COSTEL SÂRBU
LORENTZ JÄNTSCHI

Validarea şi evaluarea statistică a metodelor analitice prin studii comparative

I. Validarea metodelor analitice folosind analiza de regresie

C.Z.: 545.8

The evaluation and validation of analytical methods and instruments require comparison studies using samples of raw material for testing the accuracy and precision. Traditionally, in analytical chemistry the correlation coefficient "r" is calculated and equality is tested with the parameters of the linear model. In this paper the results from ordinary least squares method were compared with a new regression procedure that takes into account the errors in both variables (methods). After a discussion of the properties of the regression procedures, recomandation are given for carrying out a method comparison study. The robust behaviour of the regression procedure proposed is demonstrated by different sata sets.

Măsurările analitice sunt de o importanţă vitală în multe domenii de activitate incluzând diagnosticarea şi tratarea diferitelor boli, protecţia mediului, producerea şi comercializarea unor materiale utile şi sigure şi, nu în ultimul, rând performanţa multor studii ştiinţifice. De aceea, datele provenite din măsurări chimice sau fizico-chimice trebuie să fie reale, adică să fie, înainte de toate, exacte şi precise.

În general, validarea unei metode noi sau a unei metode îmbunătăţite (preţ, timp de analiză, uşurinţă în execuţie etc.) trebuie să asigure integritatea şi calitatea metodei respective din punctul de vedere al preciziei, exactităţii, limitei de detecţie, limitei de determinare, selectivitate, domeniul de liniaritate şi nu în ultimul rând transferabilitatea metodei [1-3]. Adesea, în acest scop rezultatele unui număr mare de probe pe un domeniu larg de concentraţie sunt analizate fie cu ajutorul testului-*t* pe perechi, fie folosind analiza de regresie. Analiza de regresie este metoda statistică preferată în astfel de cazuri pentru că, spre deosebire de testul *t*, este mai puţin restrictivă şi în acelaşi timp furnizează o mai mare cantitate de informaţie [4,5].

Consideraţii teoretice

Analiza de regresie obişnuită

Pentru a putea aplica analiza de regresie în vederea validării unei metode noi sau a unei metode îmbunătăţite, probe de concentraţii diferite ale speciei chimice de interes sunt analizate atât cu metoda propusă, cât şi cu o metodă standard sau de referinţă caracterizată prin performanţe ridicate şi recunoscute. Dacă se reprezintă grafic rezultatele obţinute cu metoda testată faţă de cele obţinute cu metoda standard sau de referinţă se aşteaptă obţinerea unei linii dreapte cu o pantă care să nu difere semnificativ de unitate şi cu o intercepţie care să nu difere semnificativ de zero. O pantă cu valoarea diferită semnificativ de unu indică o eroare proporţională (de exemplu interferenţe sau efecte de matrice), iar o intercepţie diferită semnificativ de zero indică o eroare constantă (de exemplu probleme privind fondul). Pentru estimarea pantei (a_1) şi a intercepţiei (a_0) în cazul unui astfel de model liniar (1) se foloseşte metoda celor mai mici pătrate obişnuită (MCMMPO).

$$y = a_0 + a_1 x \qquad (1)$$

Conf. dr. COSTEL SÂRBU şi ing. LORENTZ JÄNTSCHI, Universitatea "Babeş-Bolyai", Facultatea de Chimie şi Inginerie Chimică, Cluj-Napoca

Testarea semnificației diferențelor față de 1 pentru pantă și față de 0 pentru intercepție se realizează fie prin verificarea includerii lui 1 în intervalul de încredere al pantei, $a_1 \pm t_{\alpha/2,\, n-2} \cdot s_{a1}$, și respectiv a lui 0 în intervalul de încredere al intercepției $a_o \pm t_{\alpha/2,\, n-2} \cdot s_{ao}$, fie aplicând testul t pentru pantă (2) și respectiv intercepția (3).

$$t = \frac{|a_1 - 1|}{S_{a1}} \tag{2}$$

$$t = \frac{|a_{ao} - 0|}{S_{ao}} \tag{3}$$

unde:

a_o și a_1 reprezintă valorile estimate ale intercepției și ale pantei;

α - este nivelul de semnificație pentru un test bilateral;

n - reprezintă numărul datelor folosite pentru comparare;

$t_{\alpha/2,n-2}$ - reprezintă valoarea tabelată a lui t pentru un test bilateral cu n-2 grade de libertate;

t - este valoarea calculată care trebuie comparată cu valoarea tabelată;

s_{a1} și s_{ao} - reprezintă deviația standard a pantei și respectiv a intercepției și se calculează cu ajutorul expresiei (4) și respectiv (5).

$$s_{a1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{4}$$

$$s_{ao} = s\sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}} \tag{5}$$

x_i reprezintă o valoare individuală; $\bar{x}$ este valoarea medie a valorilor x_i, iar s reprezintă abaterea standard a reziduurilor și se calculează cu ajutorul expresiei (6):

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \tag{6}$$

unde $\hat{y}_i$ reprezintă valoarea estimată (calculată) a lui y_i, folosind ecuația (1).

În general, metoda celor mai mici pătrate în forma obișnuită poate fi folosită numai dacă sunt îndeplinite următoarele condiții:

1) variabila independentă x poate fi măsurată fără eroare;

2) erorile (y_i-y_i) sunt independente și normal distribuite cu media zero, iar varianța s^2 este constantă (cazul homoscedastic).

În practică, însă, aceste restricții sunt foarte rar întâlnite. Multe metode analitice realizează rezultate heteroscedastice cu coeficientul de variație mai degrabă constant decât cu varianța constantă așa cum o cere metoda celor mai mici pătrate obișnuită. Totuși, în asemenea cazuri, când se compară două metode analitice, cea mai mare problemă o constituie faptul că ambele metode sunt afectate de erori. De aceea metoda celor mai mici pătrate devine inaplicabilă și trebuie folosite metode de regresie care țin cont de erorile de măsurare manifestate în cadrul ambelor metode analitice comparate [6-8].

Un nou algoritm de regresie

Atunci când se compară două metode analitice deoarece ambele sunt afectate de erori, într-o măsură mai mică sau mai mare, practic este indiferent care este x și care este y. Adică, se poate scrie la fel de bine: fie y=f(x), fie x = f(y). Într-o asemenea situație, apare, astfel, mult mai înțelept să se considere forma generală a unei drepte redată de relația (7) sau (8):

$$ax + by + c = 0 \tag{7}$$

sau:

$$f(x,y) = ax + by + c \tag{8}$$

În acest caz, distanța oricărui punct (x_i, y_i) la dreapta redată prin ecuația (7) va avea următoarea expresie:

$$d_i^2 = \frac{(ax_i + by_i + c^2)}{a^2 + b^2} \tag{9}$$

Problema de ajustare rezultată poate fi definită prin expresia (10):

$$S = \frac{1}{a^2 + b^2} \sum_{i=1}^{n} (ax_i + by_i + c)^2 = min \tag{10}$$

Condiția de minimum a lui S conduce la sistemul cu ecuații normale rezultat din anularea derivatelor în raport cu a,b și respectiv c. Considerând M(x) = 0, M(y) = 0 și respectiv c=0 se obține următoarea ecuație:

$$\frac{M(y^2) - M(x^2)}{M(xy)} = \frac{b^2 - a^2}{ab} = \frac{b}{a} - \frac{a}{b} \tag{11}$$

Se observă că raportul a/b în ecuația (11) reprezintă panta ecuației liniare rezultată din (7).

$$y = mx + n \tag{12}$$

După substituirea lui m în ecuația (11) și considerând $\frac{M(y^2) - M(x^2)}{M(xy)} = w$ se obține ecuația pătratică (13),

$$m^2 - wm - 1 = 0 \tag{13}$$

a cărei valoare pozitivă m permite calcularea lui n în ecuația (12) dacă se consideră, în plus, valoarea centrului de greutate a lui x și y ce trebuie să satisfacă, de asemenea, ecuația dreptei (12), adică,

$$y = mx + \bar{y} - m\bar{x} \tag{14}$$

Comparând ecuațiile (7) și (12) se identifică ușor $c = -nb$ și $a = -mb$, iar dacă se consideră $c = 1$, rezultă imediat valoarea lui a și b, parametrii de regresie ai dreptei scrisă în forma cea mai generală așa cum se arată în ecuațiile (7) și (8). Astfel, devine posibilă compararea metodele analitice considerând fie raportul a/b sau b/a (ecuațiile sunt simetrice), fie direct valorile acestora. Considerând acestea din urmă se obseservă că în cazul ideal, când cele două metode realizează practic rezultate identice, a și b vor avea aceeași valoare absolută. În caz contrar, cu cât diferența dintre a și b va fi mai mare cu atât cele două metode vor fi mai diferite în ceea ce privește rezultatele obținute. Testarea diferenței semnificative dintre cei doi coeficienți a și b se poate realiza folosind analiza de varianță informațională bazată pe energia informațională, $E = \sum_{i=1}^{n} p_i^2$ [9,10]. În acest caz, ipoteza de nul este echivalentă cu următoarea ipoteză

H*: $p_a = p_b$, iar probabilitățile p se calculează cu ajutorul relațiilor (15).

$$p_a = \frac{a}{a+b} \quad si \ respectiv \quad p_b = \frac{b}{a+b} \tag{15}$$

Așa cum s-a arătat [9,10] ipoteza de nul se reduce practic la testarea egalității E = ε, unde,

în acest caz, E = 1/2 reprezintă energia informațională teoretică, $\varepsilon = \frac{a^2 + b^2}{(a+b)^2}$, reprezintă energia informațională empirică. Dacă E = ε, rezultă că nu există o diferență statistică semnificativă între cei doi coeficienți a și b considerați în valoare absolută, iar dimpotrivă dacă E ≠ ε atunci se poate aprecia că a și b diferă semnificativ, adică, astfel spus, cele două metode testate sunt semnificativ diferite din punctul de vedere al rezultatelor furnizate.

Rezultate şi discuţii

Pentru a ilustra modul de aplicare al analizei de regresie, precum şi avantajele algoritmului propus în această lucrare privind validarea (compararea) metodelor analitice, se fac referiri la trei cazuri mediatizate pe larg în literatura de specialitate. Primul caz (exemplul 1) [5] se referă la compararea unei metode gravimetrice (MG), deci o metodă absolută foarte exactă şi foarte precisă, cu o metodă electrochimică folosind un electrod membrană ion selectiv (EMIS), adică o metodă instrumentală relativă, mai puţin precisă şi exactă. Rezultatele privind determinarea sulfurii (mg) sunt prezentate în tabelul 1.

Al doilea caz (exemplul 2) [5] se referă la determinarea plumbului în sucul de fructe (ppb) printr-o metodă potenţiometrică de analiză prin striping (MPS) folosind un electrod de lucru din carbon sticlos şi prin tehnica de absorbţie atomică fără flacără (MAA). Diferenţa dintre rezultatele obţinute cu cele două metode (tabelul 1) a fost testată folosind analiza de regresie obişnuită. Ultimul caz discutat [11] se referă la compararea rezultatelor obţinute la determinarea mercurului (ppm) în probe reziduale solide prin spectrometrie de absorbţie atomică folosind două metode de dezagregare, dezagregarea în cuptor cu microunde (I) şi respectiv dezagregarea în baie de apă (II). Rezultatele sunt prezentate, de asemenea, în tabelul 1.

În tabelul 2 sunt prezentate rezultatele de regresie obţinute folosind metoda celor mai mici pătrate obişnuită considerând ca metodă standard pentru primul exemplu metoda gravimetrică, iar pentru exemplul al doilea s-a considerat ca metodă de referinţă metoda de absorbţie atomică; în exemplul al treilea metoda de dezagregare în cuptor cu microunde a fost aleasă ca abscisă.

Tabelul 1
DATELE ANALIZATE PRIVIND COMPARAREA METODELOR ANALITICE

Exemplul 1		Exemplul 2		Exemplul 3	
MG	EMIS	MPS	MAA	(I)	(II)
105	108	35	35	7,32	5,48
16	12	70	75	15,8	13,0
113	152	80	75	4,60	3,29
0,0	3,0	80	80	9,04	6,84
108	106	120	125	7,16	6,00
11	11	200	205	6,80	5,84
141	128	220	205	9,90	14,3
11	12	200	215	28,7	18,8
182	160	250	240		
118	128	330	350		

În acelaşi tabel este redată, de asemenea, valoarea coeficientului de corelaţie, r şi a coeficientului de determinare R^2. Analiza intervalelor de încredere pentru parametrii de regresie, intercepţie şi respectiv pantă, nu indică o diferenţă semnificativă între metodele comparate în cazul primelor două exemple discutate, deoarece intercepţiile şi respectiv pantele sunt incluse în intervalele de încredere corespunzătoare unui nivel de semnificaţie $\alpha = 0,05$. Pentru exemplul 3, practic intervalul de încredere al pantei nu include valoarea ideală 1. Rezultă, aşadar, că în acest caz cele două metode de dezagregare comparate diferă semnificativ. O analiză atentă a datelor (incluzînd, de asemenea, valoarea coeficientului de determinare, R^2 = 78,75%) ilustrează prezenţa erorilor proporţionale introduse de metoda de dezagregare cu microunde.

Tabelul 2

REZULTATELE DE REGRESIE OBȚINUTE FOLOSIND METODA CELOR MAI MICI PĂTRATE PENTRU CELE TREI EXEMPLE DISCUTATE

Parametrul statistic	Exemplul 1	Exemplul 2	Exemplul 3
r	0,9682	0,9945	0,8874
R^2	0,9404	0,9891	0,7875
a_o	4,4837	3,8667	2,2791
a_1	0,9629	0,9634	0,6193
s	16,726	10,567	2,7171
s_{ao}	8,6939	6,6431	1,7529
s_{a1}	0,0857	0,0358	0,1313
$a_o \pm t_{\alpha/2,n-2} \cdot s_{ao}$	4,48±20,12	3,87±15,34	2,28±4,31
$a_1 \pm t_{\alpha/2,n-2} \cdot s_{a1}$	0,96±0,20	0,96±0,08	0,62±0,32

Rezultatele obținute aplicând noul algoritm de regresie care iau în considerare erorile manifestate în ambele metode comparate sunt prezentate în tabelul 3. Având în vedere simetria algoritmului putem considera fie y ca funcție de *x* (*y*=f(*x*)) sau x ca funcție de *y* (*x*=f(*y*)) fără ca semnificația rezultatului final să se schimbe.

Examinarea rezultatelor din tabelul 3 indică o diferență nesemnificativă între metodele analitice discutate în exemplul 1 și respectiv exemplul 2 deoarece E = ε = 0,500. Referindu-ne la exemplul 3, deoarece E ≠ ε (E=0,500 și ε = 0,520) trebuie să se admită că cele două metode de dezagregare comparate nu conduc la rezultate statistic identice. Trebuie subliniat, că, deși, rezultatele obținute cu noul algoritm discutat în această lucrare sunt asemănătoare cu cele furnizate de metoda celor mai mici pătrate obișnuită, totuși, aceasta nu se întâmplă întotdeauna. Astfel, dacă se inversează rolul metodelor comparate, adică se schimbă axele, rezultatele obținute sunt total diferite. De exemplu, dacă se fac referiri la exemplul 3, alegând de data aceasta ca metodă de referință metoda de dezagregare în baia de apă se obține pentru intercepție valoarea a_o = 0,41, iar pentru pantă valoarea a_1 = 0,79.

Tabelul 3

REZULTATELE OBȚINUTE APLICÂND NOUL ALGORITM DE REGRESIE PENTRU DATELE DIN TABELUL 1 PRIVIND COMPARAREA METODELOR ANALITICE

Parametrul de regresie	Exemplul 1	Exemplul 2	Exemplul 3
y = mx+n			
a	-4,802	3,182	3,856
b	4,767	-3,286	-5,770
c	1,0	1,0	1,0
m	1,007	0,969	0,668
n	-2,097	3,043	1,733
E	0,500	0,500	0,500
ε	0,500	0,500	0,520
x = m*y+n*			
a	4,767	-3,286	-5,770
b	-4,802	3,182	3,856
c	1,0	1,0	1,0
m*	0,993	1,032	1,496
n*	2,082	-3,142	-2,593
E	0,500	0,500	0,500
ε	0,500	0,500	0,520

Concluzii

Un nou algoritm de regresie care ia în considerare erorile manifestate în ambele metode sau procedee de analiză comparate a fost discutat şi raportat la tehnica de regresie obişnuită bazată pe metoda celor mai mici pătrate. Este evident, că în studiile privind compararea metodelor analitice când erorile se manifestă în ambele variabile (metode), metoda celor mai mici pătrate poate conduce la estimaţii eronate ale parametrilor de regresie şi în mod implicit la decizii false privind performanţele metodei testate. Rezultatele obţinute cu noul algoritm de regresie sunt mult mai relevante, iar testarea diferenţei semnificative dintre metode prin compararea coeficienţilor de regresie cu ajutorul analizei de varianţă informaţională pare a fi, de asemenea, mult mai obiectivă.

Bibliografie

1. CARDONE, J.M., J. Assoc. Off. Anal. Chem., **66**, 1983, p. 1257
2. MILLER, J.N., Analyst, **116**, 1991, p. 3
3. ALEXANDROV, Yu. I., Analyst, **121**, 1996, p. 1137
4. MASSART, D.J., VANDEGINSTE, B. M.G., DEMING, S.N., MICHOTTE, Y. şi KAUFMAN, L., Chemometrics: A Textbook, Elsevier, Amsterdam, 1988, p. 88
5. MILLER, J.C. şi MILLER, J.N., Statistics for Analytical Chemistry, Horwood, New York, 1988, p. 101
6. PASSING, H. şi BABLOK, W., J. Clin. Chem. Clin. Biochem., **21**, 1983, p. 709
7. HARTMANN, C., SMEYERS-VERBEKE, J. şi MASSART, D.L., Analusis, **21**, 1993, p. 125
8. KALANTAR, A.H., GELB, B.R. şi ALPER, J.S., Talanta, **42**, 1995, p. 597
9. SÂRBU, C., Anal. Chim. Acta, **271**, 1993, p. 269
10. SÂRBU, C., Anal. Lett., **30**, 1997
11. MAW, R., WITRY, L. şi EMOND, T., Spectroscopy, **9**, 1994, p. 39

Intrat în redacţie: 24 I 1994