UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Regressió Lineal amb Errors en Ambdós Eixos. Aplicació a la Calibració i a la Comparació de Mètodes Analítics.

Tesi Doctoral

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

UNIVERSITAT ROVIRA I VIRGILI

Departament de Química Analítica i Química Orgànica

Àrea de Química Analítica

REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS

Memòria presentada per

JORDI RIU RUSELL

per assolir el grau de

Doctor en Química

Tarragona, 1999

UNIVERSITAT ROVIRA I VIRGILI

DEPARTAMENT DE QUÍMICA ANALÍTICA
I QUÍMICA ORGÀNICA

Plaça Imperial Tàrraco, 1
43005 Tarragona
Tel. 34 977 55 81 37
Fax 34 977 55 95 63
e-mail: secqaqo@quimica.urv.es

Dr. FRANCESC XAVIER RIUS i FERRÚS, Catedràtic del Departament de Química Analítica i Química Orgànica de la Facultat de Química de la Universitat Rovira i Virgili,

CERTIFICA: Que la present memòria que duu per títol: **"REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS"**, ha estat realitzada per en JORDI RIU i RUSELL sota la meva direcció a l'Àrea de Química Analítica del Departament de Química Analítica i Química Orgànica d'aquesta Universitat i que tots els resultats presentats són fruit de les experiències realitzades per l'esmentat doctorant.

Tarragona, desembre de 1998

Prof. F. Xavier Rius i Ferrús

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Aquests fulls intenten resumir quasi cinc anys de treball. He intentat no fer una llista exhaustiva per no oblidar a ningú, i perquè a fi de comptes, ja us tinc ben presents a tots aquells que sou importants per a mi. Amb tot, voldria expressar el meu més profund agraïment a:

Al professor F. Xavier Rius, per tota la seva confiança i ànims durant aquest llarg camí, i per totes les seves ensenyances, no només dins del camp científic.

A l'Àngel i el Javi, per la vostra inestimable col·laboració dins d'aquesta tesi.

A tota la gent del grup de quimiometria amb els qui he conviscut els darrers anys, pels bons i no tant bons moments que hem passat junts: Marisol, Pili, Itziar, Ricard, Joan, Jaume, Floren, Santi, Toni, Alicia M., Alicia P., Josep Lluís i a la gent que hi ha passat però que per diverses circumstàncies ja no estan al grup: Jaume Ramon, Justo, Alex, Cristina i Yolanda.

A tota la resta de gent de l'Àrea de Química Analítica.

Al meu pare i a la meva mare, perquè sense ells tot aquest treball no seria possible. A la meva mare i al meu germà, per tots els seus ànims i suport.

A l'Annabel, perquè ja saps tot el que signifiques i l'important que ets per a mi.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Als de casa,
perquè aquest treball
és tant vostre com meu.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

ÍNDEX

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Objecte de la tesi doctoral

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

La present tesi doctoral és una contribució als estudis de comparació de mètodes analítics considerant sempre els errors individuals associats als resultats obtinguts amb cada mètode analític. Amb aquest fi, hom ha plantejat una sèrie d'objectius individuals entre els quals poden esmentar-se:

- Una revisió crítica de les metodologies desenvolupades fins a la data per calcular els coeficients de regressió en regressió univariant.

- El desenvolupament i la validació d'un test estadístic per avaluar les probabilitats d'error β en els tests individuals per a l'ordenada en l'origen i el pendent de la recta de regressió trobada.

- El desenvolupament i la validació d'un test estadístic útil per detectar biaix en la comparació de dos mètodes analítics.

- El desenvolupament i la validació d'un test estadístic per a comparar múltiples mètodes analítics tenint en compte els errors deguts a cada un dels mètodes analítics en comparació.

- El desenvolupament i la validació de les expressions per calcular la incertesa de la variable resposta donat un valor de la variable predictora i viceversa en regressió univariant.

- La generació d'algorismes corresponents en suport informàtic per facilitar la implantació en els laboratoris d'anàlisi dels tests desenvolupats.

- L'aplicació pràctica dels tests desenvolupats a diverses problemàtiques analítiques.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Capítol 1

Introducció general

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

En aquest primer capítol, una vegada descrits els objectius de la tesi doctoral, hom introdueix el conjunt de treballs de recerca realitzats. En primer lloc, es relaciona la comparació de mètodes amb la traçabilitat del resultat analític i s'emmarca dins el procés de validació de les metodologies analítiques, i així es justifiquen els objectius plantejats. En segon lloc, la breu perspectiva històrica de l'evolució de la regressió lineal tant univariant com multivariant presenta els inconvenients dels mètodes de regressió lineal tradicionals i proporciona la base per entendre les tècniques de regressió que consideren els errors en tots els eixos. Tot seguit es posen les bases perquè el lector pugui seguir la resta de tests estadístics desenvolupats en els capítols següents d'aquesta tesi doctoral. Finalment, la darrera secció detalla com s'han portat a terme aquests treballs descrivint l'estructura del conjunt de la tesi.

1.1 Notació

Les matrius són representades en majúscula i negreta (p.e. **R**), els vectors en minúscula i negreta (p.e. **c**) i els escalars en cursiva (p.e. c_k).

Símbols començant amb una lletra de l'alfabet llatí

a	valor vertader de l'ordenada a l'origen de la recta o hiperplà de regressió
$\hat{a}$	estimació de l'ordenada a l'origen de la recta o hiperplà de regressió
b	valor vertader del pendent de la recta de regressió
$\hat{b}$	estimació del pendent de la recta de regressió
b_j	valor vertader del j-èssim pendent de l'hiperplà de regressió
$\hat{b}_j$	estimació del j-èssim pendent de l'hiperplà de regressió
b	vector amb els valors vertaders dels coeficients de regressió

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

$\hat{\mathbf{b}}$	vector estimació dels coeficients de regressió
e_i	valor residual en el punt *i*
$\mathbf{e}$	vector d'errors
$F_{\nu1,\nu2,\alpha}$	valor de la distribució *F* de Fischer per a un nivell de significança α (1 cua) amb $\nu1$ i $\nu2$ graus de llibertat
$F_{\nu1,\nu2,\alpha/2}$	valor de la distribució *F* de Fischer per a un nivell de significança α (2 cues) amb $\nu1$ i $\nu2$ graus de llibertat
m_{j_i}	resultat obtingut amb el mètode *j* en el punt *i*
$\hat{m}_{j_i}$	valor predit del resultat obtingut amb el mètode *j* en el punt *i*
n	nombre de punts experimentals
q	número de repeticions fetes sobre una determinada mostra
s^2	valor vertader de l'error experimental de la recta de regressió
$\hat{s}^2$	estimació de l'error experimental de la recta de regressió
$\hat{s}_a^2$	estimació de la variància de l'ordenada a l'origen de la recta de regressió
$\hat{s}_b^2$	estimació de la variància del pendent de la recta de regressió
$s_{x_i}^2$	variància de la variable predictora en el punt *i*
$s_{y_i}^2$	variància de la variable resposta en el punt *i*
$t_{\alpha,\nu}$	valor de la distribució *t* de Student per a un nivell de significança α (1 cua) amb ν graus de llibertat
$t_{\alpha/2,\nu}$	valor de la distribució *t* de Student per a un nivell de significança α (2 cues) amb ν graus de llibertat
$u_{\alpha/2}$	valor de la distribució normal per a un nivell de significança α
w_i	coeficient de ponderació en el punt *i*
x	variable predictora
x_i	valor mesurat de la variable predictora en el punt *i*
$\hat{x}_i$	valor predit de la variable resposta en el punt *i*

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

$\bar{x}$	valor mitjà dels valors experimentals de la variable predictora
$\bar{x}_p$	coordenada x del centroide ponderat
$\mathbf{X}$	matriu de la variable predictora
y	variable resposta
y_i	valor mesurat de la variable resposta en el punt i
$\hat{y}_i$	valor predit de la variable resposta en el punt i
$\bar{y}$	valor mitjà dels valors experimental de la variable resposta
$\bar{y}_p$	coordenada y del centroide ponderat
$\mathbf{y}$	vector d'observacions

Símbols començant amb una lletra de l'alfabet grec

α	error de primera espècie o de tipus I
β	error de segona espècie o de tipus II
Δ	biaix, màxima diferència acceptable entre el paràmetre estimat i un valor de referència
η_p	moment d'ordre p
μ_{x_i}	valor vertader de la variable predictora en el punt i
μ_{y_i}	valor vertader de la variable resposta en el punt i
λ	relació entre els errors de les variables resposta i predictora en regressió CVR
ρ_i	valor vertader de la covariància entre les variables predictora i resposta en el punt i
$\sigma^2_{x_i}$	valor vertader de la variància de la variable predictora en el punt i
$\sigma^2_{y_i}$	valor vertader de la variància de la variable resposta en el punt i

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

1.2 Comparació de mètodes. Traçabilitat

Assegurar l'exactitud d'un mètode analític és un dels passos fonamentals en el seu procés de desenvolupament. Segons Eurachem/Welac,[1,2] "la validació d'un mètode estableix, mitjançant estudis sistemàtics al laboratori, que les seves característiques compleixen les especificacions relacionades amb l'ús que es vol destinar als resultats analítics. Les característiques a determinar inclouen: selectivitat i especificitat, interval d'aplicació, linealitat, sensitivitat, límit de detecció, límit de quantificació, robustesa, exactitud i precisió."

L'exactitud, segons les normes ISO,[3] es defineix com "el grau de concordança entre el resultat d'una mesura i el valor acceptat com a referència". El terme exactitud, aplicat a un resultat analític, inclou la combinació de components aleatoris i components deguts a l'error sistemàtic o biaix. La impossibilitat de mesurar la proximitat entre cada valor individual i el valor de referència, és a dir, la impossibilitat d'assegurar l'exactitud per a cada resultat individual, ha fet que el terme exactitud es vagi substituint pel terme traçabilitat.[2] El terme traçabilitat ha estat emprat en mesures físiques durant molts anys, però tot just s'està introduint en el camp químic. La traçabilitat es pot definir com "la propietat del resultat d'una mesura que consisteix que es pugui establir el resultat previsible de la seva comparació directa amb els patrons apropiats, generalment nacionals o internacionals, mitjançant una cadena ininterrompuda de comparacions reals, totes amb les seves incerteses".[4,5] Seguint el terme traçabilitat, la filosofia per assegurar la validesa dels resultats analítics consisteix a comprovar que el mètode analític funciona adequadament i a utilitzar-lo sota garanties de qualitat. D'aquesta manera, els resultats obtinguts amb un mètode analític traçable seran considerats correctes.

El concepte de traçabilitat va guanyant força dins del camp químic, i ja les normes EN 45001/UNE 66501 [6] exigeixen que cada resultat analític vagi acompanyat de dos paràmetres de qualitat bàsics: traçabilitat i incertesa. Malgrat aquesta duplicitat

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

en les definicions, cal reconèixer que l'exactitud i la traçabilitat dels resultats analítics estan íntimament units. La traçabilitat no té sentit si no es persegueix l'exactitud del resultat que es dóna.[5]

La verificació de la traçabilitat d'un mètode analític, per tant, és un pas clau dins del seu desenvolupament i validació. Per tal de dur a terme aquesta verificació, es comparen els resultats obtinguts amb el mètode amb una referència. De referències n'hi ha de molts tipus. Com més bona sigui l'escollida (traçable a patrons metrològics més elevats), més segurs estarem de la traçabilitat del nostre mètode. La millor referència la constitueixen els que s'anomenen mètodes definitius. És a dir, la millor forma de verificar la traçabilitat del nostre mètode és comparar els resultats obtinguts de l'anàlisi d'una mostra o sèrie de mostres representatives amb el nostre mètode, amb els resultats obtinguts d'analitzar la mateixa mostra o sèrie de mostres amb un dels anomenats mètodes definitius. Els mètodes definitius són aquells mètodes que es poden traçar directament al mol i són:[7]

- Espectrometria de masses amb dilució isotòpica.
- Volumetria.
- Coulombimetria.
- Gravimetria.
- Grup de mètodes col·ligatius, incloent-hi la disminució de la pressió de vapor, augment del punt de fusió, disminució de la temperatura d'ebullició i la pressió osmòtica, tots basats en el teorema que en qualsevol solució prou diluïda el solvent es comporta idealment.

La utilització d'un d'aquests mètodes suposa la millor referència possible, sempre que siguin aplicats en condicions de garanties de qualitat. Lògicament, no en tots els casos pot ser comparat el mètode analític a validar amb un mètode definitiu, pel que s'haurà d'emprar una altra referència.

UNIVERSITAT ROVIRA·I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

La referència següent, en ordre d'importància metrològica, són els materials de referència certificats (CRM). Un material de referència (RM), segons les definicions de la ISO (*International Organization for Standardization*) en la ISO Guide 30,[8] és un material o substància que té una o diverses de les seves propietats prou ben establertes, de manera que en permet l'ús per calibrar un aparell o instrument, validar un mètode analític o assignar valors a un material o sistema. Un material de referència certificat[8] és aquell material de referència que té certificats un o diversos valors d'una o més propietats per procediments tècnicament vàlids duts a terme per un organisme competent. Aquestes definicions són acceptades per les quatre organitzacions internacionals implicades en metrologia:

- *International Bureau of Weights and Measurements* (BIPM)
- *International Electrochemical Commission* (IEC)
- *International Organization for Standardization* (ISO)
- *International Organization for Legal Metrology* (OILM)

Un material de referència certificat ha de complir una sèrie de requisits per poder-se considerar com a tal: s'ha de conèixer el valor de la concentració o del paràmetre que es vulgui determinar, ha de ser estable durant un període raonable de temps i ha de ser homogeni. A més, en la mesura que sigui possible també ha de complir dos requisits més, portar associat el valor de la precisió al valor de la concentració o paràmetre que es vulgui determinar, i ser tan semblant com sigui possible a les mostres reals que s'analitzaran amb el mètode analític.

Els inconvenients principals dels materials de referència certificats són el fet que només n'hi ha per a aproximadament el 5% de les anàlisis que es fan actualment i el preu elevat. El fet que hi hagi un nombre tan baix de materials de referència certificats és degut principalment al gran i divers nombre de matrius existent dins de les mostres reals. Així, per exemple, aquesta causa pot fer que un possible material de referència de plom en un vi Riestling tingui una matriu tan diferent

d'un vi del Penedès, que la seva utilització en el procés de validació d'un mètode per a l'anàlisi de plom en vins que després serà aplicat a vins del Penedès pugui donar lloc a resultats erronis.

La tercera referència en ordre d'importància metrològica la constitueixen els mètodes de referència. S'analitzen una sèrie de mostres representatives amb el mètode que volem validar i amb el mètode de referència, i es comparen els resultats. Lògicament, tal com comentàvem en el cas dels mètodes definitius, aquesta comparació serà vàlida sempre que els mètodes de referència siguin aplicats en condicions de garanties de qualitat.

A la figura 1.1 es pot observar una classificació de les referències segons l'ordre d'importància. Un cas extrem seria el d'algun analit que només sigui analitzat per un laboratori en tot el món. Llavors, probablement l'única referència seria el mateix laboratori al llarg del temps.

1.3 Perspectiva històrica. Definicions relacionades amb la regressió lineal

La regressió lineal és una de les operacions més freqüents en la química analítica. Dins de la regressió lineal trobem la regressió lineal univariant i la multivariant, segons el nombre de variables relacionades (una variable resposta i una variable predictora en regressió lineal univariant, una variable resposta i més d'una variable predictora en regressió lineal multivariant). La regressió lineal univariant és àmpliament emprada, moltes vegades perquè hi ha un suport teòric en forma de llei (equació de Lambert-Beer, equació d'Ilkovich, etc.) que justifica la relació lineal entre la resposta instrumental i la concentració d'analit present.

Pel que fa a la regressió lineal, ens centrarem en els anomenats mètodes de regressió no esbiaixats: mètodes de regressió lineal, univariant o multivariant, que no presenten biaix respecte als coeficients que es volen estimar. Els mètodes de regressió esbiaixats són aquells en els quals no es requereix l'absència de biaix respecte a l'estimació dels coeficients. Per exemple, s'ha suggerit l'ús d'aquests mètodes com una possible solució al problema de la col·linearitat.[9] Aquests mètodes són emprats principalment en calibració multivariant, i algun exemple pot ser la regressió per components principals (*principal components regression*, PCR) o la regressió per mínims quadrats parcials (*partial least squares*, PLS).[10]

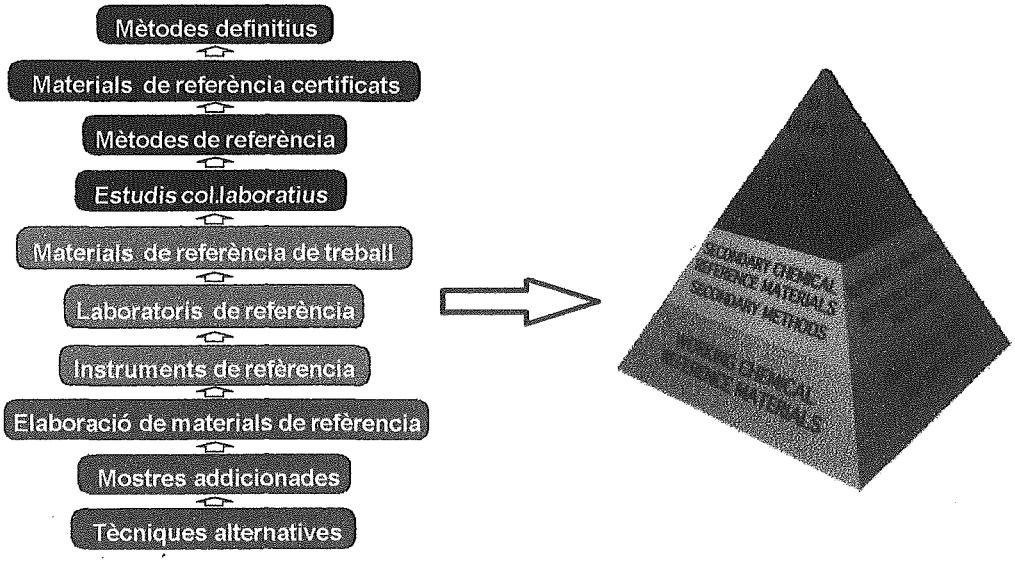


Figura 1.1. Relació entre diverses referències i la piràmide metrològica.

1.3.1 Regressió lineal univariant

Tradicionalment, i sobretot a causa de la simplicitat, s'han emprat mètodes de regressió que només consideren errors en la variable resposta. En la regressió lineal univariant, el mètode més emprat és el de mínims quadrats (*ordinary least squares*, OLS), sobre el qual encara ara es manté una disputa sobre qui va ser-ne

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

descobridor. Sembla que va ser descobert independentment per Carl Friedrich Gauss (1777-1855) i Adrien Marie Legendre (1752-1833), que Gauss el va començar a emprar abans del 1803 (segons ell el va començar a utilitzar el 1795, però no hi ha constància d'aquesta data) i que la primera referència va ser publicada per Legendre el 1805. Quan Gauss va escriure el 1809 que havia emprat aquest mètode abans de la data de publicació de Legendre, va començar la controvèrsia sobre a qui corresponia la prioritat.[11,12] El mètode de mínims quadrats, que fins a l'any 1970 va ser emprat quasi en exclusivitat en regressió lineal univariant, postula que el vertader model per a la relació entre la variable predictora (x) i la variable resposta (y) correspon a:

$$y_i = a + bx_i + e_i \tag{1.1}$$

Però com que els valors de l'ordenada a l'origen i el pendent no es poden trobar exactament, només pot obtenir-se'n una estimació i el model real correspon a:

$$\hat{y}_i = \hat{a} + \hat{b}x_i \tag{1.2}$$

El valor residual, e_i, correspon a:

$$e_i = \left|y_i - \hat{y}_i\right| = \left|y_i - \hat{a} - \hat{b}x_i\right| \tag{1.3}$$

Rigorosament, el mètode de mínims quadrats és només aplicable si es compleixen les condicions següents:[13,14]

- L'error, expressat en termes de variància, per a cada valor de la variable resposta ($s^2_{y_i}$) ha de ser molt més gran que per al valor corresponent de la variància de la variable predictora multiplicat pel quadrat del pendent ($\hat{b}^2 s^2_{x_i}$).

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

- Les variàncies dels valors de la variable resposta han de tenir valors constants al llarg de tot l'interval de linealitat (homoscedasticitat).
- Els errors de la variable resposta han de ser mútuament independents.

Si es compleixen aquestes condicions, els valors de l'ordenada a l'origen i el pendent trobats amb el mètode de mínims quadrats donen lloc a les estimacions més precises no esbiaixades dels valors vertaders de l'ordenada a l'origen i el pendent, és a dir, donen el valor mínim de l'error experimental.

Els coeficients de regressió han de ser tals que la suma de les desviacions respecte a la recta real sigui mínima, és a dir, que la suma dels quadrats dels residuals expressada en l'equació 1.3 sigui mínima:[13,15]

$$S=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}\left(y_i-\hat{a}-\hat{b}x_i\right)^2 \qquad (1.4)$$

El mètode es troba esquematitzat a la figura 1.2, on les desviacions verticals dels punts experimentals a la recta de regressió representen els residuals.

Les estimacions de l'ordenada a l'origen i el pendent es troben calculant les derivades parcials de l'equació 1.4 respecte a aquests coeficients i igualant els resultats a zero:

$$\frac{\partial}{\partial\hat{a}}\left[\sum_{i=1}^{n}\left(y_i-\hat{a}-\hat{b}x_i\right)^2\right]=0 \qquad (1.5)$$

$$\frac{\partial}{\partial\hat{b}}\left[\sum_{i=1}^{n}\left(y_i-\hat{a}-\hat{b}x_i\right)^2\right]=0 \qquad (1.6)$$

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

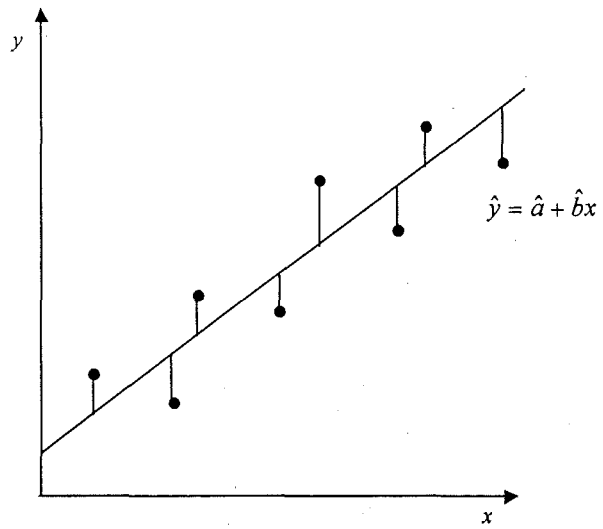


Figura 1.2 Il·lustració del mètode de mínims quadrats.

Desenvolupant les equacions 1.5 i 1.6, s'arriba a les expressions per a l'ordenada a l'origen i el pendent de la recta de regressió:

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - \left[\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i\right] / n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n} \qquad (1.7)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad (1.8)$$

La recta de regressió així trobada passa pel punt ($\bar{x}$, $\bar{y}$), anomenat centroide.

Un terme important, necessari en l'obtenció de les variàncies dels coeficients de regressió i d'altres paràmetres relacionats, útils per al desenvolupament posterior de tests estadístics, és l'error experimental (s^2). Aquest error, mesurat en termes de variància, és estimat segons:

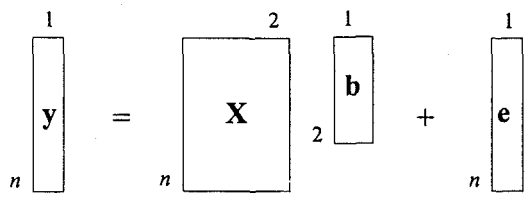
UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

$$\hat{s}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} \left(y_i - \hat{a} - \hat{b} x_i \right)^2 \tag{1.9}$$

El terme s^2 també es pot anomenar error estàndard.

El mateix model de l'equació 1.1 també pot expressar-se en forma matricial com:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1.10}$$



on el *vector d'observacions* $\mathbf{y}_{n\times 1}$ representa els valors de la variable resposta; $\mathbf{X}_{n\times 2}$ és la *matriu de la variable predictora*, formada per una columna d'1 i una columna amb els valors de la variable predictora; **b** és el *vector dels coeficients de regressió*, i **e** és el *vector d'errors*. El vector $\hat{\mathbf{b}}$ amb les estimacions de l'ordenada a l'origen i el pendent es pot trobar segons:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.11}$$

Les condicions abans esmentades perquè es compleixi el mètode de mínims quadrats no sempre es poden assegurar. En certs casos algunes de les observacions emprades en la regressió lineal són menys fiables que altres. Això significa que les variàncies associades a la variable resposta no són totes iguals al llarg de l'interval de regressió (heteroscedasticitat). En aquest cas, i continuant considerant la variable predictora com a lliure d'error, es pot aplicar el mètode de mínims quadrats ponderats (*weighted least squares*, WLS). Aquest mètode permet

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

heteroscedasticitat en la variable resposta, però els errors entre els seus diversos valors no poden estar correlacionats. En aquest mètode es minimitza la suma de residuals ponderats expressats en l'equació 1.12:

$$S=\sum_{i=1}^{n}\frac{e_i^2}{w_i}=\sum_{i=1}^{n}\frac{\left(y_i-\hat{a}-\hat{b}x_i\right)^2}{w_i} \tag{1.12}$$

on el terme w_i (factor de ponderació) correspon a la variància de cada punt de la variable resposta ($s_{y_i}^2$). En aquest mètode es dóna més importància als punts experimentals amb menys incertesa a la variable resposta (expressada en termes de variància, $s_{y_i}^2$) i l'estimació de l'error experimental pren la forma següent:

$$\hat{s}^2=\frac{1}{(n-2)}\sum_{i=1}^{n}\frac{\left(y_i-\hat{a}-\hat{b}x_i\right)^2}{w_i} \tag{1.13}$$

Similarment al mètode de mínims quadrats, les estimacions de l'ordenada a l'origen i el pendent es troben calculant les derivades parcials de l'equació 1.12 respecte a aquests coeficients i igualant els resultats a zero. Les expressions per a l'ordenada a l'origen i el pendent de la recta de regressió tenen la forma següent:

$$\hat{b}=\frac{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x}_p)(y_i-\bar{y}_p)}{w_i}\right)}{\sum_{i=1}^{n}\left(\frac{(x_i-\bar{x}_p)^2}{w_i}\right)} \tag{1.14}$$

$$\hat{a}=\bar{y}_p-b\bar{x}_p \tag{1.15}$$

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

on $\bar{x}_p$ i $\bar{y}_p$ corresponen a les mitjanes ponderades de les variables predictora i resposta respectivament:

$$\bar{x}_p = \frac{\sum_{i=1}^{n} x_i / w_i}{\sum_{i=1}^{n} 1/w_i} \tag{1.16}$$

$$\bar{y}_p = \frac{\sum_{i=1}^{n} y_i / w_i}{\sum_{i=1}^{n} 1/w_i} \tag{1.17}$$

La recta de regressió trobada pel mètode de mínims quadrats ponderats passa pel punt ($\bar{x}_p$, $\bar{y}_p$), anomenat en aquest cas centroide ponderat.

En forma matricial, el vector $\hat{\mathbf{b}}$ amb les estimacions de l'ordenada a l'origen i el pendent es pot trobar segons:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{1.18}$$

on $\mathbf{V}_{nxn}$ és una matriu diagonal en què l'element *i* de la diagonal correspon a la variància del punt *i* de la variable resposta ($s^2_{y_i}$). És important puntualitzar que si la variància de la variable resposta és igual a la unitat en tots els punts ($s^2_{y_i} = 1$), s'obtenen els mateixos resultats amb el mètode de mínims quadrats ponderats que amb el mètode de mínims quadrats. Watters i col·laboradors han proposat una modificació del mètode de mínims quadrats ponderats que consisteix a modelar l'error al llarg de l'interval de regressió i utilitzar els resultats d'aquest modelatge

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

com a factors de ponderació per tal de trobar els coeficients de la recta de regressió, en comptes d'utilitzar la variància de cada punt de la variable resposta.[16]

En el cas en què els errors en la variable resposta siguin heteroscedàstics i estiguin correlacionats, es pot aplicar el mètode de mínims quadrats generalitzats (*generalized least squares*, GLS).[9,17] L'expressió per calcular els coeficients de la recta de regressió coincideix amb l'equació 1.18, però en aquest cas $\mathbf{V}_{n\times n}$ és una matriu en què l'element i de la diagonal correspon a la variància del punt i de la variable resposta ($s^2_{y_i}$), i l'element (i,k) de fora de la diagonal correspon a la correlació entre els errors dels punts i i k de la variable resposta ($\mathrm{cov}(y_i, y_k)$). L'error experimental ara pren l'expressió matricial següent:

$$\hat{s}^2 = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)}{n-2} \qquad (1.19)$$

Cal assenyalar que si les correlacions entre els errors dels diferents punts de la variable resposta són tots zero ($\mathrm{cov}(y_i, y_k) = 0$), amb el mètode de mínims quadrats generalitzats s'obtenen els mateixos resultats que amb l'aplicació del mètode de mínims quadrats ponderats.

1.3.2 Regressió lineal multivariant

En la calibració lineal multivariant, la variable resposta (y) és funció de diverses variables predictores (x_j, j=1...p) segons el model següent:

$$y_i = a + b_1 x_{1_i} + b_2 x_{2_i} + \ldots + b_p x_{p_i} + e_i \qquad (1.20)$$

on a correspon al valor vertader de l'ordenada a l'origen, les variables b_j (j=1...p) representen els valors vertaders dels pendents i e_i és el valor residual. Com que

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

només podem tenir una estimació dels valors de l'ordenada a l'origen i els pendents, el model real correspon a:

$$y_i = \hat{a} + \hat{b}_1 x_{1_i} + \hat{b}_2 x_{2_i} + \ldots + \hat{b}_p x_{p_i} \qquad (1.21)$$

on la variable $\hat{a}$ representa l'estimació de l'ordenada a l'origen de la recta de regressió i les variables $\hat{b}_j$ (j=1...p) són les estimacions dels pendents del model. En certs casos també hi ha un suport teòric darrere de l'aplicació, com en el ja esmentat cas de la llei de Lambert-Beer, en què la resposta instrumental (y_i) és mesurada a p longituds d'ona.

El residual e_i en l'equació 1.20 correspon a:

$$e_i = \left| y_i - \hat{a} - \hat{b}_1 x_{1_i} - \hat{b}_2 x_{2_i} - \ldots - \hat{b}_p x_{p_i} \right| \qquad (1.22)$$

En aquest cas els punts experimentals s'ajusten a un pla (p+1)-dimensional. Un dels mètodes més emprats per tal de trobar els coeficients de regressió és el de regressió lineal múltiple (*multiple linear regression*, MLR), l'anàleg del mètode de mínims quadrats en regressió multivariant. Igual que en el mètode de mínims quadrats, els coeficients de regressió es busquen de tal forma que la suma de les desviacions respecte al pla sigui mínima, és a dir, la suma del quadrat dels residuals expressats en l'equació 1.22 ha de ser mínima:[9]

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{a} - \hat{b}_1 x_{1_i} - \hat{b}_2 x_{2_i} - \ldots - \hat{b}_p x_{p_i} \right)^2 \qquad (1.23)$$

El mètode es troba esquematitzat a la figura 1.3, on hi ha un exemple de regressió multivariant en un pla tridimensional. Les desviacions verticals dels punts experimentals al pla de regressió representen els residuals.

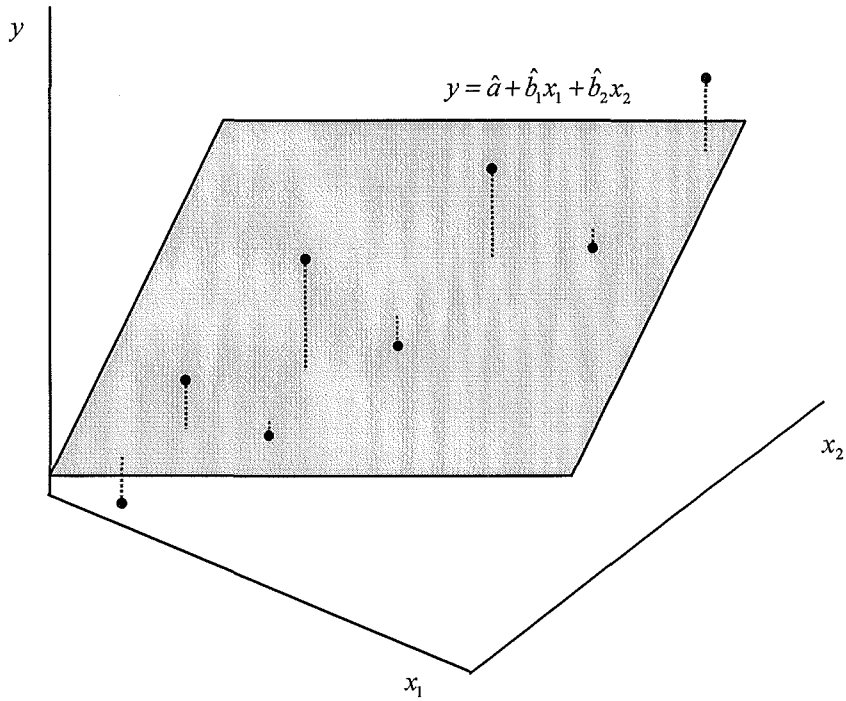
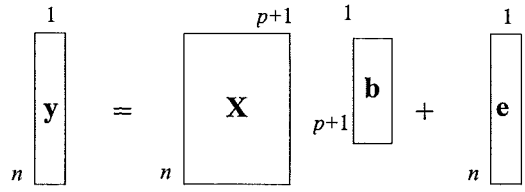


Figura 1.3 Il·lustració del mètode de regressió múltiple lineal.

La dificultat per trobar els coeficients de regressió es veu incrementada pel fet que s'augmenta el nombre de variables, per la qual cosa en calibració multivariant s'utilitza quasi exclusivament la notació matricial. L'equació 1.20 expressada en forma matricial queda com:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \qquad (1.24)$$

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008



on el vector d'observacions $\mathbf{y}_{n\times 1}$ representa els valors de la variable resposta, $\mathbf{X}_{n\times(p+1)}$ és la matriu de les variable predictores, formada per una columna d'1 i p columnes amb els valors de les p variables predictores, $\mathbf{b}$ és el vector dels coeficients de regressió i $\mathbf{e}$ és el vector d'errors. El vector $\hat{\mathbf{b}}$ amb les estimacions de l'ordenada a l'origen i els pendents es pot trobar segons l'expressió següent:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (1.25)$$

la qual és coincident amb l'equació 1.11 del mètode de mínims quadrats. L'única diferència es troba en la dimensió de les matrius $\mathbf{X}$ i $\hat{\mathbf{b}}$.

1.4 Regressió lineal considerant errors en tots els eixos

Les hipòtesis necessàries per aplicar el mètode de mínims quadrats o regressió lineal múltiple no sempre es compleixen rigorosament. Pel que fa al camp químic, dins de la calibració univariant, en el procés d'establiment de la recta que relaciona la resposta instrumental (eix y) amb el valor de les concentracions (eix x), en el qual tradicionalment s'han considerat aquestes últimes lliures d'error, la constant millora de l'instrumental químic fa que els errors en l'eix de les respostes sigui en alguns casos comparable als errors associats als valors de les concentracions. Fins i tot en algunes tècniques analítiques, com les que inclouen datació per radiocarboni, els valors de les concentracions presenten clarament uns errors d'una magnitud tal

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

que fan que no puguin ser negligibles.[18,19] Altres àmbits del camp químic on s'ha comprovat l'aplicabilitat de la regressió lineal considerant errors en dos eixos els constitueixen les determinacions de constants de reacció per reaccions en fase gasosa,[20] en l'estudi de reaccions cinètiques,[21] estimant la relació d'una mescla[22] o en alguns tipus de mesures espectroscòpiques.[23]

Mentre la regressió lineal considerant errors en dos eixos no és nova (ja el 1940 se'n troben articles[24]), pel que fa a la regressió multivariant considerant errors en tots els eixos hi ha poca bibliografia. De fet, un dels problemes principals per difondre els mètodes de regressió lineal univariant considerant errors en dos eixos ha estat la dificultat a l'hora de programar els algorismes i aplicar-los per trobar els coeficients de la recta de regressió (dificultat que actualment està superada gràcies a les tècniques informàtiques), la qual cosa s'ha multiplicat en regressió multivariant considerant errors en tots els eixos. Un altre inconvenient el presenta el fet de que la necessitat de conèixer les incerteses individuals en tots els eixos per a cada punt experimental normalment implica més temps d'anàlisi, per la qual cosa en molts casos es prefereix sacrificar l'exactitud dels resultats per obtenir un estalvi econòmic i en el temps de treball.

1.5 Aspectes conceptuals relacionats amb la regressió lineal considerant errors en tots els eixos

1.5.1 Possible falta d'adequació dels valors experimentals al model

En els tests estadístics basats en la regressió lineal, és important assegurar l'adequació dels valors experimentals al model. En cas contrari, el model pot no ser un reflex fidel de les dades experimentals i produir-se falta d'ajust. En aquests casos, l'elevat error experimental inherent a la falta d'ajust pot fer convertir errors sistemàtics en aleatoris. En el mètode de mínims quadrats, la majoria dels usuaris

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

normalment acostumen a emprar el coeficient de correlació, r (o el seu quadrat, r^2, coeficient de determinació), per detectar la falta d'ajust. Però aquest no és un test estadístic i no constitueix un paràmetre informatiu de la qualitat de l'ajust dels punts experimentals a la recta,[25] si bé la seva utilització juntament amb els gràfics de residuals fa augmentar la fiabilitat per a la detecció de la falta d'ajust. El grau d'acoblament del model lineal establert respecte als punts experimentals pot verificar-se mitjançant el test d'anàlisi de la variància (ANOVA) o un test χ^2.[13]

Fins i tot s'ha suggerit que les representacions gràfiques d'una sèrie de punts amb ajust lineal només es podrien divulgar en revistes especialitzades si estiguessin preparades adequadament per tal de demostrar-ne la linealitat d'acord amb tests rigorosos.[26] En regressió lineal univariant ponderada emprant el mètode WLS o en regressió lineal univariant considerant errors en ambdós eixos, hi ha molt poca bibliografia respecte a la detecció de la possible falta d'adequació dels valors experimentals al model.

En la regressió lineal multivariant, dins de les tècniques no esbiaixades hi ha també l'anàleg al coeficient de determinació, el coeficient de múltiple determinació (R^2).[27] Tal com en la regressió univariant, aquest terme ha de ser emprat amb precaució. Una millor opció la constitueix la utilització de l'anàlisi de la variància, tal com succeeix en la regressió univariant.

1.5.2 Existència de valors discrepants en els valors experimentals

Un valor discrepant, valor aberrant o *outlier* és un punt que és diferent de la resta d'observacions.

Un punt pot ser un valor discrepant a causa d'errors aleatoris (en aquest supòsit no caldria prendre mesures encaminades a revisar el model escollit) o perquè el punt no pertany a la mateixa població que la resta. Un punt vàlid pot aparèixer com un

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

valor discrepant perquè el model no representa adequadament la realitat (per exemple, perquè s'està intentant modelar una zona corba a una línia recta). En certs casos un valor discrepant dóna una informació que la resta de punts no poden aportar perquè un valor discrepant sol venir d'un cúmul de circumstàncies que poden ser d'interès i que solen requerir investigació posterior, més que no un rebuig immediat. Com a regla general, els valors discrepants només haurien de ser rebutjats si poden ser traçats a causes tals com errors en les observacions o la posada a punt dels instruments.[28]

No sempre que un punt tingui un residual molt més gran que la resta ha de ser necessàriament un valor discrepant. En regressió tenint en compte errors en tots els eixos, un punt individual pot tenir una incertesa tan gran en tots els eixos que la distància del punt a la recta o al pla de regressió pot ser deguda precisament a aquesta gran incertesa (la recta o el pla de regressió s'apropa més als punts experimentals amb menys incertesa i deixa més lluny els punts amb més incertesa).

En regressió lineal univariant, utilitzant el mètode de mínims quadrats hi ha diversos tests estadístics per detectar punts discrepants,[29] com pot ser per exemple el test de Cook,[30,31] però en regressió per mínims quadrats ponderats o tenint en compte errors en tots els eixos hi ha molt poca bibliografia. Dins de la calibració multivariant també hi ha diversos tests per detectar la presència de punts discrepants.[10]

En regressió lineal univariant, una altra aproximació a la detecció i el tractament de punts discrepants l'aporten les tècniques robustes. En aquest camp, robust és aplicat en el sentit que significa insensibilitat a petites desviacions en les assumpcions de les distribucions.[32] Aquesta sèrie de tècniques són més insensibles respecte als punts discrepants, per la qual cosa donen menys importància a les dades experimentals amb majors residuals. Les tècniques robustes solen minimitzar altres funcions que la suma de quadrats dels residuals per trobar els coeficients de

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

la recta de regressió.[33] Algunes tècniques robustes són per exemple la regressió de residuals absoluts mínims (*least absolute residual regression*), la regressió de la minimització del quadrat de la mediana (*least median squares regression*, LSM) o la regressió per mínims quadrats iterativament ponderats (*iteratively reweighted least squares regression*, IRWLS)

1.5.3 Tipus de distribució dels valors experimentals i dels coeficients de la recta de regressió

La majoria de mètodes de regressió lineal tenen com a assumpció la normalitat en la distribució dels resultats: els resultats en les variables resposta i predictora han de seguir la distribució normal. Els procediments de mesura químics solen estar compostos de diverses etapes, normalment independents. Aquesta és una de les característiques fonamentals que diferencien les mesures químiques de les mesures de tipus físic. És a dir, en la majoria de les anàlisis químiques el resultat és suma d'un nombre relativament gran de variables aleatòries, i en aquests casos el teorema del límit central[34] postula que la distribució de la variable final (el resultat de l'anàlisi química en el nostre cas) és molt aproximadament normal. Amb aquesta hipòtesi treballen la quasi totalitat dels analistes més prestigiosos. De fet, la majoria de tests estadístics paramètrics emprats (com els coneguts test *t* o test *F*), estan basats en les assumpcions que els valors experimentals segueixen la distribució normal.

Malgrat això, en certs casos no es pot assegurar la normalitat en la distribució dels resultats.[35] Clancey[36] va estudiar 50.000 anàlisis en mostres de metalls i aliatges i va intentar associar els resultats a una de les aproximadament 250 distribucions escollides. D'acord amb els resultats, el 10-15% de les distribucions van ser normals, el 15% eren corbes normals truncades, el 10% eren simètriques però amb el pic més alt que la distribució normal (leptocúrtiques), el 20-25% eren asimètriques, el 20-25% tenien forma de J i unes poques eren bimodals. Algunes

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

causes de la no-normalitat de les distribucions dels resultats són, per exemple, l'heterogeneïtat de les mostres o mesures pròximes al límit de detecció.

Una possible alternativa, a causa de la presència de la no-normalitat en les distribucions dels resultats químics, la constitueixen els anomenats tests d'aleatorietat (*randomization tests*). Aquests poden definir-se com tests en els quals la significança dels resultats experimentals es calcula a base de permutacions repetides de les dades.[37] Segons Box i col·laboradors,[38] els tests d'aleatorietat són els correctes i els paramètrics en són una aproximació. En els tests d'aleatorietat la probabilitat d'un resultat experimental es determina a partir de la distribució de les dades experimentals. Com que aquesta significança està basada en la distribució real de les dades, no cal que es compleixi cap hipòtesi prèvia. Per a cada test, la significança del test estadístic experimental no es compara amb un valor tabulat. L'inconvenient principal en l'aplicació d'aquest tipus de tests és que quan el nombre total de permutacions a efectuar és elevat (és a dir, quan el nombre de dades experimentals és gran), el temps de càlcul pot ser considerable.

En casos en què no es pugui obtenir la distribució normal i es vulgui evitar el procés de càlcul inherent als tests d'aleatorietat, una altra alternativa és l'ús de tècniques que no es basen en cap distribució en particular. D'aquesta manera no cal calcular paràmetres tals com la mitjana o la desviació estàndard i per això s'anomenen tècniques no paramètriques (el terme no paramètric va ser emprat per primera vegada per J. Wolfowitz el 1942).[39] Aquestes tècniques tenen l'avantatge que demanen càlculs molt simples, però són menys eficients i solen requerir més rèpliques que les tècniques paramètriques.[40,41] També es poden emprar quan la mida de la mostra és molt petita.

De la mateixa manera que els resultats de les variables resposta i predictora han de seguir la distribució normal per tal de poder-s'hi aplicar les tècniques de regressió usuals, la majoria dels tests aplicats als coeficients de la recta o hiperplà de

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

regressió (ordenada a l'origen i pendent o pendents) també estan basats en el supòsit que la distribució d'aquests coeficients és normal. En tècniques de regressió lineal que tenen en compte els errors en només la variable resposta, els paràmetres de la recta o hiperplà de regressió segueixen la distribució normal, però tenint en compte errors en tots els eixos aquest supòsit sembla que no es compleix sempre.[42] L'ús de test basats en la distribució normal poden conduir en aquestes situacions a conclusions errònies si la diferència entre la distribució real dels coeficients de regressió i la distribució normal és gran.

En regressió tenint en compte els errors en tots els eixos, ens centrarem en els tests estadístics basats en tècniques paramètriques, intentant comprovar si efectivament els coeficients de la recta o hiperplà de regressió no segueixen la distribució normal, i comprovant el grau de desviament de la distribució real respecte a la distribució normal. Si el grau de desviament no és gaire alt, una hipòtesi raonable pot ser l'acceptació de la normalitat de la distribució dels coeficients de la recta o hiperplà de regressió.

1.5.4 Existència de l'error de segona espècie, β

L'error β, de tipus II, o de segona espècie, és l'error comès en acceptar erròniament la hipòtesi nul·la. Una hipòtesi consisteix en la suposició que es realitza sobre un determinat succés. Al fer un test estadístic, cal escollir dos tipus d'hipòtesis:

1. La hipòtesi nul·la (H_0), que implica manca de diferències entre paràmetres i és la que normalment es vol comprovar. Per exemple, que el valor de la concentració de l'anàlisi d'un material de referència no difereix de forma estadísticament significativa del seu valor nominal.
2. La hipòtesi alternativa (H_1), que s'accepta quan falla la hipòtesi nul·la. Seguint l'exemple anterior, que el valor de la concentració de l'anàlisi d'un material de referència difereix de forma estadísticament significativa del seu valor nominal.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN: 978-84-691-1897-9/ D.L: T-353-2008

Per comprovar la validesa de la hipòtesi nul·la, es fa servir un test estadístic. Plantejada la hipòtesi nul·la i un cop efectuat el test, es poden adoptar dues decisions: acceptar-la o rebutjar-la. Tanmateix, pot ser que sigui certa o falsa. Això s'esquematitza en el quadre següent:

Decisió adoptada mitjançant el test

Situació real		**H_0 certa** **acceptació**	**H_0 falsa** **rebuig**
	H_0 certa	correcta	error α
	H_0 falsa	error β	correcta

on es pot observar que la decisió correcta consisteix a acceptar la hipòtesi nul·la que és certa, o rebutjar la hipòtesi nul·la que és falsa. Si es rebutja una hipòtesi nul·la certa es comet un error α, mentre que si s'accepta una hipòtesi nul·la falsa es comet un error β.

1.5.5 Relació entre els errors α i β

Els errors α (també anomenats de primera espècie o de tipus I) i β són contraposats, és a dir, si en disminuïm d'un tipus, n'augmentem de l'altre per a la resta de condicions constants.

L'error α és àmpliament emprat en els tests estadístics aplicats a l'anàlisi química. Potser per això hi ha una tendència a escollir els valors α=0.05 o α=0.01 quasi sense plantejar-se'n el significat. El fet d'escollir aquests valors es remunta en el temps, i per exemple, el 1925, Fisher ja escrivia "*If P is between 0.1 and 0.9 there*

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05 ...".[43] En canvi, l'error β és poc conegut dins de l'anàlisi química. Fins i tot hi ha certs autors que suggereixen que la utilització de l'error β és massa complexa per als químics ordinaris.[44] El que és clar és que l'usuari pot decidir entre tenir en compte l'error β o no, però el fet d'ignorar-lo no fa que desaparegui.

En certs casos pot ser més important l'error β que l'α. Per exemple, en la verificació de la traçabilitat d'un mètode analític (en què la hipòtesi nul·la és considerar com a traçable el mètode), unes probabilitats d'error β elevades impliquen que s'accepta com a bo un mètode analític que probablement és esbiaixat, per la qual cosa les mostres que s'hi analitzin poden donar lloc a resultats incorrectes. En canvi, unes probabilitats d'error α grans suposen que es rebutja un mètode que probablement és bo, per la qual cosa caldria revisar innecessàriament el procés de verificació de la traçabilitat. Segons quin tipus de mostres s'hagin d'analitzar amb el mètode, convindrà més exposar-se a repetir innecessàriament el procés de verificació de la traçabilitat que no arriscar-se a emprar un mal mètode. Per exemple, en estudis farmacològics es tracta d'evitar concloure erròniament que una substància actua com a droga, ja que això comportaria l'ús de drogues que realment no tenen cap efecte terapèutic. Per tant, en estudis farmacològics es necessitaran errors α baixos.[45] D'altra banda, en estudis de bioequivalència de drogues normalment sol ser més important no acceptar erròniament la bioequivalència de dues drogues, la qual cosa implica que aquests estudis donen més importància a l'error β.[46]

S'ha suggerit[45] que en aquells casos en què sigui més important l'error β que l'α, s'intercanviïn les hipòtesis nul·la i alternativa per així, per exemple en processos de

comparació de mètodes, controlar el risc de concloure que els dos mètodes són comparables quan en realitat no ho són.

1.6 Aplicacions de la regressió lineal considerant errors en tots els eixos

Tradicionalment l'ús més general per a la regressió lineal en l'anàlisi química és l'establiment de la relació entre la resposta instrumental i la concentració de l'analit o analits. No obstant això, aquesta no és només l'única utilitat dins del camp químic.

1.6.1 Detecció d'errors constants o proporcionals

La detecció d'errors constants o proporcionals pot ser útil per tal de saber si s'han d'aplicar correccions del blanc o en processos en què intervinguin recuperacions. Aquest càlcul pot ser dut a terme mitjançant tests individuals en l'ordenada o el pendent de la recta de regressió.

1.6.2 Comparació de mètodes

La comparació de dos o més mètodes analítics a diversos nivells de concentració pot dur-se a terme mitjançant la regressió lineal. En aquest sentit, es busca comparar si els coeficients de la recta (quan es volen comparar els resultats proporcionats per dos mètodes analítics) o hiperplà de regressió (quan es volen comparar els resultats proporcionats per més de dos mètodes analítics) no són significativament diferents dels coeficients teòrics. En general, els mètodes en comparació presenten errors del mateix ordre de magnitud, per la qual cosa cal emprar regressió lineal que tingui en compte errors en tots els eixos a l'hora de comparar-los.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

1.6.3 Predicció

En calibració lineal univariant la predicció és molt utilitzada per trobar el valor de la concentració d'una mostra desconeguda donat el valor de la resposta instrumental per a aquella mostra. Però en processos de comparació de mètodes a vegades es volen saber el valor i la incertesa d'una mostra si fos analitzada per un nou mètode a partir del valor que té aquella mostra analitzada per un mètode ja establert.

1.7 Estructura de la tesi

La complexitat en el procés de cerca dels coeficients de la recta en regressió lineal considerant els errors en les variables resposta i predictora ha fet que s'hagin desenvolupat diverses tècniques per tal de trobar-los. En el segon capítol de la present tesi doctoral es presenta una revisió crítica d'aquestes tècniques, fent especial èmfasi en l'obtenció d'altres paràmetres, com poden ser les variàncies dels coeficients de regressió o els intervals de confiança de les variables tant resposta com predictora. Aquest capítol ens servirà de punt de partida per poder desenvolupar posteriorment els tests estadístics aplicables en processos de comparació de metodologies analítiques.

Atès que la bibliografia indica que les distribucions de l'ordenada a l'origen i el pendent en regressió lineal tenint en compte els errors en dos eixos són no normals (vegeu secció 1.5.3), en el tercer capítol es verifica aquesta afirmació. Malgrat que es comprova la manca de normalitat d'aquestes distribucions, es verifica que la desviació de les distribucions reals de l'ordenada a l'origen i el pendent respecte a la normalitat no és gaire gran. A més l'acceptació de la normalitat en les distribucions d'aquests coeficients de regressió dóna millors resultats que l'ús de tècniques de regressió que només tenen en compte errors en un sol eix, malgrat que

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

en aquests casos els coeficients de regressió segueixen la distribució normal. Així mateix, s'introdueix l'error β en els tests individuals per a l'ordenada a l'origen i el pendent, útils en processos de detecció d'errors proporcionals o sistemàtics.

Tenint en compte totes les consideracions teòriques discutides a l'apartat 1.4, en el quart capítol s'entra de ple en la comparació de dos mètodes analítics. Aquest capítol es basa en la construcció de l'interval de confiança conjunt per a l'ordenada a l'origen i el pendent de la recta de regressió obtinguda de la representació dels resultats d'un mètode respecte a l'altre, tenint en compte els errors deguts a tots dos mètodes. Aquest capítol es completa amb la descripció d'un programa informàtic i la comparació de l'interval de confiança conjunt per a l'ordenada a l'origen i el pendent considerant els errors en dos eixos amb altres tests estadístics de la bibliografia aplicables en processos de comparació de mètodes.

La calibració lineal multivariant tenint en compte els errors en tots els eixos centra el cinquè capítol. Ampliant el concepte de comparació de dos mètodes analítics exposat en el quart capítol, es desenvolupa el test conjunt per als coeficients de regressió en regressió lineal multivariant tenint en compte els errors en tots els eixos. Aquest test pot ser útil en exercicis interlaboratori, on es comparen els resultats obtinguts per diversos laboratoris, tots amb incerteses associades, o en la comparació de múltiples metodologies analítiques a diversos nivells de concentració.

El sisè capítol incideix en un altre dels aspectes útils de la regressió lineal: la incertesa en els valors de predicció. Dins del camp de la comparació de mètodes, a vegades interessa saber la incertesa a l'hora de predir una mostra mitjançant un nou mètode sabent el valor i la incertesa que té aquella mostra analitzada per un mètode antic. Com que ambdós mètodes tenen incerteses associades, la regressió a utilitzar cal que tingui en compte les incerteses dels dos mètodes.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Finalment, en l'apartat de conclusions es discuteixen els avantatges i les limitacions dels tests estadístics basats en la regressió lineal univariant i multivariant tenint en compte els errors en tots els eixos presentats en aquesta tesi doctoral, i es donen una sèrie de pautes de com s'ha d'enfocar l'estudi de la seva millora en treballs futurs.

1.8 Referències

1. Eurachem/Welac Guide 1, *Accreditation of Chemical Laboratories*, Laboratory of the Government Chemist, London (1993)
2. H. Günzler (Ed.), *Accreditation and Quality Assurance in Analytical Chemistry*, Springer-Verlag, Heidelberg (1996)
3. ISO 3543-1, *International Organization for Standardization*, Geneva (1993)
4. International Vocabulary or Basic and General Terms in Metrology, *International Organization for Standardization*, Geneva (1993)
5. M. Valcárcel, A. Ríos, *La Calidad en los Laboratorios Analíticos*, Reverté, Barcelona (1992)
6. EN 45001/UNE 66501, *Criterios Generales para el Funcionamiento de los Laboratorios de Ensayo*, Aenor, Madrid (1991)
7. T.J. Quinn, *Metrologia*, **34** (1997) 61
8. ISO Guide 30, *International Organization for Standardization*, Geneva (1981)
9. J.O. Rawlings, *Applied Regression Analysis: A Research Tool*, Wadsworth & Brooks/Cole Advanced Books & Software, Belmont (1988)
10. H. Martens, T. Næs, *Multivariate Calibration*, Wiley, Chichester (1989)
11. C. Eisenhart, *Journal of the Washington Academy of Sciencies*, **54** (1964) 24
12. R.L. Plackett, *Biometrika*, **59** (1972) 239
13. N. Draper, H. Smith, *Applied Regression Analysis*, 2nd ed., Wiley, New York (1981)
14. J.A. Irvin, T.I. Quickenden, *Journal of Chemical Education*, **60** (1983) 711

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

15. R.H. Myers, *Classical and Modern Regression with Applications*, 2nd ed., Duxbury Press, Belmont (1986)
16. R.L. Watters, R.J. Carroll, C.H. Spiegelman, *Analytical Chemistry*, **59** (1987) 1639
17. M. Meloun, J. Militký, M. Forina, *Chemometrics for Analytical Chemistry. Volume 2. PC-aided Regression and Related Methods*, Ellis Horwood, London (1994)
18. R.M. Clark, *Journal of the Royal Statistical Society, Series A*, **142** (1979) 47
19. R.M. Clark, *Journal of the Royal Statistical Society, Series A*, **143** (1980) 177
20. T. Brauers, B.J. Finlayson-Pitts, *International Journal of Chemical Kinetics*, **29** (1997) 665
21. D.P. Chong, *Journal of Chemical Education*, **71** (1994) 489
22. H. Marshak, C.H. Spiegelman, *Nuclear Instruments and Methods in Physics Research*, **A234** (1985) 455
23. T. Lwin, C.H. Spiegelman, *Journal of the Royal Statistical Society, Series C*, **35** (1986) 256
24. J.H. Wald, *Annals of Mathematical Statistics*, **11** (1940) 891
25. J.S. Hunter, *Journal Association of Official Analytical Chemists*, **64** (1981) 574
26. D.C. Johnson, *Analytica Chimica Acta*, **204** (1988) 1
27. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam (1997)
28. F.J. Anscombe, *Technometrics*, **2** (1960) 123
29. V. Bartnett, T. Lewis, *Outliers in Statistical Data*, 3rd ed., John Wiley & Sons, Chichester (1994)
30. R.D. Cook, *Technometrics*, **19** (1977) 15
31. R.D. Cook, *Journal of the American Statistical Association*, **74** (1969) 169
32. P.J. Huber, *Robust Statistics*, Wiley, New York (1981)

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

33. P.J. Rousseeuw, A.M. Leroy, *Robust Regression & Outlier Detection*, John Wiley & Sons, New York (1987)
34. W.A. Fuller, *Measurement Error Models*, John Wiley & Sons, New York (1987)
35. D.L. Massart, A. Dijkstra, *Evaluation and Optimization of Laboratory Methods*, Elsevier, Amsterdam (1978)
36. V.J. Clancey, *Nature*, **159** (1947) 339
37. E.S. Edgington, *Randomization Tests*, 2nd ed., D.B. Owen Eds, Marcel Dekker Inc., New York (1987)
38. G.E.P. Box, W.G. Hunter, J.S. Hunter, *Statistics for Experimenters, An introduction to Design, Data Analysis and Model Building*, Wiley and Sons, New York (1978)
39. J. Wolfowitz, *Annals of Mathematical Statistics*, **13** (1942) 247
40. S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York (1956)
41. W.J. Conover, *Practical Nonparametric Statistics*, Wiley-Interscience, New York (1971)
42. A.H. Kalantar, *Talanta*, **42** (1995) 597
43. R.A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd, London (1925)
44. J.D. Ingle jr., *Journal of Chemical Education*, **51** (1974) 100
45. C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, Y. Vander-Heyden, P. Vankeerberghen, D.L. Massart, *Analytical Chemistry*, **67** (1995) 4491
46. V.W. Steinijans, D. Hauscke, *Clinical Research Regulatory Affairs*, **10** (1993) 203

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Capítol 2

Regressió lineal univariant considerant errors en dos eixos. Estat actual de la qüestió

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1807-9/ D.L: T-353-2008

En aquest capítol es pretén fer una revisió crítica dels diversos mètodes que calculen els coeficients de la recta en regressió univariant considerant els errors en dos eixos (també anomenada regressió bivariant o regressió de Model II, en contraposició a la regressió de Model I en la qual la variable predictora és coneguda sense error). L'objectiu és facilitar al lector el coneixement dels mètodes que hi ha i dels avantatges i inconvenients que aporta cadascun, per ser aplicats després al desenvolupament de tests estadístics útils en processos de comparació de mètodes analítics. L'exposició de l'estat de la qüestió permet tenir una visió global del problema, la qual cosa facilita la tasca de plantejar com s'ha d'abordar posteriorment l'estudi dels diversos tests associats als coeficients de la recta de regressió, i quins són els aspectes, des d'un punt de vista teòric o experimental, en què cal incidir. En primer lloc, es justificarà la necessitat d'utilitzar mètodes de regressió considerant errors en dos eixos per la incorrecció en l'estimació dels coeficients de regressió si s'utilitzen mètodes que només consideren errors en un eix. Tot seguit es passarà a comentar l'aproximació de la relació constant de variàncies, un tipus de regressió que considera errors en dos eixos però que només és aplicable sota certes restriccions, per acabar amb els mètodes de regressió que contemplen errors heteroscedàstics individuals en ambdós eixos.

El gruix de la revisió crítica es troba al final del capítol, en l'article titulat *Univariate Regression Models with Errors in Both Axes*, i que ha estat publicat a la revista *Journal of Chemometrics*.

2.1 Necessitat de la regressió univariant considerant errors en dos eixos

Mínims quadrats és el mètode de regressió univariant tradicionalment més emprat, sobretot a causa de les seves propietats matemàtiques i d'algunes característiques de caràcter pràctic (com per exemple la simplicitat, que fa que per exemple estigui incorporat a la majoria de calculadores de butxaca). Tal com s'ha comentat en la

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

apartat 1.3, sota certes condicions la seva utilització pot donar lloc a l'estimació d'uns coeficients de regressió incorrectes. Scheffé[1] i Mandel[2] van afirmar que l'error en l'eix de les x pot ser ignorat si la desviació estàndard en cada punt individual (s_{x_i}) és petita respecte a la relació $s_{y_i}/\hat{b}$. Aquesta relació representa la desviació estàndard de l'error en l'eix de les y convertit a les unitats de x.

D'altra banda, Draper i Smith[3] van exposar que l'error sistemàtic comès en l'aplicació del mètode de mínims quadrats ignorant un possible error en la variable predictora ($x_i = \mu_{x_i} + \delta_i$ on μ_{x_i} correspon al valor vertader de x_i, i δ_i a l'error aleatori de x_i) es determina amb el factor següent:

$$r = \frac{\sigma_\delta^2}{\sigma_{\mu_x}^2} \tag{2.1}$$

on σ_δ^2 representa la variància en l'error de mesura de la variable predictora i $\sigma_{\mu_x}^2$ la variància de totes les dades de la variable predictora. Com més gran sigui el valor del coeficient r de l'equació 2.1, més esbiaixades seran les estimacions de l'ordenada a l'origen i el pendent obtingudes amb el mètode de mínims quadrats (per aproximadament r>0.2, els resultats obtinguts amb el mètode de mínims quadrats comencen a tenir errors significatius). Montgomery i Peck[4] van seguir un raonament similar i digueren que per a un nombre elevat de punts l'ordenada a l'origen i el pendent estimades amb el mètode de mínims quadrats tendeixen a convergir a:[3,5]

$$\hat{b} \approx \frac{1}{1+r} b \tag{2.2}$$

$$\hat{a} \approx a + \frac{r}{1+r} \bar{x} b \tag{2.3}$$

on $\bar{x}$ és la mitjana dels valors de la variable predictora i b és el valor vertader del pendent. Però aquesta correcció en els valors de l'ordenada a l'origen i el pendent només té uns resultats acceptables si es té una bona estimació del factor r de l'equació 2.1 i l'error de mesura dels valors de la variable predictora són petits.[6]

Dins de l'àmbit de la química analítica, en alguns casos no es poden negligir les incerteses de la variable predictora, tal com s'ha comentat en la apartat 1.4, i l'ús del mètode de mínims quadrats o fins i tot dels mètodes de mínims quadrats ponderats o mínims quadrats generalitzats donarien lloc a estimacions incorrectes de l'ordenada a l'origen i el pendent i de paràmetres relacionats com poden ser les seves variàncies, útils en el desenvolupament de tests estadístics. Alguns d'aquests casos els constitueixen, com ja s'ha comentat, processos de comparació de mètodes mitjançant regressió lineal, en què cada mètode analític porta associat errors sovint dels mateixos ordres de magnitud, o processos de calibració de certs mètodes químics, com els radioquímics, en els quals la incertesa associada a la variable predictora no es pot considerar en cap manera despreciable.

2.2 Relació constant de variàncies i regressió ortogonal

L'aproximació de la relació constant de variàncies (*constant variance ratio*, CVR)[7,8], també anomenada regressió de *errors-in-variables*[9,10] és un cas particular de regressió univariant considerant errors en els dos eixos, ja que només es pot aplicar amb la condició de que la relació entre l'error associat a la variable resposta i l'error associat a la variable predictora ha de ser igual per a tots els punts: la relació entre variàncies (λ) es manté constant.

Com que ambdues variables estan afectades per una relació constant d'errors, la regressió CVR es pot considerar un primer pas dins dels mètodes de regressió que tenen en compte errors en dos eixos. Hi ha diverses aproximacions per tal de

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

calcular els coeficients de la recta de regressió tenint en compte les assumpcions inherents a aquest mètode. L'aproximació de Mandel[2] consisteix en una transformació de les variables originals (x,y) en unes altres (u,v) que compleixen les condicions del mètode de mínims quadrats. Els coeficients de regressió de les variables originals es troben llavors mitjançant una altra transformació de les noves variables. S'ha demostrat[11] que el procediment de Mandel evita una subestimació del valor del pendent de la recta de regressió. El pendent i l'ordenada a l'origen de la recta de regressió es calculen segons les següents expressions (on s'assumeix que la correlació entre la variable dependent i la independent és zero, el que té sentit sobretot en processos de comparació de mètodes):

$$\hat{b} = \frac{S_{yy} - \lambda S_{xx} + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \tag{2.4}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{2.5}$$

on:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.6}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.7}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2.8}$$

i el paràmetre λ de l'equació 2.4 correspon a la relació entre les variàncies de la variable resposta i la variable predictora. Mandel va derivar també les expressions per al càlcul de les desviacions estàndard de l'ordenada a l'origen i el pendent.[2] El mètode és invariant respecte a un canvi d'eixos.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

En el cas particular de la regressió CVR en el què s'assumeix el mateix error en les dues variables (λ=1), gràficament es minimitza la suma de distàncies de cada punt individual en direcció perpendicular a la recta de regressió, cosa que es veu representada a la figura 2.1, i s'anomena regressió ortogonal (*orthogonal regression*, OR). La regressió ortogonal també és coneguda com regressió de la distància ortogonal (*orthogonal distance regression*, ODR)[10] o com regressió per mínims quadrats totals (*total least squares*, TLS).[12]

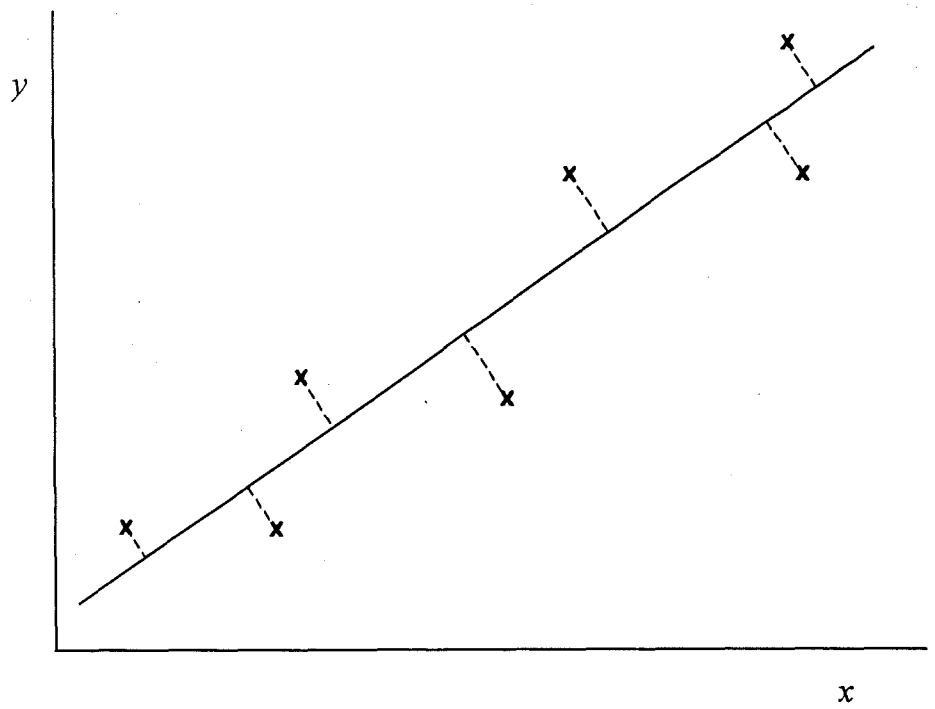


Figura 2.1. Representació de la regressió ortogonal.

L'aproximació de Mandel no és l'única per tal de solucionar el problema de l'aproximació CVR en algun dels seus casos particulars. La tècnica de l'anàlisi dels components principals (*principal component analysis*, PCA),[13] malgrat que té com a principal utilitat l'estudi de l'estructura d'un conjunt de dades multivariant, també pot actuar com a mètode de regressió ortogonal. En aquest cas, només és útil el

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

càlcul del primer component principal, que coincideix amb la recta de regressió trobada mitjançant l'aproximació de Mandel i per a un valor de λ en l'equació 2.4 igual a 1.

Alguns autors[6,14] asseguren que el fet d'emprar la regressió ortogonal en casos en què els errors en les mesures siguin heteroscedàstics i diferents per a les variables resposta i predictora, no produeix un canvi significatiu en el càlcul dels coeficients de regressió. Malgrat aquestes afirmacions, el lector podrà comprovar en la secció *Results and discussion* de l'article que es troba al final del capítol que el mètode de PCA utilitzat com a tècnica de regressió aplicat al conjunt de dades que es discuteix en l'article produeix resultats que són lluny dels valors vertaders. Per tant, sembla clar que l'aproximació CVR (o el seu cas particular de regressió ortogonal) pot ser útil, sobretot a causa de la simplicitat de càlcul, com es pot observar en les equacions 2.4-2.8, però en casos amb errors heteroscedàstics individuals i diferents per a les variables resposta i predictora caldrà aplicar altres tècniques que considerin errors en dos eixos.

Wakkers i col·laboradors[15] van proposar les expressions basades en aquest model per ser aplicades en la comparació de mètodes clínics i van demostrar que els resultats obtinguts són més fiables que els aconseguits amb la tècnica usual de mínims quadrats.

Fins a la data, l'aproximació CVR (i principalment el seu cas particular de regressió ortogonal) ha estat tractada àmpliament[16-24] i ha estat aplicada per exemple als intervals de confiança individuals de la ordenada a l'origen i el pendent, al càlcul de les probabilitats d'error β associades amb aquests intervals de confiança individuals, a trobar el número de punts que ha de tenir la recta de regressió per tal de que els intervals de confiança individuals de l'ordenada a l'origen i el pendent tinguin unes probabilitats fixades d'errors α i β,[6] i per calcular

els intervals de confiança associats a la predicció de la variable dependent per a l'anàlisi d'una mostra futura.[2]

2.3 Regressió amb errors heteroscedàstics en ambdós eixos

El problema de trobar els coeficients de la recta de regressió considerant errors en dos eixos no és nou. El 1940 ja es troben articles sobre el tema[25] i el 1949 Bartlett[26] va proposar un mètode molt simple per calcular els coeficients de la recta de regressió tenint en compte els errors en dos eixos, amb l'avantatge addicional que no es necessiten els errors individuals associats a cada punt experimental. Aquest mètode consisteix a dividir el conjunt de dades en tres subconjunts de mida tan semblant com sigui possible de la manera següent:

- Aquells amb els valors de la variable predictora més baixos. Calcular la mitjana d'aquest conjunt per a les variables predictora i resposta: $P_1 = (\overline{x}_1, \overline{y}_1)$.
- Aquells amb els valors de la variable predictora més alts. Calcular la mitjana d'aquest conjunt per a les variables predictora i resposta: $P_3 = (\overline{x}_3, \overline{y}_3)$.
- Descartar els valors intermedis.

La recta de regressió correspondrà a la resultant d'unir els punts P_1 i P_3, i tindrà com a coeficients de regressió:

$$\hat{b} = \frac{\overline{y}_3 - \overline{y}_1}{\overline{x}_3 - \overline{x}_1} \tag{2.9}$$

$$\hat{a} = \overline{y} - \hat{b}\overline{x} \tag{2.10}$$

on $\bar{x}$ i $\bar{y}$ corresponen als valors mitjans de la totalitat del conjunt de dades. Com el lector ja pot intuir, aquest mètode té importància històrica però condueix a resultats incorrectes en l'estimació dels coeficients de la recta de regressió.

Els mètodes principals per calcular els coeficients de la recta de regressió tenint en compte els errors en dos eixos es divideixen en dos grans grups: mètodes que necessiten la variància exacta de les variables predictora i resposta en tots els punts experimentals, i mètodes que no la necessiten. En aquesta tesi doctoral ens centrarem en els primers, ja que creiem que el fet de considerar les variàncies de cada punt individual fa que s'abordi el problema des del punt de vista tan semblant com sigui possible a les dades experimentals, tenint en compte els errors individuals en cada punt. D'aquesta manera la recta de regressió trobada s'acostarà més als punts amb menys incertesa i no s'aproparà tant als punts amb més incertesa. D'altra banda, la complexitat matemàtica i de càlcul en els mètodes que no necessiten la variància dels punts experimentals normalment sol ser més gran (com per exemple es podrà comprovar en els mètodes de calibració robusta en l'article del final del capítol), per la qual cosa la divulgació i l'aplicació a la comunitat científica pot ser més dificultosa, i en alguns casos els resultats obtinguts amb la seva aplicació difereixen dels resultats reals (com per exemple en el cas anteriorment comentat de la regressió ortogonal).

Els mètodes de regressió lineal univariant considerant errors en els dos eixos que tenen en compte les incerteses individuals en cada punt troben els coeficients de la recta de regressió d'una manera semblant a com ho fan els mètodes de mínims quadrats i de mínims quadrats ponderats. Si en aquests dos últims es minimitza la suma de distàncies verticals (mínims quadrats) o la suma ponderada de distàncies verticals (mínims quadrats ponderats) a la recta de regressió, considerant errors en els dos eixos es minimitza la suma ponderada de distàncies verticals i de distàncies horitzontals a la recta de regressió (S en l'equació 2.11):

UNIVERSITÀT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

$$S = \sum_{i=1}^{n} \left(\frac{(x_i - \hat{x}_i)^2}{s_{x_i}^2} + \frac{(y_i - \hat{y}_i)^2}{s_{y_i}^2} \right) \qquad (2.11)$$

El procés de regressió lineal considerant errors en dos eixos es veu representat en la figura 2.2. Suposant que el model expressat és el correcte, cada punt experimental (representat per un cercle a la figura) es veu afectat per un error en cada un dels dos eixos. L'error en cada eix, expressat en unitats de variància, es troba representat segons $\sigma_{x_i}^2$ i $\sigma_{y_i}^2$ per a la variable predictora i resposta respectivament del punt experimental *i*, i només se'n pot tenir una estimació segons $s_{x_i}^2$ i $s_{y_i}^2$. La llargària de les línies verticals o horitzontals de la figura és proporcional a la magnitud de l'error en cada eix. La minimització es du a terme en la direcció de les línies discontínues de la figura, però cada component de l'eix *x* i *y* d'aquestes línies es veu ponderat per la variància individual de cada punt experimental (equació 2.11).

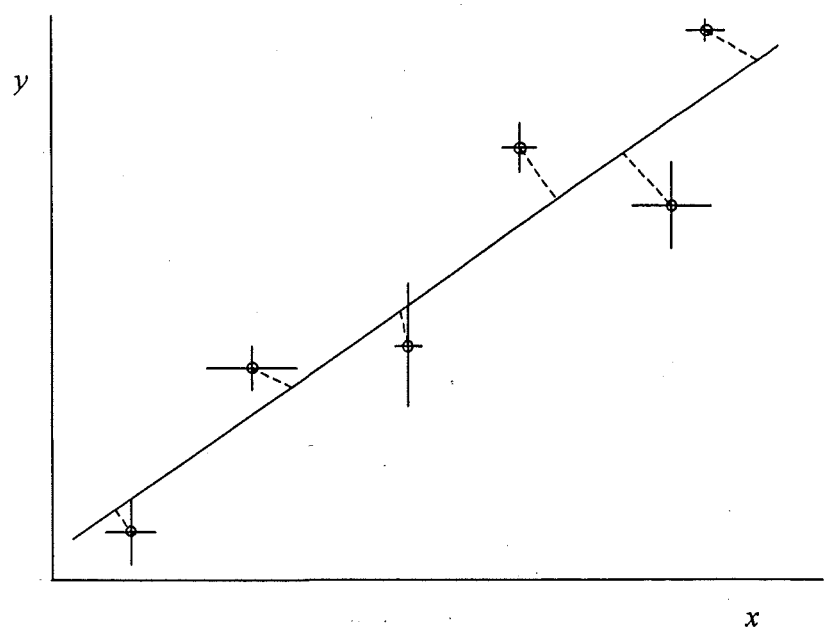


Figura 2.2. Representació de la regressió lineal considerant errors heteroscedàstics individuals en dos eixos.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Si es tornés a repetir l'anàlisi de cada punt experimental, la probabilitat d'obtenir un cert valor a cada punt és màxima a (μ_{x_i}, μ_{y_i}), els valors vertaders de cada punt experimental. Els punts experimentals (x_i,y_i) es veuen afectats pels errors en dos eixos, per la qual cosa rarament coincidiran amb els punts teòrics (μ_{x_i}, μ_{y_i}). Cada punt teòric segueix una distribució normal bivariant, amb funció de probabilitat definida segons:[10]

$$f(x_i, y_i) = \frac{1}{2\pi\sigma_{x_i}\sigma_{y_i}\sqrt{1-\rho_i^2}} \exp \left[-\frac{1}{2(1-\rho_i^2)} \left\{ \left(\frac{x_i - \mu_{x_i}}{\sigma_{ix}} \right)^2 + \left(\frac{y_i - \mu_{y_i}}{\sigma_{y_i}} \right)^2 - 2\rho_i \left(\frac{x_i - \mu_{x_i}}{\sigma_{x_i}} \right) \left(\frac{y_i - \mu_{y_i}}{\sigma_{y_i}} \right) \right\} \right] \quad (2.12)$$

La representació gràfica de la distribució normal bivariant per a un parell qualsevol de dades es veu representada a la figura 2.3.

Ambdues variables agafades separadament presenten una distribució normal i si es fessin talls paral·lels al pla (x,y), s'obtindrien les anomenades el·lipses o cercles d'isoprobabilitat. La inclinació de l'el·lipse ve donada pel paràmetre ρ_i a l'equació 2.12. Si ρ_i=0 (és a dir, no hi ha correlació entre els valors individuals de les variables predictora i resposta), els semieixos major i menor de l'el·lipse són paral·lels als eixos de coordenades, i per tant, l'el·lipse no presenta inclinació. Si, a més, els semiexos major i menor són iguals, es té un cercle. Si, pel contrari, $\rho_i \neq 0$ (és a dir, hi ha correlació entre els valors individuals de les variables predictora i resposta), els semieixos major i menor de l'el·lipse no són paral·lels als eixos de coordenades i l'el·lipse presenta algun tipus d'inclinació. La distribució normal bivariant té el seu centre al punt (μ_{x_i}, μ_{y_i}) i la longitud de cada eix és proporcional a les desviacions estàndards de les variables predictora i resposta.

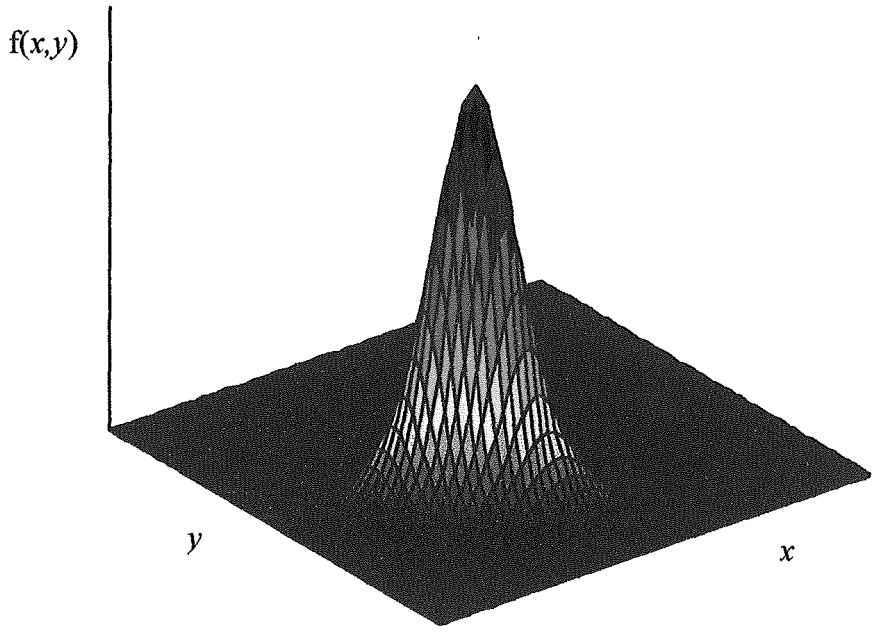


Figura 2.3. Distribució normal bivariant per a un parell de dades experimentals.

La minimització de l'equació 2.11 per tal de trobar els coeficients de regressió no és una tasca senzilla ja que les equacions obtingudes són no lineals en els paràmetres. Per tant, com es veurà més endavant, hi ha diverses aproximacions per tal de dur a terme aquest procés. Una de les aproximacions més simples va ser desenvolupada per Wald[25,27] basada en l'agrupació de punts experimentals. El pendent i l'ordenada a l'origen es calculen segons les expressions següents (on el número de punts, n, s'assumeix que és parell per simplicitat):

$$\hat{b} = \frac{(y_1 + y_2 + \ldots + y_m) - (y_{m+1} + y_{m+2} + \ldots + y_n)}{(x_1 + x_2 + \ldots + x_m) - (x_{m+1} + x_{m+2} + \ldots + x_n)} \qquad (2.13)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad (2.14)$$

on $m=n/2$. En la majoria de les altres aproximacions cal un procés iteratiu per tal de trobar els coeficients de la recta de regressió. Amb el desenvolupament actual de

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

les tècniques informàtiques, aquest procediment iteratiu es resol en poques dècimes de segon per a la majoria de conjunts de dades. Precisament aquesta facilitat és la que fa que aquesta sèrie de tècniques puguin considerar-se avui com una alternativa a l'ús tradicional de les tècniques de mínims quadrats o mínims quadrats ponderats.

L'inconvenient principal en la utilització de les tècniques de regressió que consideren errors dos eixos és la necessitat d'obtenir les variàncies de la variable predictora i resposta per a cada punt experimental de la recta de regressió. L'obtenció d'aquestes variàncies normalment implica repeticions i, per tant, augment del cost de l'anàlisi i del temps emprat. Una altra solució pot ser l'estimació d'aquestes variàncies. Però per tal d'obtenir una estimació adequada l'usuari hauria de tenir un gran coneixement sobre les variables que intervenen en el procés. I fins i tot en aquest cas, sempre és difícil estar segur que l'estimació feta és correcta, per la qual cosa augmentaria la incertesa sobre els valors trobats dels coeficients de regressió.

De totes les tècniques de regressió que consideren els errors individuals en dos eixos revisades a l'article que es troba al final del capítol, per al posterior desenvolupament de tests estadístics associats als coeficients de la recta de regressió desenvolupats en aquesta tesi doctoral s'escollirà el mètode de Lisý i col·laboradors, també anomenat a partir d'ara mètode de mínims quadrats bivariants (*bivariate least squares*, BLS), degut a la rapidesa en l'obtenció de resultats correctes dels coeficients de la recta de regressió (malgrat que com es veurà a continuació no és l'únic que arriba als resultats correctes dels coeficients de la recta de regressió), degut a la facilitat de la programació del seu algorisme de càlcul, i degut a que aquest també proporciona la matriu de variància-covariància dels coeficients de regressió, útil per al posterior desenvolupament de tests estadístics relacionats.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

2.4 Referències

1. H. Scheffé, *Annals Of Statistics*, **1** (1973) 1
2. J. Mandel, *Journal of Quality Technology*, **16** (1984) 1
3. N. Draper, H. Smith, *Applied Regression Analysis*, 2nd ed., Wiley, New York (1981)
4. D.C. Montgomery, E.A. Peck, *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York (1982)
5. R.J. Carroll, C.H. Spiegelman, *Journal of Quality Technology*, **18** (1986) 170
6. C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, D.L. Massart, *Analytica Chimica Acta*, **338** (1997) 19
7. R.L. Anderson, *Practical Statistics for Analytical Chemists*, Van Nostrand Reinhold, New York (1987)
8. M.A. Creasy, *Journal of the Royal Statistical Society, Series B*, **18** (1956) 65
9. W.A. Fuller, *Measurement Error Models*, John Wiley & Sons, New York (1987)
10. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam (1997)
11. C. Hartmann, J. Smeyers-Verbeke, D.L. Massart, *Analusis*, **21** (1993) 125
12. S. Van Huffel, J. Vandewalle, *The Total Least Squares Problems. Computational Aspects and Analysis*, Siam, Philadelphia (1991)
13. S. Wold, K. Esbensen, P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, **2** (1987) 37
14. P.J. Cornbleet, N. Gochman, *Clinical Chemistry*, **25** (1979) 4326
15. P.J.M. Wakkers, H.B.A. Hellendoorn, G.J. Op de Weegh, W. Heerspink, *Clinica Chimica Acta*, **64** (1975) 173
16. T.C. Koopmans, *Linear Regression Analysis of Economic Time Series*, DeErven F. Bohn, Haarlem (1937)
17. G. Tintner, *Annals of Mathematical Statistics*, **16** (1945) 304

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

18. D.V. Lindley, *Journal of the Royal Statistical Society Suplements*, **9** (1947) 218
19. A. Madansky, *Journal of American Statistical Society*, **54** (1959) 173
20. T.W. Anderson, *Annals of Mathematical Statistics*, **22** (1951) 327
21. V.D. Barnett, *Biometrika*, **54** (1967) 670
22. P.A.P. Moran, *Journal of Multivariate Analysis*, **1** (1971) 232
23. M.G. Kendall, A. Stuart, *The Advanced Theory of Statistics*, Vol. II, 4th ed., Hafner, New York (1979)
24. W.A. Fuller, *Annals of Statistics*, **8** (1980) 407
25. J.H. Wald, *Annals of Mathematical Statistics*, **11** (1940) 891
26. M.S. Bartlett, *Biometrics*, **5** (1949) 207
27. M.A. Sharaf, D.L. Illman, B.R. Kowalski, *Chemometrics*, John Wiley & Sons, New York (1986)

Journal of Chemometrics 9 (1995) 343-362

Univariate regression models with errors in both axes

J. Riu* and F.X. Rius

Departament de Química. Universitat Rovira i Virgili.

Pl. Imperial Tàrraco, 1. 43005-Tarragona. CATALONIA, SPAIN.

ABSTRACT

Calibration is a fundamental step in the calculation of the unknown concentration of analyte in most analytical methods. It is known that for certain methodologies, if only the errors in the independent variable are taken into account, there may be considerable errors in the estimation of the value of the regression coefficients, the derived statistical parameters and, in some cases, the sought for response and concentration values. This paper reviews the calibration methods including some references to procedures for the detection of outliers and robust regression when there are errors in both axes. The advantages and limitations of the different approaches are discussed and a comparative study is made of the approaches of several techniques for which computer programs have been developed based on the algorithms put forward by the different authors. Finally, some trends of future development in this field are envisaged.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

INTRODUCTION

Most of analytical methods incorporate the calibration stage as a fundamental step in the calculation of the concentration of the unknown analyte. The most widely used mathematical technique to calculate the regression coefficients, least squares, only takes into account the error in the dependent variable (y), considering that the error associated to the independent variable (x) is negligible in comparison to the former. Nonetheless, there are other errors in the calibration step, both in the process of measuring the signal or the response and in the preparation of the calibration samples.

The least squares technique can be applied whenever the following conditions are fulfilled:[1,2] (i) the variance for each value of the dependent variable ($s_{y_i}^2$) must be much larger than for the corresponding value of the independent variable multiplied by the slope ($b \cdot s_{x_i}^2$).[3,4] (ii) the variances of the dependent variable have uniform values throughout the linear range (homoscedasticity). If this is not the case, the weighted least squares method,[5-7] or non-parametric methods[8,9] can be applied. (iii) the errors of the dependent variable must be mutually independent, and if subsequent inferences are to be drawn, they must follow a normal distribution for each calibration point.[10,11]

It has been shown that for certain methodologies the first of these conditions is not fulfilled, giving rise to considerable errors in the estimation of the value of the regression coefficients and of the derived statistical parameters, even if the weighted least squares method is used.[12] In these cases, the error associated to the independent variable must be taken into account, a procedure which is known as calibration with errors in both axes.

There are numerous examples of its application in the field of chemical analysis. One of them consists of several radiochemical techniques in which the calibration samples must be prepared by means of a process in which the errors in the concentrations are far from negligible. Another example is the determination of alkaline elements by flame emission spectrometry (FES)[13] where the instrumental responses are so stable that the errors of the independent variable, comparable to those of the responses, have to be taken into account.

Although the latter example formally consists of the same problem of having errors in both coordinates, it is apparent that in this case the errors associated with the regression coefficients will be very small compared to the former example. An exception is the so-called Berkson case, in which least squares regression can be applied in cases where the x variable can be set to pre-assigned target values (the x_i's are fixed and predetermined by the experimenter), even when there may be errors in the independent variable.[14,15] Another very important field in chemical analysis in which calibration methods which take errors in both axes into account must be used, is in the procedure of method validation. In particular, when two different methods are compared by analyzing samples which contain different concentrations of analyte, the responses of each of the methodologies are plotted both on the abscissa and the ordinate axes, sometimes with comparable errors.[4] In this case, the interpretation of variables as dependent or independent is purely arbitrary, and the method used should give the same relationship whichever way the two variables are designated.

Several authors have partially reviewed the subject of linear calibration considering errors in both axes,[3,4,16] but the literature in existence is far from plentiful and certainly disperse, especially that which may be of interest to the analyst. This paper reviews the regression techniques with errors in both axes and discusses the advantages and limitations of the different approaches. At the same time a comparative study is made of several methods, for which computer programs have

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

been developed based on the algorithms provided by the different authors. Finally, some trends of future development in this field are ventured.

The paper is structured on the basis of the data-model-method triplet. Firstly, the procedures for regression diagnostics that check the structure of the data are briefly examined. The study of the model is restricted to the straight line that fits the experimental data containing errors in both axes. Finally, as far as the method is concerned, there is a review of a number of algorithms that have been developed to determine the least-squares estimates of the regression coefficients and derived statistical parameters.

NOTATION

When dealing with regression methods where both variables have uncertainties, the terms independent or explanatory variable for x_i-values and dependent or response variable for y_i might lose the meaning that they have in ordinary least squares regression. This would be the case when comparing the results of two different methods although this terminology is valid when regressing the instrumental response for the concentration values for a method which also has abscissa errors. The data estimated from the model are denoted as $(\hat{x}_i, \hat{y}_i)$; $s_{x_i}^2$ and $s_{y_i}^2$ are the respective estimated variances of x_i and y_i.

THE DATA

When a regression technique is applied to data pairs (x_i, y_i), it is important to investigate the data in order to check that it complies with the statistical conditions. Regression diagnostic tests should be applied to detect the presence of influential

points, i.e. outliers and high leverages. To do so, several test can be applied[17-20] such as Huber's M-estimator,[21] Jackknife residuals,[22] standardized residuals, normalized residuals, predicted residuals,[23] studentized residuals,[24,25] recursive residuals or diagonal elements of the Hat matrix.[26,27] But these tests can only be found in literature applied to models which incorporate errors in only one axis. Therefore, some modifications should be introduced in order to use them in cases in which there are errors in both axes. However, some other procedures have been developed specifically to detect outliers in the field of calibration with errors in both axes.[28-34] It is well known that the effect of outliers can be avoided by using robust calibration.[17,28-36] In this latter case, the residuals are not only useful to examine influential points, but the transformation to uncorrelated residuals also has a certain intuitive appeal for some special purposes such as developing formal tests for normality,[37] change points,[38] serial correlation[39] or heteroscedasticity.[40-42]

THE MODEL

The model to be studied is the straight line that best fits the experimental data containing errors in both variables. The well known relationship between each experimental data pair (x_i, y_i) is

$$y_i = a + bx_i + \varepsilon_i \tag{1}$$

The main objective is to calculate the regression coefficients, a and b and the error, in terms of variances, associated to each of them. In a further step, statistical tools to test the performance of analytical methods and procedures should be developed.

In addition to this model, there are some authors that consider different modifications when there are errors in both axes. Christian *et al.*[43] use a linear

model with several independent variables, but they do not describe the method used to optimize the values of the regression coefficients. They only state that the problem of optimizing functions with any number of parameters and with variances which vary from point to point can be reduced to a one-parameter non-linear optimization problem.

As well as their contribution to the straight line model (which will be seen below), Lisý *et al.* developed a method that can be applied in cases in which calibration with a polynomial model whose degree is higher than 1 is needed.[44] An expression for the coefficients of the polynomial and the variance-covariance matrix is also obtained.

Jefferys has developed GaussFit, a computer program for solving least squares and robust estimation problems.[45] GaussFit allows the user to work not only with linear calibration, but also with a wide range of models (linear, non-linear, simple, complex).

Neri *et al.*[46] have developed a straightforward generalization of their method for calculating the regression coefficients of a straight line when there are errors in both axes[47] (reviewed below). This is useful when the polynomial model that has to be used has a degree which is higher than 1.

THE METHOD

The best straight line that fits data when both variables contain errors should minimize the expression

$$S = \sum_{i=1}^{n} \left[\frac{(x_i - \hat{x}_i)^2}{s_{x_i}^2} + \frac{(y_i - \hat{y}_i)^2}{s_{y_i}^2} \right] \tag{2}$$

However, the difficulties of deriving the regression coefficients a and b in the previous expression have led to the generalised use of the following basic expression involving only y-residuals:[48]

$$S = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{w_i} \tag{3}$$

This expression is obtained by introducing into equation (2) the value corresponding to x_i obtained from the derivative of equation (2) relative to the independent variable. The weight[49]

$$w_i = s_{\varepsilon_i}^2 = s_{y_i}^2 + b^2 s_{x_i}^2 \tag{4}$$

has its origin in the estimated variance of the i-th residual, obtained by applying the error propagation law.

Therefore, the weighted sum of squares to be minimized becomes:

$$S = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{s_{y_i}^2 + b^2 s_{x_i}^2} = \sum_{i=1}^{n} \frac{(y_i - a - bx_i)^2}{s_{y_i}^2 + b^2 s_{x_i}^2} \tag{5}$$

Numerous approaches discussed here deal with different mathematical ways of solving equation (5). All the methods that minimize equation (5) get the same results with a precision of 8-9 decimal places. Other authors have used slightly different expressions and different strategies that will be mentioned in each case.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Journal of Chemometrics 9 (1995) 343-362

THEORETICAL APPROACHES TO REGRESSION WITH ERRORS IN BOTH AXES

The different approaches reviewed are divided into two groups: methods in which the exact variance of both dependent and independent variables at each data point must be known, and other not so stringent methods, in which the variances can be either estimated or not taken into account.

Methods which need the exact point variances in each axis

The York Method and the Reed solution

One of the first rigorous attempts to solve the problem of univariate regression with errors in both axes was undertaken by York,[50] who suggested the best solution to the equation (5) when he developed the iterative method called "least-squares cubic", consisting of estimating an initial value for the slope of the calibration line and then solving the resulting cubic equation

$$f(b)=b^3-3ab^2+3\beta b-\gamma=0 \qquad (6)$$

so as to obtain a better result for the slope until convergence is achieved.

This equation gives three real solutions, one of which is presumably the correct one, arrived at after an iterative process. York mistakenly concluded that the correct solution is the third root of the cubic equation, but very often this method does not lead to the correct solution (despite the fact that York's method arrives at the correct solution for the data set described in Table 1). York's method yields a slope estimate that is invariant upon switching axes.[51]

Reed[49,52] has re-examined the York method and has tried to clarify some of its points. The problem lies in the fact that equation (6) is not really cubic because of the implicit dependence of α, β and γ on the slope. Having proved these conditions, the iterative process used does not always lead to the correct root. Reed even found that in some cases the first iterations with the York method lead to a complex third root from which the real part must be chosen in order to continue with the next iterations according to the procedure described by York. Reed suggested another approach, which consisted of examining the real roots of equation (6), cancelling the cubic factors and clarifying some factors, changing equation (6) into a least squares quadratic given by

$$g(b) = Ab^2 + Bb - C = 0 \qquad (7)$$

Bearing in mind that the coefficients A, B and C are functions of the slope, an iterative process is needed to find the best-fit slope, which is usually found after a few iterations. The solutions arrived at using this method for several sets of data match the ones arrived at with other methods reviewed. Reed[49] provides the expressions for the calculation of equation (7) together with the "correct" variances of the slope and the intercept, since the ones given by York[50] were not correct.

When only the weighting factor corresponding to the dependent variable is taken into account in term B of equation (7), this equation is the same as the one given by Deming.[53] Deming's method can be applied only when the errors are constant over the range of the x and y values, otherwise it fails even in simple cases.

The Williamson Method

Williamson[54] showed that equation (6) in York's method is redundant and that it can be written as a linear equation, which is simpler and gives more accurate

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

answers. Williamson developed an easily programmable iterative method to determine the intercept and the slope of the calibration straight line which minimize the weighted sum S described by equation (5). Other authors have suggested different weighting factors.[55]

It should be pointed out that $(\hat{x}_i, \hat{y}_i)$ do not correspond to a theoretical point through which the regression line would pass, but that $\hat{x}_i$ corresponds to the value of the independent variable given by the model on introducing an experimental result of the dependent variable, and $\hat{y}_i$ corresponds to the value of the dependent variable given by the model on introducing an experimental result of the independent variable.

Apart from the expressions for finding the value of the slope and the intercept, the method also allows an expression to be found for their variances, in terms of the known point variances of x_i and y_i. One of the advantages of this method is that the condition of homoscedasticity does not have to be fulfilled either at the dependent variable or the independent one. Apart from obtaining more accurate answers than with the York method, the number of iterations needed and the number of calculations to be performed is smaller, which means that they are obtained faster.

To measure the goodness of the fit, the χ^2 test is used on the S sum with n-2 degrees of freedom. The fit is considered to be a good one when the value of S in equation (5) obtained by substituting values a and b is close to the number of degrees of freedom (i.e. the number of data points minus the number of parameters, n-2). Therefore, if the individual variances in x_i and y_i and the regression coefficients a and b are known, tables of the χ^2 distribution may be used to determine whether or not the method provides an adequate fit of the data set. The expression S is not limited and can take any value, in contrast to the least squares correlation coefficient, which, although its use in regression does not have a

mathematical base, is limited to between -1 and 1. The χ^2 test is used in most of the methods reviewed, as long as there is a normal distribution of errors in both variables.

From the analyst's point of view, one drawback of the method is that it needs the values of the variances for each experimental point, and these are not always easily obtainable. One has to be aware that to find the variance of the independent variable, x, several repetitions of the calibration samples have to be prepared and measured with a well established method. Frequently, this family of methods may furnish incorrect results if outliers are present in the data. In these cases, robust regression methods or methods for detecting outliers have to be resorted to.

Williamson's[54] mathematical development to determine the expressions for the slope and the intercept is not very detailed and, in the case of the variances, is reduced to the final expressions. Williamson's approach leads to a symmetrical regression as regards the coordinate axes. Thus, the confidence interval of y for a given value of x (direct calibration), is the same as the confidence interval of x, given one value of the signal (inverse interpolation). The variance of the slope should also be invariant upon switching the axes and the value of the quantity s_b/b^2 should be constant when x and y designations are exchanged.[51]

Other authors have continued this line of research using the mathematical process developed by Williamson and developing the algorithm for computers so as to be able to calculate the slope, intercept,[2,56] and variances.[56] The results obtained by applying the Williamson method to a set of data,[2,56] differ by 10% with regard to the least squares method in the variances of the slope and the intercept, and by 1% in the values of the slope and the intercept. Despite these results, there can be a difference up to 30% in these values for a set of data that is no correlated so well. For the set of data studied there had to be 12 iterations to get accuracy up to 13 decimal places.[2]

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

A similar approach, carried out by González *et al.*,[16] consists of minimizing equation (5). This results in normal equations that are non-linear in their parameters, and instead of using a complicated numerical procedure to solve them, an iterative procedure is used which involves the resolution of a cubic equation for b to obtain the expressions for the slope and the intercept. The main difference between the methods of Williamson and González *et al.* is that while González *et al.* use a cubic equation to find the regression coefficients of the straight line, Williamson states that the equation can be constructed as a polynomial in b of any degree, and he writes it in linear form. González *et al.* do not include expressions for the variances and covariances of the different coefficients in their paper.

Neri *et al.*[47] developed a method which is very similar to González's *et al.* This method consists of minimizing the sum of the shortest distances (the squared perpendicular distances) from each experimental point to the theoretical straight line (N-minimization)[28]:

$$S = \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{\sqrt{b^2 - 1}} \right)^2 \tag{8}$$

By applying the error propagation law with no approximation, equation (8) becomes equation (5). To carry out the minimization, the authors use an iterative method to find the root of the expression $\delta S/\delta b = 0$, using a cubic equation in the slope similar to González *et al.* Neri *et al.*[46] claim that in all the fitting procedures the most important feature is the choice of the weighting factor and not the direction in which the minimization is performed.

Another method that minimizes equation (5) is the method developed by Press and Teukolsky,[57] but they use standard numerical algorithms to minimize one-dimensional functions. In their method, they find the uncertainties for the slope and

the intercept by looking at the respective projections onto the intercept and slope axes of the "confidences region boundary". The expressions to find the slope, the intercept and their variances are not given, and the theoretical background to interpret this method is more difficult than most of the other methods reviewed.

Another method following these premises is the method of effective variance by Irvin and Quickenden among others.[10,58-60] This method is based on the minimization of equation (5) as well, but it does not achieve accurate results because all the partial derivatives are evaluated in relation to the values of the independent variable obtained according to the model,[59] instead of being evaluated in relation to experimental data. Furthermore, Irvin's development[10] fails when doing the derivation relative to b since they neglect the term dw_i/db.[11] The errors obtained by this approach can be considerable. The use of the standard least squares algorithm to reevaluate the weights in a new iteration does not ensure that the weights will converge on a solution that minimizes S.[61] Several authors have stated[2,11,61] that the method of effective variance does not lead to a correct solution in the case of linear calibration when the two variables are subject to error. Only a slight improvement can be achieved with respect to the least squares method.

The Lisý method et al. and related methods

The method developed by Lisý *et al.*[62] consists of minimizing the sum of the weighted residuals expressed as equation (3), but this method uses the variance of residuals as the weighting factor:

$$w_i = s_{\varepsilon_i}^2 = s_{y_i}^2 + b^2 s_{x_i}^2 - 2b\,\mathrm{cov}(x_i, y_i) \qquad (9)$$

They can be expressed using the Taylor series, even when the covariance between each point of the dependent and independent variables is not zero.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

The fact that the covariance between each point of the dependent and independent variable is not zero means that the method can be applied in situations where the errors between both variables are correlated. By minimizing the sum of weighted residuals in relation to the slope and the intercept two non-linear equations are obtained and by putting in the partial derivatives of the squared residuals the following can be written in matrix form:

$$\mathbf{Rb} = \mathbf{g} \tag{10}$$

$$\begin{vmatrix} \sum_{i=1}^{n} \frac{1}{s_{\varepsilon_i}^2} & \sum_{i=1}^{n} \frac{x_i}{s_{\varepsilon_i}^2} \\ \sum_{i=1}^{n} \frac{x_i}{s_{\varepsilon_i}^2} & \sum_{i=1}^{n} \frac{x_i^2}{s_{\varepsilon_i}^2} \end{vmatrix} \times \begin{vmatrix} a \\ b \end{vmatrix} = \begin{vmatrix} \sum_{i=1}^{n} \left[\frac{y_i}{s_{\varepsilon_i}^2} + \frac{1}{2} \left(\frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right)^2 \frac{\partial s_{\varepsilon_i}^2}{\partial a} \right] \\ \sum_{i=1}^{n} \left[\frac{x_i y_i}{s_{\varepsilon_i}^2} + \frac{1}{2} \left(\frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right)^2 \frac{\partial s_{\varepsilon_i}^2}{\partial b} \right] \end{vmatrix} \tag{11}$$

To determine the slope and the intercept, which are the components of vector **b** in equations (10) and (11), it is only necessary to carry out an iterative process on the following matrix form:

$$\mathbf{b} = \mathbf{R}^{-1}\mathbf{g} \tag{12}$$

where very small values have been assigned as starting guesses of the regression coefficients. With this method the variance-covariance matrix of the calibration straight line coefficients are obtained without having to use additional expressions, only by multiplying the final matrix $\mathbf{R}^{-1}$ by $S/(n\text{-}2)$.[63,64]

It should be pointed out that if one were to be in the situation in which $s_{\varepsilon_i}^2 = s_{y_i}^2$ (all the errors are due to the independent variable), the expressions obtained are the same as if the least squares method were to be applied.

The method still has the drawback that the uncertainties of each experimental point, represented by their variance, have to be known and that incorrect results might be obtained if outliers are present. In a similar manner as in previous approaches, neither do the authors deal with the confidence intervals of the regression coefficients. Despite this, the method is quick, it gets correct results with few iterations and it has the advantage of being able to find the variance-covariance matrix with the same iterative process, with no additional expressions.

The authors state the mathematical expressions corresponding to the matrix but they do not give the algorithm to find the values of the slope and the intercept by an iterative process.

Another variation was introduced by Brooks *et al.*[65] and consists of minimizing the sum of the lines between each experimental point and the regression line in the direction proportional to the ratio of errors in y_i and x_i. The slope of the line between each experimental point and the regression line is $\alpha_i = -s_{y_i} / s_{x_i}$, and each of these lines is represented by

$$\hat{y}_i - y_i = -\frac{s_{y_i}}{s_{x_i}}(\hat{x}_i - x_i) \tag{13}$$

The introduction of equation (13) into equation (2) finally gives

$$S = \sum_{i=1}^{n} \frac{(y_i - a - bx_i)^2}{w_i} = \sum_{i=1}^{n} \frac{(y_i - a - bx_i)^2}{(s_{y_i} + bs_{y_i})^2} \tag{14}$$

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

The minimization is carried out by deriving equation (14) relative to the slope and the intercept, resulting in a pseudo-quadratic equation on the slope, which is solved iteratively. The authors give the expressions for the errors in the slope and in the intercept as well. Brooks' method *et al.* has the advantage that it has a more consistent theoretical background since the minimization is carried out on distances that form an angle with the regression straight line which is proportional to the uncertainty of the errors in the independent variable divided by the uncertainty of the errors in the dependent variable. In contrast, in most of the other methods reviewed the uncertainties are only taken into account in the weighting factor and the methods minimize the sum of *y*-residuals (or at most minimize the sum of the residuals perpendicular to the calibration line). But in deriving the expression corresponding to the weighting factor, the authors introduce the term $s_{y_i} s_{x_i}$ that should correspond to the covariance between the errors in the dependent and independent variables at each individual point when the errors are correlated. This term should be omitted in cases when the errors are not correlated (most of the cases). The introduction of this term in the algorithm as the simple multiplication of the standard deviations of the errors of each individual point lead to erroneous values of the slope and the intercept. Without this factor, the method leads to correct results for several sets of data with errors in both axes. In cases in which the errors between each point of the dependent and independent variable are uncorrelated, equation (14) can be expressed the same as equation (5).

Other Methods

Some authors[13,66] also use a type of weighted regression. In these cases, the process is carried out taking into account the errors in only one axis, and then the necessary corrections are applied so as to introduce the errors in both axes. Because of the heteroscedasticity present in many sets of data at the dependent variable, a modelling is carried out of the variances along the calibration line with an iterative

process. The iteration is done by estimating a function of the standard deviations along the range of linearity (the function may be linear, quadratic ... it depends on each individual case). The standard deviation is found for each point by repeating the measurements and it is fitted to the model using least squares, so obtaining a new estimation for the standard deviation. These new values are again adjusted to the model, but this time using weighted least squares, where the weighting factor can be $1/s_{y_i}^2$ or $1/s_{y_i}^4$. Each iteration uses weighting factors calculated from the values predicted in the previous step. The procedure finishes when the difference between two values is equal to a pre-established value. The variance at each point can be estimated like this, and these values are taken as weighting factors. These authors carry out a more thorough data processing procedure, including the detection of outliers and the establishment of some confidence intervals for the independent variable, when only the errors associated with the dependent variable are considered.

Considering the errors in both axes, Lwin and Spiegelman[67] put forward a procedure appropriate for widening the confidence interval when the errors in the independent variable are small but no negligible when compared to the dependent variable. Using experimental values it is assumed that errors are generally limited to a value that is taken at most to be 0.5% of the value of each measurement. Because of this, the confidence intervals that were deduced by only taking the errors in one axis into account are increased by a certain quantity which varies along the range of linearity, depending on the value of the independent variable chosen, but which tends to be around 0.5%. If the error associated to each measurement is higher than 0.5% of the value of each measurement, the confidence intervals will increase by a higher value. Despite this, the effect of taking errors in the independent variable into account does not alter the values of the slope and the intercept obtained when taking only the errors in one axis into account. Only the confidence intervals are widened.

This method is useful when an associate estimator which requires assumptions about the error variances cannot be applied to the error of the independent variable. The method used is a non-parametric method, which does not consider certain distributions, usually Gaussian, of measurement errors.

Lybanon[68] uses a method developed by Jefferys[69,70] which considers an n degree polynomial, finding its coefficients by iterations. It is based on minimizing the sum of terms $\frac{1}{2}D_i^2$, where D_i is the perpendicular that joins the estimated curve with each experimental point. The minimization is subject to a series of restrictions, so Lagrange's multiplying method is used to find the solution. A series of equations, for the most part non-linear, are obtained and so the Newton method is immediately applied to linearize them. These solutions make up the first iteration. Successive iterations are necessary until a criterion of convergence is arrived at. Jeffery's method is the standard reduction method in GaussFit, a computer program for solving least squares and robust estimation problems.[45] The program is written in C, and when forming the equations of condition, partial derivatives with regard to the coefficients and the data are computed using an algebraic manipulator. Then, the solution algorithm performs the matrix calculations to iteratively obtain the regression coefficients. When a 1 degree polynomial is used (straight line model), the method gives accurate results, varying slightly when a higher-degree polynomial is used.

Lybanon[71] gives the mathematical processing of the method and the algorithm to program it on a computer in Applesoft BASIC language. Despite everything, the author states that the method is formally identical to least squares, with the provisos inherent to the method, so it can be easily implemented if some minor modifications are made to the software. The method is also applicable if heteroscedasticity is present in the data, both in the dependent and independent variable, and the program gives the values for the intercept, the slope and their variances.

Jefferys's method, mentioned above, is very similar to the one developed by Britt and Luecke.[72] Although Britt and Luecke's method is prior to his, it seems that Jefferys was not aware of it, which corroborates the impression that the methods of linear calibration taking errors in both axes into account are not very familiar. The method developed by Britt and Luecke is valid for linear and non-linear functions in variables and coefficients, and requires the user to introduce both the function to be adjusted and the first partial derivatives. In contrast, the Powell-Macdonald[73] method calculates the derivatives numerically and only requires the function of regression from the user.[74,75]

Mandel[3] developed a method which consists of constructing a set of new variables (u and v) related to x and y, but in such a way that the least squares conditions are at least approximately complied with for (u,v). The new variables are constructed as follows: $u_i = x_i - ky_i$, $v_i = y_i - b_1 x_i$.

The results of applying ordinary least squares are converted back into the (x,y) original variables. The application of the least squares conditions to the new variables gives a slope of zero and, after transforming them back, gives an iterative expression for the slope of the original variables. This method can be applied even when errors are correlated. The author also gives the expression for the intercept and the variances for the slope, the intercept, and for the value of a variable (either the dependent or the independent) given the value of the other one. Probably the main drawback of the method is that the relation between the variance of the error of x to that of the error of y needs to be known and that the expressions obtained for the slope and the intercept are only strictly correct if b_1 and k are given constants. This latter assumption is not correct since they are only estimators of the true values. However, Mandel states that a Monte Carlo experiment has verified that they closely match the true values.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Cumming *et al.*[76] developed another approach for fitting a straight line to a set of independent points subject to correlated errors and applied it to determining stability constants from the potentiometrically determined formation curve. The authors redefine the straight line given in equation (1), as

$$px+qy+t=0 \tag{15}$$

and minimize equation (16), in which the uncertainties in both variables and the correlation between errors are taken into account:

$$S=\sum_{i=1}^{n}\left[W_{x_i}(\hat{x}_i-x_i)^2+W_{y_i}(\hat{y}_i-y_i)^2+W_{xy_i}(\hat{x}_i-x_i)^2(\hat{y}_i-y_i)^2\right] \tag{16}$$

where

$$W_{x_i}=\left[\frac{s_{y_i}^2}{s_{x_i}^2 s_{y_i}^2-\mathrm{cov}(x_i,y_i)^2}\right] \tag{17}$$

$$W_{y_i}=\left[\frac{s_{x_i}^2}{s_{x_i}^2 s_{y_i}^2-\mathrm{cov}(x_i,y_i)^2}\right] \tag{18}$$

$$W_{xy_i}=\left[\frac{\mathrm{cov}(x_i,y_i)}{s_{x_i}^2 s_{y_i}^2-\mathrm{cov}(x_i,y_i)^2}\right] \tag{19}$$

When there is no correlation between errors, equation (19) and all the terms involving $\mathrm{cov}(x_i,y_i)$ are zero. Equation (16) is minimised taking into account the following conditions:

$$p^2+q^2=1 \tag{20}$$

$$px_i+qy_i+t=0 \quad \text{for all } i \tag{21}$$

Equation (21) is used so that neither p nor q become infinite for any slope. The minimization is carried out iteratively taking p and q as the elements of the eigenvector which corresponds to the smallest eigenvalue of the resulting (2x2) matrix, normalized so that $p^2 + q^2 = 1$. After p, q and t have been found, the slope and the intercept of the calibration line are easily found by comparing equation (15) and (1). The authors give the expression for the calculation of the variance of the slope of the calibration line as well. However, in our experience this method only gives correct results for some data sets (e.g. the data shown in Table 1), and in certain cases gives worse results than the line obtained with least squares.

i	x	$1/s^2_x$	y	$1/s^2_y$
1	0.0	1000	5.9	1
2	0.9	1000	5.4	1.8
3	1.8	500	4.4	4
4	2.6	800	4.6	8
5	3.3	200	3.5	20
6	4.4	80	3.7	20
7	5.2	60	2.8	70
8	6.1	20	2.8	70
9	6.5	1.8	2.4	100
10	7.4	1	1.5	500

Table 1. Pearson's data and York's weights.

Methods which do not need the exact variance

The methods looked at up to now have the drawback that the variances of the points in each axis have to be known exactly. If the variances of the points are not known, some assumptions have to be made for some of the special cases which have been reviewed.[77,78] However, there are some methods for which this information is not needed, at least in exact terms. These methods can be subdivided into two groups: robust methods (which are not sensitive to the presence of outliers) and non-robust methods (which are affected by the presence of outliers and can lead to incorrect results). Fuzzy calibration and least median structural regression (LSR) can be classified among robust methods, while principal component analysis (PCA) and Bartlett's three groups are non-robust methods.

Fuzzy Calibration

The fuzzy calibration method is a non-statistical calibration method developed by Otto and Bandemer[79-81] based on the theory of fuzzy sets.[82] This theory was born from the necessity to work with ambiguous information or terms which can include a wide range of a certain property. Zadeh's idea is to define a membership function, which allows one item of data to be assigned different degrees of belonging to a set.[83]

Applied to univariate calibration with uncertainties in both axes, it essentially involves finding the straight line that cuts most surface area of the tridimensional domains of influence, or membership functions, of the experimental points. These tridimensional domains (spheres, ellipsoids, ... or parabolas if only errors in one axis are taken into account) have a structure which is appropriate to the mathematical functions called supports of the fuzzy set (circles, ellipses ... or lines if errors in only one axis are taken into account) which describe in the best way possible the uncertainty of these experimental values (Figure 1). The uncertainty

Journal of Chemometrics 9 (1995) 343-362

values of the experimental points are specified thanks to the relative knowledge of the measuring process and any subjective information that one might have about the calibration system. It is very important to take care when choosing the type and size of the supports, as this will affect the sensitivity with which they detect outliers. Once the coefficients of the calibration straight line have been obtained it can be used both for direct and inverse calibration.

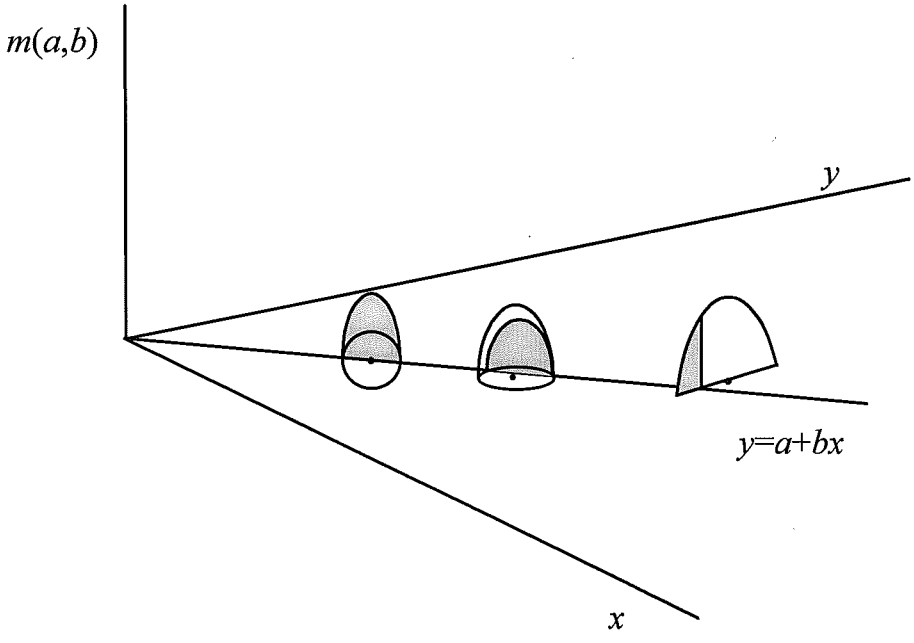


Figure 1. Graphical representation of fuzzy calibration: circle, ellipse and line are used as supports. The membership function $m(a,b)$ is represented in the ordinate axis.

To obtain the slope and the intercept of the straight line which cuts most surface area of the tridimensional domains of the experimental points, it is advisable to plot the experimental points and their domains of influence, and then calculate a first approximation of the coefficients of the calibration line according to a linear model or, for example, a cubic spline method. Immediately afterwards, the membership function for each point ($m_i(a,b)$) is calculated, so obtaining the area of

tridimensional domain crossed by the calibration line. For each experimental point the so called local approximation ($m_i(f)$) can also be found, which defines the maximum area of tridimensional domain which can be cut by the calibration line. If this is done for each experimental point a relative membership function can be found according to

$$m(a,b) = \frac{\sum_{i=1}^{n} m_i(a,b)}{\sum_{i=1}^{n} m_i(f)} \qquad (22)$$

As can be seen, $m(a,b)$ will be between 0 and 1 and reflects the degree of approximation with which the calibration line fits the experimental points. A value of zero would mean that the point does not belong to the experimental universe (from which it can be deduced that the point is almost certainly an outlier).

Another series of coefficients a and b are taken and the process is repeated, calculating the degree of approximation again. By tridimensionally plotting the relative membership function ($m(a,b)$) in relation to a and b, and by modelling the response area, the values of a and b which have the highest degree of approximation are obtained. These are the results of the calibration process. Another alternative is to obtain the relative membership function ($m(a,b)$) for three series of (a,b) and optimizing the regression coefficients obtained using the modified Simplex method.[84-86]

In his work, Otto and Bandemer[79] describe the steps to take to find the slope and the intercept, but he does not give the final algorithm. Neither does he give the expressions for the calculation of the a- and b-variances, nor any sort of confidence interval. Hu *et al.*[84] give the algorithm of their fuzzy calibration program, CAC, written in Fortran 77 for IBM PC. Apart from the estimation of the slope and the

intercept, the residuals, the standardized residuals and the confidence intervals for a significance level of 95% for the regression line and individual values are also given, but only in those cases in which there is error in only one axis (a line as fuzzy support) or there is the same error in both axes (a circle as fuzzy support).

The main advantage of the fuzzy set theory applied to linear calibration is that the resulting regression method is robust to the presence of outliers.[87] So, it minimises the influence of outliers in the calculation of the regression coefficients and, at the same time, can be used to detect[28] their presence. It has the advantage that it works both for traditional calibration and for inverse calibration. Otto[79] claims that a feature of the fuzzy methods is their usefulness when the sample size is small. In this respect there are no limitations as far as the number of experiments are concerned. As mentioned above, an important aspect is that with this method the uncertainty of each experimental point does not have to be accurately known since subjective appreciations can be used. In cases of linear regression in which there are no criteria to discern the size of the supports, it can be used as an explorative method to designate the points that probably are outliers.

The fact that knowledge can be gained about the uncertainty of the observations without needing to assume a statistical model on the properties of the residuals is also important.[84] Fuzzy calibration overcomes the limitations of parametric methods in the construction of confidence intervals by using domains of influence. However, when applying this method, it is very useful to have computerised graphical support.

Least Median Structural Regression (LSR)

Feldmann[34] developed a robust method for data that incorporates errors in both axes which is a version of the so-called least median of squares (LMS) introduced

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

by Rousseeuw and Leroy.[33] In this method, instead of minimizing a sum of squares, the median

$$\psi_r^2(\mu_x;\mu_y;b) = \mathrm{med}_i\left[|b|^{-1}(y_i-\mu_y)^2+|b|(x_i-\mu_x)^2\right] \tag{23}$$

is considered and minimized relative to the model parameters; μ_x and μ_y are the respective expectations of x and y, and (μ_x, μ_y) defines the so-called robust focal point. Equation (23) is the basis of the least median structural regression (LSR). In this regression, the uncertainties associated to each variable are not needed. The calculation of the regression coefficients is more complicated than in the LMS regression since the LSR determines three model parameters: a robust slope and a robust focal point. To minimize equation (23), the author uses the simplex algorithm, but no expressions for the slope or the intercept are given. Furthermore, the author states that tools to investigate statistical properties of the estimates are not available in the framework of LSR. To avoid numerical difficulties and statistical shortcomings, an approximate estimator, called the absolute median structural regression (ASR), is proposed. In this different approach, the slope and the intercept are given by

$$b = \pm\,\mathrm{med}_i\left[\frac{|y_i - \mathrm{med}_k(y_k)|}{|x_i - \mathrm{med}_k(x_k)|}\right], \qquad a = med_i(y_i - bx_i) \tag{24}$$

and the sign of b corresponds to the sign of

$$\tilde{b} = \mathrm{med}_i\left[\frac{y_i - \mathrm{med}_k(y_k)}{x_i - \mathrm{med}_k(x_k)}\right] \tag{25}$$

The absolute median structural regression yields an appropriate robust fit with no minimization procedure, and it allows confidence intervals to be determined by applying the cumulative binomial distribution.[88] In addition to the development of finding the coefficients of the regression line, the author gives a procedure for the detection of outliers.

Bartlett's three groups method

This very simple approach[89] was one of the first to deal with the subject of linear calibration with errors in both axes. In this method, which needs no information about the error in the variables, the data is ranked according to the size of x and then divided into three groups where at least the first and the last should be of equal sizes. Then the means of the first and third group are calculated $((\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_3, \bar{y}_3))$ and the coefficients of the regression line are obtained as

$$b = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1}, \quad a = \bar{y} - b\bar{x} \tag{26}$$

where $\bar{x}, \bar{y}$ are the mean values of the total data. This method is important historically, but leads to erroneous results both in the slope and in the intercept.

Principal Components Analysis (PCA)

The main use of the technique of PCA is not that of finding a regression model, but of acting as a basic technique for studying the structure of sets of multivariate data.[6]

In its function as a straight line calibration technique, only the calculation of the first principal component is relevant. As is well known, this first linear

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

combination of the original variables is positioned towards the maximum dispersion of points and best fits the original bivariate data. This dispersion can be expressed as the sum of the squares of the distance from each data point to the centroid. This sum can be separated into two factors: the sum of the distances between the coordinate centre and the projections of each point on the principal component (scores), which represents the variance throughout the component and which ideally should be as high as possible (M in Figure 2); and the sum of the distances of each point from its projection on the principal component, which represents the variance around the component and which ideally should be minimal (m in Figure 2). In the PCA technique each point is projected perpendicularly onto the principal component, so taking into account uncertainty in both variables, since this projection contemplates the uncertainties of the dependent and the independent variable. In this way, the defined ratio $s_{y_i}^2 / s_{x_i}^2 = 1$.

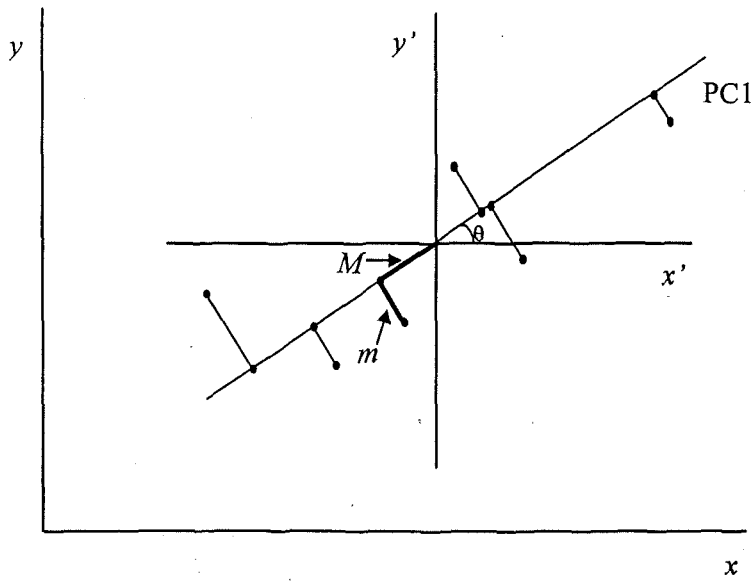


Figure 2. PCA applied to straight line calibration.

The two principal components which can be calculated from bivariate data are nothing more than a translation followed by a rotation of the original coordinate axes. Therefore, if the angle θ in which the coordinate axes have rotated is known, the slope of the regression straight line can be easily calculated, since $b = \tan\theta$. The expressions to find the value of 2 have been reported:[5,6]

$$\theta = \begin{cases} \frac{1}{2}\tan^{-1}\left[\frac{2\,\mathrm{cov}(x,y)}{s_x^2 - s_y^2}\right] & \text{if } s_x^2 > s_y^2 \\ 90° + \frac{1}{2}\tan^{-1}\left[\frac{2\,\mathrm{cov}(x,y)}{s_x^2 - s_y^2}\right] & \text{if } s_x^2 < s_y^2 \end{cases} \tag{27}$$

where

$$s_x^2 = \left[(x_i - \bar{x})^2/(n-1)\right]^{1/2} \tag{28}$$

$$s_y^2 = \left[(y_i - \bar{y})^2/(n-1)\right]^{1/2} \tag{29}$$

The intercept can be found if the slope and any point on the line are known, according to $y_i = a + bx_i$. One point which can be used is the one corresponding to the coordinate centroid. As the data are mean centered, to obtain results in relation to the original variables, they have to be recuperated by adding to each coordinate centroid the mean of its original column of data. When centering the data the mean of the variables was zero, so the point that we will obtain will simply be the mean value of the original dependent and independent variables. It should be pointed out that PCA is scale invariant, and looking at equation (27) it can be seen that PCA is symmetrical upon the axes, since switching variables results in complementary angles for the slope.

The method does have a drawback, since the individual uncertainties of each point are not taken into account. It is a sort of non-weighted regression. What is more,

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

the fact of minimising the sum of the perpendicular distances from the calibration straight line presupposes that the errors are of the same size in the dependent variable as in the independent variable, which is not always the case. The PCA slope is also the major axis (also known as ordinary major axis, orthogonal regression or model II regression) and can be considered as a particular case of the linear bivariate structure relationship (BSR).[77,90] The uncertainty associated to the slope and the intercept of the first principal component can be found in literature.[90-92] Another particular case of the bivariate structure relationship could be for example the standardized major axis (SMA) or the constant variance ratio approach, for which $s_{y_i}^2 / s_{x_i}^2$ [51,77,90] do not need to be unity. This set of methods needs no iterations to find the regression coefficients of the calibration line.

COMPARATIVE STUDY OF THE VARIOUS CALCULATION METHODS

Software

Home-made computer programs containing the algorithms from Williamson, Lisý *et al.*, González *et al.*, Neri *et al.*, Brooks *et al.*, Reed and Cumming *et al.* have been developed in Matlab for Microsoft Windows v. 4.0 (The MathWorks, Inc.). With these programs, which can be used in PCs, results which are accurate up to twelve decimal places can be obtained with only 3-9 iterations in the methods of Williamson, Lisý *et al.*, Brooks *et al.* and Reed and with 10-20 iterations for the methods of González *et al.*, Neri *et al.* and Cumming et al. for several sets of data. With processors of the type 486 DX2/66, calculation time is less than a second for most of the methods and data sets tested. The programs are available for those who should like to order them.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Results and Discussion

The regression coefficients when fitting a calibration line for the data set given by Pearson[93] in 1901 and weights given by York's paper[50] (Table 1), have been obtained by several methods: least squares, weighted least squares, York, Reed, Williamson, González *et al.*, Neri *et al.*, effective variance, Lisý *et al.*, Brook *et al.*, Lybanon, Cumming *et al.*, PCA and Bartlett. The results obtained by these methods are compared with the published exact generalized least-squares results.[68,94] Table 2 shows these results. The goodness of the fit is given as the value S, the weighted sum of the residuals described in equation (3). It should be pointed out, however, that several methods use different weighting factors, w_i, when calculating the value of S.

As can be seen, the weighted least squares method gives results that are erroneous for the slope by 27% for this data set. By applying PCA to the linear regression we find results that in this particular case improve on the least squares method but which are still a long way from the correct values. The effective variance method improves on these results, but also has an associated error. However, the methods of Williamson, Lisý *et al.*, Cumming *et al.*, Reed, Brooks *et al.*, González *et al.* and Neri *et al.* get the exact result. Lybanon states that the correct result can also be obtained by using Jeffery's polynomial, although in this case the last decimals may be different because of the accuracy of calculation. The characteristics of the most significant methods reviewed in this paper are summarised in Table 3.

A feature that is common to all the methods looked at in this paper which consider errors in both axes and which give correct results for the coefficients of regression, is that for a sufficiently correlated set of data, such as the one given by Pearson,[93] results which are appreciably different from each other are obtained for their variances of the regression coefficients.[70,95-97] This is probably due to the fact that

in the development of the expressions only the lowest-order terms in the Taylor-series expansion have been taken into account, some times even with approximations.[98] A notable exception is Williamson's method[54] which derived the variances without considering approximations. It is worth pointing out that the variance expressions for the slope and the intercept are often multiplied by $S/(n\text{-}2)$ to give a standardized variance. In fact, if the standard deviations for the slope and the intercept given by Williamson's method (s_a = 0.2919335, s_b = 0.057617) are multiplied by the value of $S/(n\text{-}2)$, with S given by Table 2, they give the approximate standard deviations for the slope and the intercept given by the Lisý *et al.* method (s_a = 0.3618719, s_b = 0.0719964).

Calculation technique	*a*	*b*	*S*
Least squares	5.76118519	-0.53957727	----------
Weighted least squares	6.10010933	-0.610812958	34.345211629
York's method	5.47991022	-0.480533407	11.8663539487
Reed's method	5.47991022	-0.480533407	11.8663539487
Williamson's method	5.47991022	-0.480533407	11.8663531941
González's *et al.* method	5.47991022	-0.480533407	11.8663539487
Neri's *et al.* method	5.47991022	-0.480553403	11.8663555842
Effective variance	5.39605212	-0.46344885	11.956449080
Lisý's *et al.* method	5.47991022	-0.480533407	11.8663531941
Brooks' *et al.* method	5.47991022	-0.480533407	11.8663539487
Lybanon	5.47991025	-0.480533415	11.8663531941
Cumming's *et al.* method	5.47991022	-0.480533407	11.8663539487
Bartlett's three groups	5.68728323	-0.52023121	----------
Principal Components	5.78404377	-0.545561197	13.8079868422
Exact solution	5.47991022	-0.480533407	11.8663531941

Table 2. Comparison of various techniques of linear calibration for the data set of Table 1.

Conclusions

Of all the linear calibration methods which consider errors in both axes, the methods of Williamson[54] and Lisý *et al.*[62] stand out. Their respective algorithms are easily programmed and they lead to accurate results. Furthermore, the Lisý *et al.* method allows sets of data with correlated errors to be worked with and gives the variance-covariance matrix. It should to be pointed out that despite taking into account the uncertainties of both the dependent and independent variables, most of the methods continue to minimize the weighted sum of the distances between each point and their vertical projection on the regression straight line and only use these uncertainties in the calculation of the weighting factor. Similarly, neither the expression for the variance of the intercept nor for the variance of the slope are taken into account in the minimization process of any expression reviewed as should be required according to the theory of error propagation.

Most of the methods reviewed lead to very similar results for the data set tested. Therefore, obtaining reliable results together with the ease of programming the corresponding algorithm are the main criteria for selecting a specific method to handle a data set containing errors in both axes. One of the main difficulties of extending the use of these methods, the lack of suitable software, has been overcome nowadays.

It should be pointed out that in some of these methods (Williamson, Lisý *et al.*, González *et al.*, Neri *et al.*, Brooks *et al.*, Cumming *et al.*, Reed and PCA) the estimate for b is invariant upon switching axes. In the case of Williamson's method, this symmetry also includes the variances of the slope and the intercept. The fact that the variances are symmetrical is particularly important, since the confidence intervals for the direct and inverse calibration coincide and it is easier to find the

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

errors associated to each of the variables in the process of interpolation that might be useful for the analyst.

Despite the effectiveness and the simplicity of some of these methods, statistical tests such as the joint confidence region for the slope and intercept, useful for the validation of methods which contain uncertainties in both axes have still not been reported.[99] Once the expressions for the slope and the intercept, dependent variable and independent variable have been calculated, as well as their respective confidence intervals, it would be interesting to apply these expressions to the calculation of important quality parameters of analytical methods such as detection or quantification limits which have not been found in the reviewed bibliography.

Acknowledgements

We would like to thank the Spanish Ministry of Education and Science (DGICyT project n1 BP90-0453) for their financial support. The authors thank A. Kalantar for his many useful comments on the manuscript.

Method	Initial equation	Equation to minimise	Weights known	Outlier resistant	Algorithm	Variances of the regression coefficients	Confidence intervals for y-predicted
York[50]	Equation (2)	Equation (5)	Yes. Heteroscedasticity	No	Iterative	Yes	No
Reed[49]	Equation (2)	Equation (5)	Yes. Heteroscedasticity	No	Iterative	Yes	No
Williamson[54]	Equation (2)	Equation (5)	Yes. Heteroscedasticity	No	Iterative	Yes	No
González *et al.*[16]	Equation (3)	Equation (5)	Yes. Heteroscedasticity	No	Iterative	No	No
Neri *et al.*[47]	Equation (12)	Equation (5)	Yes. Heteroscedasticity	No	Iterative	No	No
Press and Teukolsky[57]	Equation (5)	Equation (5)	Yes. Heteroscedasticity	No	Numerical	No	No
Lisý *et al.*[62]	$S = \frac{(y_i - a - bx_i)^2}{s_{y_i}^2 + b^2 s_{x_i}^2 - 2b\mathrm{cov}(x_i, y_i)}$	Equation (12)	Yes. Heteroscedasticity	No	Iterative	Yes	No
Brooks et al.[65]	$S = \frac{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2}{s_{x_i}^2 + s_{y_i}^2}$	Equation (14)	Yes. Heteroscedasticity	No	Iterative	Yes	No
Lybanon[68]	$S = \frac{1}{2}\hat{v}^T \sigma^{-1} \hat{v}$ (σ, covariance matrix; v, vector of residuals)	$S = \frac{1}{2}\hat{v}^T \sigma^{-1} \hat{v}$	Yes. Heteroscedasticity	No	Iterative	Yes	No
Mandel[3]	Constants of transformation of the variables into new ones	Constants of transformation of the variables	Yes. Relation between error variances constant	No	Iterative	Yes	Yes
Cumming[76]	Equation (16)	Equation (16)	Yes. Heteroscedasticity	No	Iterative	Yes	No
Fuzzy[79-82,84]	Equation (22)	Equation (22)	Estimation	Yes	Several	Yes	Yes
LSR[34]	Equation (23)	Equation (23)	No	Yes	Simplex method	No	No
PCA[6]	Perpendicular from each point to the calibration line	Perpendicular to the calibration line	No	No	Numerical	Yes	Yes
SMA[51,77,90]	Perpendicular from each point to the calibration line	Perpendicular to the calibration line	Yes. Relation between variances constant	No	Numerical	Yes	Yes

Table 3. Comparison of the different calibration methods.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Journal of Chemometrics 9 (1995) 343-362

REFERENCES

1. N. Draper and H. Smith, *Applied Regression Analysis*, 2nd ed., John Wiley & Sons, New York (1981).
2. A.H. Kalantar, *Trends Anal. Chem.* **9**, 149 (1990).
3. J. Mandel, *J. Quality Tech.* **16**, 1 (1984).
4. J. Hartmann, J. Smeyers-Verbeke and D.L. Massart, *Analusis* **21**, 125 (1993).
5. M. Sharaf, D. Illman and B.R. Kowalski, *Chemometrics*, John Wiley & Sons, New York (1986).
6. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a Textbook*, Elsevier, Amsterdam (1988).
7. R. De Levie, *J. Chem. Educ.* **63**, 10 (1986).
8. F.C. Garner and G.L. Robertson, *Chemom. Intell. Lab. Syst.* **3**, 53 (1988).
9. M.H. Feinberg, *J. Chemom.* **3**, 103 (1988).
10. J. Irvin and T. Quickenden, *J. Chem. Educ.* **60**, 711 (1983).
11. A.H. Kalantar, *J. Chem. Educ.* **64**, 28 (1987).
12. L. Meites, H.C. Smit, and G. Kateman, *Anal. Chim. Acta* **164**, 287 (1984).
13. C.H. Spiegelman, R.L. Watters and L. Hungwu, *Chemom. Intell. Lab. Syst.* **11**, 121 (1991).
14. R.R. Sokal and F.J. Rohlf, *Biometry*, W.H. Freeman and Co., San Francisco (1969).
15. J. Berkson, *J. Am. Statis. Assoc.* **45**, 164 (1950).
16. A. Gustavo González, A. Márquez and J. Fernández Sanz, *Computers Chem.* **16**, 25 (1992).
17. A. Wald, *Ann. Math. Statist.* **11**, 284 (1940).
18. R.D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman and Hall, London (1982).
19. D.A. Belsley, E. Kuh and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York (1980).

20. M. Meloun, J. Militky and M. Forina, *Chemometrics for Analytical Chemistry* Vol. II, Ellis Horwood, London (1994).
21. P.J. Huber, *Robust Statistics*, John Wiley & Sons, New York (1981).
22. K. Linnet, *Statistics in Medicine* **9**, 1463 (1990).
23. D.M. Allen, *Technometrics* **16**, 125 (1974).
24. P. Prescott, *Technometrics* **17**, 129 (1975).
25. D.F. Andrews and D. Pregibon, *J. Roy. Statist. Soc., Ser. B.* **40**, 85 (1978).
26. D.C. Hoaglin and R. Welsch, *Amer. Statistician* **32**, 17 (1978).
27. R.D. Cook and S. Weisberg, *Technometrics* **22**, 495 (1980).
28. Y. Hu, J. Smeyers-Verbeke and D.L. Massart, *Chemom. Intell. Lab. Syst.* **9**, 31 (1990).
29. M.O. Moen, K.F. Griffin and A.H. Kalantar, *Anal. Chim. Acta* **277**, 477 (1993).
30. D.L. Massart, L. Kaufman, P.J. Rousseeuw and A. Leroy, *Anal. Chim. Acta* **187**, 171 (1986).
31. A. Leroy and P.J. Rousseeuw, *PROGRESS: A Program for Robust Regression Analysis*, Technical Report 201, Center for Statistics and O.R., University of Brussels, Belgium (1984).
32. P.J. Rousseeuw, *J. Am. Statist. Assoc.* **79**, 871 (1984).
33. P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York (1987).
34. U. Feldmann, *Eur. J. Clin. Chem. Clin. Biochem.* **30**, 405 (1992).
35. J.D. Emerson and D.C. Hoagland in D.C. Hoagland, F. Mosteller and J.W. Tukey (Editors), *Understanding Robust and Exploratory Data Analysis*, Wiley, New York (1983).
36. F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stakel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York (1986).
37. C.J. Huang and B.W. Bloch, *J. Amer. Statist. Assoc.* **69**, 330 (1974).

38. R.L. Brown, J. Durbin and J.M. Evans, *J. Roy. Statist. Soc., Ser. B.* **37**, 149 (1975).
39. G.D.A. Phillips and A.C. Harvey, *J. Amer. Statist. Assoc.* **69**, 935 (1974).
40. A. Hedayat, B. Ratkoe and P. Telwar, *Communications in Statistics* **A6**, 497 (1977).
41. A. Hedayat and D.S. Robson, *J. Amer. Statist. Assoc.* **65**, 1573 (1970).
42. A.C. Harvey and G.D.A. Phillips, *Econometrics* **2**, 307 (1974).
43. S.D. Christian, E.H. Lane and F. Garland, *J. Chem. Educ.* **51**, 475 (1974).
44. J.M. Lisý, A. Cholvadová and B. Drobná, *Computers Chem.* **15**, 135 (1991).
45. W.H. Jefferys, M.J. Fitzpatrick and B.E. McArthur, *Cel. Mech.* **41**, 39 (1988).
46. F. Neri, S. PatanP and G. Saitta, *Meas. Sci. Technol.* **1**, 1007 (1990).
47. F. Neri, G. Saitta and S. Chiofalo, *J. Phys. E. Sci. Instrum.* **22**, 215 (1989).
48. B.D. Ripley and M. Thompson, *Analyst* **112**, 377 (1987).
49. B.C. Reed, *Am. J. Phys.* **60**, 59 (1992).
50. D. York, *Can. J. Phys.* **44**, 1079 (1966).
51. D.L. MacTaggart and S.O. Farwell, *J. of AOAC Intnl.* **75**, 608 (1992).
52. B.C. Reed, *Am. J. Phys.* **57**, 642 (1989).
53. W.E. Deming, *Statistical Adjustment of Data*, Dover, New York (1964).
54. J.A. Williamson, *Can. J. Phys.* **46**, 1845 (1968).
55. A.G. Asuero and A.G. González, *Microchem. J.* **40**, 216 (1989).
56. P.J. Ogren, J.R. Norton, *J. Chem. Educ.* **69**, A130 (1992).
57. W.H. Press and S.A. Teukolsky, *Comput. in Phys.* **6**, 274 (1992).
58. M. Clutton-Brock, *Technometrics* **9**, 261 (1967).
59. D.R. Barker and L.M. Diana, *Am. J. Phys.* **42**, 224 (1974).
60. J. Orear, *Am. J. Phys.* **50**, 912 (1982). *Erratum* **52**, 278 (1984).
61. M. Lybanon, *Am. J. Phys.* **52**, 276 (1984).
62. J.M. Lisý, A. Cholvadová and J. Kutej, *Computers Chem.* **14**, 189 (1990).
63. W.E. Wentworth, *J. Chem. Educ.* **42**, 96 (1965).
64. W.E. Wentworth, *J. Chem. Educ.* **42**, 162 (1965).
65. C. Brooks, I. Went and W. Harre, *J. Geophys. Res.* **73**, 6071 (1968).

66. R.L. Watters, R.J. Carroll and C.H. Spiegelman, *Anal. Chem.* **59**, 1639 (1987).
67. T. Lwin and C.H. Spiegelman, *J. Royal Statist. Soc. Series C* **35**, 256 (1986).
68. M. Lybanon, *Am. J. Phys.* **52**, 22 (1984).
69. W.H. Jefferys, *Astron. J.* **85**, 177 (1980).
70. W.H. Jefferys, *Astron. J.* **86**, 149 (1981).
71. M. Lybanon, *Comp. & Geosc.* **11**, 501 (1985).
72. H.I. Britt and R.H. Luecke, *Technometrics* **15**, 233 (1973).
73. D.R. Powell and J.R. Macdonald, *Comput. J.* **15**, 148 (1972); *Ibid.* **16**, 51 (1973).
74. J.R. Macdonald, *Am. J. Phys.* **43**, 372 (1975).
75. J.R. Macdonald and W.J. Thompson, *Am. J. Phys.* **60**, 66 (1992).
76. G.L. Cumming, J.S. Rollett, F.J.C. Rossotti and R.J. Whewell, *J. Chem. Soc. Dalton Trans.* **23**, 2652 (1972).
77. W.E. Ricker, *Can. J. Zool.* **62**, 1897 (1984).
78. T.A. Jones, *Math. Geol.* **11**, 1 (1979).
79. M. Otto and H. Bandemer, *Chemom. Intell. Lab. Syst.* **1**, 71 (1986).
80. H. Bandemer, *Fuzzy Sets and Systems* **16**, 41 (1985).
81. H. Bandemer and M. Otto, *Mikrochim. Acta* **II**, 93 (1986).
82. L.A. Zadeh, *Information and Control* **8**, 338 (1965).
83. M. Otto, *Anal. Chem.* **14**, 797A (1990).
84. Y. Hu, J. Smeyers-Verbeke and D.L. Massart, *Chemom. Intell. Lab. Syst.* **8**, 1485 (1990).
85. J.A. Nelder and R. Mead, *Comput. J.* **7**, 308 (1965).
86. W. Spendley, G.R. Hext and F.R. Himsworth, *Technometrics* **4**, 441 (1962).
87. G.R. Philips and E.M. Eyring, *Anal. Chem.* **55**, 1134 (1983).
88. I.M. Jonhstone and P.F. Velleman, *J. Am. Stat. Assoc.* **80**, 1041 (1985).
89. M.S. Bartlett, *Biometrics* **5**, 207 (1949).
90. P. Jolicoeur, *J. Theor. Biol.* **144**, 275 (1990).
91. N.M. Faber, L.M.C. Buydens and G. Kateman, *J. Chemometrics* 7, 495 (1993).
92. P. Jolicoeur and G. Ducharme, *J. Theor. Biol.* **154**, 35 (1992).

93. K. Pearson, *Philos. Mag.* **2**, 559 (1901).

94. A. Celmins, *Ballistic Research Laboratories Report No. 1658*, Aberdeen, Proving Ground, Maryland (1973).

95. G.C. Cecchi, *Meas. Sci. Techno.* **2**, 1127 (1991). *Ibid* **4**, 906 (1993).

96. C. Moreno and H. Bruzzone, *Meas. Sci. Technol.* **4**, 635 (1993).

97. A. Kalantar, *Meas. Sci. Technol.* **3**, 1113 (1992).

98. D.P. Chong, *Am. J. Phys.* **59**, 472 (1991).

99. J. Riu and F.X. Rius (In preparation).

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Capítol 3

Introducció de l'error de segona espècie en els tests individuals per a l'ordenada en l'origen i el pendent en regressió lineal univariant considerant errors en dos eixos

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Un cop examinades les limitacions dels mètodes tradicionals de regressió dins del camp de la química analítica, i després de revisar els mètodes de regressió univariant que consideren errors en els dos eixos (com ja s'ha comentat en el capítol anterior, a partir d'ara ens centrarem en el mètode de regressió que considera errors en dos eixos de Lisý i col·laboradors, que l'anomenarem mètode BLS, *bivariate least squares*), entrem directament en el desenvolupament de tests estadístics aplicables en processos de comparació de mètodes analítics.

Una de les característiques en els mètodes de regressió que tenen en compte els errors en dos eixos és el fet que les distribucions de l'ordenada a l'origen i el pendent de la recta de regressió no són normals per a la majoria de conjunts de dades. Aquest aspecte és important, ja que del tipus de distribució seguida se'n deriven posteriorment el tipus de tests estadístics a aplicar. És important intentar quantificar la desviació de la normalitat de les distribucions de l'ordenada a l'origen i el pendent trobats amb el mètode BLS, ja que si aquesta desviació no fos important, probablement es podrien aplicar tests estadístics basats en la hipòtesi de la normalitat de la distribució dels coeficients. Dit d'una altra manera, és important comprovar en quin dels dos casos es comet més error: ignorant els errors existents en la variable predictora (la qual cosa implica la utilització dels mètodes de regressió OLS o WLS, en què els coeficients de regressió trobats amb la seva aplicació sí que segueixen la distribució normal) o suposar com a vàlida la hipòtesi de la normalitat en les distribucions dels coeficients de regressió trobats mitjançant el mètode BLS.

Un cop comprovada la validesa de considerar la hipòtesi de la normalitat en les distribucions dels coeficients de regressió trobats mitjançant el mètode BLS, s'entra directament en la descripció dels intervals de confiança individuals dels coeficients de la recta de regressió considerant errors en els dos eixos. Aquests intervals són útils, per exemple, per comprovar si un mètode analític presenta errors sistemàtics proporcionals o constants, o en processos de calibració per comprovar si són

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

necessaris processos de correcció del blanc o si hi ha efectes de matriu. En aquests tests estadístics s'introdueixen les expressions per calcular les probabilitats d'error β associades tenint en compte la diferència màxima permesa (biaix) fixada per l'analista i per calcular el nombre de punts necessari per construir la recta de regressió, fixades unes determinades probabilitats d'error α i β.

El gruix de la revisió crítica es troba al final del capítol, en l'article titulat *Detecting Proportional and Constant Bias in Method Comparison Studies by Using Linear Regression With Errors in Both Axes*, que ha estat enviat per a la seva publicació a la revista *Chemometrics and Intelligent Laboratory Systems*.

3.1 Comprovació de la normalitat en les distribucions de l'ordenada a l'origen i el pendent en regressió lineal considerant errors en dos eixos

Tal com s'ha comentat en la secció 1.5.3, en les tècniques de regressió lineal que consideren errors en els dos eixos no es compleix el supòsit de la normalitat en les distribucions associades als coeficients de la recta de regressió. La comprovació d'aquest fet és molt important, ja que si no es verifica la normalitat en les distribucions dels coeficients de la recta de regressió, rigorosament no es podrien aplicar els tests estadístics usuals basats en paràmetres estadístics com *t* de Student o *F* de Fischer.

La comprovació de la normalitat o de la no-normalitat en la distribució de l'ordenada a l'origen i el pendent de la recta de regressió considerant errors en els dos eixos, es durà a terme de tres formes: el mètode de Cetama, el test de Kolmogorov i les gràfiques de probabilitat normal (*normal probability plots*). En cas que la distribució resultant no sigui normal, el mètode de Cetama, a més, permet obtenir el tipus de la distribució i l'expressió. En tots tres casos el procediment serà el mateix i es troba esquematitzat a la figura 3.1: a partir d'un

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

conjunt de dades inicial que presenta errors en els dos eixos (al gràfic de l'esquerra de la figura 3.1 les línies verticals i horitzontals representen les incerteses individuals de cada punt experimental, de les quals podem tenir una estimació mitjançant les desviacions estàndards de cada parell de punts experimental x_i,y_i), el mètode de Monte Carlo[1] ens proporciona k nous conjunts de dades (on k en aquest cas pot arribar fins a 200.000) resultants d'afegir a cada punt experimental del conjunt de dades inicial un error aleatori basat en les pròpies incerteses individuals. De cada un dels k nous conjunts generats mitjançant el procés de simulació de Monte Carlo es pot trobar la seva recta de regressió; es tindran, per tant, k ordenades a l'origen i k pendents amb els quals se'n podrà estudiar la normalitat per a cada un dels procediments estadístics abans esmentats, que es comentaran a continuació.

3.1.1 Mètode de Cetama

El mètode de comprovació de la normalitat de Cetama[2] permetrà obtenir el tipus de distribució de l'ordenada a l'origen i el pendent d'un conjunt de dades que presenti errors en els dos eixos. El fet d'indicar no només si la distribució seguida és normal o no, sinó de proporcionar també l'expressió de la distribució, servirà per quantificar el tipus d'error comès en acceptar la hipòtesi de la normalitat en la distribució dels coeficients de la recta de regressió.

Aquest mètode es basa en la utilització dels moments centrats de segon, tercer i quart ordre. Els anomenats moments centrats d'ordre p (η_p) són els valors probables de $(X-\overline{X})^p$, on $\overline{X}$ és la mitjana de la població de la variable X:

$$\eta_p = E(X-\overline{X})^p \qquad (3.1)$$

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

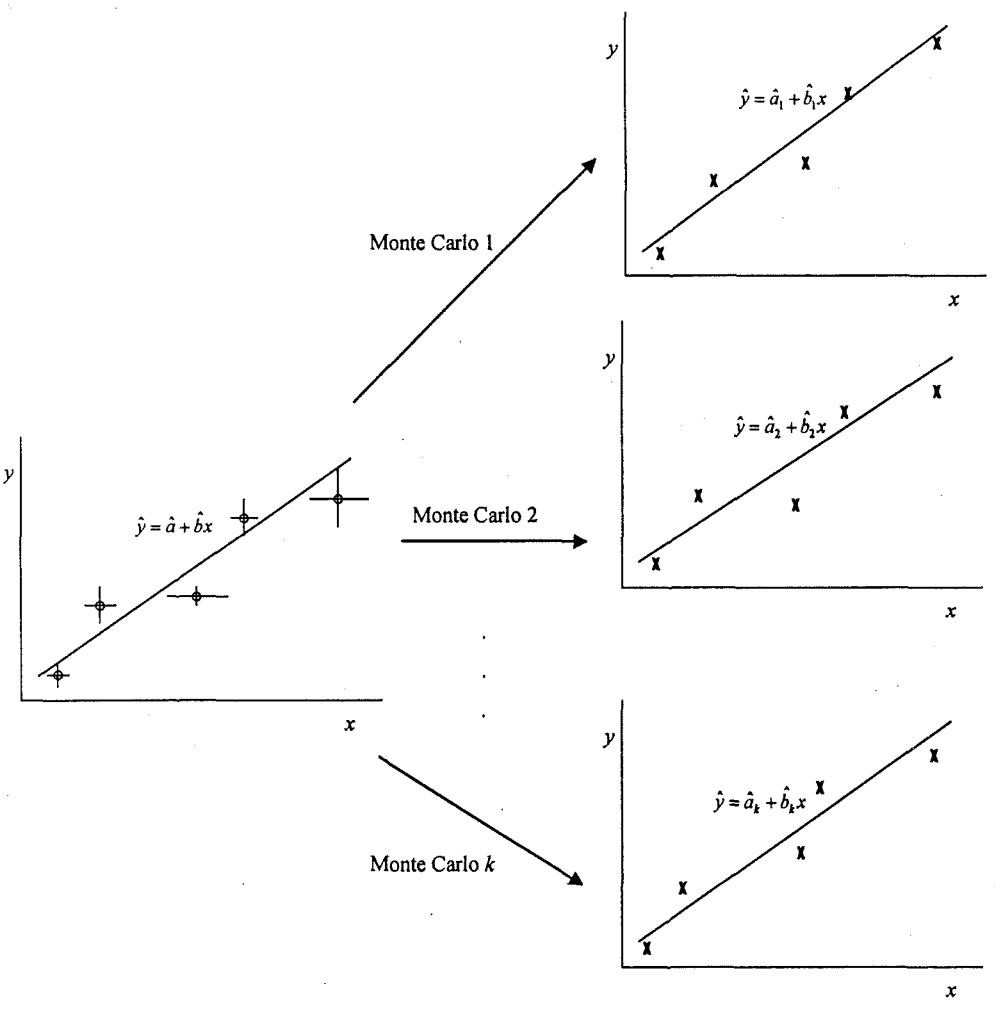


Figura 3.1. Procés d'obtenció de les k ordenades a l'origen i k pendents per estudiar la normalitat en les seves distribucions.

El moment d'ordre 1 correspon al valor mitjà, el moment d'ordre 2, a la variància de la població (σ^2); el moment d'ordre 3, al coeficient d'asimetria (*skewness*, η_3/σ^3), i el moment d'ordre 4, a la curtosi (*kurtosis*, η_4/σ^4).[3] La variància de la població és una mesura de la dispersió de la distribució, el coeficient d'asimetria caracteritza el grau d'asimetria d'una distribució respecte a la seva mitjana (l'asimetria positiva indica una distribució unilateral que s'estén cap a valors més positius, mentre l'asimetria negativa indica una distribució unilateral que s'estén

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

cap a valors més negatius), i la curtosi representa el grau d'apuntament d'una distribució estadística, és a dir, el grau de convexitat o aplatament de la corba representativa d'una distribució estadística (una curtosi positiva indica una distribució relativament elevada, mentre que una curtosi negativa indica una distribució relativament plana). Els diferents tipus de distribucions poden classificar-se mitjançant els coeficients següents, basats en els moments de diferent ordre:

$$\gamma_1 = \frac{\eta_3}{\eta_2^{3/2}} = \frac{\eta_3}{\sigma^3} \tag{3.2}$$

$$\gamma_2 = \frac{\eta_4}{\eta_2^2} - 3 \tag{3.3}$$

on els coeficients γ_1 i γ_2 són els anomenats coeficients de Fisher; el primer és el coeficient d'asimetria i el segon, el coeficient d'aixafament. Aquests coeficients de Fisher poden ser reemplaçats pels coeficients de Pearson:

$$\beta_1 = \frac{\eta_3^2}{\eta_2^3} = \gamma_1^2 \tag{3.4}$$

$$\beta_2 = \frac{\eta_4}{\eta_2^2} = \gamma_2 + 3 \tag{3.5}$$

on β_1 varia entre 0 i ∞, i $\beta_2 > \beta_1 + 1$. Només podem tenir estimacions tant dels coeficients de Fisher com dels de Pearson. Els coeficients de Fisher són calculats segons:

$$g_1 = \frac{k_3}{k_2^{3/2}} \tag{3.6}$$

$$g_2 = \frac{k_4}{k_2^2} \tag{3.7}$$

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

i els coeficients de Pearson segons:

$$b_1 = g_1^2 = \frac{k_3^2}{k_2^3} \tag{3.8}$$

$$b_2 = g_2 + 3 \tag{3.9}$$

Els coeficients k_2, k_3 i k_4 es troben definits a l'apèndix.

Una corba de distribució simètrica ve caracteritzada per uns valors nuls del primer coeficient de Fisher i del primer coeficient de Pearson. En aquest cas, el signe del segon coeficient de Fisher indica el tipus de distribució. Les possibilitats es veuen reflectides a la taula 3.1.

$\gamma_1 = \beta_1 = 0$ corba simètrica	$\gamma_2 = 0$ $\beta_2 = 3$	Corba normal
	$\gamma_2 < 0$ $\beta_2 < 3$	Corba limitada pel domini $\bar{x} \pm d_1$
	$\gamma_2 > 0$ $\beta_2 > 3$	Corba il·limitada $-\infty,+\infty$

Taula 3.1. Diferents distribucions quan els primers coeficients de Fisher i Pearson són nuls.

Com que dels coeficients de Fisher i Pearson exposats a la taula 3.1 només en podem conèixer les estimacions, en la comprovació de la simetria de la corba, la qual es du a terme amb la comprovació de la nul·litat de g_1, cal també tenir en compte la seva variància:

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

$$s_{g_1}^2 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \tag{3.10}$$

Si el valor $u_1 = g_1 / s_{g_1}$ es troba comprès dins de l'interval de confiança definit pels valors $-u_{\alpha/2}$ i $u_{\alpha/2}$, es dirà que el valor g_1 no difereix significativament de zero pel nivell de significança α escollit, i que per tant la corba és simètrica.

Una vegada la distribució es pot considerar simètrica, serà considerada normal si el segon paràmetre de Fisher és nul. Com que del segon paràmetre de Fisher només en tenim l'estimació (g_2), també s'ha de tenir en compte la variància:

$$s_{g_2}^2 = \frac{24n(n-1)^2}{(n-2)(n-3)(n+3)(n+5)} \tag{3.11}$$

Si el valor $u_2 = g_2 / s_{g_2}$ es troba comprès dins de l'interval de confiança definit pels valors $-u_{\alpha/2}$ i $u_{\alpha/2}$, es dirà que el valor g_2 no difereix significativament de zero pel nivell de significança α escollit, i que per tant la distribució de la variable és normal.

Si el segon coeficient de Fisher no pot ser considerat nul i la seva estimació (g_2), és negativa, la distribució segueix una corba simètrica limitada pel domini $(\bar{x} - d_1, \bar{x} + d_1)$, amb el coeficient d_1 definit a l'apèndix. Aquest tipus de distribucions estan definides per la funció següent:

$$f(x) = f_0 \left[1 - \frac{(x-\bar{x})^2}{d_1^2} \right]^{m_1} \tag{3.12}$$

on els paràmetres de l'equació 3.12 es troben a l'apèndix.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Si el segon coeficient de Fisher no pot ser considerat nul i la seva estimació (g_2), és positiva, la distribució és il·limitada pels dos costats. En aquests casos, la distribució està definida per la funció següent:

$$f(x) = \frac{f_0}{\left[1 + \frac{(x-\bar{x})^2}{m_2^2}\right]^{m_3^2}} \tag{3.13}$$

on els diferents paràmetres de l'equació 3.13 es troben a l'apèndix.

Si, pel contrari, s'arriba a la conclusió que la corba no és simètrica (g_1 és estadísticament diferent de zero), la forma de la distribució ve donada en funció del paràmetre següent:

$$k = \frac{b_1(b_2+3)^2}{4(4b_2-3b_1)(2b_2-3b_1-6)} \tag{3.14}$$

El valor del paràmetre k només pot ser nul per a una distribució simètrica ($b_1 = 0$), ja que el coeficient b_2 és positiu per definició. Segons els valors que agafi el paràmetre k, es poden tenir els tipus de distribucions següents:

$k<0$	Corba limitada pels dos costats	$d_2 \ldots d_3$
$0<k<1$	Corba il·limitada pels dos costats	$-\infty \ldots +\infty$
$1 \leq k$	Corba limitada per un costat	$g_1>0$ $d_4 \ldots +\infty$ $g_1<0$ $-\infty \ldots d_5$

Taula 3.2. Diferents possibilitats quan la corba presenta una distribució no simètrica.

Si el valor del paràmetre k és negatiu, la corba és limitada als dos costats pels paràmetres d_2 i d_3, els quals es troben definits a l'apèndix. Aquest tipus de corbes vénen definides per la funció següent:

$$f(x) = f_0 (x - d_2)^{q_1 - 1} \cdot (d_3 - x)^{q_2 - 1} \tag{3.15}$$

on els diferents paràmetres de l'equació 3.15 es troben a l'apèndix.

Si el valor del paràmetre k expressat en l'equació 3.14 està comprès entre 0 i 1, la distribució és il·limitada pels dos costats (taula 3.2). En aquest cas, aquestes corbes estan definides per la funció següent:

$$f(x) = f_0 \left(\cos\left(\frac{x - c_1}{v} \right) \right)^{2q} e^{p \cdot \frac{x - c_1}{v}} \tag{3.16}$$

on els paràmetres de l'equació 3.16 es troben a l'apèndix.

Si el valor del paràmetre k és igual o superior a 1, la distribució es troba limitada per un costat (taula 3.2). Segons el valor que adopti el paràmetre g_1 (equació 3.6), la corba estarà limitada a l'esquerra o a la dreta. Si g_1 pren un valor positiu, la corba es trobarà limitada per l'esquerra pel paràmetre d_4 definit a l'apèndix i la distribució segueix la funció següent:

$$f(x) = f_0 \frac{(x - d_4)^{q_3 - 1}}{(x - d_4 + c_2)^{q_4 - 1}} \tag{3.17}$$

Si juntament amb el fet que el valor del paràmetre k expressat en l'equació 3.14 sigui superior o igual a 1, el paràmetre g_1 és negatiu, la corba serà limitada per la

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

dreta pel paràmetre d_5 definit a l'apèndix. En aquest cas la corba segueix la distribució següent:

$$f(x) = f_0 \frac{(d_5 - x)^{q_3 - 1}}{(d_5 - x + c_2)^{q_4 - 1}} \tag{3.18}$$

3.1.2 Test de Kolmogorov

El test de Kolmogorov té dues variants, la gràfica i la numèrica.[3] Ens hem decantat per la variant numèrica per la seva rapidesa i simplicitat i perquè les gràfiques de probabilitat normal, també emprades per comprovar la normalitat de la distribució d'una variable, ja constitueixen un mètode gràfic. El test de Kolmogorov numèric consisteix a ordenar les n dades experimentals de la variable de la qual es vol comprovar la normalitat en ordre ascendent, i calcular per a cada una l'expressió següent:

$$D_i = S(x_i) - i/n \tag{3.19}$$

$S(x_i)$ correspon al valor de la distribució normal acumulativa per a cada valor experimental i. Se selecciona el valor màxim entre tots els n valors D_i (anomenat D_m), i aquest es compara amb els valors tabulats que depenen del nombre de dades i del nivell de significança α escollit. Si el valor D_m és superior al valor tabulat, es pot concloure que per a aquell nivell de significança la distribució pot ser considerada no normal.

3.1.3 Gràfiques de probabilitat normal

Les gràfiques de probabilitat normal, també anomenades comunament test de Rankit, són una de les anomenades eines gràfiques per comprovar la normalitat de la distribució d'una variable. Aquesta eina consisteix en la representació gràfica de

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

les dades experimentals respecte als valors de la distribució normal acumulativa. Aquesta representació permet classificar la distribució d'una variable segons el grau d'asimetria, curtosi i llargària de les cues de la distribució. Una gràfica còncava o convexa indica una distribució asimètrica. Una forma sigmoïdal indica que la llargària de les cues de la distribució difereix de la llargària de la distribució normal. Les gràfiques de probabilitat normal són un cas particular de les gràfiques Q-Q,[4] les quals serveixen per comparar la distribució de les dades experimentals amb una distribució teòrica. A la figura 3.2 s'hi troba representada la gràfica de probabilitat normal per a una sèrie simulada de 50 dades aleatòries que provenen d'una distribució normal amb mitjana 10 i desviació estàndard 1. Les dades simulades i la gràfica de probabilitat normal han estat obtingudes mitjançant la utilització de Matlab 4.0 per a Windows 3.1 o superior.[5] Cal indicar que amb aquesta eina la solució és visual, per la qual cosa en determinades ocasions pot ser difícil arribar a conclusions definitives amb la seva aplicació. A la figura es pot observar com els punts es distribueixen al voltant de la línia contínua que indica la distribució normal teòrica, i que amb només 50 punts hi ha dificultats per decidir si les dades experimentals segueixen la distribució normal o no (ja que alguns punts s'allunyen de la recta teòrica que indica el valor que haurien de tenir si la distribució fos normal). Per aquesta raó, a l'hora de comprovar la distribució dels coeficients de la recta de regressió considerant errors en els dos eixos, s'utilitzarà un nombre més alt de dades experimentals (entre 10.000 i 200.000), no només amb els gràfics de probabilitat normal, sinó amb tots els tests estadístics emprats per tal de comprovar la distribució real dels coeficients de la recta de regressió.

Com es podrà comprovar en la secció *Results and Discussion* de l'article que es troba al final del capítol, l'aplicació a una sèrie de conjunts de dades reals dels tres mètodes exposats en aquest apartat per comprovar-ne la normalitat arriba a la conclusió que els coeficients de la recta de regressió trobats mitjançant el mètode BLS no segueixen la distribució normal. Malgrat aquesta conclusió, els resultats indiquen que per a tots els conjunts de dades estudiats la desviació de la normalitat

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

no és gaire elevada, i a més s'ha comprovat que la utilització de les tècniques de regressió OLS i WLS en conjunts de dades que presenten errors en els dos eixos dóna un error més gran que la utilització de la tècnica BLS acceptant la normalitat dels coeficients de la recta de regressió. És a dir, es comet més error negligint els errors en la variable predictora i en la variable resposta quan aquests hi són presents (utilitzant les tècniques OLS i WLS quan hi ha errors en dos eixos), que acceptant la hipòtesi de la normalitat en els coeficients de regressió trobats emprant la tècnica BLS.

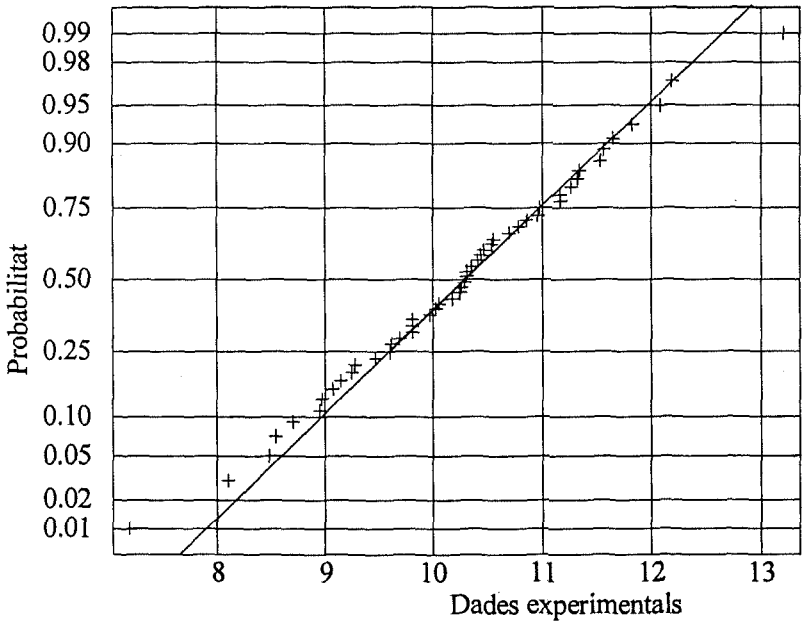


Figura 3.2. Gràfica de probabilitat normal per a un conjunt simulat de 50 dades.

3.2 Intervals de confiança individuals per a l'ordenada a l'origen i el pendent

En estudis de comparació de metodologies analítiques, a vegades és interessant comprovar l'existència d'errors sistemàtics proporcionals o errors sistemàtics constants. Per comprovar-ho al llarg d'un interval de concentracions s'analitzen una

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

sèrie de mostres a diversos nivells de concentració mitjançant els dos mètodes, i es representen els resultats obtinguts amb el mètode amb què es vol comprovar la presència o absència d'errors respecte als resultats obtinguts amb un mètode del qual se sap que no presenta aquests tipus d'errors. Com que els dos mètodes en comparació solen portar associats errors del mateix ordre de magnitud, la recta de regressió s'hauria de trobar emprant tècniques que tinguessin en compte els errors en els dos eixos. Si el mètode a comprovar no presenta errors sistemàtics constants, l'ordenada a l'origen de la recta de regressió no ha de ser significativament diferent de zero. Això es comprova mitjançant l'interval de confiança individual de l'ordenada a l'origen de la recta de regressió. Si el nou mètode no presenta errors sistemàtics proporcionals, el pendent de la recta de regressió no ha de ser significativament diferent d'1. Això es pot comprovar mitjançant l'interval de confiança individual del pendent de la recta de regressió. Altres aplicacions dels intervals de confiança individuals dels coeficients de la recta de regressió poden ser la constatació de si cal efectuar correccions del blanc (comprovant si l'ordenada a l'origen de la recta de regressió difereix significativament d'un determinat valor establert), o l'aplicació en processos de recuperació (amb l'interval de confiança individual del pendent).

Cal tenir molt present, però, que en cas que es vulgui comprovar si els resultats dels dos mètodes en comparació no difereixen estadísticament entre si -no només verificar que no hi hagi errors sistemàtics proporcionals o constants-, s'ha de comprovar que l'ordenada a l'origen de la recta de regressió no difereixi estadísticament de zero i que simultàniament el pendent de la recta de regressió no difereixi estadísticament d'1. Aquesta operació es pot dur a terme mitjançant el test conjunt per a l'ordenada a l'origen i el pendent que es troba desenvolupat al següent capítol.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

L'expressió de l'interval de confiança individual de l'ordenada a l'origen, suposant que es pugui acceptar com a vàlida la hipòtesi de la normalitat en la seva distribució, correspon a:

$$\hat{a} \pm t_{\alpha/2,n-2} \cdot \hat{s}_a \tag{3.20}$$

i l'interval de confiança individual del pendent és:

$$\hat{b} \pm t_{\alpha/2,n-2} \cdot \hat{s}_b \tag{3.21}$$

Les expressions de les desviacions estàndard de l'ordenada a l'origen i el pendent trobades mitjançant la tècnica de regressió BLS (capítol 2) corresponen a:

$$\hat{s}_a = \sqrt{\frac{\sum_{i=1}^{n} \frac{x_i^2}{w_i}}{\sum_{i=1}^{n} \frac{1}{w_i} \times \sum_{i=1}^{n} \frac{x_i^2}{w_i} - \left[\sum_{i=1}^{n} \frac{x_i}{w_i}\right]^2} \times \hat{s}^2} \tag{3.22}$$

$$\hat{s}_b = \sqrt{\frac{\sum_{i=1}^{n} \frac{1}{w_i}}{\sum_{i=1}^{n} \frac{1}{w_i} \times \sum_{i=1}^{n} \frac{x_i^2}{w_i} - \left[\sum_{i=1}^{n} \frac{x_i}{w_i}\right]^2} \times \hat{s}^2} \tag{3.23}$$

Aquestes expressions es troben fàcilment i ràpida a través de la matriu variància-covariància obtinguda amb el procés iteratiu BLS.

Es dirà que un mètode no presenta errors sistemàtics constants si:

$$\left|\hat{a} - \textit{valor de referència}\right| \leq t_{\alpha/2,n-2} \cdot \hat{s}_a \tag{3.24}$$

on el valor de referència normalment sol ser zero. Similarment, es dirà que un mètode no presenta errors sistemàtics proporcionals si:

$$\left|\hat{b} - valor\ de\ referència\right| \leq t_{\alpha/2,n-2} \cdot \hat{s}_b \qquad (3.25)$$

on el valor de referència normalment sol ser 1.

3.3 Error β aplicat als intervals de confiança individuals per a l'ordenada a l'origen i el pendent tenint en compte l'error màxim fixat

Cometre un error β en l'aplicació dels intervals de confiança individuals de l'ordenada a l'origen o el pendent correspon a afirmar que les diferències entre el valor del coeficient de regressió que s'està comprovant i el valor de referència (normalment zero en el cas de l'ordenada a l'origen i 1 en el cas del pendent) no són superiors a un valor màxim (anomenat biaix) fixat per l'analista en cada cas, quan en realitat són superiors a aquest valor màxim fixat. A les seccions 1.5.4 i 1.5.5 ja s'ha introduït la importància de l'error β aplicat als problemes químics i de la poca atenció que normalment se li dóna dins del camp analític. Unes probabilitats d'error β elevades podrien portar com a conseqüència la no-detecció d'errors proporcionals o sistemàtics, la qual cosa pot comportar que per exemple, no s'apliqui cap correcció del blanc en processos de calibració quan en realitat s'hauria d'aplicar, o que en processos de recuperació no es detecti que la recuperació sigui significativament diferent del valor prefixat.

Les probabilitats d'error β fixant l'error màxim que es vol cometre corresponen a la part de la distribució centrada en el valor del biaix escollit (Δ) que se solapa amb la distribució associada al valor de referència limitada segons el nivell de

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

significança α escollit.[2,6,7] És a dir, a les probabilitats de concloure que el coeficient de regressió experimental pertany a la distribució associada al valor de referència quan en realitat pertany a la distribució associada al biaix.

Cal dir que només té sentit pràctic calcular les probabilitats d'error β un cop s'ha arribat a la conclusió que no hi ha diferències significatives entre el coeficient de regressió experimental i el valor de referència (emprant les equacions 3.24 o 3.25). Segons el tipus de mostra o mètode analític utilitzat, potser serà més important intentar minimitzar les probabilitats d'error α o les probabilitats d'error β.[8] Les probabilitats d'error α es poden escollir fixant el nivell de significança α desitjat i calculant les probabilitats d'error β associades mantenint el nivell α prèviament escollit constant. Si es volen minimitzar les probabilitats d'error β, s'escollirà un nivell de significança α tal que el coeficient de regressió experimental estigui situat just a un extrem de l'interval de confiança per al valor de referència (sempre que aquesta probabilitat d'error α sigui acceptable per l'analista), i es calcularà llavors la probabilitat d'error β a partir d'aquest valor d'α màxim a partir del qual no es trobarien diferències significatives entre el coeficient de regressió experimental i el valor de referència. Les probabilitats d'error α i β en un test individual segons aquestes dues alternatives es veuen representades a la figura 3.3.

Per tal de calcular les probabilitats d'error β, el primer problema sorgeix a l'hora de considerar la distribució associada al valor de referència i al biaix escollit. Per tal de resoldre això, tant al valor de referència com al biaix se li associen la mateixa distribució del coeficient de regressió experimental, ja sigui l'ordenada a l'origen o el pendent (com que ja s'ha comentat que s'acceptava com a vàlida la hipòtesi de la normalitat en la distribució dels coeficients de la recta de regressió, tant al valor de referència com al biaix se'ls associarà una distribució *t* de Student). Però el fet d'associar la mateixa distribució tant al valor de referència com al biaix comporta

alguns problemes, principalment en els tests individuals per al pendent. Hi ha una relació directa entre la desviació estàndard del pendent i el pendent via el coeficient

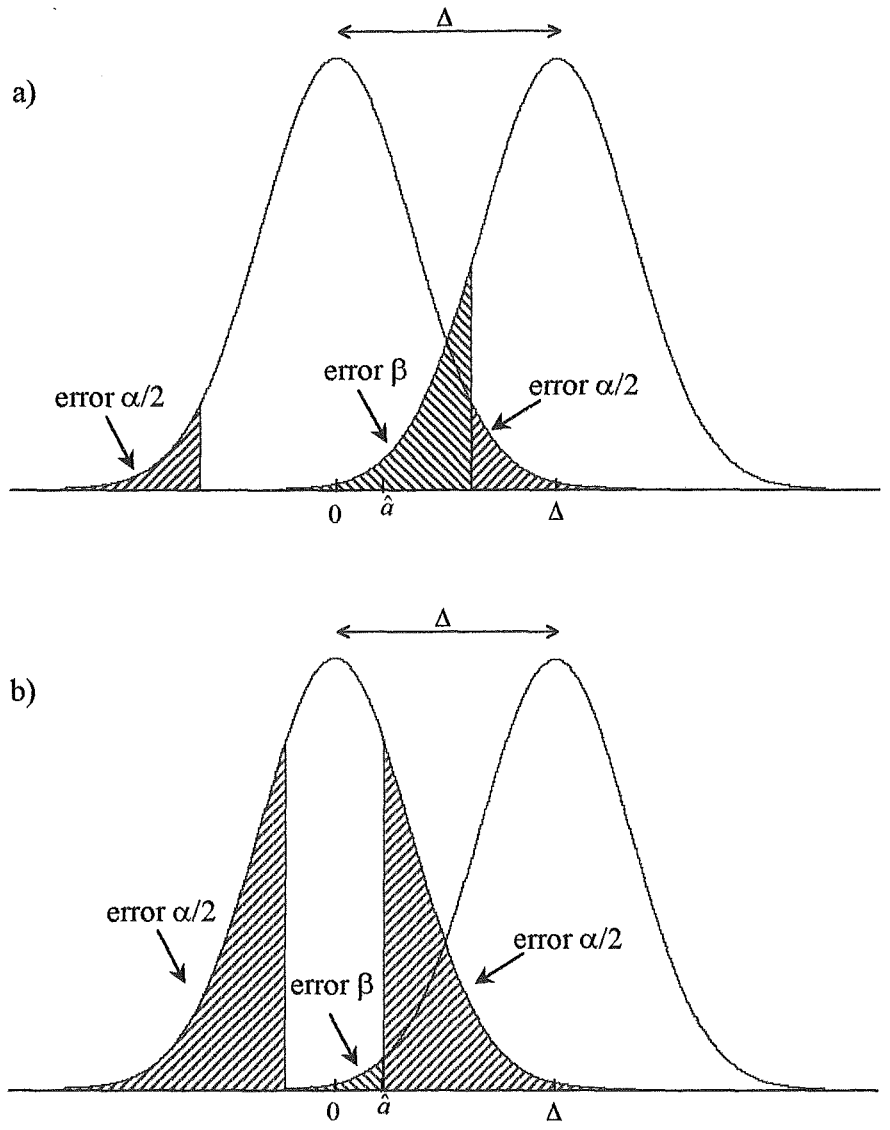


Figura 3.3. Representació de les probabilitats d'error α i β associades a l'interval de confiança individual de l'ordenada a l'origen o el pendent. En la figura es representen les probabilitats d'error α i β per a l'ordenada a l'origen, un valor de referència de 0 i un biaix Δ determinat, a) fixant el nivell de significança α i calculant les probabilitats d'error β associades, b) agafant el nivell de significança α màxim per tal de concloure que no hi ha diferències significatives entre el coeficient de regressió experimental i el valor de referència i calculant les probabilitats d'error β associades. El procediment és anàleg per al pendent.

de ponderació de cada punt individual (w_i a l'equació 3.23): com més alt és el valor del pendent, més alt és el valor de la seva desviació estàndard, tal com es podrà comprovar en la secció *Results and Discussion* de l'article situat al final del capítol. Tenint en compte la relació directa entre el valor de la desviació estàndard i la distribució *t* de Student, es pot deduir que en aquells casos en què el valor del pendent sigui més gran que el valor de referència, la distribució associada al biaix serà més gran de la que probablement li correspondria, i les prediccions de les probabilitats d'error β dutes a terme seran més grans que els seus valors reals. De la mateixa manera, en aquells casos en què el valor del pendent sigui més petit que el valor de referència, les prediccions de les probabilitats d'error β calculades seran més petites que els valors reals. Aquest fet s'arregla en part introduint al valor del pendent, inclòs dins del factor de ponderació de cada punt individual (w_i a l'equació 3.23), el valor corresponent al biaix escollit o el valor de referència, segons de quina distribució es tracti, i no el valor del pendent trobat experimentalment.

Aquest fet no es produeix en l'interval de confiança individual de l'ordenada a l'origen, ja que el valor de la desviació estàndard de l'ordenada a l'origen no depèn del valor que prengui l'ordenada a l'origen, com a conseqüència de la no-dependència directa entre aquests dos valors (equació 3.22).

Per a l'experimentador també pot ser important el fet de poder predir el nombre de punts del conjunt experimental (n), per tal de detectar una certa diferència màxima permesa Δ, de tal forma que hi hagi una probabilitat d'error α de concloure que hi ha diferències entre els valors de referència i experimental quan en realitat no hi són, i una probabilitat d'error β de no detectar aquestes diferències quan en realitat hi són presents. En els mètodes de regressió OLS i OR, la predicció del nombre de punts n necessaris per construir la recta de regressió per tal de detectar una certa diferència Δ donades unes certes probabilitats d'error α i β, es calcula d'una manera bastant ràpida i senzilla, a causa de la simplicitat en les expressions per

calcular els coeficients de regressió en OLS (secció 1.3.1) i OR (apartat 2.2). Així, en les expressions per predir el nombre de punts en OLS només cal considerar l'estimació de l'error experimental $\hat{s}^2$, la diferència Δ i paràmetres relatius a la variable predictora, com la mitjana de la variable predictora dels punts experimentals o la seva desviació estàndard.[9]

En OR les expressions es compliquen una mica més, i cal calcular també paràmetres com el pendent de la recta de regressió i l'error de mesura en la variable predictora.

L'estimació del nombre de punts de la recta de regressió per tal de detectar una certa diferència Δ donades unes certes probabilitats d'error α i β emprant el mètode BLS és més complicada. Això és degut sobretot a la presència del factor de ponderació (w_i) en les expressions de la desviació estàndard de l'ordenada a l'origen i el pendent, que fa que aquestes expressions no es puguin descompondre en factors més simples que ajudin a interpretar i predir el nombre de punts necessaris per construir la recta de regressió. El camí seguit per calcular el nombre de punts de la recta de regressió és considerar inicialment que les incerteses individuals en els dos eixos són constants per a tots els punts experimentals (encara que hi hagi heteroscedasticitat tant en la variable predictora com en la resposta), i per tant el factor de ponderació també roman constant. A partir de les estimacions de la desviació estàndard del coeficient de regressió experimental desitjat i del l'error experimental, obtingudes amb un conjunt inicial de dades experimentals, mitjançant un procés iteratiu es pot arribar a l'estimació del nombre de punts per tal de detectar-hi una certa diferència Δ donades unes certes probabilitats d'error α i β emprant el mètode BLS. Amb aquesta aproximació s'han obtingut bons resultats en l'estimació del nombre de punts, tant en conjunts homoscedàstics com heteroscedàstics, tal com es podrà comprovar a l'article que es troba al final del capítol.

3.4 Referències

1 P.C. Meier, R.E. Zünd, *Statistical Methods in Analytical Chemistry*, John Wiley & Sons, New York (1993)

2. Commission d'Établissement des Méthodes d'Analyses du Commissariat à l'Énergie Atomique (Cetama), *Statistique Appliquée a l'exploitation des Mesures*, Masson, Paris (1986)

3. G. Kateman, F.W. Pijpers, *Quality Control in Analytical Chemistry*, John Wiley & Sons, New York (1981)

4. M. Meloun, J. Militký, M. Forina, *Chemometrics for Analytical Chemistry. Volume 1: PC-aided statistical data analysis*, Ellis Horwood, Chichester (1992)

5. Matlab, *The MathWorks, Inc.*, Natick, Massachussets

6. M.R. Spiegel, *Theory and problems of statistics*, McGraw-Hill, New York (1988)

7. W. Pennincks, P. Vankeerberghen, D.L. Massart, J. Smeyers-Verbeke, *Journal of Analytical Atomic Spectrometry*, **207** (1995) 207

8. C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, Y. Vander Heyden, P. Vankeerberghen, D.L. Massart, *Analytical Chemistry*, **67** (1995) 4491

9. C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, D.L. Massart, *Analytica Chimica Acta*, **338** (1997) 19

Detecting proportional and constant bias in method comparison studies by using linear regression with errors in both axes

Àngel Martínez, F. Javier del Río, Jordi Riu* and F. Xavier Rius

Departament de Química Analítica i Química Orgànica.

Universitat Rovira i Virgili.

Pl. Imperial Tàrraco, 1. 43005-Tarragona. CATALONIA, SPAIN.

ABSTRACT

Constant or proportional bias in method comparison studies using linear regression can be detected by an individual test on the intercept or the slope of the line regressed from the results of the two methods to be compared. Since there are errors in both methods, a regression technique that takes into account the individual errors in both axes (bivariate least squares, BLS) should be used. In this paper we demonstrate that the errors made in estimating the regression coefficients by the BLS method are fewer than with the OLS or WLS regression techniques and that the coefficient can be considered normally distributed. We also present expressions for calculating the probabilities of committing a β error in individual tests under BLS conditions and theoretical procedures for estimating the sample size in order to obtain the desired probabilities of α and β errors made when testing each of the BLS regression coefficients individually. Simulated data were used for the validation process. Examples for the application of the theoretical expressions developed are given using real data sets.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

INTRODUCTION

Linear regression is widely used in the validation of analytical methodologies. In method comparison studies, for example, a set of samples of different concentration levels are analysed by the two methods to be compared, and the results are regressed on each other. Ordinary least-squares (OLS), or weighted least-squares (WLS), which considers heteroscedasticity in the response variable, are the most widely used regression techniques. However, these techniques have a limited scope, since they consider the x-axis to be free of error. OLS and WLS should not usually be applied, for instance, in method comparison studies, since the uncertainties associated with the methods to be compared are usually of the same order of magnitude. An alternative is the errors-in-variables regression [1], also called CVR approach [2-4], which considers the errors in both axes. It does not take into account the individual uncertainties of each experimental point but considers the ratio of the variances of the response to predictor variables to be constant for every experimental point ($\lambda=s_y^2/s_x^2$). A particular case of the CVR approach is the orthogonal regression (OR) [5], in which the errors are of the same order of magnitude in the response and predictor variable (i.e. $\lambda=1$). Another option is a bivariate least squares (BLS) regression technique [6,7], which takes into account individual non-constant errors in both axes to calculate the regression coefficients.

Despite the recent development of a joint confidence interval test for the BLS regression method [8], no statistical test to individually assess the presence of bias in the regression coefficients which takes into account the individual uncertainties in every experimental point has yet been described. For this reason, we present expressions for the application of the individual tests which take into account individual errors in both axes. Although the distributions of the BLS slope and intercept have been reported to be nongaussian [9], in this paper we show that the results of applying statistical tests based on the assumption of normality of the

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

BLS regression coefficients do not show significant errors and that these errors are fewer than those obtained with the OLS or WLS regression techniques.

Of the two types of error associated with the statistical tests (α and β), the β error, related to the probability of not detecting an existing proportional or constant bias is seldom considered. However, the theoretical background and the expressions which enable its calculation in the individual tests which use the OLS method have already been developed [5]. In this paper we describe the expressions for estimating the probabilities of β error when performing an individual test on one of the regression coefficients to detect a set proportional or constant bias based on the BLS regression technique. These expressions take into account the different distributions that may be associated to the reference and to the selected biased regression coefficient values. These estimates are compared with the ones from the OLS and the WLS techniques for several real data sets. Finally, we describe the procedure for estimating the sample size, i.e. the number of experimental data pairs necessary for detecting the specific selected bias when performing an individual test with set probabilities of making α and β errors when the BLS regression method is used. Simulated data sets have been used to validate the theoretical expressions.

BACKGROUND AND THEORY

Notation

The true values of the BLS regression coefficients are represented by a (intercept) and b (slope), while their respective estimates are denoted as $\hat{a}$ and $\hat{b}$. The estimates of the standard deviation of the slope and the intercept for the BLS regression line, are symbolised as $\hat{s}_b$ and $\hat{s}_a$ respectively. The experimental error,

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

expressed in terms of variance for the n experimental data pairs (x_i,y_i), is referred to as s^2, while its estimate is $\hat{s}^2$. By analogy, $\hat{y}_i$ represents the estimated value for the y_i predicted. The variance-covariance matrix of the regression coefficients related to the BLS regression technique is denoted as **B**.

In the individual tests, the terms a_{H_0}, a_{H_1}, b_{H_0} and b_{H_1} represent the values of the theoretical regression coefficients from which the null (H_0) and the alternative hypothesis (H_1) are assumed. The distance between a_{H_0} and a_{H_1} or between b_{H_0} and b_{H_1}, known as bias, is denoted by Δ and represents the value of the systematic error that the experimenter wants to check. By analogy, the values of the standard deviations of the theoretical regression coefficients defining H_0 and H_1 are denoted as $\hat{s}_{a_{H_0}}$ (or $\hat{s}_{b_{H_0}}$) and $\hat{s}_{a_{H_1}}$ (or $\hat{s}_{b_{H_1}}$).

Bivariate Least-Squares Regression (BLS)

BLS is the generic name given to a set of regression techniques applied to data which contain errors in both axes. From all the different existing approaches for calculating the regression coefficients, Lisý's method [6] was found to be the most suitable [7]. It minimises the sum of the weighted residuals, S, expressed in Eq. (1):

$$S = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{w_i} = (n-2)\hat{s}^2 \qquad (1)$$

The weighting factor w_i is expressed as the variance of the ith residual ($s^2_{\varepsilon_i}$) and takes into consideration the variances of any individual point in both axes ($s^2_{x_i}$ and $s^2_{y_i}$), as well as the covariance between the variables for each (x_i,y_i) data pair, which is normally assumed to be zero:

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

$$s_{\varepsilon_i}^2 = w_i = s_{y_i}^2 + \hat{b}^2 s_{x_i}^2 - 2\hat{b}\,\mathrm{cov}(x_i, y_i) \qquad (2)$$

For this reason, the BLS regression technique assigns higher weights to those data pairs with larger $s_{x_i}^2$ and $s_{y_i}^2$ values, i.e. the most imprecise data pairs. By minimising the sum of the weighted residuals (Eq. (1)), two non-linear equations are obtained, from which the regression coefficients $\hat{a}$ and $\hat{b}$ can be estimated by an iterative process [8].

Characterisation of the distribution of the BLS regression coefficients

The distribution functions of the regression coefficients $\hat{a}$ and $\hat{b}$ found by the BLS regression technique have been reported to be nongaussian [9]. This influences the individual tests on the regression coefficients, since they are usually performed under the assumption of normality. To determine the degree of non-normality of the distributions of the BLS coefficients, three different statistical tests were used: Cetama [10] (which also allows the actual probability function to be characterised), the Kolmogorov test [11] and the normal probability plot (or Rankit test) [12]. These tests were applied to different types of real data sets to find a relationship between their structure and the degree of non-normality. Furthermore, to characterise their distribution, the real distributions and some theoretical distributions were compared. These comparisons were carried out with the quantile-quantile graphic method (Q-Q plot) [12].

β error in the individual tests for the BLS regression coefficients

According to the theory of hypothesis testing, when an individual test is applied on a regression coefficient, the null hypothesis H_0 is defined as the one that considers the estimated regression coefficient to belong to the distribution of a hypothetical

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

regression coefficient (a_{H_0} or b_{H_0}) equal to the reference value, or in other words, that there are no proportional or constant systematic errors in the method being tested. On the other hand, the alternative hypothesis H_1 considers that the estimated regression coefficient belongs to the distribution of a hypothetical regression coefficient (a_{H_1} or b_{H_1}) with a given value. This value, which has to be set by the experimenter according to the systematic error one wants to detect in the analytical method being tested, defines the distance between a_{H_0} (or b_{H_0}) and a_{H_1} (or b_{H_1}), or in other words the so-called bias [13]. The standard deviations $\hat{s}_{a_{H_0}}$ (or $\hat{s}_{b_{H_0}}$) and $\hat{s}_{a_{H_1}}$ (or $\hat{s}_{b_{H_1}}$) can be calculated for a given data set with the values of a_{H_0} (or b_{H_0}) and a_{H_1} (or b_{H_1}).

The expressions developed for estimating the probabilities of committing a β error in the application of an individual test to one of the regression coefficients calculated by using the OLS regression technique are established [5]. Analogous expressions can be adapted for the BLS technique by considering the appropriate standard deviation values:

$$\Delta_b = t_{\alpha/2} \cdot \hat{s}_{b_{H_0}} + t_\beta \cdot \hat{s}_{b_{H_1}} \quad \Rightarrow \quad t_\beta = \frac{\Delta_b - t_{\alpha/2} \cdot \hat{s}_{b_{H_0}}}{\hat{s}_{b_{H_1}}} \tag{3}$$

$$\Delta_a = t_{\alpha/2} \cdot \hat{s}_{a_{H_0}} + t_\beta \cdot \hat{s}_{a_{H_1}} \quad \Rightarrow \quad t_\beta = \frac{\Delta_a - t_{\alpha/2} \cdot \hat{s}_{a_{H_0}}}{\hat{s}_{a_{H_1}}} \tag{4}$$

The probability of committing a β error under the assumption of normality is finally given by the Student's *t* value for *n*-2 degrees of freedom for a fixed α level of significance. The standard deviations $\hat{s}_{a_{H_0}}$ (or $\hat{s}_{b_{H_0}}$) and $\hat{s}_{a_{H_1}}$ (or $\hat{s}_{b_{H_1}}$) can be estimated in a similar way to the standard deviations of the intercept and the slope,

and are easily obtained from the **B** variance-covariance matrix [8] calculated while estimating the regression coefficients with the BLS technique:

$$\hat{s}_a = \sqrt{\frac{\sum_{i=1}^{n} \frac{x_i^2}{s_{\varepsilon_i}^2}}{\sum_{i=1}^{n} \frac{1}{s_{\varepsilon_i}^2} \times \sum_{i=1}^{n} \frac{x_i^2}{s_{\varepsilon_i}^2} - \left[\sum_{i=1}^{n} \frac{x_i}{s_{\varepsilon_i}^2}\right]^2}} \times \hat{s} \qquad (5)$$

$$\hat{s}_b = \sqrt{\frac{\sum_{i=1}^{n} \frac{1}{s_{\varepsilon_i}^2}}{\sum_{i=1}^{n} \frac{1}{s_{\varepsilon_i}^2} \times \sum_{i=1}^{n} \frac{x_i^2}{s_{\varepsilon_i}^2} - \left[\sum_{i=1}^{n} \frac{x_i}{s_{\varepsilon_i}^2}\right]^2}} \times \hat{s} \qquad (6)$$

To calculate the values of $\hat{s}_{a_{H_0}}$ (or $\hat{s}_{b_{H_0}}$) and $\hat{s}_{a_{H_1}}$ (or $\hat{s}_{b_{H_1}}$) it is only necessary to recalculate the value of the weighting factor (Eq. (2)) according to the new slope value. Due to the dependence of the weighting factor on the slope, the values of $\hat{s}_{a_{H_0}}$ and $\hat{s}_{a_{H_1}}$ will be equal to the standard deviation obtained for the estimated regression coefficient ($\hat{s}_a = \hat{s}_{a_{H_0}} = \hat{s}_{a_{H_1}}$), which is not true for the slope. The experimental error $\hat{s}^2$ remains unchanged.

Estimating the sample size

Relating Eqs. (5-6) with the number of data pairs *n* it is possible to estimate the number of data pairs required to detect certain bias with set probabilities of committing α and β errors. This can only be achieved if the individual

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

uncertainties, and hence the weighting factors are considered constant for all the data pairs ($\hat{s}^2_{\varepsilon_{a_{H_0}}}$, $\hat{s}^2_{\varepsilon_{b_{H_0}}}$ or $\hat{s}^2_{\varepsilon_{b_{H_1}}}$ = ct):

$$\hat{s}_{a_{H_0}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \cdot \hat{s}^2_{\varepsilon_{a_{H_0}}}}{n_a \cdot \sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \cdot \hat{s}^2} \tag{7}$$

$$\hat{s}_{b_{H_0}} = \sqrt{\frac{n_b \cdot \hat{s}^2_{\varepsilon_{b_{H_0}}}}{n_b \cdot \sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \cdot \hat{s}^2} \quad \text{or} \quad \hat{s}_{b_{H_1}} = \sqrt{\frac{n_b \cdot \hat{s}^2_{\varepsilon_{b_{H_1}}}}{n_b \cdot \sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \cdot \hat{s}^2} \tag{8}$$

Introducing these two expressions in Eqs. (3-4) respectively it is possible to isolate n in terms of the desired variables α, β and Δ:

$$n_a = \frac{(t_{\alpha/2} + t_\beta)^2 \cdot \hat{s}^2_{\varepsilon_{a_{H_0}}}}{\Delta_a^2} \cdot \hat{s}^2 + \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{\sum_{i=1}^{n} x_i^2} \tag{9}$$

$$n_b = \frac{\Delta_b^2 \cdot \left[\sum_{i=1}^{n} x_i\right]^2}{\Delta_b^2 \cdot \sum_{i=1}^{n} x_i^2 - \left(t_{\alpha/2} \cdot \hat{s}_{\varepsilon_{a_{b_{H_0}}}} + t_\beta \cdot \hat{s}_{\varepsilon_{b_{H_1}}}\right)^2 \cdot \hat{s}^2} \tag{10}$$

Initial estimates of the terms $\hat{s}^2_{\varepsilon_{a_{H_0}}}$ or $\hat{s}^2_{\varepsilon_{b_{H_0}}}$ and $\hat{s}^2_{\varepsilon_{b_{H_1}}}$, $\hat{s}^2$ and both sums involving x data coordinates can be set from an initial data set containing few data pairs. After an iterative process (due to the dependence of the $t_{\alpha/2}$ and t_β values on the

number of data pairs), it is important to recalculate the sample size by adding more data to the initial data set, as the estimates of the terms mentioned in Eqs. (9-10) are likely to change. The process ends when the differences between two consecutive n_a or n_b values are below a set threshold value.

Validation

The objective of the validation process is twofold. Firstly, to show that, despite the non-normal distribution of the BLS regression line coefficients, the confidence interval computed using the *t*-distribution can generally be accepted without committing relevant errors. Secondly, to assess whether the theoretical estimate of either the β error and the number of data pairs required to perform the individual tests, based on BLS under defined statistical conditions, provides correct results.

To show the degree of non-normality of the intercept and the slope distributions, six real data sets with errors in both axes were studied. The Monte Carlo method [14] was applied to generate 200,000 data sets from each of the six initial ones. In this way, a random error based on the individual uncertainties in both axes was added to each data pair. This gave rise to 200,000 regression lines, to which the three selected tests for assessing the normality of the distributions were applied. The error made in estimating the BLS regression coefficients when their respective distributions were assumed to be normal (when in fact they are not) was quantified and compared with the error made in estimating the regression coefficients by OLS and WLS techniques. Figure 1 illustrates the comparison procedure. Once the distribution of the regression coefficients corresponding to the real data set is obtained by the Cetama method, we can determine its left (x_{lr}) and right (x_{rr}) limits for a chosen α level of significance. The shaded areas in Figure 1 represent the errors made by estimating the regression coefficients with each of the three regression techniques studied.

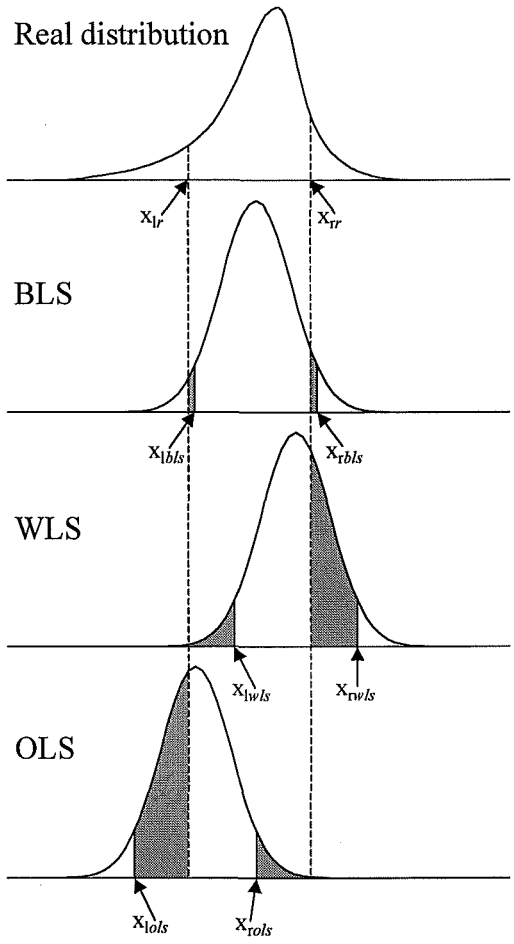


Figure 1. Error made in estimating the BLS regression coefficients assuming normal distributions. Comparison with errors made using OLS and WLS regression techniques.

To validate the expressions for the estimated of the probabilities of β error, 24 initial simulated data sets were used with all the data pairs perfectly fit to an straight line with either biased slope or intercept values. From each of these initial data sets, 100,000 simulated new ones were generated using the Monte Carlo method. An individual test was then applied on one of the regression coefficients for every one of these data sets to check whether H_0 could be accepted in each case for a fixed α level of significance. So every time H_0 was accepted, a β error was being committed because the data set had been generated from an initial biased

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

one. Due to the application of random errors by the Monte Carlo method, however, the bias could not be detected. The value of the bias was chosen to provide a probability of β error similar to the α level of significance in each of the four cases. In this way, if the estimate of the probabilities of β error from the theoretical expressions was similar to the one from the simulation process, we may conclude that the stated expressions provide correct results.

Once the estimates of the probabilities of β error were proved to be correct, the expressions to estimate the sample size were validated. The probabilities of β error estimated for the different α levels of significance, the calculated standard deviations and the experimental error from the iterative process (terms t_β , $t_{\alpha/2}$, $\hat{s}^2_{\varepsilon_{\sigma H_0}}$, $\hat{s}_{\varepsilon_{b H_0}}$ or $\hat{s}_{\varepsilon_{b H_1}}$ and $\hat{s}^2$ respectively) for each of the initial data sets in the validation process were introduced in expressions 9 and 10. If the estimated sample size required to achieve the chosen α and β probabilities of error was similar to the number of data pairs in each data set, results were considered correct. To show the applicability of the procedure, a real data set was used as a case study.

EXPERIMENTAL SECTION

Data sets and software

Six real data sets with different characteristics (such as number of data points, heteroscedasticity or position within the experimental domain) were used to check the distribution of the BLS regression coefficients. Twenty-four different simulated data sets were considered to validate the expressions for the estimates of the probabilities of β error (Eqs. (3-4)). Finally, one of the six former real data sets was used to show the different estimates of the probabilities of β error between BLS,

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

OLS and WLS regression techniques and provide an example of the sample size estimation procedure using data with errors in both axes.

Data Set 1 [15]. Data set obtained from the study of the supercritical fluid extraction (SFE) recoveries of policyclic aromatic hydrocarbons (PAHs) from railroad bed soil using two different modifiers; CO_2 (on the *x*-axis) and a mixture of CO_2 with 10% of toluene (on the *y*-axis). The data set is composed of seven data pairs. The standard deviations were the result of a triplicate supercritical fluid extraction. The units are expressed in terms of μg/g of soil. The data set and the regression lines obtained by the OLS, WLS and BLS regression techniques are shown in Figure 2a.

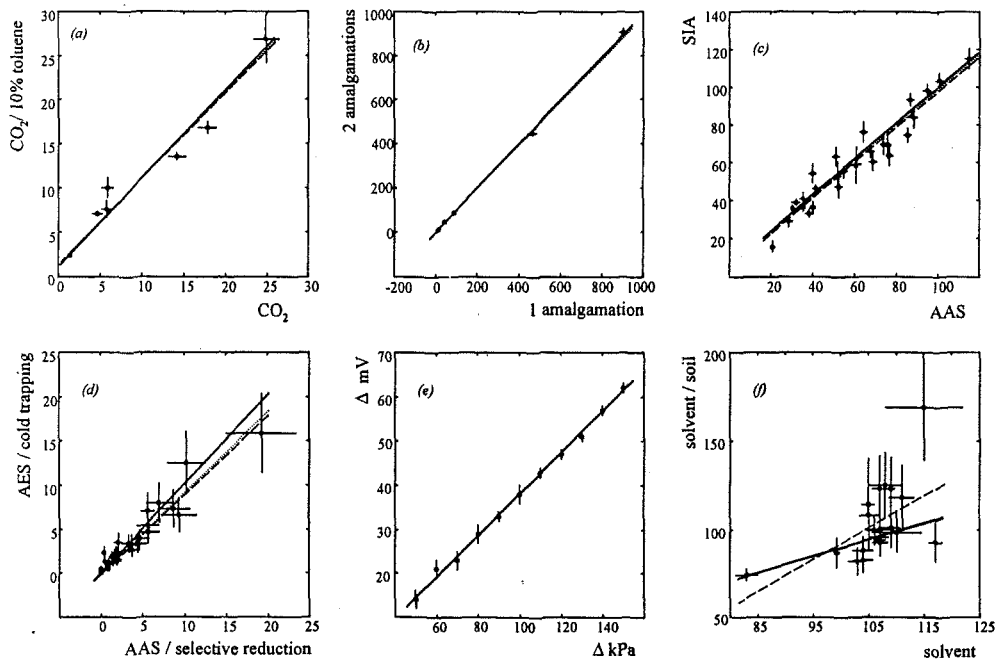


Figure 2. OLS (dashed line), WLS (dotted line) and BLS (solid line) regression lines obtained for the six real data sets.

Data Set 2 [16]. Comparative study of mercury determination using gas chromatography coupled to a cold vapour atomic fluorescence spectrometer following derivatization with sodium tetraethylborate. One (*x*-axis) and two (*y*-

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

axis) amalgamation steps were used to obtain five data pairs with their respective uncertainties generated from six replicates performed at each point. Units are expressed in terms of pg of recovered mercury. The data set and the regression lines generated by the three regression techniques are shown in Figure 2b.

Data Set 3 [17]. Twenty-seven data pairs obtained from a method comparison study which analysed Ca(II) in water by atomic absorption spectroscopy (AAS), taken as the reference method (x-axis), and sequential injection analysis (SIA), taken as the tested method (y-axis). The data set and the regression lines generated by OLS, WLS and BLS regression techniques are shown in Figure 2c. Units are expressed in mg/l. The uncertainties associated with the AAS method were derived from the analytical procedure, including the linear calibration step [18]. The uncertainties of the SIA results were calculated with a multivariate regression model and the PLS technique using the Unscrambler program (Unscrambler-Ext, ver. 4.0, Camo A/S, Trondheim, Norway).

Data Set 4 [19]. Comparative study for determining arsenic in natural waters from two techniques: continuous selective reduction and atomic absorption spectrometry (AAS) as the reference method (x-axis) and non-selective reduction, cold trapping and atomic emission spectrometry (AES) as the tested method (y-axis). Thirty experimental data pairs were obtained. The units are expressed in terms of μg/l. The data set and the regression lines obtained using all three regression techniques are shown in Figure 2d.

Data Set 5 [20]. Data set obtained by measuring the CO_2 Joule-Thompson coefficient. The data was acquired from thermocouple-measured voltage differences (ΔmV, on the y-axis) as a function of pressure increments (ΔkPa, on the x-axis). Eleven equally-distributed data pairs were obtained with estimated unity x-axis uncertainties. The y-axis uncertainties were estimated to be between one and

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

two units. The data set and the three regression lines found by using the stated regression techniques are shown in Figure 2e.

Data Set 6 [21]. Comparative study of the average recoveries for organochlorine pesticides present in solvent (on the *x*-axis) or in solvent/soil suspension (on the *y*-axis) after microwave-assisted extraction (MAE) analysis. Twenty-one data pairs were used in the analysis. The uncertainties were obtained from triplicate MAE analysis at each point. The data set and the straight lines regressed by the three regression techniques are shown in Figure 2f.

To validate the estimates of the probabilities of β error, twenty-four different initial data sets showing different values of bias in the intercept or in the slope were built to cover several analytical situations; different linear ranges, number of data pairs and uncertainty patterns.

Linear Ranges: Two linear ranges were considered during validation, a short one for values from 0 to 10 units, and a large one for values from 0 to 100 units.

Number of data pairs: Data sets containing five, fifteen, thirteen and a hundred data pairs were selected. In all cases the data pairs were randomly distributed throughout the two different linear ranges.

Uncertainties: Homoscedastic and heteroscedastic data sets were considered. The homoscedastic data sets were comprised of data pairs with constant standard deviations on both *x* and *y* values. In the short linear ranges the standard deviations presented half unity values, whereas in the large linear ranges they showed unity values. The heteroscedastic data sets were divided into two other different types. On one hand those with increasing standard deviations and on the other hand, those which presented random standard deviations. In both cases however, the standard deviation values were never higher than the 10% of each individual *x* and *y* value.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

For every one of the twenty four different simulated data sets, four α levels of significance were considered: 10, 5, 1 and 0.1%. Depending on the regression coefficient being tested and on the level of significance, the slope (b_{H_1}) or the intercept value (a_{H_1}) of the selected bias changed in such a way that the probabilities of β error from the iterative process were similar to the specified α values. In this way the accuracy of estimates of different magnitudes from Eqs. (3-4) was also tested.

All the computational work was performed with home-made Matlab subroutines (Matlab for Microsoft Windows ver. 4.0, The Mathworks, Inc., Natick, MA).

RESULTS AND DISCUSSION

Distribution of the regression coefficients

The results of studying the distributions of the slope ($\hat{b}$) and the intercept ($\hat{a}$) using the three tests to check normality are summarised in Table 1. The variation in the number of iterations needed to achieve non-normality can be used to identify the degree of normality. The more iterations needed to achieve non-normality (if finally achieved) the more normal the distribution is.

Data set 1 presents non-normal distributions mainly due to the high lack of fit of the data pairs to the regression line. Data sets 2 and 5 present the best goodness of fit of all the sets, which helps the distribution of the regression coefficients to be normal. In data set 3, the data structure and the errors in both axes make the regression line mainly change the intercept value, which leaves the slope almost unmodified. In this way the intercept value shows a major uncertainty which leads

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

to a non-normal distribution, whereas a much lower uncertainty is associated to the slope value.

				Kolmogorov							
		Cetama		α=1%		α=5%		α=10%		Rankit Plot	
data set	Iterations	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
1	10.000	NSNL	NSLRL	N	NN	NN	NN	NN	NN	NN	NN
	30.000	NSNL	NSNL	NN	NN	NN	NN	NN	NN	NN	NN
	50.000	NSNL	NSNL	NN	NN	NN	NN	NN	NN	NN	NN
	100.000	NSNL	NSNL	NN	NN	NN	NN	NN	NN	NN	NN
	200.000	NSNL	NSNL	NN	NN	NN	NN	NN	NN	NN	NN
2	10.000	N	NSNL	N	N	N	N	N	N	NN	NN
	30.000	N	NSLRL	N	N	N	N	N	N	N	N
	50.000	N	NSNL	N	N	N	N	N	N	N	N
	100.000	NSNL	NSNL	N	N	N	N	N	N	N	N
	200.000	NSLRL	NSLL	N	N	N	N	N	N	N	N
3	10.000	NSNL	NSLRL	N	N	N	N	N	N	NN	NN
	30.000	NSNL	NSLRL	N	N	N	N	N	N	NN	N
	50.000	NSNL	NSNL	NN	N	NN	N	NN	N	NN	N
	100.000	NSLRL	NSLRL	NN	N	NN	N	NN	N	NN	N
	200.000	NSNL	NSNL	NN	N	NN	N	NN	N	NN	N
4	10.000	N	NSNL	N	N	N	N	N	N	N	N
	30.000	N	NSNL	N	NN	N	NN	N	NN	N	NN
	50.000	N	NSNL	N	NN	N	NN	N	NN	N	NN
	100.000	N	NSNL	N	NN	N	NN	N	NN	N	NN
	200.000	N	NSNL	N	NN	N	NN	N	NN	N	NN
5	10.000	N	N	N	N	N	N	N	N	N	N
	30.000	N	N	N	N	N	N	N	N	N	N
	50.000	N	N	N	N	N	N	N	N	N	N
	100.000	N	N	N	N	N	N	N	N	N	N
	200.000	N	N	N	N	N	N	N	N	N	N
6	10.000	NSNL	NSNL	N	NN	N	NN	N	NN	NN	NN
	30.000	NSNL	NSNL	N	NN	NN	NN	NN	NN	NN	NN
	50.000	NSNL	NSNL	NN	NN	NN	NN	NN	NN	NN	NN
	100.000	NSNL	NSNL	NN	NN	NN	NN	NN	NN	NN	NN
	200.000	NSNL	NSNL	NN	NN	NN	NN	NN	NN	NN	NN

N: Normal distribution.
NN: Non-normal distribution.
NSNL: Non-symmetric and non-limited.
NSLRL: Non-symmetric and left and right limited.
NSLL: Non-symmetric and left limited.

Table 1. Normality study results for the BLS regression coefficients.

In data set 4, the slope of the regression line does not follow a normal distribution since the remarkable heteroscedasticity along the experimental range causes the regression line to move along a conical-shaped region when considering errors in both axes. This varies the slope and leaves the intercept almost unmodified.

Finally, data set 5 has normal distributions and data set 6 presents non-normal ones due to the irregular disposition of the points in the space and the high heteroscedasticity. The more similar the error pattern to OLS conditions (i.e. larger errors in the *y* axis than in the *x* axis, homoscedasticity) and the better the goodness of fit, the more normal the distribution is. It has to be pointed out that the Cetama method was the most sensitive in detecting deviations from normality.

Table 2 shows the quantification of the error made in estimating the BLS regression coefficients when normality in their distributions is assumed, and the comparison with the analogous results from OLS and WLS regression techniques. The error is calculated according to the shaded areas in Figure 1 (where the error is considered to be the part that belongs to the OLS, WLS or BLS distribution for a fixed α level and which does not belong to the real distribution, and the part that does not belong to the OLS, WLS or BLS distribution for the same α level and belongs to the real one). This table shows that the error made from assuming normality for the BLS regression technique is low, and significantly lower than the ones obtained for the OLS and WLS regression methods for all the data sets. The data sets that present BLS regression coefficients as normally distributed have errors equal to zero. We can also see that the error committed when using the WLS method is usually lower than when using OLS.

Once the BLS regression coefficients have been found, in most cases, to be non-normally distributed, their distributions were compared with some theoretical ones (beta, binomial, chi-squared, exponential, *F*, gamma, geometric, hypergeometric, normal, Poisson, *t*-Student, uniform, uniform discrete and Weibull distributions) using the quantile-quantile plot graphic method (Q-Q plot) [12]. As the results provided by the Cetama method (Table 1) indicate that the regression coefficients that do not follow a normal distribution are mainly non-symmetric and non-limited, it seems reasonable to suppose that the regression coefficient distributions follow some kind of constant pattern. However, the results given by the Q-Q plot indicate

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

that the theoretical distributions that are most similar to the real ones are the chi-squared, normal and *t*-Student since their differences are very difficult to appreciate.

		% Error		
data set	Coefficient	BLS	WLS	OLS
1	$\hat{a}$	4.69	26.84	58.29
	$\hat{b}$	4.46	14.59	16.43
2	$\hat{a}$	0	9.81	44.35
	$\hat{b}$	0	5.51	3.66
3	$\hat{a}$	0.53	1.37	11.42
	$\hat{b}$	0.58	6.20	11.03
4	$\hat{a}$	0	5.11	88.50
	$\hat{b}$	2.79	14.97	25.28
5	$\hat{a}$	0	0.26	0.62
	$\hat{b}$	0	0.25	3.28
6	$\hat{a}$	2.48	2.31	6.60
	$\hat{b}$	2.48	3.75	6.45

Table 2. Difference between the theoretical and estimated regression coefficients by the three regression techniques (normal distributions assumed).

β error and sample size validation

Tables 3 and 4 summarise the results from 100,000 iterations using the Monte Carlo method for the four levels of significance in the twenty four simulated data sets. Columns a_{H_1} and b_{H_1} show the regression coefficient values which define the chosen bias (distance between H_0 and H_1). The values in the β_{exp} column are those from the simulation process, whereas the values shown in the β_{pred} column are the ones obtained with the theoretical expressions to be validated (Eqs. (3-4)). Finally, the values in the column n_{pred} are the estimated sample sizes of the different simulated data sets for the different levels of significance.

n	Uncertainty	α(%)	a_{H_1}	$\hat{s}_{a_{H_0}}$	$\beta_{exp.}$	$\beta_{pred.}$	$n_{pred.}$
5	homo.	10	2.4	0.641	9.97	12.91	5
		5	3.2		5.02	8.39	5
		1	5.2		2.22	5.38	5
		0.1	10.5		0.13	2.03	5
	hetero.	10	0.7	0.189	10.11	13.67	5
		5	0.95		4.32	8.26	5
		1	1.5		2.75	6.53	5
		0.1	3		0.74	3.14	5
	heter. rnd.	10	1	0.261	8.36	11.77	5
		5	1.3		4.80	8.48	5
		1	2.1		2.23	5.71	5
		0.1	4.3		0.11	2.59	5
15	homo.	10	1	0.341	13.24	13.34	15
		5	1.3		5.73	6.14	15
		1	1.9		0.93	1.19	15
		0.1	2.6		0.10	0.24	15
	hetero.	10	5e-2	1.69e-2	12.02	12.99	15
		5	6.5e-2		4.98	4.9	15
		1	9.5e-2		0.57	1.11	15
		0.1	0.125		0.10	0.28	15
	heter. rnd.	10	2.5e-2	8.79e-3	13.95	15.12	15
		5	3.4e-2		4.39	5.56	15
		1	4.5e-2		1.81	2.75	15
		0.1	6.4e-2		0.13	0.45	15
30	homo.	10	0.75	0.262	12.93	12.82	30
		5	1		4.36	4.43	30
		1	1.3		1.74	1.84	30
		0.1	1.8		0.12	0.17	30
	hetero.	10	5.5e-3	1.92e-3	12.19	12.62	30
		5	7e-3		5.53	5.99	30
		1	9.5e-3		1.43	1.84	30
		0.1	1.2e-2		0.54	0.76	30
	heter. rnd.	10	1.9e-2	6.48e-3	11.07	11.46	30
		5	2.4e-2		4.97	5.47	30
		1	3.2e-2		1.50	1.92	30
		0.1	4.3e-2		0.16	0.31	30

Table 3. Estimated and experimentally obtained probabilities of β error for individual tests on the intercept. Predicted sample size to achieve the α and β probabilities of error for each data set.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

n	Uncertainty	α(%)	a_{H_1}	$\hat{s}_{a_{H_0}}$	$\beta_{exp.}$	$\beta_{pred.}$	$n_{pred.}$
100	homo.	10	0.4	0.142	12.78	12.68	100
		5	0.5		6.61	6.51	100
		1	0.68		1.77	1.70	100
		0.1	0.88		0.35	0.32	100
	hetero.	10	1.5e-5	5.37e-6	12.89	12.98	100
		5	1.9e-5		6.02	6.16	100
		1	2.6e-5		1.41	1.45	100
		0.1	3.4e-5		0.19	0.20	100
	heter. rnd.	10	1.9e-4	6.41e-5	9.49	9.76	100
		5	2.4e-4		3.86	4.07	100
		1	3e-4		1.91	2.13	100
		0.1	4.2e-4		0.07	0.10	100

Table 3 (cont). Estimated and experimentally obtained probabilities of β error for individual tests on the intercept. Predicted sample size to achieve the α and β probabilities of error for each data set.

n	Uncertainty	α (%)	b_{H_1}	$\hat{s}_{b_{H_0}}$	$\hat{s}_{b_{H_1}}$	$\beta_{exp.}$	$\beta_{pred.}$	$n_{pred.}$
5	homo.	10	1.45	0.118	0.147	10.39	16.44	5
		5	1.6		0.157	5.87	12.60	5
		1	2		0.187	3.09	9.87	5
		0.1	3.1		0.272	0.62	6.37	5
	hetero.	10	1.27	7.48e-2	8.55e-2	12.67	17.64	5
		5	1.36		9.02e-2	4.56	10.70	5
		1	1.65		0.102	1.11	6.42	5
		0.1	2.3		0.132	0.22	4.36	5
	heter. rnd.	10	1.27	7.59e-2	9.07e-2	14.41	19.44	5
		5	1.4		9.80e-2	3.76	10.24	5
		1	1.67		0.113	1.19	6.99	5
		0.1	2.35		0.153	0.26	4.78	5

Table 4. Estimated and experimentally obtained probabilities of β error for individual tests on the slope. Predicted sample size to achieve the α and β probabilities of error for each data set.

n	Uncertainty	α (%)	b_{H_1}	$\hat{s}_{b_{H_0}}$	$\hat{s}_{b_{H_1}}$	$\beta_{exp.}$	$\beta_{pred.}$	$n_{pred.}$
15	homo.	10	0.8	6.92e-2	6.26e-2	10.84	11.91	15
		5	0.75		6.11e-2	5.11	6.21	15
		1	0.68		5.86e-2	3.75	2.14	15
		0.1	0.55		5.58e-2	0.48	0.71	15
	hetero.	10	0.93	2.49e-2	2.41e-2	14.59	15.2	15
		5	0.91		2.39e-2	6.98	7.73	15
		1	0.87		2.34e-2	1.14	1.78	15
		0.1	0.83		2.29e-2	0.35	0.72	15
	heter. rnd.	10	0.965	1.19e-2	1.16e-2	11.77	12.72	15
		5	0.955		1.153e-	5.07	5.98	15
		1	0.94		1.19e-2	1.98	2.74	15
		0.1	0.915		1.12e-2	0.15	0.42	15
30	homo.	10	1.12	4.27e-2	4.53e-2	14.92	15.22	30
		5	1.16		4.62e-2	5.44	6.38	30
		1	1.23		4.78e-2	0.99	1.32	30
		0.1	1.32		4.99e-2	0.10	0.14	30
	hetero.	10	1.02	7.18e-3	7.25e-3	14.22	14.61	30
		5	1.026		7.27e-3	5.92	6.59	30
		1	1.036		7.31e-3	1.29	1.77	30
		0.1	1.05		7.36e-3	0.082	0.17	30
	heter. rnd.	10	1.037	1.26e-2	1.28e-2	10.79	11.62	30
		5	1.047		1.29e-2	4.82	5.48	30
		1	1.065		1.30e-2	0.95	1.35	30
		0.1	1.085		1.31e-2	0.14	0.31	30
100	homo.	10	0.93	2.41e-2	2.32e-2	10.39	9.94	100
		5	0.951		2.30e-2	5.81	5.47	100
		1	0.89		2.28e-2	2.35	2.13	100
		0.1	0.85		2.23e-2	0.16	0.14	100
	hetero.	10	0.995	1.89e-3	1.88e-3	15.92	16.16	100
		5	0.993		1.88e-3	4.17	4.31	100
		1	0.991		1.87e-3	1.56	1.68	100
		0.1	0.988		1.87e-3	0.16	0.18	100
	heter. rnd.	10	0.986	4.85e-3	4.82e-3	11.02	11.07	100
		5	0.983		4.81e-3	6.45	6.48	100
		1	0.979		4.80e-3	4.39	4.48	100
		0.1	0.972		4.79e-3	0.81	0.90	100

Table 4 (cont). Estimated and experimentally obtained probabilities of β error for individual tests on the slope. Predicted sample size to achieve the α and β probabilities of error for each data set.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

To detect significant differences between the estimated probabilities of β error and the values from the simulation process, paired *t*-tests [22] (with α=1%) were applied on the β error values obtained for the different number of data points (since it is the most critical factor for achieving good predictions of β probabilities of error) at the same level of significance. In this way significant differences between the values in the β_{exp} and β_{pred} columns were found only in the data sets with five data pairs for the slope and intercept at the four levels of significance. The possible sources of error and some important observations concerning the results from the simulation process can be summarised as follows:

(i) In most cases the predicted probabilities of β error from Eqs. (3-4) are higher than the experimental values from the simulation process. This overestimation may be due to a lack of information, since the overestimation is higher in those data sets with fewer data pairs (where the experimental error, and thus the uncertainty of the regression coefficient is higher [23]), and lower in those data sets with a larger number of points. In this latter case however, small disagreements still exist due to the assumption of the normality of the regression coefficients. Figure 3 plots the differences between the experimentally-obtained probabilities of β error (from the simulation process) and the predicted probabilities against the number of data pairs of each data set for the slope and intercept with a 5% α level of significance.

(ii) Results for the intercept show a higher agreement than the ones for the slope (Figure 3). This may be because estimating the slope is more complex since two different distributions have to be considered for b_{H_0} and b_{H_1}, whereas only one is needed when the probabilities of β error are estimated for the intercept, as $\hat{s}_{a_{H_0}} = \hat{s}_{a_{H_1}}$.

(iii) There is no clear relationship between the uncertainty patterns and the error made in predicting the β error (in percent) for the different simulated data sets. As

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Figure 3 shows, the three lines depicting the three patterns of uncertainty do not maintain a constant relative position as they cross each other. Results for the intercept seem to follow a steadier pattern for the different uncertainties. As previously stated, the number of data pairs on the regression line is the key factor for obtaining a better estimate of the β error.

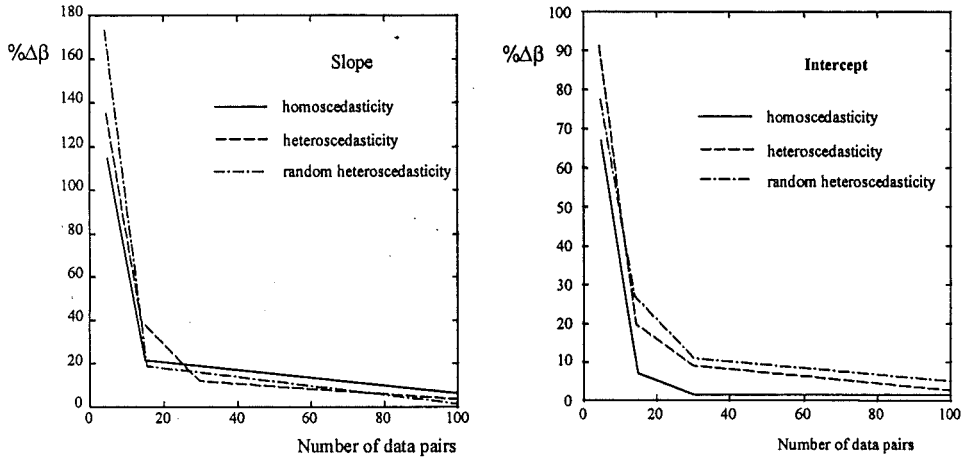


Figure 3. Difference between the experimentally-obtained probabilities (simulation process) and the predicted probabilities of β error for the slope and the intercept (in percent) in relation to the number of data pairs for each data set.

(iv) Results from the predicting the probabilities of β error (Eqs. (3-4)) and sample size for data sets with a high linear range were identical to the ones with a low linear range. Results shown in Tables 3 and 4 correspond to the low linear range, while the ones from the high linear range have been omitted. These results can be explained because the distribution of the data pairs in data sets (for a given uncertainty and number of data pairs) with different linear ranges is identical. So the only difference between data sets with different linear ranges is that the values of the individual data pairs and their respective uncertainties (taken as standard deviations) are ten times higher in the high linear range than in the low linear

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

range. Only the standard deviation values for the intercept were exactly ten times higher in the high linear range than the ones in the low linear range. This is due to the direct dependence of the standard deviation for the intercept on the sum of the x-axis values (Eq. (5)).

If we look at the results of estimating the sample size in Tables 3 and 4 (n_{pred} columns), we can see that the predicted results in all cases provide the number of data pairs of the different initial data sets considered. From these results we can conclude that the expressions for estimating the sample size provide correct results for the three kinds of distribution of uncertainties considered.

Procedure for β error and estimation of sample size in a real data set

Table 5 summarises the results of estimating the probabilities of committing a β error in the individual tests for the BLS slope and intercept at a 5% α level of significance (β column, in percent) for data set 3. Columns $\left|a_{H_0} - \hat{a}\right|$ and $\left|b_{H_0} - \hat{b}\right|$ show the distance between the estimated regression coefficients and the reference values ($a_{H_0} = 0$ and $b_{H_0} = 1$). The columns $t \cdot \hat{s}_{a_{H_0}}$ and $t \cdot \hat{s}_{b_{H_0}}$ (α=5%) show the values of the confidence intervals associated to the reference values. Columns a_{H_1} and b_{H_1} represent the bias that the experimenter wants to check in the regression coefficient being tested. Bias is detected in the regression coefficient whenever the difference $\left|a_{H_0} - \hat{a}\right|$ and $\left|b_{H_0} - \hat{b}\right|$ is higher than its associated confidence interval. Probabilities of β error are not calculated if bias is detected.

Table 5 shows that neither constant nor proportional bias are found in the SIA methodology in the analysis of Ca(II) in water according to the results from the three regression techniques. The highest probabilities of β error are estimated at 62.5% for the OLS technique, due to the highest standard deviation value. On the

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L.: T-353-2008

other hand, the probabilities of β error for BLS and WLS are lower and similar to each other although the WLS intercept value is nearer the upper confidence interval limit. This means that the results are less reliable, although this is not reflected in the estimated probabilities of β error. Results for the slope show that the estimated probabilities of β error in the three cases are very similar, despite the differences in the slope values from the three regression methods. However, if we look at the slope values we can be more confident about the accuracy of the one estimated by the BLS method as it is the closest to the reference value b_{H_0}.

	$\lvert a_{H_0} - \hat{a} \rvert$	$t \cdot \hat{s}_{a_{H_0}}$	a_{H_1}	β
BLS	2.94	5.35		40.2
WLS	4.38	5.19	6	37.6
OLS	3.97	7.11		62.5
	$\lvert b_{H_0} - \hat{b} \rvert$	$t \cdot \hat{s}_{b_{H_0}}$	b_{H_1}	β
BLS	0.0364	0.0991		2.77
WLS	0.0571	0.100	1.2	2.60
OLS	0.0656	0.110		5.30

Table 5. Results obtained in estimating the β error in the individual tests for the intercept and the slope in data set 3.

The iterative process for estimating the sample size to achieve the calculated probabilities of β error in the slope (2.77%) and intercept (40.2%) for a 5% α level of significance is shown in Table 6. For the intercept, starting with an initial data set of five data pairs (n_{a_0} column), thirteen iterations were needed to end up with twenty-seven data pairs. For the slope, twenty-six data pairs were needed to achieve convergence and there was no estimate of the data pairs until 13 had been considered (n_{b_0} column) since, according to the denominator of Eq. (10), high

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

experimental errors may produce negative estimates of sample size for the slope (denoted by <0 in Table 6).

iteration	n_{b_0}	$\hat{s}_{b_{H_0}}$	$\hat{s}_{b_{H_1}}$	n_{b_f}	n_{a_0}	$\hat{s}_{a_{H_0}}$	n_{a_f}
1	5	0.0974	0.0992	<0	5	6.369	9
2	9	0.131	0.134	<0	9	3.694	11
3	13	0.0753	0.0769	18	11	3.511	13
4	18	0.0666	0.0678	22	13	3.728	16
5	22	0.0609	0.0622	24	16	3.403	18
6	24	0.0530	0.0542	25	18	3.391	20
7	25	0.0511	0.0522	26	20	3.199	22
8	26	0.0492	0.0502	26	22	3.103	23
9					23	3.103	24
10					24	2.954	25
11					25	2.887	26
12					26	2.838	27
13					27	2.657	27

Table 6. Iterations during estimation of the sample size for the slope and the intercept performed in data set 3.

CONCLUSIONS

The results of this paper show that, in spite of the non-normality of the distributions of the BLS regression coefficients, the errors made in calculating the confidence intervals for the BLS regression coefficients are lower than the ones made with OLS or WLS techniques for data with uncertainties in both axes. Thus, the probabilities of β error in the individual tests on the BLS regression coefficients can be estimated under the hypothesis of normality.

We have also demonstrated that the expressions for estimating the probabilities of committing a β error when testing an individual regression coefficient with the BLS regression technique and considering different distributions for the reference

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

(a_{H_0} or b_{H_0}) and for the biased (a_{H_1} or b_{H_1}) regression coefficients, provide correct results. Some sources of error have also been detected and identified to explain the disagreements produced in validating the results. The number of data pairs of the regression line appear to be crucial for better estimating the probabilities of β error. In addition, results in real data show that in some cases it may be interesting to calculate the probabilities of β error not with the set α threshold value, but with the maximum α level of significance for which no bias is detected in the regression coefficient. One would be more confident of the regression coefficient value being accurate than when it falls near one of the boundaries of the confidence interval (in this way the α probabilities of error would be higher but the β probabilities of error would be lower than in the usual way).

Finally, we found that it is advisable to estimate the sample size, since it allows the experimenter to control the probabilities of committing α and β errors that they consider reasonable for the analytical problem in question. The iterative process for estimating the sample size guaranteed the chosen probabilities of making α and β errors when an individual test is applied to one of the estimated BLS coefficients and produced correct results for those data sets with moderate heteroscedasticity, but not for those with high heteroscedasticity. The experimenter also has to weigh up the pros and cons of performing the discontinuous series of experiments that this iterative procedure requires.

ACKNOWLEDGMENTS

We would like to thank the DGICyT (project no. BP96-1008) for financial support, and the Rovira i Virgili University for providing a doctoral fellowship to A. Martínez and F. J. del Río.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

BIBLIOGRAPHY

1.- W.A. Fuller, Measurement Error Models, John Wiley & Sons, New York, 1987.

2.- R.L. Anderson, Practical Statistics for Analytical Chemists, Van Nostrand Reinhold, New York, 1987.

3.- M.A. Creasy, Confidence limits for the gradient in linear in the linear functional relationship, J. Roy. Stat. Soc. B 18 (1956) 65-69.

4.- J. Mandel, Fitting straight lines when both variables are subject to error, J. Qual. Tech. 16 (1984) 16 1-14.

5.- C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, D.L. Massart, Detection of bias in method comparison by regression analysis, Anal. Chim. Acta 338 (1997) 19-40.

6.- J.M. Lisý, A. Cholvadová, J. Kutej, Multiple straight-line least-squares analysis with uncertainties in all variables, Comput. Chem. 14 (1990) 189-192.

7.- J. Riu, F.X. Rius, Univariate regression models with errors in both axes, J. Chemom. 9 (1995) 343-362.

8.- J. Riu, F.X. Rius, Assessing the accuracy of analytical methods using linear regression with errors in both axes, Anal. Chem. 68 (1996) 1851-1857.

9.- A.H. Kalantar, R.I. Gelb, J.S. Alper, Biases in summary statistics of slopes and intercepts in linear regression with errors in both variables, Talanta 42 (1995) 597-603.

10.- Cetama, Statistique appliquée à l'exploitation des mesures, 2nd ed., Masson, Paris, 1986.

11.- G. Kateman and L. Buydens, Quality Control in Analytical Chemistry, 2nd ed., John Wiley & Sons, New York, 1993.

12.- M. Meloun, J. Militký and M. Forina, Chemometrics for Analytical Chemistry. Volume 1: PC-aided statistical data analysis, Ellis Horwood ltd., Chichester, 1992.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

13.- M.R. Spiegel, Theory and Problems of Statistics; McGraw-Hill, New York, 1988.

14.- O. Güell, J.A. Holcombe, Analytical applications of Monte Carlo techniques, Anal Chem. 60 (1990) 529A - 542A.

15.- J.J. Langenfeld, S.B. Hawthorne, D.J. Miller, J. Pawliszyn, Role of modifiers for analytical-scale supercritical fluid extraction of environmental samples, Anal. Chem. 66 (1994) 909-916.

16.- I. Saouter, B. Blattmann, Analyses of organic and inorganic mercury by atomic fluorescence spectrometry using a semiautomatic analytical system, Anal. Chem. 66 (1994) 2031-2037.

17.- I. Ruisánchez, A. Rius, M.S. Larrechi, M.P. Callao, F.X. Rius, Automatic simultaneous determination of Ca and Mg in natural waters with no interference separation, Chemom. Intell. Lab. Syst. 24 (1994) 55-63.

18.- R. Boqué, F.X. Rius, D.L. Massart, Straight line calibration: something more than slopes, intercepts and correlation coefficients, J. Chem. Educ. (Comput. Ser.) 71 (1994) 230-232.

19.- B.D. Ripley, M. Thompson, Regression techniques for the detection of analytical bias, Analyst 112 (1987) 337-383.

20.- P.J. Ogren, J.R. Norton, Applying a simple linear least-squares algorithm to data with uncertainties in both variables, J. Chem. Educ. 69 (1992) 130-131.

21.- V. López-Ávila, R. Young, F.W. Beckert, Microwave-assisted extraction of organic compounds from standard reference soils and sediments, Anal. Chem. 66 (1994) 1097-1106.

22.- D. L. Massart, B.M.G. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.

23.- G.J. Hahn, W. Q. Meeker. Statistical Intervals, a guide for practitioners, John Wiley & Sons, New York, 1991.

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Capítol 4

Comparació de dos mètodes analítics mitjançant regressió lineal considerant errors en dos eixos

UNIVERSITAT ROVIRA I VIRGILI
REGRESSIÓ LINEAL AMB ERRORS EN AMBDÓS EIXOS. APLICACIÓ A LA CALIBRACIÓ I A LA COMPARACIÓ DE MÈTODES ANALÍTICS
Jordi Riu Rusell
ISBN:978-84-691-1897-9/ D.L: T-353-2008

Un cop comprovada la no-normalitat en la distribució dels coeficients de regressió trobats mitjançant el mètode BLS (malgrat que aquesta hipòtesi pot ser acceptada en les condicions normals de treball), i després del desenvolupament i de la caracterització dels intervals de confiança individuals per a l'ordenada a l'origen i el pendent de la recta de regressió, útils en la detecció d'errors sistemàtics proporcionals o constants, en aquest capítol es presenta un test estadístic per comparar dos mètodes analítics a diversos nivells de concentració: el test conjunt per a l'ordenada a l'origen i el pendent. Aquest test es basa en la comparació simultània de l'ordenada a l'origen i el pendent de la recta de regressió obtinguda representant els resultats proporcionats pels dos mètodes en comparació, amb els seus valors teòrics. És una adaptació del test desenvolupat el 1957 per Mandel i Linnig, però considerant els errors presents en els dos mètodes analítics en comparació, és a dir, emprant els coeficients de la recta de regressió trobada amb el mètode BLS. En aquest capítol es detallarà aquesta aplicació i el desenvolupament teòric es pot trobar a l'article *Assessing the Accuracy of Analytical Methods Using Linear Regression with Errors in Both Axes*, publicat a la revista *Analytical Chemistry*, que és el primer dels tres que es troben en aquest capítol. Per tal de facilitar-ne la implantació en la comunitat científica, es va desenvolupar un programa d'ordinador que està descrit a l'article *Method Comparison Using Regression with Uncertainties in Both Axes*, publicat a la revista *Trends in Analytical Chemistry*, que és el segon article exposat en aquest capítol. Per últim, per tal de comparar el test conjunt per a l'ordenada a l'origen i el pendent considerant errors en dos eixos amb altres eines estadístiques també aplicades en processos de comparació de mètodes analítics, es va comparar el primer amb el test conjunt per a l'ordenada a l'origen i el pendent -abans esmentat, desenvolupat per Mandel i Linnig-, amb els tests individuals de De la Guardia i col·laboradors, i amb el test simultani d'una hipòtesi composta (*simultaneous test of a composite hypothesis*) aplicats a la comparació de dos models multivariants per a la determinació del percentatge d'etilè en copolímers de poli (propilè-etilè). Aquest treball està descrit a l'article *Detection of bias in method-comparison studies*,