -a \exp(bx)) > \phi(\sqrt{3}) \approx 0{,}96$, azaz, hogy legalább 96 %-os valószinüséggel teljesül a transzformáció végrehajthatóságának feltétele./ Erre a modellre a Q-módszer (10-12) szerint a

$$\sum_{i=1}^{n} \tilde{y}_i^2 (\log \tilde{y}_i - \log a - bx_i)^2 \qquad (15)$$

kifejezés minimalizálását jelenti.

A másik tekintett linearizálható modell a biokémiában alapvető jelentőségü Michaelis-Menten-féle enzimkinetikai modell, melynek alakja:

$$\tilde{y} = \frac{ax}{x + b} + \varepsilon , \quad a,b,x > 0 . \qquad (16)$$

E modell reciprok-transzformációval linearizálható:

$$\frac{1}{\tilde{y}} = \frac{1}{a} + \frac{b}{a} \frac{1}{x} + \eta , \qquad (17)$$

ahol η ismét a linearizált modell hibája.

Megjegyezzük, hogy a linearizálás nem egyértelmü, ugyanis ha a fenti egyenletet valamilyen $g(x)$ függvény-

nyel beszorozzuk, az nem változtat a $\underline{\beta}$ paraméterek lineáris voltán. Igy pl. a Michaelis-Menten-egyenletnek többféle linearizálása nevezetes /Lineweaver-Burk-, Hanes-Woolf-, Eadie-Hofstee-féle [4]/, köztük olyanok is, ahol a függő változó az egyenlet mindkét oldalán előfordul. Belátható, hogy a (17) linearizált modell η hibájának még a várható értéke sem létezik; azonban ettől még az L-módszerrel kapott becslések elfogadhatóak lehetnek. Bár ismeretes /[4], [10]/, hogy a (14) szerinti linearizálás nem a legszerencsésebb a becslések pontossága szempontjából, mégis azt választottuk, mert kíváncsiak voltunk arra, hogy mennyit javit rajta a Q-módszer. Utóbbi alakja most

$$\sum_{i=1}^{n} \tilde{y}_i^4 \left(\frac{1}{\tilde{y}_i} - \frac{1}{a} - \frac{b}{a}\frac{1}{x_i}\right)^2 \longrightarrow \min , \qquad (18)$$

ami tulajdonképpen a Michaelis-Menten-egyenlet

$$y = \frac{1}{a} y^2 - \frac{b}{a}\frac{y^2}{x} \qquad (19)$$

implicit alakjának illesztését jelenti.

A Michaelis-Menten-féle egyenlettel kapcsolatban Fajszi és Endrenyi [4] is végeztek szimulációs vizsgálatot, de ők a különféle linearizálások után a "meredekség becslésének" differenciahányadosokon alapuló módszerét elemezték. Mi elsősorban arra a kérdésre keresünk

választ, hogy az itt leirt három módszerrel kapott paraméterbecslések pontossági viszonyai körülbelül hogyan változnak a σ hibaszórás függvényében, ill. hogy alkalmazhatók-e az egyszerübb transzformációs módszerek a nemlineáris legkisebb négyzetek módszere helyett a pontosság különösebb feláldozása nélkül.

A szimuláció eredményeinek bemutatása

A szimuláció R-10-es számitógépen történt; ennek során több száz szimulált mintát dolgoztunk fel. A normális eloszlású hibákat az inverz eloszlásfüggvény segitségével, transzformáció útján állitottuk elő egyenletes eloszlású pszeudo-véletlenszámokból. Az egyenletes eloszlású véletlenszámokat viszont a VIDEOTON BPSS véletlenszámgeneráló szubrutinjának [9] felhasználásával, a Lehmer-féle kongruenciamódszer /lásd [8]/ alapján generáltuk. A független változó értékeit a bemutatásra kerülő ábrák esetében mindig ekvidisztáns beosztásúnak vettük /az [1, 20] intervallumon/, mivel a gyakorlatban az esetek legnagyobb százalékában ilyen beosztás fordul elő. Minden egyes modell esetén /azonos paraméterek és hibaszórás mellett/ 20-szor ismételtük meg a szimulációt és a táblázatok a 20-as ismétlések átlagait mutatják. A paraméterbecslések pontosságát a

$$\varrho_B = \frac{\sqrt{(a_B-a)^2+(b_B-b)^2}}{\sqrt{a^2+b^2}} \qquad (20)$$

relativ /négyzetes/ hibával mértük, ahol a B index a megfelelő becslési módszerre utal.

Az a,b paraméterek becsléseinek együttes φ relativ hibáinak alakulását a (13) exponenciális modellre néhány paraméterpár esetén az 1. ábrán tüntettük fel. Az ábráról a következő összefüggések láthatók.

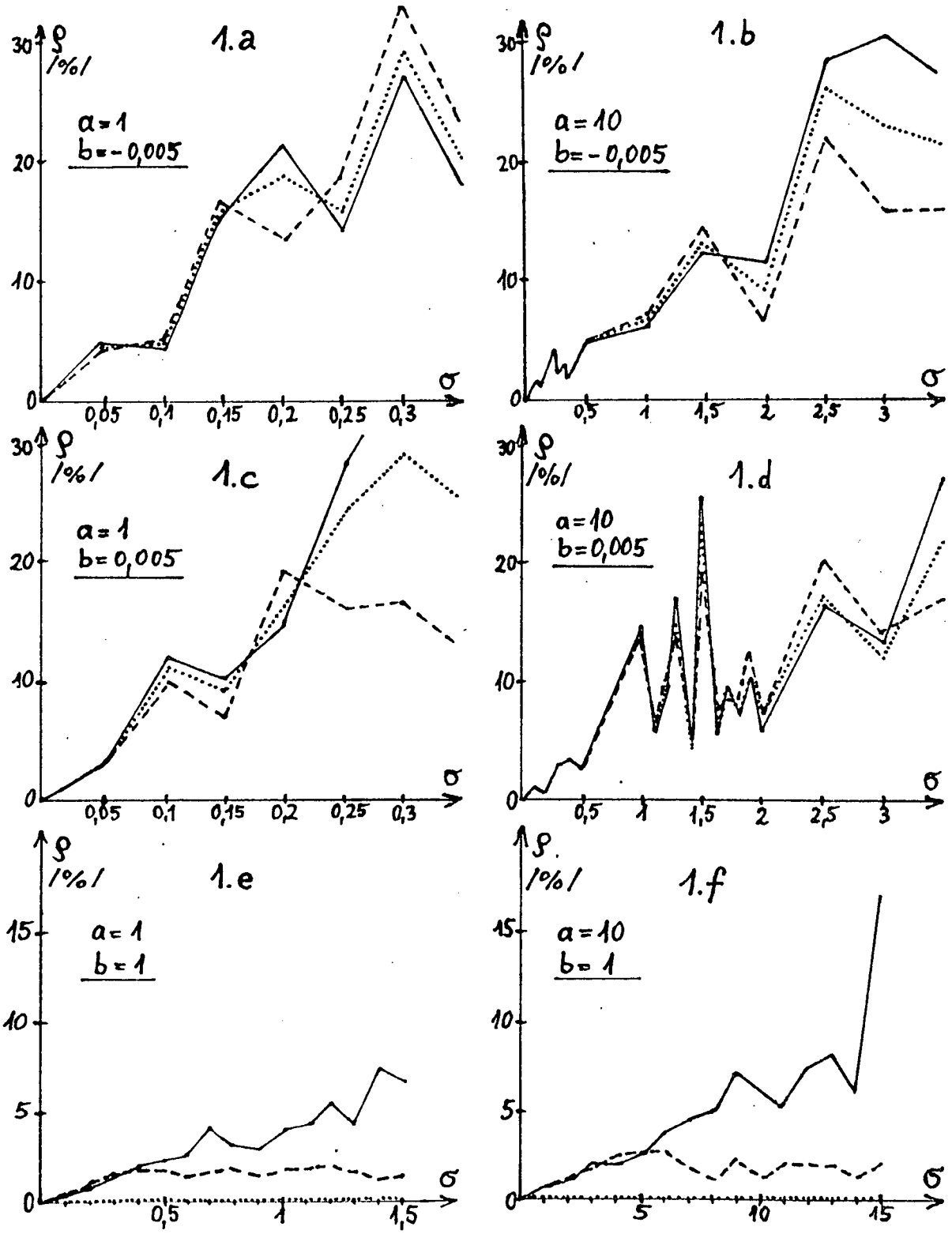
/i/ Negativ /1.a-b. ábrák/, ill. kis pozitiv /1.c-d. ábrák/ kitevőbeli b paraméter esetén a három módszer között lényeges különbség nincsen; tehát nem követünk el nagy hibát, ha a nemlineáris regresszió helyett a linearizáló transzformáció technikáját alkalmazzuk.

/ii/ Pozitiv kitevőbeli b paraméter esetén /1.e-f. ábrák/ a relativ paraméterhibák viselkedése a három módszernél nagyon eltérő. Mig a (7) L-módszerrel kapott becslések relativ hibái egy határozottan növekedő tendenciát követnek, a σ hibaszórás növekedésével, addit a (12) Q-módszerrel nyert becslések relativ hibája bizonyos stabilitást mutat a vizsgált intervallumon; a nemlineáris regressziós becslés pedig mindvégig rendkivül pontosnak bizonyult. Exponenciális növekedési folyamatoknál tehát a Q-módszer lényeges javitást eredményez az

L-módszerhez képest és gyakorlatilag elfogadható pontosságú becsléshez vezet. A nemlineáris regresszió viszont még az előzőnél is lényegesen pontosabb, ezért érdemes iterációs módszerhez folyamodni nagy pontosság igénye esetén.

A 2. ábrán az 1.f. ábrához tartozó exponenciális függvény a=10 és b=1 paraméterei becsléseinek egyedi relativ hibáit ábrázoltuk. Látható, hogy az a paraméter hibája lényegesen nagyobb, mint b-é. Ennek alapján a következő heurisztikus eljárás javasolható: Becsüljük b-t linearizáló transzformáció útján pl. a Q-módszerrel, és utána b igy nyert becslését rögzitve, minimalizáljuk az eredeti (3) S hibanégyzetösszeget a szerint. Ez az eljárás igen egyszerü, kevés számolást igényel, de várhatóan lényegesen javit a becslésén.

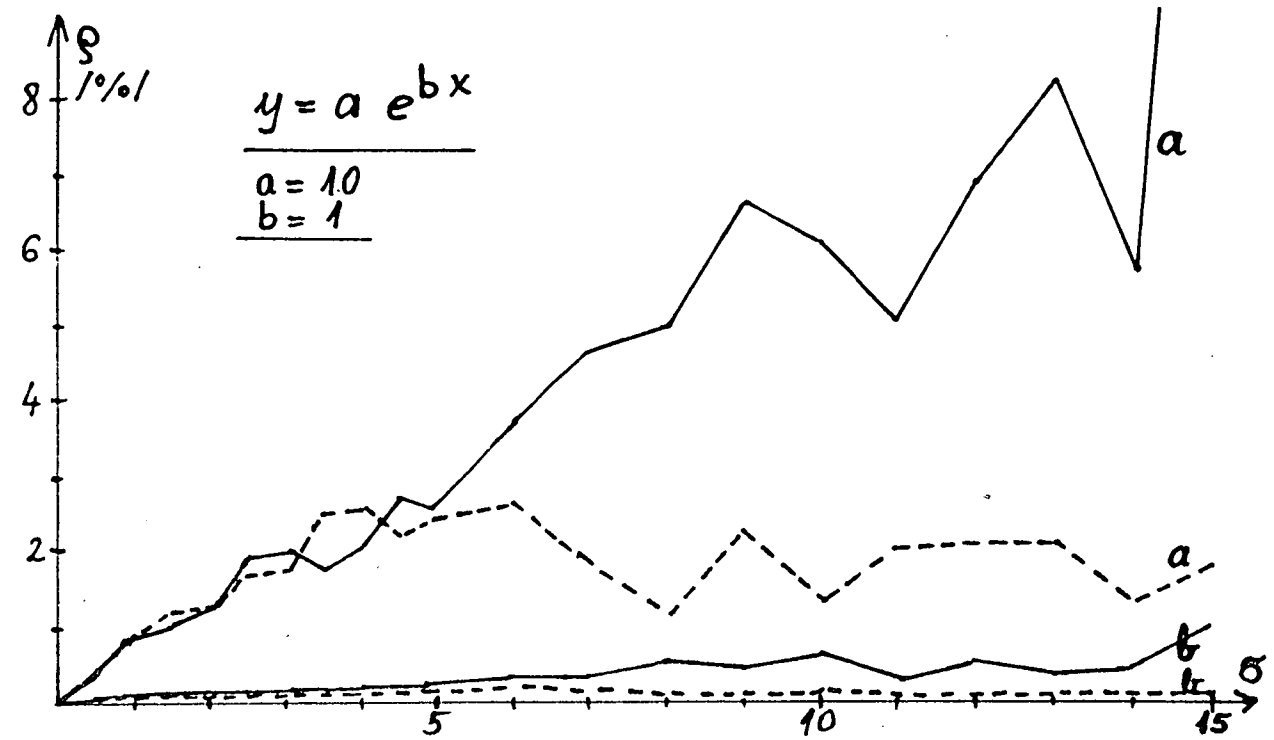
A Michaelis-Menten-féle egyenletnél csak az a=1, b=1 "dimenziómentes" [4] esetet modelleztük. Itt a legkisebb függvényérték $\frac{1}{2}$, a legnagyobb közel 1 volt. Mint azt a 3. ábra is mutatja, már relative elég kicsiny /6 %-os/ hibaszórásnál is óriási /20-30 %-os/ relativ paraméterhibák jelentkeznek; e problémával magyarázható, hogy a Michaelis-Menten-egyenlet paraméterbecslési feladatával cikkek tucatjai foglalkoznak. Ami az általunk vizsgált kérdést illeti, a számitógépes kisérlet eredmé-



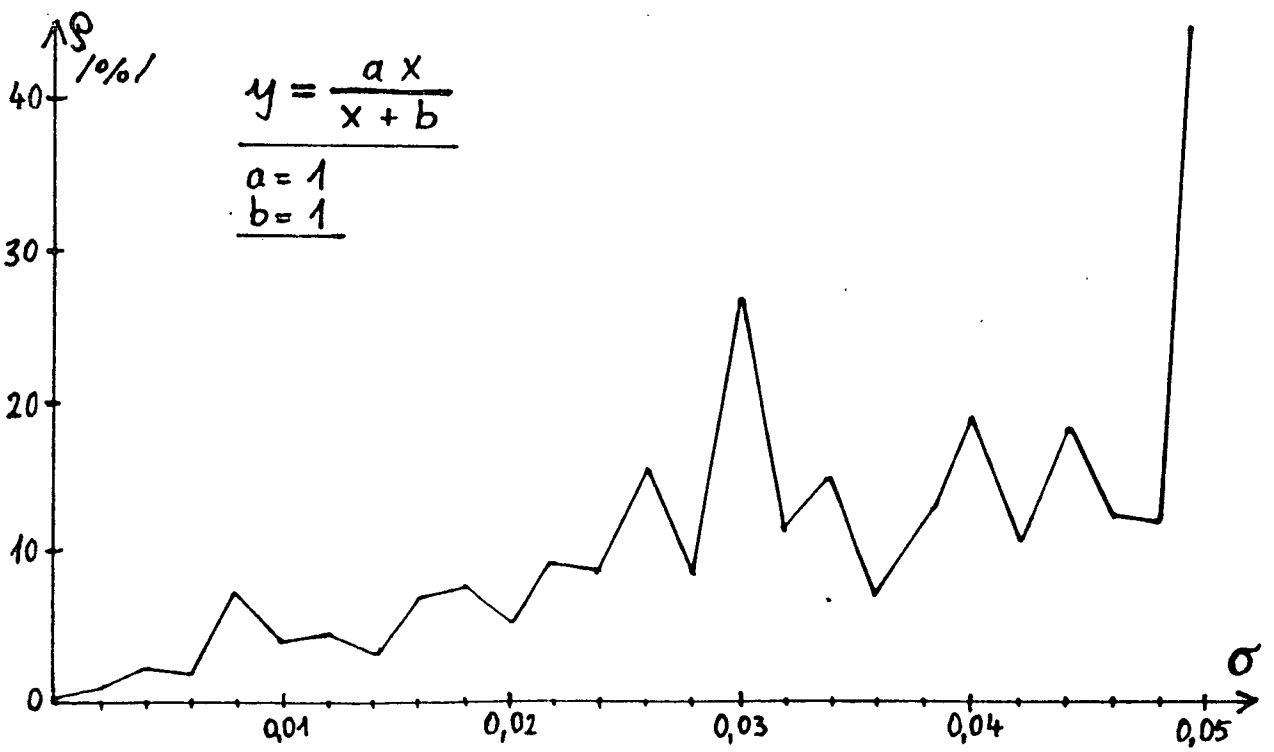
1. ábra

Együttes relativ paraméterhibák az y=aexp(bx) függvény paraméterbecslésénél. Jelölés:

L-módszer: ————, Q-módszer: --------, nemlin.regr.: ·····



2. ábra Egyedi relativ paraméterhibák az y=10exp(x) függvény paraméterbecslésénél. /Jelölés az 1. ábra szerinti/



3. ábra Együttes relativ hibák a Michaelis-Menten egyenlet paraméterbecslésénél. Mindhárom módszerre szinte ugyanez a görbe adódott.

nyei szerint ebben az esetben egyik módszer sem jobb lényegesen a másiknál, sőt a három m módszer relativ hibája olyan közel van egymáshoz, hogy az ábrán nem tudtuk őket elkülöniteni.

I r o d a l o m

[1] Bailey, N. T. J.: The Mathemathical Approach to Biology and Medicine, /2. fejezet/, Wiley, New York, (1967).

[2] Dennis, J. E. Jnr.: Non-linear least squares and equations. lásd: The State of the Art in Numerical Analysis, /szerk. D. A. H. Jacobs/, 269-312. Academic Press, London - New York, (1977).

[3] Eller J.: Nemlineáris legkisebb négyzetek módszerek linearizálható regressziós függvény esetén. "Operációkutatás a gyakorlatban - 1978" konferencia előadás kiv., 95-97., Szeged, (1978).

[4] Fajszi Cs., Endrenyi, L.: New linear plots for the separate estimation of Michaelis-Menten parameters. FEBS Letters 44: 240-246, (1974).

[5] Kendall, M. G., Stuart, A.: The Advanced Theory of Statistics, 2. kötet, 3. kiadás, /18-19. fejezet/, Griffin, London, (1973).

[6] Kubicek, M., Marek, M., Eckert, E.: Quasilinearized regression. Technometrics 13: 601-608, (1971).

[7] Rubinow, S. I.: Introduction to Mathematical Biology, /1-2. fejezet/, Wiley, New York, (1975).

[8] Srejgyer, Ju. A. /szerk./: Monte-Carlo módszerek, Müszaki Kiadó, Budapest, (1965).

[9] VIDEOTON BPSS Fortran nyelvü matematikai statisztikai szubrutinok, 1. kötet, 4-7, (1976).

[10] Wilkinson, G. N.: Statistical estimation in enzyme kinetics. Biochem. J. 80: 324-332, (1961).