

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

6 1656-12960  
0130-38660

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

TESI-RIU

Tesi Doctoral

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI  
BIBLIOTECA



1700205818



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 RUSSELL**

per assolir el grau de

Doctor en Química

Tarragona, 1999

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

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 signifiqués 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.

## ÍNDEX

|   |    |
|---|----|
| Objecte de la Tesi Doctoral.  | 1  |
| Capítol 1. Introducció general.   | 5  |
| 1.1 Notació.  | 7  |
| 1.2 Comparació de mètodes. Traçabilitat.  | 10 |
| 1.3 Perspectiva històrica. Definicions relacionades amb la regressió lineal.                        | 13 |
| 1.3.1 Regressió lineal univariant.  | 14 |
| 1.3.2 Regressió lineal multivariant.  | 21 |
| 1.4 Regressió lineal considerant errors en tots els eixos.  | 24 |
| 1.5 Aspectes conceptuals relacionats amb la regressió lineal considerant errors en tots els eixos.  | 25 |
| 1.5.1 Possible falta d'adequació dels valors experimentals al model.                                | 25 |
| 1.5.2 Existència de valors discrepants en els valors experimentals.                                 | 26 |
| 1.5.3 Tipus de distribució dels valors experimentals i dels coeficients de la recta de regressió.   | 28 |
| 1.5.4 Existència de l'error de segona espècie, $\beta$ .  | 30 |
| 1.5.5 Relació entre els errors $\alpha$ i $\beta$ .   | 31 |
| 1.6 Aplicacions de la regressió lineal considerant errors en tots els eixos.                        | 33 |
| 1.6.1 Detecció d'errors constants o proporcionals.  | 33 |
| 1.6.2 Comparació de mètodes.  | 33 |
| 1.6.3 Predicció.  | 34 |
| 1.7 Estructura de la Tesi.  | 34 |
| 1.8 Referències.  | 36 |
| Capítol 2. Regressió lineal univariant considerant errors en dos eixos. Estat actual de la qüestió. | 39 |
| 2.1 Necessitat de la regressió univariant considerant errors en dos eixos.                          | 41 |
| 2.2 Relació constant de variàncies i regressió ortogonal.   | 43 |
| 2.3 Regressió amb errors heteroscedàstics en ambdós eixos.  | 47 |
| 2.4 Referències.  | 53 |
| Article: <i>Univariate Regression Models with Errors in Both Axes.</i>                              | 55 |

---

|  |     |
|--|-----|
| 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. | 95  |
| 3.1 Comprovació de la normalitat en les distribucions de la ordenada a l'origen i el pendent en regressió lineal considerant errors en dos eixos.                                      | 98  |
| 3.1.1 Mètode de Cetama.  | 99  |
| 3.1.2 Test de Kolmogorov.  | 106 |
| 3.1.3 Gràfiques de probabilitat normal.  | 106 |
| 3.2 Interval de confiança individuals per a l'ordenada a l'origen i el pendent.  | 108 |
| 3.3 Error $\beta$ aplicat als interval de confiança individuals per a la ordenada a l'origen i el pendent tenint en compte l'error màxim fixat.  | 111 |
| 3.4 Referències.   | 116 |
| Article: <i>Detecting Proportional and Constant Bias in Method Comparison Studies by Using Linear Regression with Errors in both Axes.</i>   | 117 |
| Capítol 4. Comparació de dos mètodes analítics mitjançant regressió lineal considerant errors en dos eixos.  | 147 |
| 4.1 Regressió lineal aplicada a la comparació de mètodes analítics.  | 150 |
| 4.2 Test conjunt per a l'ordenada en l'origen i el pendent de la recta de regressió.   | 156 |
| 4.3 Referències.   | 161 |
| Article: <i>Assessing the Accuracy of Analytical Methods Using Linear Regression with Errors in Both Axes.</i>   | 162 |
| Article: <i>Method Comparison Using Regression with Uncertainties in Both Axes.</i>  | 184 |
| Article: <i>Detection of Bias in Method-Comparison Studies.</i>  | 194 |
| Capítol 5. Regressió lineal multivariant considerant errors en tots els eixos. Aplicació a la comparació de múltiples mètodes.   | 207 |
| 5.1 Regressió multivariant amb errors heteroscedàstics individuals en tots els eixos.  | 210 |
| 5.2 Test conjunt per a l'ordenada a l'origen i la suma de pendents de l'hiperplà de regressió.   | 213 |
| 5.3 Referències.   | 217 |

---

|   |     |
|---|-----|
| Article: <i>Multiple Analytical Method Comparison Using Regression with Uncertainties in All Axes.</i>            | 218 |
| Capítol 6. Predicció de les variables resposta i predictor a en regressió lineal considerant errors en dos eixos. | 245 |
| 6.1 Intervals de confiança considerant només errors en la variable resposta.                                      | 248 |
| 6.2 Intervals de confiança considerant errors en les variables predictor a i resposta.                            | 258 |
| 6.3 Referències.  | 261 |
| Article: <i>Confidence Intervals in Linear Regression Taking into Account Uncertainties in Both Axes.</i>         | 264 |
| Capítol 7. Conclusions.   | 287 |
| 7.1 Conclusions generals.   | 289 |
| 7.2 Conclusions del capítol 3.  | 294 |
| 7.3 Conclusions del capítol 4.  | 294 |
| 7.4 Conclusions del capítol 5.  | 295 |
| 7.5 Conclusions del capítol 6.  | 296 |
| 7.6 Perspectives de futur.  | 296 |
| Apèndix   | 299 |
| Glossari  | 307 |

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**

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  $\beta$  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

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<sub>k</sub>*).

### *Símbols començant amb una lletra de l'alfabet llatí*

|             |   |
|-------------|---|
| <i>a</i>    | 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      |
| $\bar{b}$   | valor vertader del pendent de la recta de regressió                         |
| $\hat{b}$   | estimació del pendent de la recta de regressió                              |
| $b_j$       | valor vertader del <i>j</i> -èssim pendent de l'hiperplà de regressió       |
| $\hat{b}_j$ | estimació del <i>j</i> -èssim pendent de l'hiperplà de regressió            |
| <b>b</b>    | vector amb els valors vertaders dels coeficients de regressió               |

|                      |   |
|----------------------|---|
| $\hat{\mathbf{b}}$   | vector estimació dels coeficients de regressió  |
| $e_i$                | valor residual en el punt $i$   |
| $\mathbf{e}$         | vector d'errors   |
| $F_{v1,v2,\alpha}$   | valor de la distribució $F$ de Fischer per a un nivell de significança $\alpha$ (1 cua) amb $v1$ i $v2$ graus de llibertat  |
| $F_{v1,v2,\alpha/2}$ | valor de la distribució $F$ de Fischer per a un nivell de significança $\alpha$ (2 cues) amb $v1$ i $v2$ 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 $i$  |
| $t_{\alpha,v}$       | valor de la distribució $t$ de Student per a un nivell de significança $\alpha$ (1 cua) amb $v$ graus de llibertat          |
| $t_{\alpha/2,v}$     | valor de la distribució $t$ de Student per a un nivell de significança $\alpha$ (2 cues) amb $v$ graus de llibertat         |
| $u_{\alpha/2}$       | valor de la distribució normal per a un nivell de significança $\alpha$   |
| $w_i$                | 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 $i$   |

|              |   |
|--------------|---|
| $\bar{x}$    | 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*

|                  |   |
|------------------|---|
| $\alpha$         | error de primera espècie o de tipus I   |
| $\beta$          | error de segona espècie o de tipus II   |
| $\Delta$         | biaix, màxima diferència acceptable entre el paràmetre estimat i un valor de referència   |
| $\eta_p$         | moment d'ordre $p$  |
| $\mu_{x_i}$      | valor vertader de la variable predictora en el punt $i$                                   |
| $\mu_{y_i}$      | valor vertader de la variable resposta en el punt $i$                                     |
| $\lambda$        | relació entre els errors de les variables resposta i predictora en regressió CVR          |
| $\rho_i$         | valor vertader de la covariància entre les variables predictora i resposta en el punt $i$ |
| $\sigma_{x_i}^2$ | valor vertader de la variància de la variable predictora en el punt $i$                   |
| $\sigma_{y_i}^2$ | valor vertader de la variància de la variable resposta en el punt $i$                     |

## 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,<sup>1,2</sup> “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, sensibilitat, límit de detecció, límit de quantificació, robustesa, exactitud i precisió.”

L'exactitud, segons les normes ISO,<sup>3</sup> 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.<sup>2</sup> 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”.<sup>4,5</sup> 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<sup>6</sup> exigeixen que cada resultat analític vagi acompanyat de dos paràmetres de qualitat bàsics: traçabilitat i incertesa. Malgrat aquesta duplicitat

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 dona.<sup>5</sup>

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:<sup>7</sup>

- 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.

La referència següent, en ordre d'importància metro lògica, 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,<sup>8</sup> 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<sup>8</sup> 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 metroològica 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.<sup>9</sup> 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).<sup>10</sup>

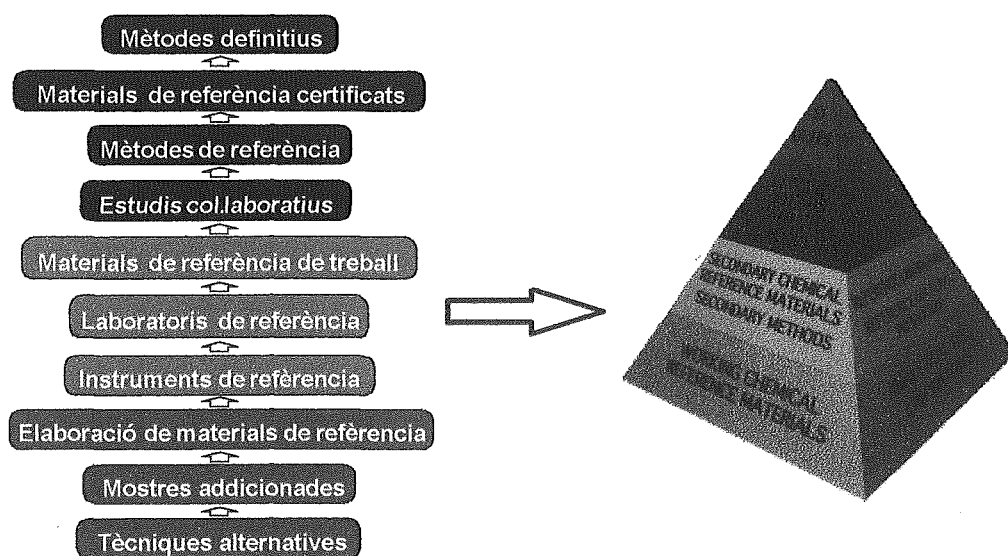


Figura 1.1. Relació entre diverses referències i la piràmide metrollògica.

### 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

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.<sup>11,12</sup> 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:

$$y_i = a + bx_i + e_i \quad (1.1)$$

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:

$$\hat{y}_i = \hat{a} + \hat{b}x_i \quad (1.2)$$

El valor residual,  $e_i$ , correspon a:

$$e_i = |y_i - \hat{y}_i| = |y_i - \hat{a} - \hat{b}x_i| \quad (1.3)$$

Rigorosament, el mètode de mínims quadrats és només aplicable si es compleixen les condicions següents:<sup>13,14</sup>

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

- 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:<sup>13,15</sup>

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

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:

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

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

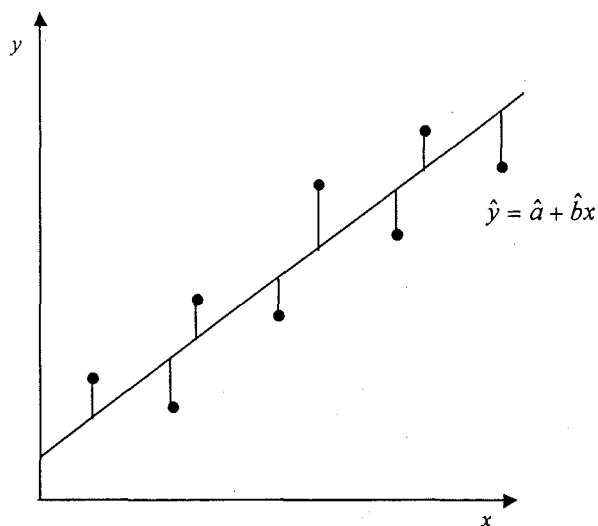


Figura 1.2 Il·lustració del mètode de mínims quadrats.

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:

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

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

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:

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

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:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (1.10)$$

$$\begin{array}{c} 1 \\ \boxed{\mathbf{y}} \\ n \end{array} = \begin{array}{c} 2 \\ \boxed{\mathbf{X}} \\ n \end{array} \begin{array}{c} 1 \\ \boxed{\mathbf{b}} \\ 2 \end{array} + \begin{array}{c} 1 \\ \boxed{\mathbf{e}} \\ n \end{array}$$

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 predictor*, formada per una columna d'1 i una columna amb els valors de la variable predictor;  $\mathbf{b}$  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 el pendent es pot trobar segons:

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

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 predictor com a lliure d'error, es pot aplicar el mètode de mínims quadrats ponderats (*weighted least squares*, WLS). Aquest mètode permet

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:

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

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:

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

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:

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

$$\hat{a} = \bar{y}_p - b\bar{x}_p \quad (1.15)$$

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

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

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

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

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

on  $\mathbf{V}_{nn}$  és una matriu diagonal en què l'element  $i$  de la diagonal correspon a la variància del punt  $i$  de la variable resposta ( $s_{y_i}^2$ ). És important puntualitzar que si la variància de la variable resposta és igual a la unitat en tots els punts ( $s_{y_i}^2 = 1$ ), 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

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.<sup>16</sup>

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).<sup>9,17</sup> 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 ( $s_{y_i}^2$ ), 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 ( $\text{cov}(y_i, y_k)$ ). L'error experimental ara pren l'expressió matricial següent:

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

Cal assenyalar que si les correlacions entre els errors dels diferents punts de la variable resposta són tots zero ( $\text{cov}(y_i, y_k) = 0$ ), 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 ( $x_j, j=1 \dots p$ ) segons el model següent:

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \dots + b_p x_{pi} + e_i \quad (1.20)$$

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

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

$$y_i = \hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \dots + \hat{b}_p x_{pi} \quad (1.21)$$

on la variable  $\hat{a}$  representa l'estimació de l'ordenada a l'origen de la recta de regressió i les variables  $\hat{b}_j$  ( $j=1\dots p$ ) 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:

$$e_i = \left| y_i - \hat{a} - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i} - \dots - \hat{b}_p x_{pi} \right| \quad (1.22)$$

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:<sup>9</sup>

$$S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left( y_i - \hat{a} - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i} - \dots - \hat{b}_p x_{pi} \right)^2 \quad (1.23)$$

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.

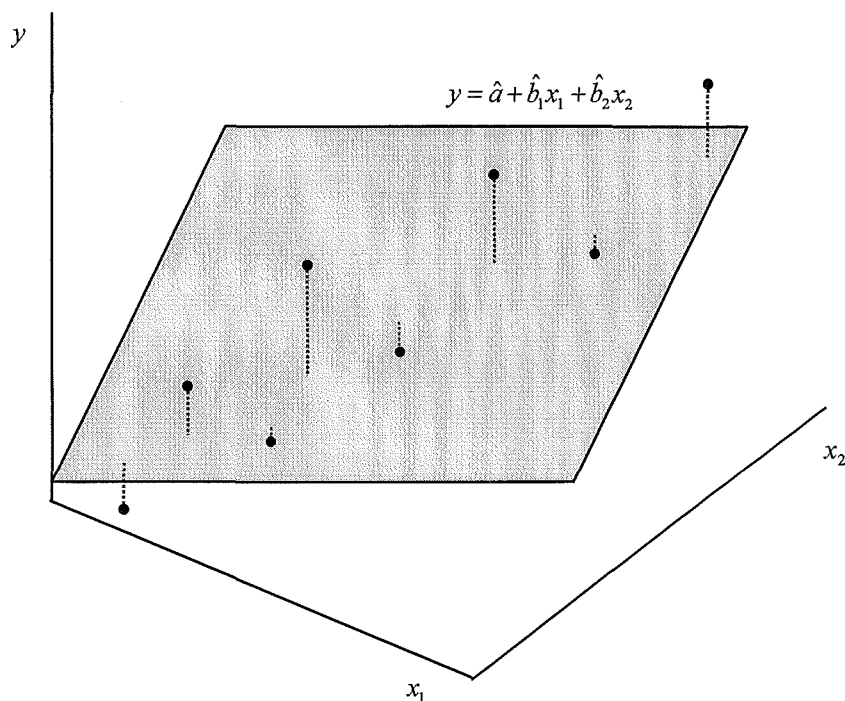


Figura 1.3 Il·lustració del mètode de regressió múltiple lineal.

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:

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

$$\begin{array}{c} 1 \\ \boxed{\mathbf{y}} \\ n \end{array} = \begin{array}{c} p+1 \\ \boxed{\mathbf{X}} \\ n \end{array} \begin{array}{c} 1 \\ \boxed{\mathbf{b}} \\ p+1 \end{array} + \begin{array}{c} 1 \\ \boxed{\mathbf{e}} \\ n \end{array}$$

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:

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

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

que fan que no puguin ser negligibles.<sup>18,19</sup> 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,<sup>20</sup> en l'estudi de reaccions cinètiques,<sup>21</sup> estimant la relació d'una mescla<sup>22</sup> o en alguns tipus de mesures espectroscòpiques.<sup>23</sup>

Mentre la regressió lineal considerant errors en dos eixos no és nova (ja el 1940 se'n troben articles<sup>24</sup>), 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

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,<sup>25</sup> 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  $\chi^2$ .<sup>13</sup>

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.<sup>26</sup> 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 ( $R^2$ ).<sup>27</sup> 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

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.<sup>28</sup>

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,<sup>29</sup> com pot ser per exemple el test de Cook,<sup>30,31</sup> 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.<sup>10</sup>

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.<sup>32</sup> 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

la recta de regressió.<sup>33</sup> 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 predictor 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<sup>34</sup> 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.<sup>35</sup> Clancey<sup>36</sup> 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

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.<sup>37</sup> Segons Box i col·laboradors,<sup>38</sup> 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).<sup>39</sup> Aquestes tècniques tenen l'avantatge que demanen càlculs molt simples, però són menys eficients i solen requerir més rèpliques que les tècniques paramètriques.<sup>40,41</sup> 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 predictor 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

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.<sup>42</sup> 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, $\beta$*

L'error  $\beta$ , 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.

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*

|                      |             | $H_0$ certa   | $H_0$ falsa    |
|----------------------|-------------|---------------|----------------|
|                      |             | acceptació    | rebuig         |
| <i>Situació real</i> | $H_0$ certa | correcta      | error $\alpha$ |
|                      | $H_0$ falsa | error $\beta$ | 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  $\alpha$ , mentre que si s'accepta una hipòtesi nul·la falsa es comet un error  $\beta$ .

### 1.5.5 Relació entre els errors $\alpha$ i $\beta$

Els errors  $\alpha$  (també anomenats de primera espècie o de tipus I) i  $\beta$  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  $\alpha$  és àmpliament emprat en els tests estadístics aplicats a l'anàlisi química. Potser per això hi ha una tendència a escollir els valors  $\alpha=0.05$  o  $\alpha=0.01$  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

*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 ...*”<sup>43</sup> En canvi, l’error  $\beta$  és poc conegut dins de l’anàlisi química. Fins i tot hi ha certs autors que suggereixen que la utilització de l’error  $\beta$  és massa complexa per als químics ordinaris.<sup>44</sup> El que és clar és que l’usuari pot decidir entre tenir en compte l’error  $\beta$  o no, però el fet d’ignorar-lo no fa que desaparegui.

En certs casos pot ser més important l’error  $\beta$  que l’ $\alpha$ . 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  $\beta$  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  $\alpha$  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  $\alpha$  baixos.<sup>45</sup> 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  $\beta$ .<sup>46</sup>

S’ha suggerit<sup>45</sup> que en aquells casos en què sigui més important l’error  $\beta$  que l’ $\alpha$ , 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.

### *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 predictor 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 predictor. 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 dona millors resultats que l'ús de tècniques de regressió que només tenen en compte errors en un sol eix, malgrat que

en aquests casos els coeficients de regressió segueixen la distribució normal. Així mateix, s'introdueix l'error  $\beta$  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.

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 of 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 Sciences*, **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

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. Barnett, 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)

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

## Capítol 2

---

**Regressió lineal univariant considerant errors en dos eixos.**

**Estat actual de la qüestió**

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 contempen 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

apartat 1.3, sota certes condicions la seva utilització pot donar lloc a l'estimació d'uns coeficients de regressió incorrectes. Scheffé<sup>1</sup> i Mandel<sup>2</sup> 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<sup>3</sup> 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  $\mu_{x_i}$  correspon al valor vertader de  $x_i$ , i  $\delta_i$  a l'error aleatori de  $x_i$ ) es determina amb el factor següent:

$$r = \frac{\sigma_{\delta}^2}{\sigma_{\mu_x}^2} \quad (2.1)$$

on  $\sigma_{\delta}^2$  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<sup>4</sup> 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:<sup>3,5</sup>

$$\hat{b} \approx \frac{1}{1+r} b \quad (2.2)$$

$$\hat{a} \approx a + \frac{r}{1+r} \bar{x} b \quad (2.3)$$

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.<sup>6</sup>

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)<sup>7,8</sup>, també anomenada regressió de *errors-in-variables*<sup>9,10</sup> 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 ( $\lambda$ ) 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

calcular els coeficients de la recta de regressió tenint en compte les assumpcions inherents a aquest mètode. L'aproximació de Mandel<sup>2</sup> 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<sup>11</sup> 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):

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

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad (2.5)$$

on:

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

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

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

i el paràmetre  $\lambda$  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.<sup>2</sup> El mètode és invariant respecte a un canvi d'eixos.

En el cas particular de la regressió CVR en el què s'assumeix el mateix error en les dues variables ( $\lambda=1$ ), 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)<sup>10</sup> o com regressió per mínims quadrats totals (*total least squares*, TLS).<sup>12</sup>

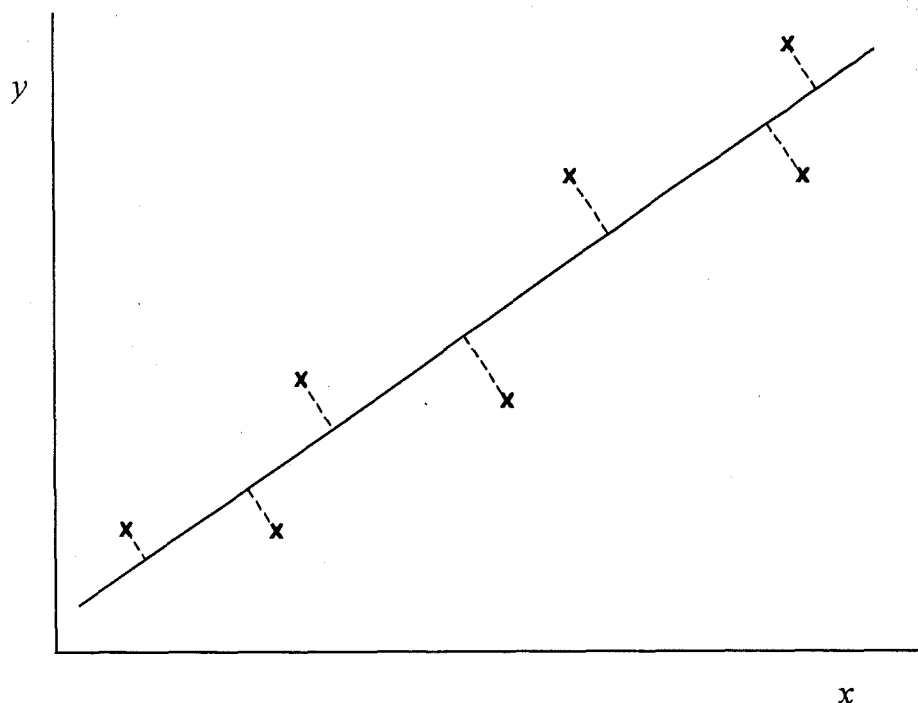


Figura 2.1. Representació de la regressió ortogonal.

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),<sup>13</sup> 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

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

Alguns autors<sup>6,14</sup> 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 predictor, 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 predictor caldrà aplicar altres tècniques que considerin errors en dos eixos.

Wackers i col·laboradors<sup>15</sup> 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<sup>16-24</sup> 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  $\beta$  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  $\alpha$  i  $\beta$ ,<sup>6</sup> i per calcular

els intervals de confiança associats a la predicció de la variable dependent per a l'anàlisi d'una mostra futura.<sup>2</sup>

## 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<sup>25</sup> i el 1949 Bartlett<sup>26</sup> 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:  $P_1 = (\bar{x}_1, \bar{y}_1)$ .
- Aquells amb els valors de la variable predictora més alts. Calcular la mitjana d'aquest conjunt per a les variables predictora i resposta:  $P_3 = (\bar{x}_3, \bar{y}_3)$ .
- 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:

$$\hat{b} = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1} \quad (2.9)$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad (2.10)$$

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 predictor 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 difícil, 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):

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

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 discontinües 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).

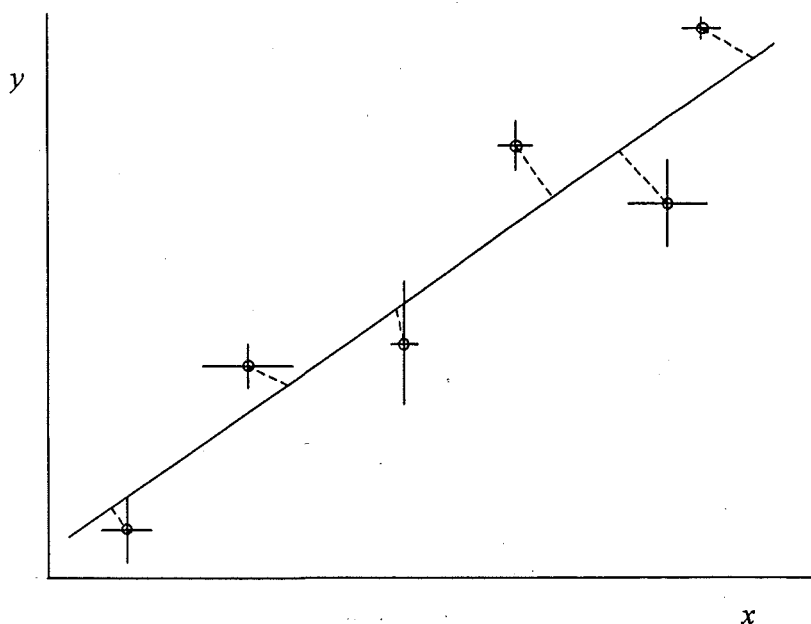


Figura 2.2. Representació de la regressió lineal considerant errors heteroscedàstics individuals en dos eixos.

Si es tornés a repetir l'anàlisi de cada punt experimental, la probabilitat d'obtenir un cert valor a cada punt és màxima a  $(\mu_{x_i}, \mu_{y_i})$ , 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 coincidirán amb els punts teòrics  $(\mu_{x_i}, \mu_{y_i})$ . Cada punt teòric segueix una distribució normal bivariant, amb funció de probabilitat definida segons:<sup>10</sup>

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

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 separatament presenten una distribució normal i si es fessin talls paral·lels al pla  $(x, y)$ , s'obtidrien les anomenades el·lipses o cercles d'isoprobabilitat. La inclinació de l'el·lipse ve donada pel paràmetre  $\rho_i$  a l'equació 2.12. Si  $\rho_i=0$  (és a dir, no hi ha correlació entre els valors individuals de les variables predictor 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 semieixos 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 predictor 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  $(\mu_{x_i}, \mu_{y_i})$  i la longitud de cada eix és proporcional a les desviacions estàndards de les variables predictor i resposta.

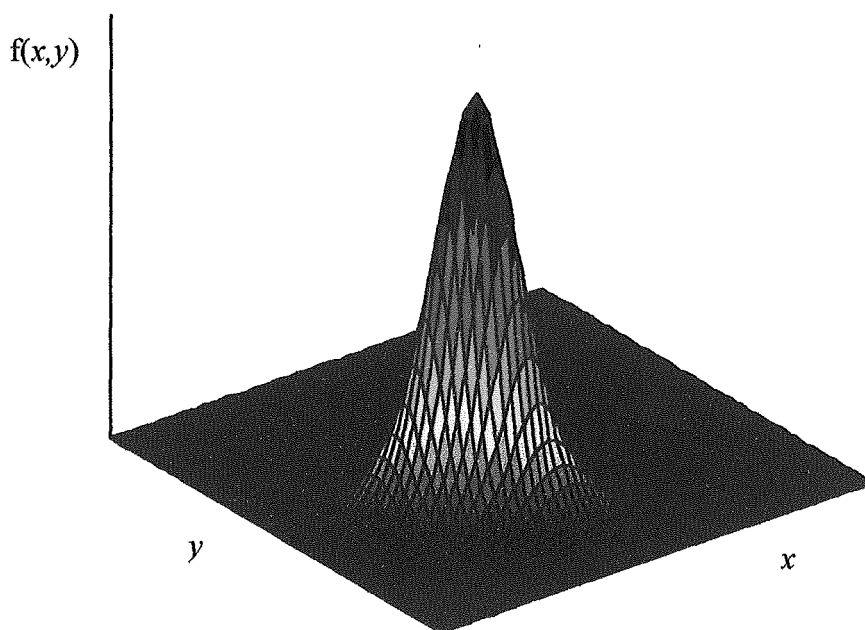


Figura 2.3. Distribució normal bivariant per a un parell de dades experimentals.

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<sup>25,27</sup> 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):

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

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

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

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.

## 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, *Analisis*, **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

18. D.V. Lindley, *Journal of the Royal Statistical Society Supplements*, **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)

## 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.

## 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:<sup>1,2</sup> (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 ( $b \cdot s_x^2$ ).<sup>3,4</sup> (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,<sup>5-7</sup> or non-parametric methods<sup>8,9</sup> 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.<sup>10,11</sup>

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.<sup>12</sup> 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)<sup>13</sup> 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.<sup>14,15</sup> 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.<sup>4</sup> 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,<sup>3,4,16</sup> 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

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<sup>17-20</sup> such as Huber's M-estimator,<sup>21</sup> Jackknife residuals,<sup>22</sup> standardized residuals, normalized residuals, predicted residuals,<sup>23</sup> studentized residuals,<sup>24,25</sup> recursive residuals or diagonal elements of the Hat matrix.<sup>26,27</sup> 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.<sup>28-34</sup> It is well known that the effect of outliers can be avoided by using robust calibration.<sup>17,28-36</sup> 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,<sup>37</sup> change points,<sup>38</sup> serial correlation<sup>39</sup> or heteroscedasticity.<sup>40-42</sup>

## 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

$$y_i = a + bx_i + \varepsilon_i \quad (1)$$

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.*<sup>43</sup> 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.<sup>44</sup> 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.<sup>45</sup> 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.*<sup>46</sup> have developed a straightforward generalization of their method for calculating the regression coefficients of a straight line when there are errors in both axes<sup>47</sup> (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

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

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:<sup>48</sup>

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

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<sup>49</sup>

$$w_i = s_{e_i}^2 = s_{y_i}^2 + b^2 s_{x_i}^2 \quad (4)$$

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:

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

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.

## 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,<sup>50</sup> 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

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

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.<sup>51</sup>

Reed<sup>49,52</sup> 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  $\alpha$ ,  $\beta$  and  $\gamma$  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

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

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<sup>49</sup> 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<sup>50</sup> 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.<sup>53</sup> 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<sup>54</sup> 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

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.<sup>55</sup>

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  $\chi^2$  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  $\chi^2$  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  $\chi^2$  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<sup>54</sup> 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.<sup>51</sup>

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,<sup>2,56</sup> and variances.<sup>56</sup> The results obtained by applying the Williamson method to a set of data,<sup>2,56</sup> 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.<sup>2</sup>

A similar approach, carried out by González *et al.*,<sup>16</sup> 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.*<sup>47</sup> 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)<sup>28</sup>:

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

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.*<sup>46</sup> 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,<sup>57</sup> 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.<sup>10,58-60</sup> 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,<sup>59</sup> instead of being evaluated in relation to experimental data. Furthermore, Irvin's development<sup>10</sup> fails when doing the derivation relative to  $b$  since they neglect the term  $dw_i/db$ .<sup>11</sup> 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  $S$ .<sup>61</sup> Several authors have stated<sup>2,11,61</sup> 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.*<sup>62</sup> 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:

$$w_i = s_{e_i}^2 = s_{y_i}^2 + b^2 s_{x_i}^2 - 2b \text{cov}(x_i, y_i) \quad (9)$$

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

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:

$$\mathbf{Rb} = \mathbf{g} \quad (10)$$

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

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

$$\mathbf{b} = \mathbf{R}^{-1} \mathbf{g} \quad (12)$$

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  $S/(n-2)$ .<sup>63,64</sup>

It should be pointed out that if one were to be in the situation in which  $s_{e_i}^2 = s_{y_i}^2$  (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.*<sup>65</sup> 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

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

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

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

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<sup>13,66</sup> 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<sup>67</sup> 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<sup>68</sup> uses a method developed by Jefferys<sup>69,70</sup> 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.<sup>45</sup> 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<sup>71</sup> 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.<sup>72</sup> 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<sup>73</sup> method calculates the derivatives numerically and only requires the function of regression from the user.<sup>74,75</sup>

Mandel<sup>3</sup> 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:  $u_i = x_i - ky_i$ ,  $v_i = y_i - bx_i$ .

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.

Cumming *et al.*<sup>76</sup> 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

$$px + qy + t = 0 \quad (15)$$

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

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

where

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

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

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

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

$$p^2 + q^2 = 1 \quad (20)$$

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

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.

| $i$ | $x$ | $1/s_x^2$ | $y$ | $1/s_y^2$ |
|-----|-----|-----------|-----|-----------|
| 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.<sup>77,78</sup> 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<sup>79-81</sup> based on the theory of fuzzy sets.<sup>82</sup> 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.<sup>83</sup>

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

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.

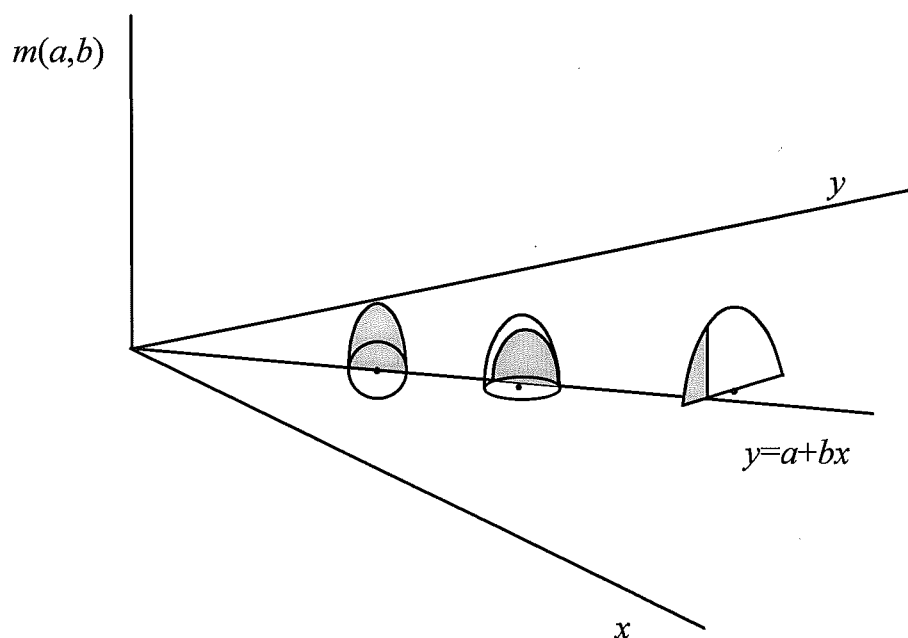


Figure 1. Graphical representation of fuzzy calibration: circle, ellipse and line are used as supports. The membership function  $m(a,b)$  is represented in the ordinate axis.

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

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

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.<sup>84-86</sup>

In his work, Otto and Bandemer<sup>79</sup> 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.*<sup>84</sup> 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.<sup>87</sup> So, it minimises the influence of outliers in the calculation of the regression coefficients and, at the same time, can be used to detect<sup>28</sup> their presence. It has the advantage that it works both for traditional calibration and for inverse calibration. Otto<sup>79</sup> 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.<sup>84</sup> 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<sup>34</sup> 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

by Rousseeuw and Leroy.<sup>33</sup> In this method, instead of minimizing a sum of squares, the median

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

is considered and minimized relative to the model parameters;  $\mu_x$  and  $\mu_y$  are the respective expectations of  $x$  and  $y$ , and  $(\mu_x, \mu_y)$  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

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

and the sign of  $b$  corresponds to the sign of

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

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.<sup>88</sup> 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<sup>89</sup> 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

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

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.<sup>6</sup>

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

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  $s_{y'}^2 / s_{x'}^2 = 1$ .

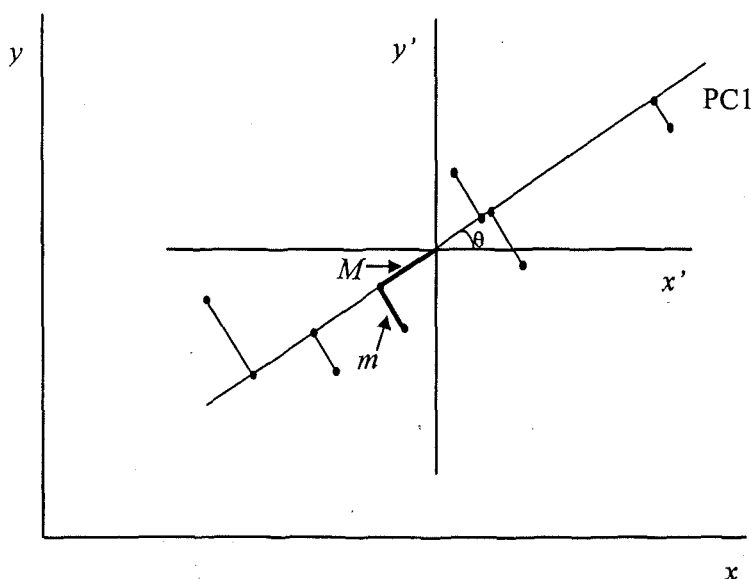


Figure 2. PCA applied to straight line calibration.

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  $\theta$  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  $\theta$  have been reported.<sup>5,6</sup>

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

where

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

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

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,

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).<sup>77,90</sup> The uncertainty associated to the slope and the intercept of the first principal component can be found in literature.<sup>90-92</sup> 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  $s_{y_i}^2 / s_{x_i}^2$ <sup>51,77,90</sup> 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.

## Results and Discussion

The regression coefficients when fitting a calibration line for the data set given by Pearson<sup>93</sup> in 1901 and weights given by York's paper<sup>50</sup> (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.<sup>68,94</sup> 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,<sup>93</sup> results which are appreciably different from each other are obtained for their variances of the regression coefficients.<sup>70,95-97</sup> 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.<sup>98</sup> A notable exception is Williamson's method<sup>54</sup> 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  $S/(n-2)$  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  $S/(n-2)$ , 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           | <i>a</i>   | <i>b</i>     | <i>S</i>      |
|---------------------------------|------------|--------------|---------------|
| 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 <i>et al.</i> method | 5.47991022 | -0.480533407 | 11.8663539487 |
| Neri's <i>et al.</i> method     | 5.47991022 | -0.480553403 | 11.8663555842 |
| Effective variance              | 5.39605212 | -0.46344885  | 11.956449080  |
| Lisý's <i>et al.</i> method     | 5.47991022 | -0.480533407 | 11.8663531941 |
| Brooks' <i>et al.</i> method    | 5.47991022 | -0.480533407 | 11.8663539487 |
| Lybanon                         | 5.47991025 | -0.480533415 | 11.8663531941 |
| Cumming's <i>et al.</i> 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<sup>54</sup> and Lisý *et al.*<sup>62</sup> 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 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

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.<sup>99</sup> 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 <sup>20</sup>                   | Equation (2)   | Equation (5)                                 | Yes. Heteroscedasticity                        | No                | Iterative      | Yes                                      | No                                   |
| Reed <sup>49</sup>                   | Equation (2)   | Equation (5)                                 | Yes. Heteroscedasticity                        | No                | Iterative      | Yes                                      | No                                   |
| Williamson <sup>54</sup>             | Equation (2)   | Equation (5)                                 | Yes. Heteroscedasticity                        | No                | Iterative      | Yes                                      | No                                   |
| González <i>et al.</i> <sup>16</sup> | Equation (3)   | Equation (5)                                 | Yes. Heteroscedasticity                        | No                | Iterative      | No                                       | No                                   |
| Neri <i>et al.</i> <sup>7</sup>      | Equation (12)  | Equation (5)                                 | Yes. Heteroscedasticity                        | No                | Iterative      | No                                       | No                                   |
| Press and Teukolsky <sup>7</sup>     | Equation (5)   | Equation (5)                                 | Yes. Heteroscedasticity                        | No                | Numerical      | No                                       | No                                   |
| Lisy <i>et al.</i> <sup>62</sup>     | $S = \frac{(y_i - a - bx_i)^2}{s_x^2 + b^2 s_y^2 - 2bcov(x_i, y_i)}$   | Equation (12)                                | Yes. Heteroscedasticity                        | No                | Iterative      | Yes                                      | No                                   |
| Brooks <i>et al.</i> <sup>65</sup>   | $S = \frac{(x_i - \bar{x}_i)^2 + (y_i - \bar{y}_i)^2}{s_x^2 + s_y^2}$  | Equation (14)                                | Yes. Heteroscedasticity                        | No                | Iterative      | Yes                                      | No                                   |
| Lybanof <sup>68</sup>                | $S = \frac{1}{2} \hat{v}^T \sigma^2 \hat{v}$   | $S = \frac{1}{2} \hat{v}^T \sigma^2 \hat{v}$ | Yes. Heteroscedasticity                        | No                | Iterative      | Yes                                      | No                                   |
| Mandel <sup>3</sup>                  | ( $\sigma$ , covariance matrix; $v$ , vector of residuals)<br>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 <sup>6</sup>                 | Equation (16)  | Equation (16)                                | Yes. Heteroscedasticity                        | No                | Iterative      | Yes                                      | No                                   |
| Fuzzy <sup>79-82, 84</sup>           | Equation (22)  | Equation (22)                                | Estimation                                     | Yes               | Several        | Yes                                      | Yes                                  |
| LSR <sup>4</sup>                     | Equation (23)  | Equation (23)                                | No   | Yes               | Simplex method | No                                       | No                                   |
| PCA <sup>6</sup>                     | Perpendicular from each point to the calibration line  | Perpendicular to the calibration line        | No   | No                | Numerical      | Yes                                      | Yes                                  |
| SMA <sup>51, 71, 90</sup>            | 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.

## 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, *Analisis* **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 Intl.* **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. Himsforth, *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. Technol.* **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).

## 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**

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 vàlidesa 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

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  $\beta$  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  $\alpha$  i  $\beta$ .

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

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<sup>1</sup> 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<sup>2</sup> 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  $p$  ( $\eta_p$ ) són els valors probables de  $(X - \bar{X})^p$ , on  $\bar{X}$  és la mitjana de la població de la variable  $X$ :

$$\eta_p = E(X - \bar{X})^p \quad (3.1)$$

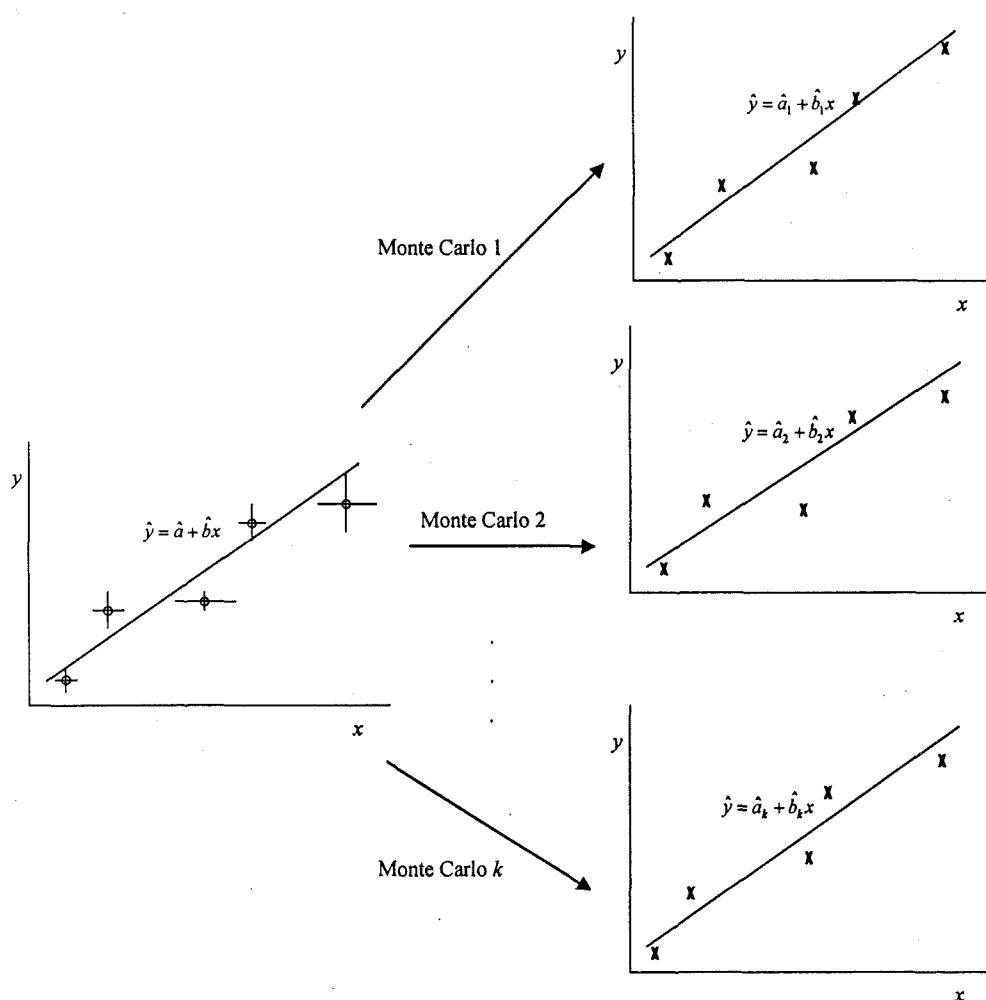


Figura 3.1. Procés d'obtenció de les  $k$  ordenades a l'origen i  $k$  pendents per estudiar la normalitat en les seves distribucions.

El moment d'ordre 1 correspon al valor mitjà, el moment d'ordre 2, a la variància de la població ( $\sigma^2$ ); el moment d'ordre 3, al coeficient d'asimetria (*skewness*,  $\eta_3/\sigma^3$ ), i el moment d'ordre 4, a la curtosi (*kurtosis*,  $\eta_4/\sigma^4$ ).<sup>3</sup> 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

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:

$$\gamma_1 = \frac{\eta_3}{\eta_2^{3/2}} = \frac{\eta_3}{\sigma^3} \quad (3.2)$$

$$\gamma_2 = \frac{\eta_4}{\eta_2^2} - 3 \quad (3.3)$$

on els coeficients  $\gamma_1$  i  $\gamma_2$  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:

$$\beta_1 = \frac{\eta_3^2}{\eta_2^3} = \gamma_1^2 \quad (3.4)$$

$$\beta_2 = \frac{\eta_4}{\eta_2^2} = \gamma_2 + 3 \quad (3.5)$$

on  $\beta_1$  varia entre 0 i  $\infty$ , 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:

$$g_1 = \frac{k_3}{k_2^{3/2}} \quad (3.6)$$

$$g_2 = \frac{k_4}{k_2^2} \quad (3.7)$$

i els coeficients de Pearson segons:

$$b_1 = g_1^2 = \frac{k_3^2}{k_2^3} \quad (3.8)$$

$$b_2 = g_2 + 3 \quad (3.9)$$

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.

|   |                                 |   |
|---|---------------------------------|---|
| $\gamma_1 = \beta_1 = 0$<br>corba simètrica | $\gamma_2 = 0$<br>$\beta_2 = 3$ | Corba normal                                |
|   | $\gamma_2 < 0$<br>$\beta_2 < 3$ | Corba limitada pel domini $\bar{x} \pm d_1$ |
|   | $\gamma_2 > 0$<br>$\beta_2 > 3$ | 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:

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

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  $\alpha$  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:

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

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  $\alpha$  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:

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

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

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:

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

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:

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

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

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:

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

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:

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

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:

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

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

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

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

### 3.1.2 Test de Kolmogorov

El test de Kolmogorov té dues variants, la gràfica i la numèrica.<sup>3</sup> 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:

$$D_i = S(x_i) - i/n \quad (3.19)$$

$S(x_i)$  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  $\alpha$  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

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,<sup>4</sup> 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.<sup>5</sup> 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

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 dona 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.

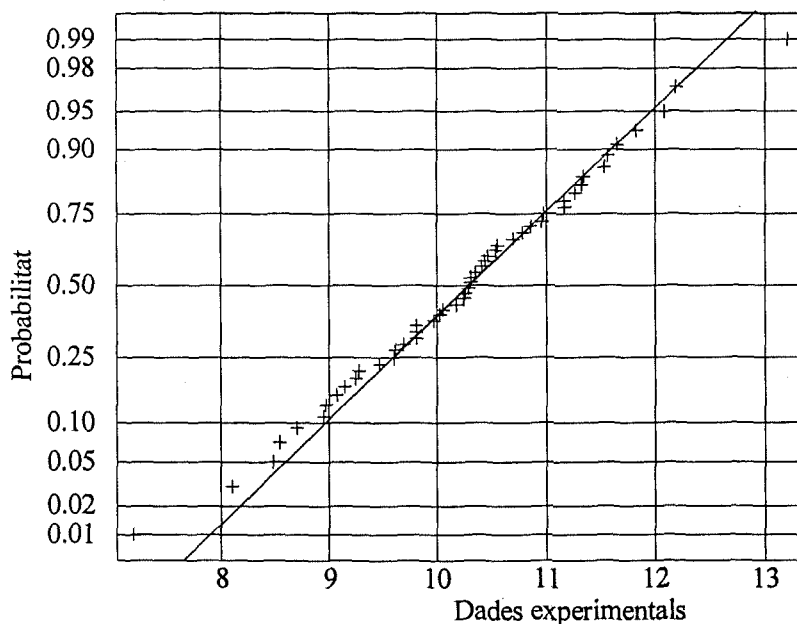


Figura 3.2. Gràfica de probabilitat normal per a un conjunt simulat de 50 dades.

### 3.2 Interval 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

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.

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:

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

i l'interval de confiança individual del pendent és:

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

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:

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

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

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:

$$|\hat{a} - \text{valor de referència}| \leq t_{\alpha/2, n-2} \cdot \hat{s}_a \quad (3.24)$$

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

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

on el valor de referència normalment sol ser 1.

### 3.3 Error $\beta$ 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  $\beta$  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  $\beta$  aplicat als problemes químics i de la poca atenció que normalment se li dóna dins del camp analític. Unes probabilitats d'error  $\beta$  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  $\beta$  fixant l'error màxim que es vol cometre corresponen a la part de la distribució centrada en el valor del biaix escollit ( $\Delta$ ) que se solapa amb la distribució associada al valor de referència limitada segons el nivell de

significança  $\alpha$  escollit.<sup>2,6,7</sup> 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  $\beta$  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  $\alpha$  o les probabilitats d'error  $\beta$ .<sup>8</sup> Les probabilitats d'error  $\alpha$  es poden escollir fixant el nivell de significança  $\alpha$  desitjat i calculant les probabilitats d'error  $\beta$  associades mantenint el nivell  $\alpha$  prèviament escollit constant. Si es volen minimitzar les probabilitats d'error  $\beta$ , s'escollirà un nivell de significança  $\alpha$  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  $\alpha$  sigui acceptable per l'analista), i es calcularà llavors la probabilitat d'error  $\beta$  a partir d'aquest valor d' $\alpha$  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  $\alpha$  i  $\beta$  en un test individual segons aquestes dues alternatives es veuen representades a la figura 3.3.

Per tal de calcular les probabilitats d'error  $\beta$ , 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

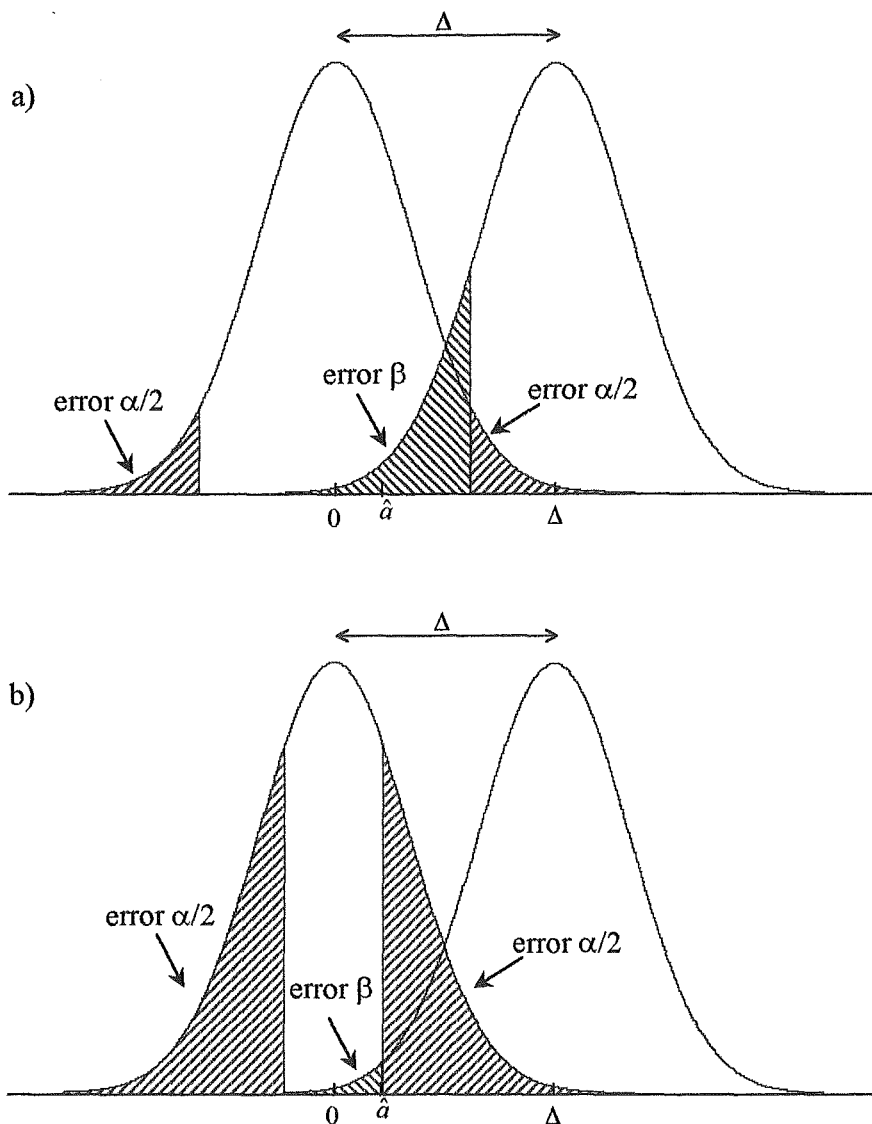


Figura 3.3. Representació de les probabilitats d'error  $\alpha$  i  $\beta$  associades a l'interval de confiança individual de l'ordenada a l'origen o el pendent. En la figura es representen les probabilitats d'error  $\alpha$  i  $\beta$  per a l'ordenada a l'origen, un valor de referència de 0 i un biaix  $\Delta$  determinat, a) fixant el nivell de significança  $\alpha$  i calculant les probabilitats d'error  $\beta$  associades, b) agafant el nivell de significança  $\alpha$  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  $\beta$  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  $\beta$  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  $\beta$  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  $\Delta$ , de tal forma que hi hagi una probabilitat d'error  $\alpha$  de concloure que hi ha diferències entre els valors de referència i experimental quan en realitat no hi són, i una probabilitat d'error  $\beta$  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  $\Delta$  donades unes certes probabilitats d'error  $\alpha$  i  $\beta$ , 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  $\Delta$  i paràmetres relatius a la variable predictora, com la mitjana de la variable predictora dels punts experimentals o la seva desviació estàndard.<sup>9</sup>

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  $\Delta$  donades unes certes probabilitats d'error  $\alpha$  i  $\beta$  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  $\Delta$  donades unes certes probabilitats d'error  $\alpha$  i  $\beta$  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  $\beta$  error in individual tests under BLS conditions and theoretical procedures for estimating the sample size in order to obtain the desired probabilities of  $\alpha$  and  $\beta$  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.

## 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

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 ( $\alpha$  and  $\beta$ ), the  $\beta$  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  $\beta$  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  $\alpha$  and  $\beta$  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,

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  $\Delta$  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):

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

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

$$s_{e_i}^2 = w_i = s_{y_i}^2 + \hat{b}^2 s_{x_i}^2 - 2\hat{b} \text{cov}(x_i, y_i) \quad (2)$$

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].

### $\beta$ 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

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  $\beta$  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:

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

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

The probability of committing a  $\beta$  error under the assumption of normality is finally given by the Student's  $t$  value for  $n-2$  degrees of freedom for a fixed  $\alpha$  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:

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

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

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  $\alpha$  and  $\beta$  errors. This can only be achieved if the individual

uncertainties, and hence the weighting factors are considered constant for all the data pairs ( $\hat{s}_{\epsilon_{aH_0}}^2$ ,  $\hat{s}_{\epsilon_{bH_0}}^2$  or  $\hat{s}_{\epsilon_{bH_1}}^2 = ct$ ):

$$\hat{s}_{aH_0} = \sqrt{\frac{\sum_{i=1}^n x_i^2 \cdot \hat{s}_{\epsilon_{aH_0}}^2}{n_a \cdot \sum_{i=1}^n x_i^2 - \left[ \sum_{i=1}^n x_i \right]^2}} \cdot \hat{s}^2 \quad (7)$$

$$\hat{s}_{bH_0} = \sqrt{\frac{n_b \cdot \hat{s}_{\epsilon_{bH_0}}^2}{n_b \cdot \sum_{i=1}^n x_i^2 - \left[ \sum_{i=1}^n x_i \right]^2}} \cdot \hat{s}^2 \quad \text{or} \quad \hat{s}_{bH_1} = \sqrt{\frac{n_b \cdot \hat{s}_{\epsilon_{bH_1}}^2}{n_b \cdot \sum_{i=1}^n x_i^2 - \left[ \sum_{i=1}^n x_i \right]^2}} \cdot \hat{s}^2 \quad (8)$$

Introducing these two expressions in Eqs. (3-4) respectively it is possible to isolate  $n$  in terms of the desired variables  $\alpha$ ,  $\beta$  and  $\Delta$ :

$$n_a = \frac{(t_{\alpha/2} + t_{\beta})^2 \cdot \hat{s}_{\epsilon_{aH_0}}^2}{\Delta_a^2} \cdot \hat{s}^2 + \frac{\left[ \sum_{i=1}^n x_i \right]^2}{\sum_{i=1}^n x_i^2} \quad (9)$$

$$n_b = \frac{\Delta_b^2 \cdot \left[ \sum_{i=1}^n x_i \right]^2}{\Delta_b^2 \cdot \sum_{i=1}^n x_i^2 - \left( t_{\alpha/2} \cdot \hat{s}_{\epsilon_{aH_0}} + t_{\beta} \cdot \hat{s}_{\epsilon_{bH_1}} \right)^2 \cdot \hat{s}^2} \quad (10)$$

Initial estimates of the terms  $\hat{s}_{\epsilon_{aH_0}}^2$  or  $\hat{s}_{\epsilon_{bH_0}}^2$  and  $\hat{s}_{\epsilon_{bH_1}}^2$ ,  $\hat{s}^2$  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  $t_{\beta}$  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  $\beta$  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 ( $x_{tr}$ ) limits for a chosen  $\alpha$  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.

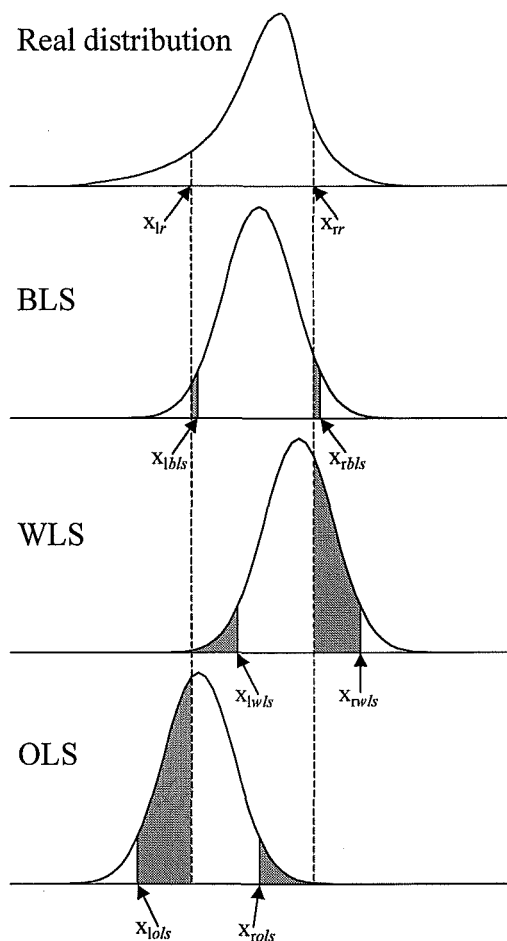


Figure 1. Error made in estimating the BLS regression coefficients assuming normal distributions. Comparison with errors made using OLS and WLS regression techniques.

To validate the expressions for the estimated of the probabilities of  $\beta$  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  $\alpha$  level of significance. So every time  $H_0$  was accepted, a  $\beta$  error was being committed because the data set had been generated from an initial biased

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  $\beta$  error similar to the  $\alpha$  level of significance in each of the four cases. In this way, if the estimate of the probabilities of  $\beta$  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  $\beta$  error were proved to be correct, the expressions to estimate the sample size were validated. The probabilities of  $\beta$  error estimated for the different  $\alpha$  levels of significance, the calculated standard deviations and the experimental error from the iterative process (terms  $t_\beta$ ,  $t_{\alpha/2}$ ,  $\hat{s}_{\varepsilon_{aH_0}}^2$ ,  $\hat{s}_{\varepsilon_{bH_0}}$  or  $\hat{s}_{\varepsilon_{bH_1}}$  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  $\alpha$  and  $\beta$  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  $\beta$  error (Eqs. (3-4)). Finally, one of the six former real data sets was used to show the different estimates of the probabilities of  $\beta$  error between BLS,

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 polycyclic aromatic hydrocarbons (PAHs) from railroad bed soil using two different modifiers; CO<sub>2</sub> (on the  $x$ -axis) and a mixture of CO<sub>2</sub> 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  $\mu\text{g/g}$  of soil. The data set and the regression lines obtained by the OLS, WLS and BLS regression techniques are shown in Figure 2a.

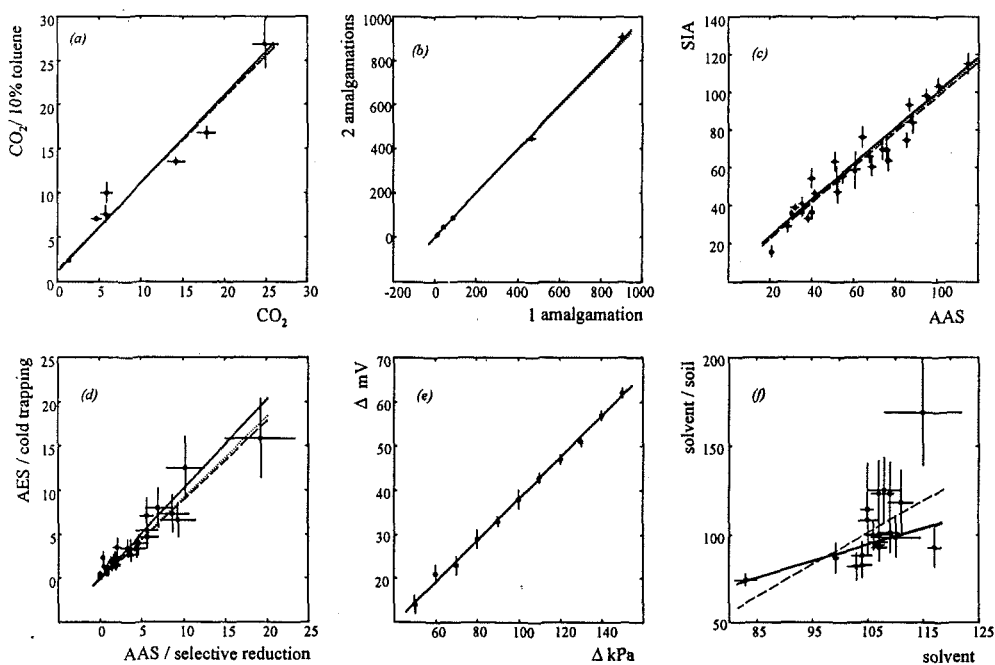


Figure 2. OLS (dashed line), WLS (dotted line) and BLS (solid line) regression lines obtained for the six real data sets.

**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 ( $y$ -

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  $\mu\text{g}$  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  $\text{mg/l}$ . 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  $\mu\text{g/l}$ . 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  $\text{CO}_2$  Joule-Thompson coefficient. The data was acquired from thermocouple-measured voltage differences ( $\Delta mV$ , on the  $y$ -axis) as a function of pressure increments ( $\Delta kPa$ , 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

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  $\beta$  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.

For every one of the twenty four different simulated data sets, four  $\alpha$  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 ( $b_{H_i}$ ) or the intercept value ( $a_{H_i}$ ) of the selected bias changed in such a way that the probabilities of  $\beta$  error from the iterative process were similar to the specified  $\alpha$  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

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

| data set | Iterations | Kolmogorov |           |              |    |              |    |               |    |             |           |
|----------|------------|------------|-----------|--------------|----|--------------|----|---------------|----|-------------|-----------|
|          |            | Cetama     |           |              |    |              |    |               |    | Rankit Plot |           |
|          |            | $\hat{a}$  | $\hat{b}$ | $\alpha=1\%$ |    | $\alpha=5\%$ |    | $\alpha=10\%$ |    | $\hat{a}$   | $\hat{b}$ |
| 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  $\alpha$  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  $\alpha$  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

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.

| data set | Coefficient | % Error |       |       |
|----------|-------------|---------|-------|-------|
|          |             | 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).

### $\beta$ 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  $\alpha_{H_1}$  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  $\beta_{\text{exp}}$  column are those from the simulation process, whereas the values shown in the  $\beta_{\text{pred}}$  column are the ones obtained with the theoretical expressions to be validated (Eqs. (3-4)). Finally, the values in the column  $n_{\text{pred}}$  are the estimated sample sizes of the different simulated data sets for the different levels of significance.

| $n$ | Uncertainty | $\alpha(\%)$ | $a_{H_1}$ | $\hat{s}_{a_{H_0}}$ | $\beta_{\text{exp.}}$ | $\beta_{\text{pred.}}$ | $n_{\text{pred.}}$ |
|-----|-------------|--------------|-----------|---------------------|-----------------------|------------------------|--------------------|
| 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  $\beta$  error for individual tests on the intercept. Predicted sample size to achieve the  $\alpha$  and  $\beta$  probabilities of error for each data set.

| $n$ | Uncertainty | $\alpha(\%)$ | $a_{H_1}$ | $\hat{s}_{a_{H_0}}$ | $\beta_{\text{exp.}}$ | $\beta_{\text{pred.}}$ | $n_{\text{pred.}}$ |
|-----|-------------|--------------|-----------|---------------------|-----------------------|------------------------|--------------------|
| 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  $\beta$  error for individual tests on the intercept. Predicted sample size to achieve the  $\alpha$  and  $\beta$  probabilities of error for each data set.

| $n$ | Uncertainty | $\alpha(\%)$ | $b_{H_1}$ | $\hat{s}_{b_{H_0}}$ | $\hat{s}_{b_{H_1}}$ | $\beta_{\text{exp.}}$ | $\beta_{\text{pred.}}$ | $n_{\text{pred.}}$ |
|-----|-------------|--------------|-----------|---------------------|---------------------|-----------------------|------------------------|--------------------|
| 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  $\beta$  error for individual tests on the slope. Predicted sample size to achieve the  $\alpha$  and  $\beta$  probabilities of error for each data set.

| $n$ | Uncertainty | $\alpha$ (%) | $b_{H_1}$ | $\hat{s}_{b_{H_0}}$ | $\hat{s}_{b_{H_1}}$ | $\beta_{\text{exp.}}$ | $\beta_{\text{pred.}}$ | $n_{\text{pred.}}$ |
|-----|-------------|--------------|-----------|---------------------|---------------------|-----------------------|------------------------|--------------------|
| 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  $\beta$  error for individual tests on the slope. Predicted sample size to achieve the  $\alpha$  and  $\beta$  probabilities of error for each data set.

To detect significant differences between the estimated probabilities of  $\beta$  error and the values from the simulation process, paired  $t$ -tests [22] (with  $\alpha=1\%$ ) were applied on the  $\beta$  error values obtained for the different number of data points (since it is the most critical factor for achieving good predictions of  $\beta$  probabilities of error) at the same level of significance. In this way significant differences between the values in the  $\beta_{\text{exp}}$  and  $\beta_{\text{pred}}$  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  $\beta$  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  $\beta$  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%  $\alpha$  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  $\beta$  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  $\beta$  error (in percent) for the different simulated data sets. As

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  $\beta$  error.

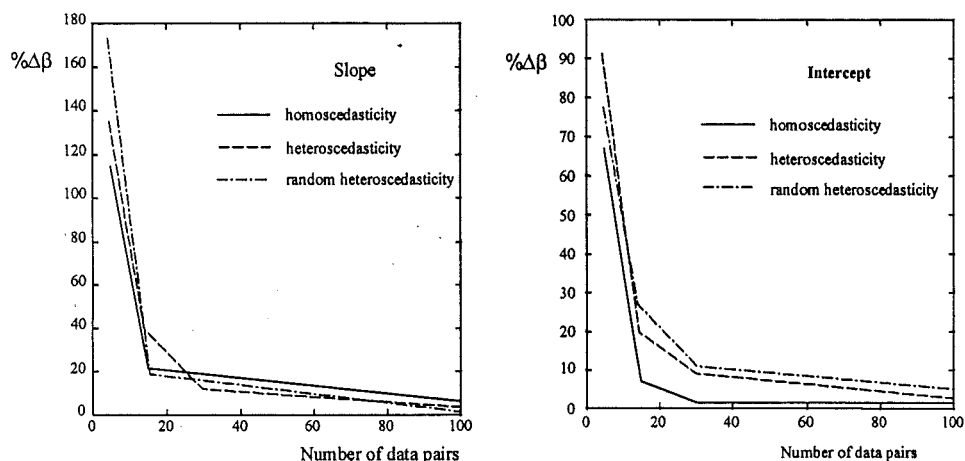


Figure 3. Difference between the experimentally-obtained probabilities (simulation process) and the predicted probabilities of  $\beta$  error for the slope and the intercept (in percent) in relation to the number of data pairs for each data set.

(iv) Results from the predicting the probabilities of  $\beta$  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

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 ( $n_{\text{pred}}$  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 $\beta$ error and estimation of sample size in a real data set

Table 5 summarises the results of estimating the probabilities of committing a  $\beta$  error in the individual tests for the BLS slope and intercept at a 5%  $\alpha$  level of significance ( $\beta$  column, in percent) for data set 3. Columns  $|a_{H_0} - \hat{a}|$  and  $|b_{H_0} - \hat{b}|$  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  $t \cdot \hat{s}_{b_{H_0}}$  ( $\alpha=5\%$ ) 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  $|a_{H_0} - \hat{a}|$  and  $|b_{H_0} - \hat{b}|$  is higher than its associated confidence interval. Probabilities of  $\beta$  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  $\beta$  error are estimated at 62.5% for the OLS technique, due to the highest standard deviation value. On the

other hand, the probabilities of  $\beta$  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  $\beta$  error. Results for the slope show that the estimated probabilities of  $\beta$  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  $b_{H_0}$ .

|     | $ a_{H_0} - \hat{a} $ | $t \cdot \hat{s}_{a_{H_0}}$ | $a_{H_1}$ | $\beta$ |
|-----|-----------------------|-----------------------------|-----------|---------|
| BLS | 2.94                  | 5.35                        |           | 40.2    |
| WLS | 4.38                  | 5.19                        | 6         | 37.6    |
| OLS | 3.97                  | 7.11                        |           | 62.5    |
|     | $ b_{H_0} - \hat{b} $ | $t \cdot \hat{s}_{b_{H_0}}$ | $b_{H_1}$ | $\beta$ |
| 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  $\beta$  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  $\beta$  error in the slope (2.77%) and intercept (40.2%) for a 5%  $\alpha$  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

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  $\beta$  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  $\beta$  error when testing an individual regression coefficient with the BLS regression technique and considering different distributions for the reference

( $a_{H_0}$  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  $\beta$  error. In addition, results in real data show that in some cases it may be interesting to calculate the probabilities of  $\beta$  error not with the set  $\alpha$  threshold value, but with the maximum  $\alpha$  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  $\alpha$  probabilities of error would be higher but the  $\beta$  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  $\alpha$  and  $\beta$  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  $\alpha$  and  $\beta$  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.

## 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.

- 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.

## Capítol 4

---

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

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*,

publicat a la revista *Química Analítica*, i és l'últim article exposat en aquest capítol.

#### 4.1 Regressió lineal aplicada a la comparació de mètodes analítics

En processos de comparació de dos mètodes analítics mitjançant regressió lineal, es representen els resultats procedents de l'anàlisi d'una sèrie de mostres a diferents nivells de concentració pels dos mètodes en comparació: el mètode sobre el qual es fa la comparació (que se sol situar sobre l'eix de les  $y$ ) i normalment un mètode ja establert (que se sol situar sobre l'eix de les  $x$ ). Per dur a terme la comparació de dos mètodes analítics, no cal que les mostres analitzades pels dos mètodes siguin materials de referència certificats (CRM), però la seva utilització permet traçar els mètodes a referències metrològiques més elevades. A més, la utilització de CRM permet establir *a priori* l'interval de concentracions en el qual es durà a terme el procés de comparació de mètodes, la qual cosa no sempre es pot establir amb la utilització de mostres reals.

Com ja s'ha comentat en el capítol anterior, en el cas hipotètic que els dos mètodes en comparació donessin exactament els mateixos resultats, la regressió lineal entre els dos mètodes seria una línia recta d'ordenada a l'origen 0 i pendent 1. A causa dels errors aleatoris, es dirà que dos mètodes analítics produeixen resultats que no difereixen estadísticament de forma significativa si alhora l'ordenada a l'origen no difereix estadísticament de 0 i el pendent no difereix estadísticament d'1 per a un nivell de significança  $\alpha$  escollit. Cal tenir present que normalment els mètodes analítics en comparació porten associats errors del mateix ordre de magnitud, per la qual cosa s'hauria de recórrer a tècniques que consideressin errors en els dos mètodes per tal de buscar els coeficients de la recta de regressió.

ISBN: 978-84-691-1897-9/D.L.: T-353-2008

En els últims anys s'han desenvolupat diversos tests estadístics per tal de dur a terme la comparació de dues metodologies analítiques mitjançant regressió lineal. Averdunk i Borner,<sup>1</sup> en el camp de la química clínica, van proposar l'observació del pendent obtingut amb la mitjana geomètrica de dues rectes de regressió trobades emprant OLS: la regressió de  $x$  sobre  $y$  i la regressió de  $y$  sobre  $x$  vistes des de l'eix de les  $x$ . El pendent d'aquesta línia coincideix amb el pendent del primer component principal estandarditzat.<sup>2,3</sup>

En el camp de la bioquímica, s'ha proposat una alternativa a la comparació de mètodes sense emprar la regressió lineal, però sí aplicable a múltiples nivells de concentració. Es basa a representar per a cada nivell de concentració la mitjana dels resultats obtinguts pels dos mètodes: (resultat mètode A + resultat mètode B)/2 respecte a les seves diferències: (resultat mètode A – resultat mètode B).<sup>4,5</sup> Si els resultats dels dos mètodes no difereixen entre si, les diferències haurien de ser simètriques al voltant de zero, i si fos possible, sense tenir cap relació amb la mitjana de resultats. La concordança entre els resultats dels dos mètodes es pot comprovar utilitzant la mitjana i la desviació estàndard de les diferències: l'exactitud mitjançant un test estadístic per comprovar si la mitjana de les diferències no difereix significativament de zero i la precisió mitjançant els intervals de confiança per les diferències individuals entre els dos mètodes. Si les diferències depenen de la mitjana de resultats, es proposa una transformació logarítmica<sup>6</sup> per tractar d'eliminar aquesta relació. L'inconvenient principal surt si algun dels dos mètodes en comparació presenta una precisió pitjor que l'altre, ja que el procés de comparació es veu afectat per aquestes diferències. El fet de no tenir en compte les incerteses individuals en cada mostra analitzada fa que la comparació pugui arribar a resultats incorrectes. En aquest cas, es recomana comparar les precisions relatives mitjançant duplicats a diferents nivells de concentració.

També s'ha proposat la comparació de mètodes analítics establint dues rectes de regressió: una recta que situï els resultats obtinguts amb un mètode a l'eix de les ordenades i els resultats obtinguts amb l'altre mètode a l'eix de les abscisses, i l'altra recta que canviï els mètodes d'eixos.<sup>7</sup> Els coeficients de les dues rectes de regressió es busquen amb el mètode OLS i llavors s'escull el pendent ( $\hat{b}_a$ ) de la recta de millor ajust als punts experimentals. Després es compara si és significativament diferent a la unitat mitjançant un test d'hipòtesi:

$$t_{\text{cal}} = \left( \frac{\hat{b}_a - 1}{\sqrt{1 - r^2}} \right) \times \sqrt{n - 2} \quad (4.1)$$

on  $r$  és el coeficient de correlació de la recta de regressió de millor ajust. El valor  $t_{\text{cal}}$  es compara amb el valor  $t$  tabulat per a un nivell  $\alpha$  de significança i  $n-2$  graus de llibertat. També s'utilitza un altre test d'hipòtesi per comprovar si el valor de l'ordenada a l'origen ( $\hat{a}_a$ ) de la recta amb millor ajust difereix significativament de 0:

$$t_{\text{cal}} = (\hat{a}_a - 0) / \hat{s}_a^2 \quad (4.2)$$

Aquest últim valor  $t_{\text{cal}}$  també es compara amb el valor  $t$  tabulat per a un nivell  $\alpha$  de significança i  $n-2$  graus de llibertat. Si el valor del pendent no difereix significativament d'1 i el valor de l'ordenada a l'origen no difereix significativament de 0, s'arriba a la conclusió que els dos mètodes donen resultats comparables.

Una altra forma similar però més visual de dur a terme la comparació de dos mètodes és la que es basa en la superposició dels intervals de confiança individuals per a l'ordenada a l'origen i el pendent de la recta de regressió trobada representant

els resultats d'un mètode respecte als resultats de l'altre mètode. Els intervals de confiança individuals es troben expressats en les equacions següents:

$$\hat{a} \pm t_{\alpha/2, n-2} \cdot \hat{s}_a \quad (4.3)$$

$$\hat{b} \pm t_{\alpha/2, n-2} \cdot \hat{s}_b \quad (4.4)$$

La superposició dels dos intervals de confiança individuals es pot veure a la figura 4.1, amb el centre corresponent al punt definit pels valors experimentals de l'ordenada a l'origen i el pendent de la recta de regressió. Si el punt teòric ordenada a l'origen = 0 i pendent = 1 es troba situat dins de la superfície definida per la superposició dels dos intervals de confiança individuals, es dirà que els dos mètodes analítics en comparació donen resultats que no difereixen estadísticament entre si per al nivell de significança  $\alpha$  escollit en els intervals de confiança individuals.

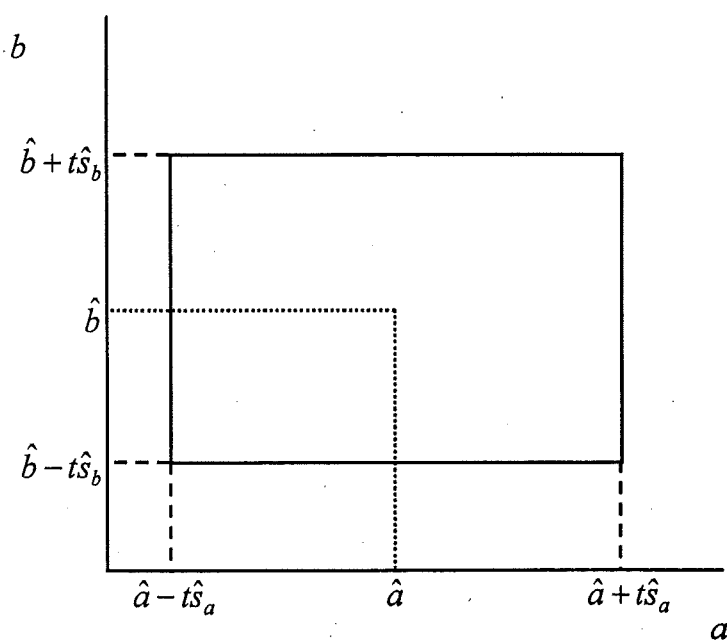


Figura 4.1. Superposició dels intervals de confiança individuals per a l'ordenada a l'origen i el pendent de la recta de regressió.

Però el fet d'emprar per a la comparació de dos mètodes els intervals de confiança individuals fa que es depreciï la correlació entre la ordenada a l'origen i el pendent (la qual ve donada per la covariància entre aquests dos coeficients), que pràcticament sempre és negativa. Aquest fet significa que en analitzar una sèrie de mostres reals amb els dos mètodes en comparació per tal d'obtenir-ne la recta de regressió, si es tornés a repetir l'anàlisi de les mostres reals per tal d'obtenir una altra sèrie de punts experimentals i ajustar-los a una nova recta de regressió, si la ordenada a l'origen d'aquesta nova recta de regressió fos més gran que l'ordenada a l'origen inicial, el pendent de la nova recta de regressió seria més petit que el pendent inicial (o si l'ordenada a l'origen d'aquesta nova recta de regressió fos més petita que la inicial, el pendent de la nova recta de regressió seria més gran que el pendent inicial). Com es veurà en l'apartat següent, el fet de no considerar la correlació existent entre l'ordenada a l'origen i el pendent de la recta de regressió pot comportar que el procés de comparació de mètodes doni conclusions errònies.<sup>8</sup>

Molt recentment, s'ha proposat un altre procediment per tal de comparar dos mètodes analítics emprant regressió lineal amb errors en dos eixos.<sup>9</sup> Aquest es basa a considerar la forma implícita de la recta de regressió trobada representant els resultats dels dos mètodes analítics en comparació:

$$Ax + By + C = 0 \quad (4.5)$$

on els coeficients  $A$ ,  $B$  i  $C$  de l'equació 4.5 es calculen mitjançant la tècnica de regressió ILF (funció lineal implícita), de la qual els autors no donen cap referència ni presenten l'algorisme o cap característica, malgrat afirmar que consideren els errors deguts als dos mètodes. Un cop trobats els coeficients  $A$ ,  $B$  i  $C$ , i fent que  $C$  sigui igual a la unitat dividint els tres coeficients per  $C$ , els autors afirmen que si els dos mètodes en comparació produeixen resultats pràcticament iguals, els nous valors de  $A$  i  $B$  tindran quasi exactament el mateix valor absolut. Per tal de verificar

ISBN: 978-84-691-1897-9/D.L: T-353-2008

la diferència entre els valors absoluts d'aquests dos coeficients, utilitzen l'anàlisi de la informació de la variància, IANOVA,<sup>10,11</sup> basada en l'energia d'informació ( $E = \sum_{i=1}^n p_i^2$ ). Segons els autors, el IANOVA és un procediment que no depèn de cap tipus de distribució estadística i és vàlid sota unes mínimes assumpcions. En el cas concret de comparació de dos mètodes analítics, les probabilitats  $p$  de l'energia d'informació es calculen segons:

$$p_A = \frac{A}{A+B} \quad \text{i} \quad p_B = \frac{B}{A+B} \quad (4.6)$$

La hipòtesi nul·la en aquest cas correspon a  $H_0 : p_A = p_B$ , i aquesta s'acceptarà si  $E = \epsilon$ , on  $E = 1/2$  representa l'energia d'informació teòrica i  $\epsilon = (A^2 + B^2)/(A+B)^2$  representa l'energia d'informació empírica. Si  $E \neq \epsilon$ , es rebutja la hipòtesi nul·la i, per tant, s'arriba a la conclusió que els dos mètodes en comparació produeixen resultats no comparables. Com ja s'ha comentat, els coeficients de la recta de regressió es busquen amb el mètode ILF, i la recta obtinguda en forma implícita té la característica que la seva expressió explícita  $y=f(x)$  coincideix amb  $x=f(y)$  (que també s'obté amb l'aplicació del mètode BLS). No obstant això, el mètode ILF no considera les incerteses individuals de cada punt, motiu pel qual els resultats de la recta de regressió no coincideixen amb els obtinguts de l'aplicació del mètode BLS, sinó que s'assemblen molt als obtinguts amb l'aplicació del mètode OR. A més, el fet que la comprovació de la concordança entre els resultats dels dos mètodes es faci comprovant si el paràmetre  $\epsilon$  és igual a  $1/2$ , crea una certa confusió a l'hora de decidir si realment  $\epsilon$  equival a  $1/2$ , ja que rarament  $\epsilon$  serà exactament igual a  $1/2$ . Probablement una aproximació millor seria trobar l'interval de confiança al voltant de  $\epsilon$  i comprovar si el valor teòric de  $1/2$  està inclòs dins d'aquest interval, evitant així possibles ambigüitats.

## 4.2 Test conjunt per a l'ordenada a l'origen i el pendent de la recta de regressió

El test conjunt per a l'ordenada a l'origen i el pendent consisteix en la comparació simultània de l'ordenada a l'origen i el pendent de la recta de regressió (tenint en compte la seva correlació, indicada per la covariància entre els dos coeficients de regressió), amb els seus valors de referència. El van aplicar per primera vegada a dades químiques J. Mandel i F.J. Linnig.<sup>12</sup> Aquest test defineix una superfície el·líptica amb el punt corresponent als coeficients de la recta de regressió  $(\hat{a}, \hat{b})$  com a centre de l'el·lipse. Els límits de l'el·lipse estan determinats per la magnitud de l'error experimental, per les variàncies dels coeficients de regressió i pel nivell de significança  $\alpha$  escollit, i la seva inclinació és funció de la correlació entre l'ordenada a l'origen i el pendent.

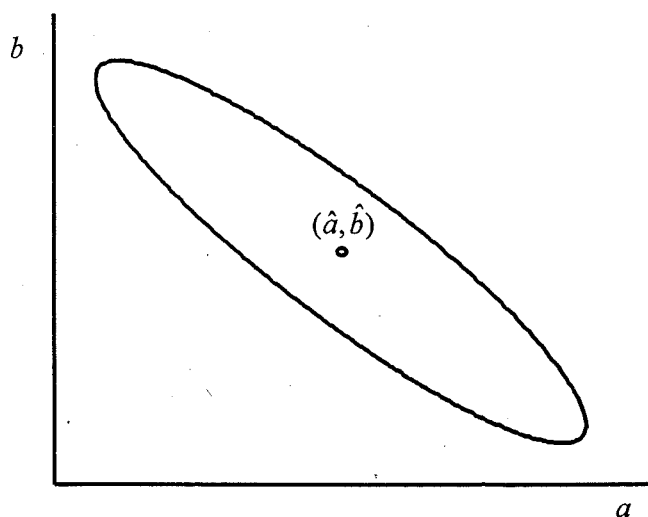


Figura 4.2. Interval de confiança conjunt per a l'ordenada a l'origen i el pendent de la recta de regressió.

Com a resultat de la inclinació, l'interval de confiança conjunt afavoreix rectes amb més pendent i menys ordenada a l'origen que els corresponents als valors

ISBN: 978-84-691-1897-9/D.L.: T-353-2008

experimentals, i rectes amb menys pendent i més ordenada a l'origen que els corresponents als valors experimentals (figura 4.2), ja que en processos de comparació de mètodes la covariància entre l'ordenada a l'origen i el pendent de la recta de regressió sol ser sempre negativa.

El test conjunt per a l'ordenada a l'origen i el pendent desenvolupat per Mandel i Linnig està basat en el mètode de regressió OLS, i l'el·lipse de confiança es defineix segons l'equació següent:

$$n(b - \hat{b})^2 + 2 \sum_{i=1}^n x_i (b - \hat{b})(a - \hat{a}) + \sum_{i=1}^n x_i^2 (a - \hat{a})^2 = 2F_{(\alpha, n-2, 2)} \hat{s}^2 \quad (4.7)$$

on el valor  $\hat{s}^2$  està definit a l'equació 1.9. Cal tenir present que la utilització del mètode OLS per tal de trobar els coeficients de la recta de regressió i paràmetres relacionats com poden ser les seves variàncies, dona resultats diferents segons quin mètode es troba situat a l'eix de les ordenades o de les abscisses. Això no té gaire sentit en processos de comparació de mètodes, on s'haurien d'obtenir els mateixos resultats encara que els mètodes es canviessin d'eixos.

El fet d'utilitzar els intervals de confiança individuals esmentats en l'apartat anterior en comptes del test conjunt per a l'ordenada a l'origen i el pendent pot comportar situacions com la reflectida en la figura 4.3, on s'arribaria erròniament a la conclusió que els dos mètodes en comparació produeixen resultats que no difereixen significativament en el nivell de significança  $\alpha$  escollit, quan l'ús del test conjunt per a l'ordenada a l'origen i el pendent indicaria que els dos mètodes en comparació produeixen resultats que difereixen estadísticament entre ells al mateix nivell de significança  $\alpha$  escollit.

Cal tenir present que fins i tot en casos on la covariància entre l'ordenada a l'origen i el pendent fos igual a zero, la superposició dels intervals de confiança individuals

ISBN:978-84-691-1897-9/D.L: T-353-2008

no es correspondria amb l'interval de confiança conjunt, ja que mentre els primers inclouen el paràmetre estadístic  $t$ , el segon, al considerar-los conjuntament, està basat en el paràmetre estadístic  $F$ . Aquest cas particular es veu representat a la figura 4.4. Aquesta representació correspondria a un procés de comparació de dos mètodes analítics en els quals s'hagués centrat cada resultat individual, és a dir, s'hagués restat a cada resultat individual el valor mitjà o el valor mitjà ponderat (segons el mètode de regressió emprat, per exemple WLS o BLS) de tots els resultats obtinguts amb aquell mètode.

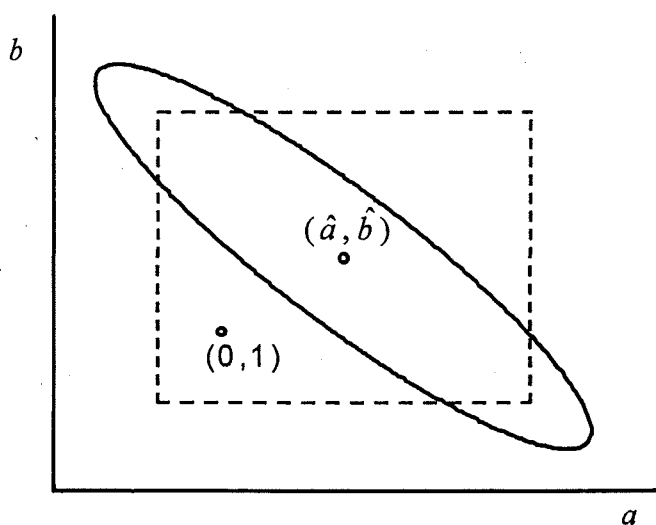


Figura 4.3 Comparació de l'aplicació del test conjunt per a l'ordenada a l'origen i el pendent (línia contínua) amb els intervals individuals de confiança per a l'ordenada a l'origen i el pendent (línia discontinua).

L'interval de confiança conjunt per a l'ordenada a l'origen i el pendent desenvolupat per Mandel i Linnig aplicat a comparació de mètodes té l'inconvenient principal que només considera errors constants en un mètode, el situat sobre l'eix de les  $y$  (normalment el nou mètode). El test conjunt per a l'ordenada a l'origen i el pendent considerant errors heteroscedàstics individuals en

ISBN:978-84-691-1897-9/D.L: T-353-2008

dos eixos -descrit en el primer dels tres articles que es troben al final del capítol- soluciona aquesta limitació. L'inconvenient principal d'aquest procediment és que es necessita la variància individual per a cada punt experimental analitzat pels dos mètodes en comparació, la qual cosa normalment sol implicar repeticions i més temps d'anàlisi. No obstant això, cal recordar que la sèrie de normes EN 45000/UNE 66500 especifiquen clarament que cada resultat analític ha d'anar acompanyat del valor de la seva incertesa. Una altra opció és estimar les variàncies de cada punt experimental per cada un dels dos mètodes en comparació, però òbviament les conclusions obtingudes amb l'aplicació del test conjunt seran més fiables com més exacta sigui la informació proporcionada.

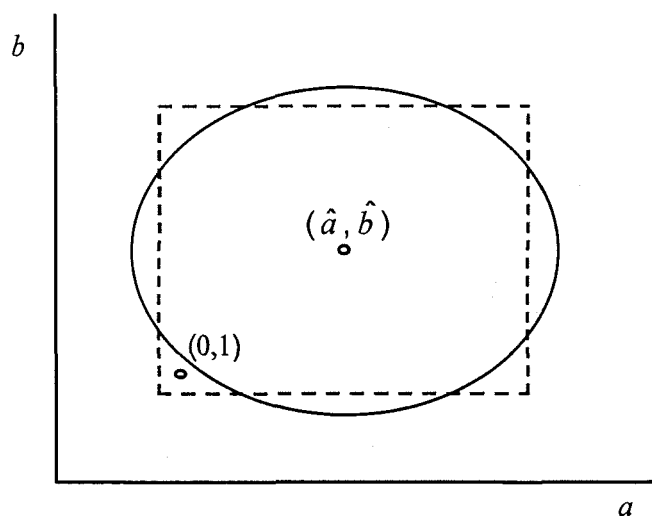


Figura 4.4 Test conjunt per a l'ordenada a l'origen i el pendent (línia contínua) i intervals individuals de confiança per a l'ordenada a l'origen i el pendent (línia discontinua) en el cas particular que la covariància entre l'ordenada a l'origen i el pendent sigui zero.

El test conjunt per a l'ordenada a l'origen i el pendent considerant errors en els dos eixos es basa en la hipòtesi que l'ordenada a l'origen i el pendent de la recta de

ISBN:978-84-691-1897-9/D.L: T-353-2008

regressió segueixen la distribució normal. Com s'ha demostrat en el capítol anterior, aquesta hipòtesi no és rigorosament certa, malgrat que l'error comès en acceptar-la no és significativament gran. De fet, com es podrà comprovar en l'article publicat a la revista *Analytical Chemistry*, els processos de simulació per tal de validar el test conjunt per a l'ordenada a l'origen i el pendent considerant errors en els dos eixos van donar resultats molt pròxims als teòrics, podent-se assimilar la lleugera discrepància entre els resultats obtinguts i els valors teòrics a l'acceptació de la normalitat en la distribució dels coeficients de regressió experimentals.

Per acabar de completar el test conjunt per a l'ordenada a l'origen i el pendent de la recta de regressió considerant errors en dos eixos, caldria desenvolupar les expressions teòriques per calcular les probabilitats d'error  $\beta$  en processos de comparació de dos mètodes analítics. Avui s'està treballant en aquestes expressions,<sup>13</sup> i similarment amb el càlcul de les probabilitats d'error  $\beta$  en els tests individuals per l'ordenada a l'origen i el pendent (capítol 3), les probabilitats d'error  $\beta$  en el test conjunt correspondrien al volum de la distribució associada al punt teòric ordenada a l'origen = 0 i pendent = 1 que se superposa amb la distribució associada al biaix de l'ordenada a l'origen i el pendent definit per l'analista pel nivell de significança  $\alpha$  escollit. El procés de càlcul en aquest cas es complica considerablement, ja que el fet de tenir distribucions de probabilitat tridimensionals (perquè es consideren conjuntament l'ordenada a l'origen i el pendent) implica que els càlculs es duen a terme amb volums i no amb àrees, com passava amb els tests individuals per a l'ordenada a l'origen i el pendent. Els resultats obtinguts fins a la data, però, mostren una bona concordança entre els valors de les probabilitats d'error  $\beta$  obtingudes experimentalment i els valors teòrics obtinguts amb processos de simulació mitjançant el mètode de Monte Carlo.<sup>14</sup>

### 4.3 Referències

1. R. Averdunk, K. Borner, *Zeitschrift fur Klinische Chemie und Klinische Biochemie*, **8** (1970) 263
2. U. Feldmann, *European Journal of Clinical Chemistry and Clinical Biochemistry*, **30** (1992) 405
3. U. Feldmann, B. Schneider, H. Klinkers, R. Haeckel, *Journal of Clinical Chemistry and Clinical Biochemistry*, **8** (1970) 263
4. D.G. Altman, J.M. Bland, *Statistician*, **32** (1983) 307
5. J.M. Bland, D.G. Altman, *Lancet*, **i** (1986) 307
6. S. Hollis, *Annals of Clinical Biochemistry*, **33** (1996) 1
7. S.J. Haswell (ed.), *Practical Guide to Chemometrics*, Marcel Dekker Inc., New York (1992)
8. P.D. Lark, *Analytical Chemistry*, **26** (1954) 1712
9. H.W. Zwanziger, C. Sârbu, *Analytical Chemistry*, **70** (1998) 1277
10. C. Sârbu, *Analitica Chimica Acta*, **271** (1993) 269
11. C. Sârbu, *Analytical Letters*, **30** (1997) 1051
12. J. Mandel, F.J. Linnig, *Analytical Chemistry*, **29** (1957) 743
13. A. Martínez, J. Riu, F.X. Rius, En preparació
14. P.C. Meier, R.E. Zünd, *Statistical Methods in Analytical Chemistry*, John Wiley & Sons, New York (1993)

# Assessing the accuracy of analytical methods using linear regression with errors in both axes

Jordi Riu\* and F. Xavier Rius

*Departament de Química. Universitat Rovira i Virgili.*

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

## ABSTRACT

In this paper a new technique for assessing the accuracy of analytical methods using linear regression is reported. The results of newly developed analytical methods are regressed against the results obtained using reference methods. The new test is based on the joint confidence interval for the slope and the intercept of the regression line, which is calculated taking the uncertainties in both axes into account. The slope, intercept and variances which are associated to the regression coefficients are calculated with Bivariate Least Squares regression, BLS. The new technique was validated using three simulated and five real data sets. The Monte Carlo method was applied to obtain 100,000 data sets for each of the initial simulated data sets to show the correctness of the new technique. The application of the new technique to five real data sets enables differences to be detected between the results of the joint confidence interval based on the BLS method and the results of the commonly used tests based on Ordinary Least Squares or Weighted Least Squares regression.

## INTRODUCTION

Assessing accuracy is a fundamental step in the method validation process. The analyte concentration value obtained with the new method is often compared with the reference method by replicating measurements and applying a *t*-test or an *F*-test. But if the validity of the new method is checked with a range of analyte concentrations, the linear regression also gives additional statistical information such as the presence of proportional errors, the need of including a blank correction and the calculation of confidence intervals for the regression coefficients.

The statistical test which compares the intercept and slope values obtained by linear calibration with the theoretical values of zero and unity, bearing in mind the correlation between the two regression coefficients, was applied by Mandel and Linnig<sup>1</sup> to analytical results. This procedure, extensively used up to now, is based on the linear regression hypotheses being fulfilled by Ordinary Least Squares (OLS) or, whenever heteroscedasticity is present in the dependent variable, by Weighted Least Squares, WLS.<sup>2</sup> But when applied to method comparison, this procedure has the drawback that it regards the reference method (usually represented on the abscissa axis) as not only being free of systematic errors but of random ones as well. This reference method often includes random errors of the same order of magnitude as the new method to be validated and, as a result, the bibliography is full of methods which are considered to be correct but which may contain systematic errors.

There are other approaches<sup>3</sup> which apply a type of weighted regression to calculate the regression coefficients taking into account errors in both axes. The individual confidence intervals of the regression coefficients are then applied to assess the accuracy. This approach, as well as leading to biased values of the regression coefficients, does not take into account the covariance between the slope and the intercept and, consequently, leads to erroneous final results.

This study proposes applying a joint confidence test for the intercept and the slope to assess the accuracy of new analytical methods. These regression coefficients are calculated here by applying calibration methods which consider errors in both axes, (Bivariate Least Squares, BLS), and which, therefore, take into account the uncertainty in the results which both methods may have.

To show the goodness of the new validation method, three simulated and five real data sets have been used. Random errors were added to the simulated data sets using the Monte Carlo method. The values obtained with the new approach are shown to agree with the theoretical results expected, while the results of using the joint confidence test based on OLS, WLS or BLS for the real data sets can lead to different conclusions about the correctness of the validation.

The main limitation of calibration methods which include errors in both axes is that the uncertainty associated to the values on both the  $x$  and the  $y$ -axis has to be estimated, which often means that data analysis will take longer.

## BACKGROUND AND THEORY

**Notation.** The estimated regression coefficients of the calibration line will be denoted as  $\hat{a}$  (intercept) and  $\hat{b}$  (slope), while the coefficients corresponding to the experimental calibration data pairs are denoted as  $a$  and  $b$ . The column matrix  $\hat{\mathbf{b}}$  is defined by the estimated regression coefficients, while the column matrix  $\mathbf{b}$  is defined by the experimental values. The error, measured in terms of variance, for the set of  $n$  experimental data points  $(x_i, y_i)$  will be denoted as  $s^2$ , and its estimated value is defined by eq 1:<sup>4-6</sup>

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

where  $w_i$  is the weighting factor for each data point and  $\hat{y}_i$  is the estimated value for the  $y$  predicted.

**Bivariate Least Squares (BLS).** Bivariate Least Squares is the generic name for a set of techniques used for regressing bivariate data, i.e. whenever a regression method is applied to data containing errors in both axes. Of all the different existing approaches for calculating regression coefficients and related statistical parameters considering errors in both axes,<sup>7</sup> Lisý's method<sup>8</sup> was selected because of its speed in estimating the correct results for the regression coefficients, because the variance-covariance matrix (which is useful in the subsequent development of statistical tests for assessing accuracy) is obtained with no extra effort, and because of the simplicity of programming its algorithm. The method consists of minimizing the sum of the weighted residuals,  $S$ , expressed in eq 2:

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

This method uses the variance of the residuals ( $s_{e_i}^2$ ), which can be expressed using the Taylor series, as a weighting factor even when the covariance between variables for each data pair is not zero, eq 3:

$$s_{e_i}^2 = w_i = s_{y_i}^2 + \hat{b}^2 s_{x_i}^2 - 2\hat{b} \text{cov}(x_i, y_i) \quad (3)$$

where  $s_{x_i}^2$  and  $s_{y_i}^2$  respectively stand for the variances of each  $(x_i, y_i)$  individual data point. By minimizing the sum of the weighted residuals in relation to the slope and

the intercept, two non-linear equations are obtained and by including the partial derivatives of the squared residuals, eq 4 and the equivalent eq 5 can be written in matrix form:

$$\mathbf{R}\hat{\mathbf{b}} = \mathbf{g} \quad (4)$$

$$\begin{pmatrix} \sum_{i=1}^n \frac{1}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{x_i}{s_{\varepsilon_i}^2} \\ \sum_{i=1}^n \frac{x_i}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{x_i^2}{s_{\varepsilon_i}^2} \end{pmatrix} \times \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \left[ \frac{y_i}{s_{\varepsilon_i}^2} + \frac{1}{2} \left( \frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right)^2 \frac{\partial s_{\varepsilon_i}^2}{\partial a} \right] \\ \sum_{i=1}^n \left[ \frac{x_i y_i}{s_{\varepsilon_i}^2} + \frac{1}{2} \left( \frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right)^2 \frac{\partial s_{\varepsilon_i}^2}{\partial b} \right] \end{pmatrix} \quad (5)$$

The slope and the intercept, which are components of vector  $\hat{\mathbf{b}}$  in eq 4 and eq 5, can be determined by carrying out an iterative process on the following matrix form, eq 6:

$$\hat{\mathbf{b}} = \mathbf{R}^{-1} \mathbf{g} \quad (6)$$

With this method, and assuming that the straight line model is the correct one, the variance-covariance matrix of the calibration straight line coefficients,  $\mathbf{B}$ , is obtained by multiplying the final matrix  $\mathbf{R}^{-1}$  by the experimental error,<sup>5-6</sup>  $s^2$ . As the experimental error is unknown, the estimated value,  $\hat{s}^2$ , expressed in eq 1 should be used.<sup>2</sup>

It should be pointed out that if the situation were to be  $s_{\varepsilon_i}^2 = w_i = 1$  (i.e, all errors are due to the experimental measurement in the ordinate axis) the expressions obtained would be the same as if the Ordinary Least Squares method were to be applied.

**Joint confidence interval.** When two analytical methodologies are compared using linear regression, the plot of the values obtained from samples which have different concentrations of the analyte and which are analysed by the two methods, should give a straight line of approximately unity slope and zero intercept if the results are not statistically different at a given level of significance. In general<sup>9</sup> the confidence region -joint confidence interval- for the straight line regression coefficients corresponds to their quadratic distribution, and is given by eq 7:

$$\sum_{i=1}^2 \sum_{j=1}^2 r_{ij} (\hat{b}_i - b_i)(\hat{b}_j - b_j) = s^2 \chi_{(1-\gamma)}^2 \quad (7)$$

where  $\chi_{(1-\gamma)}^2$  is the  $1-\gamma$  level of the  $\chi^2$  distribution with two degrees of freedom and  $r_{ij}$  are the elements of the  $\mathbf{R}$  matrix.

Eq 7 gives an ellipsoid which is defined by the regression coefficients and which has its centre at  $(\hat{a}, \hat{b})$ .

The experimental error,  $s^2$ , which is only known through its estimate,  $\hat{s}^2$ , appears together with the  $\chi^2$  distribution in eq 7. From the expression of the experimental error, eq 1, it can be shown<sup>10</sup> that:

$$\sum_{i=1}^n w_i^{-1} (y_i - \hat{y}_i)^2 = s^2 \chi_{(1-\gamma)}^2 \quad (8)$$

where eq 8 is independent of eq 7.  $w_i$  stands for the weighting factor used, and  $\chi_{(1-\gamma)}^2$  is the  $1-\gamma$  level of the  $\chi^2$  distribution with  $n-2$  degrees of freedom.

The corrected ratio of the two  $\chi^2$ -distributions has an  $F$  distribution with two and  $n-2$  degrees of freedom:<sup>9</sup>

$$F = \frac{\sum_{i=1}^2 \sum_{j=1}^2 r_{ij} (\hat{b}_i - b_i)(\hat{b}_j - b_j)}{\sum_{i=1}^n w_i^{-1} (y_i - \hat{y}_i)^2} \times \frac{n-2}{2} = \frac{\sum_{i=1}^2 \sum_{j=1}^2 r_{ij} (\hat{b}_i - b_i)(\hat{b}_j - b_j)}{2\hat{s}^2} \quad (9)$$

The joint confidence test applied by Mandel and Linnig<sup>1</sup> consists of checking the presence of the theoretical point (0,1) within the limits of the joint confidence region spanned by eq 9 when the parameters of the regression line are calculated using the OLS method. However, as has been pointed out earlier, this expression does not take into account the fact that in the comparison of analytical methodologies the errors in both axes are comparable and so OLS cannot be applied without the possibility of committing considerable errors when assessing accuracy.

The estimated variance-covariance matrix obtained using the Lisý et al. method enables the  $r_{ij}$  coefficients to be determined taking into account the errors in both methodologies. The  $r_{ij}$  coefficients are the elements of the  $\mathbf{R}$  matrix in eqs 4 and 5. By introducing these coefficients into eq 7 and by taking into account the corrected ratio of the two  $\chi^2$  distributions, eq 9, it is possible to develop the joint confidence distribution for the intercept and the slope taking into account the errors in both methodologies:

$$\left| \hat{a} - a \quad \hat{b} - b \right|_x \begin{vmatrix} \sum_{i=1}^n \frac{1}{s_{e_i}^2} & \sum_{i=1}^n \frac{x_i}{s_{e_i}^2} \\ \sum_{i=1}^n \frac{x_i}{s_{e_i}^2} & \sum_{i=1}^n \frac{x_i^2}{s_{e_i}^2} \end{vmatrix} \left| \hat{a} - a \right| \left| \hat{b} - b \right| = 2\hat{s}^2 F_{1-\alpha}(2, n-2) \quad (10)$$

$$\sum_{i=1}^n \frac{1}{s_{\epsilon_i}^2} (\hat{a} - a)^2 + 2 \sum_{i=1}^n \frac{x_i}{s_{\epsilon_i}^2} (\hat{a} - a)(\hat{b} - b) + \sum_{i=1}^n \frac{x_i^2}{s_{\epsilon_i}^2} (\hat{b} - b)^2 = 2\hat{s}^2 F_{1-\alpha}(2, n-2) \quad (11)$$

where  $F_{1-\alpha}(2, n-2)$  is the tabulated  $F$ -value at a significance level of  $\alpha$  with two and  $n-2$  degrees of freedom.

The limits of the ellipse depend on the experimental errors and on the significance level chosen, and its tilt, as in the case of OLS, is the result of the well known correlation between the slope and the intercept. The ellipsoid which defines the region of the space for the joint confidence test has its axes orientated in the direction of each of the two  $V_i$  eigenvectors of the  $\mathbf{R}$  matrix. The length of the individual semiaxes is the same as  $\sqrt{2\hat{s}^2 F_{1-\alpha}(2, n-2)} x \sqrt{\lambda_i}$ , where  $\lambda_i$  is each of the two eigenvalues of the  $\mathbf{R}$  matrix.<sup>11</sup>

**Validation process.** The objective of the validation process is to assess whether the joint confidence interval test based on the BLS regression technique provides correct results, i.e. new methodologies which show no statistical differences with respect to the reference method at the level of significance chosen must be accepted and new methods which provide results that differ statistically from the results obtained using the reference method must be rejected. It will also be shown that, for several cases, the joint confidence interval test based on ordinary least squares (OLS) or weighted least squares (WLS) methods provides results which significantly differ from the ones obtained with the joint confidence interval test based on the BLS technique.

In order to assess the correctness of the test, a study will be made of three simulated and five real data sets for which uncertainties are considered in both axes and data pairs are produced with and without homoscedasticity which are

differently distributed throughout the calibration range. The first two simulated data sets give rise to a calibration line characterized by a zero intercept and a unity slope. The third set gives a line in which there is a significant difference between the result of the method on the abscissa and the result on the ordinate axis, the slope being 1.1 and the intercept zero. The Monte Carlo method<sup>12,13</sup> was used to generate 100,000 different data sets for each original regression data set. So, a random error was added to each value of the data pairs giving rise every time to 100,000 regression lines. The validation involves checking whether the method for obtaining the joint confidence interval with errors in both axes, BLS, for an  $\alpha$  significance level gives results such that the theoretical point zero intercept and unity slope falls into the joint confidence region in  $(1-\alpha)\%$  of cases. These results are compared with the ones obtained in a similar way using OLS and WLS methods.

The joint confidence interval tests based on the OLS, WLS and BLS methods were applied to the original data sets and it was confirmed that, in some cases, results are different.

## EXPERIMENTAL SECTION

**Data sets and software.** The three simulated and five real data sets below were used to validate the methodology. In all cases, the three calibration lines obtained from applying the BLS, OLS and WLS methods are shown in Figures 1 and 2.

**Data set 1:** Homoscedastic data set containing 20 pairs of  $x,y$  values equally distributed within the range 2-40 on both axes. Standard deviations of 1 were considered for all  $x$  and  $y$  values (Figure 1a).

ISBN:978-84-691-1897-9/D.L: T-353-2008

**Data set 2:** 20 data pairs randomly distributed throughout the range between 100 and 900 units. Heteroscedasticity with random standard deviation is present, in such a way that the standard deviation of any one point is higher than the standard deviation of the previous point. The standard deviation is at most 25% of the individual point (Figure 1b).

**Data set 3:** 20 data pairs with the  $x$ -values equally distributed within the range 1-20 and the  $y$ -values 10% higher than each corresponding  $x$ -value. The theoretical values of the regression coefficients are  $a=0$ ,  $b=1.1$ . Heteroscedasticity with random standard deviation is present, in such a way that each standard deviation is at most 9% of each individual value (Figure 1c).

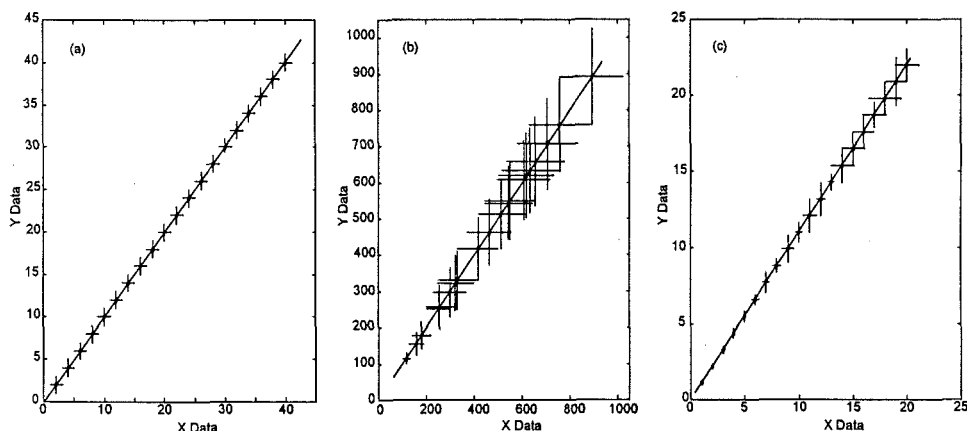


Figure 1. BLS, OLS, and WLS coincident calibration lines for the three simulated data sets used in the text, together with the individual points and their uncertainties.

**Data set 4:** a comparative study of methods for analysing  $\text{Ca}^{2+}$  in waters<sup>14</sup> using atomic absorption spectroscopy (AAS, reference method) and the technique of sequential injection analysis (SIA, new method). The uncertainties in AAS are derived from the uncertainties associated to the calculated linear regression using a univariate linear calibration computer program.<sup>15</sup> As can be seen in Figure 2a there

is considerable heteroscedasticity in this axis. The uncertainties in SIA, which are practically constant at all points, are calculated with a multivariate regression model and the PLS technique using the Unscrambler programme (Unscrambler-Ext. ver. 4.0, Camo A/S, Trondheim, Norway).

**Data set 5:** the resistance of the thicknesses (in  $\mu\text{m}$ ) of various films of commercially available photoresists and silicon dioxide substrates is measured.<sup>16</sup> An ellipsometer and a Nanospec/AFT are used to get the results. The uncertainties for all measurements represent the variations in the film thicknesses across the wafer surface (Figure 2b).

**Data set 6:** concentrations of polycyclic aromatic hydrocarbons (PAHs) recovered from railroad bed soil after supercritical fluid extraction (SFE) using two different modifiers.<sup>17</sup> The standard deviations are based on triplicate supercritical fluid extraction at each point (Figure 2c).

**Data set 7:** percentage of recovery for several organochlorine pesticides after microwave-assisted extraction (MAE) using solvent (hexane/acetone 1:1) and solvent/soil suspensions spiked with the target compounds.<sup>18</sup> The standard deviations are the average of three determinations at each point (Figure 2d).

**Data set 8:** comparative study of atomic absorption spectroscopy (AAS) and emission spectrometry using emulsion formation (ES) in the determination of Ca in lubricant oils.<sup>3</sup> The standard deviations are the average of three determinations in each condition (Figure 2e).

Applying the OLS method to the different data sets does not take into account the uncertainties of each point in the  $x$  and  $y$ -axes, and only homoscedasticity with a unity standard deviation in the  $y$ -variable is considered. When applying the WLS method, the uncertainties in the  $x$ -axis are neglected and only the uncertainties in

ISBN:978-84-691-1897-9/D.L.: T-353-2008

the  $y$ -axis are taken into account. In the first three data sets four levels of significance are used with  $\alpha$  values of 0.001, 0.01, 0.05 and 0.1.

All computations were done with home-made Matlab subroutines (Matlab for Microsoft Windows ver. 4.0, The MathWorks, Inc., Massachusetts, USA). A computer program will be available from the authors shortly.<sup>19</sup>

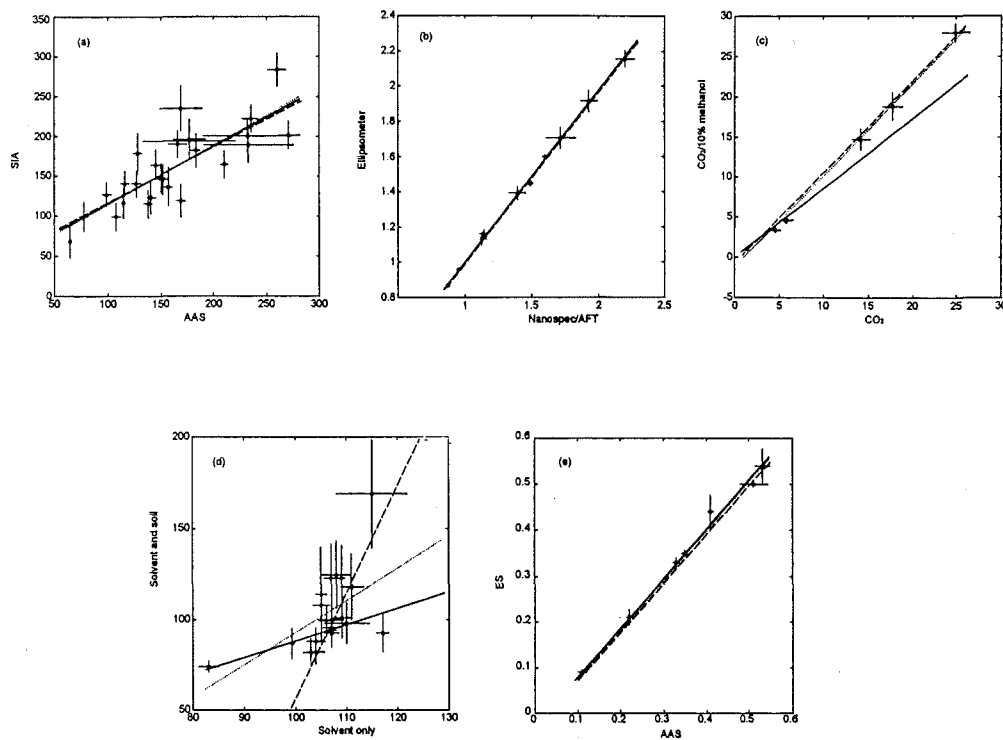


Figure 2. BLS, OLS, and WLS calibration lines for the five simulated data sets used in the text, together with the individual points and their uncertainties. BLS, solid lines; OLS, dashed lines; WLS, dotted lines.

## RESULTS AND DISCUSSION

**Data set 1.** Table 1 summarizes the results of applying the joint confidence interval test derived from BLS and OLS regression techniques to the 100,000 data sets

obtained by applying the Monte Carlo simulation to data set 1. It can be seen that, very approximately, the percentage of data sets for which the theoretical point of zero intercept and unity slope falls within the region defined by the joint confidence intervals based on the BLS or OLS methods agrees with the theoretical values given by the level of significance. In this case it is evident that the same results are obtained with the OLS and WLS methods

Due to the fact that the uncertainties in both axes are homoscedastic and of the same size ( $s = 1$  in this case) the data structure on the  $x$ -axis meets the theoretical OLS conditions. Therefore, in spite of there being uncertainties in the  $x$ -axis, the results obtained by applying the OLS method are very close to the ones obtained using BLS, with no significant differences in comparison to the theoretical values. It should be pointed out, however, that the results obtained using BLS are closer to the theoretical results than the ones obtained by OLS.

| Data set | $\alpha$ -value (%) | BLS (%) | OLS (%) | WLS (%) |
|----------|---------------------|---------|---------|---------|
| 1        | 10                  | 90.00   | 89.35   | 89.35   |
|          | 5                   | 95.01   | 94.66   | 94.66   |
|          | 1                   | 98.94   | 98.86   | 98.86   |
|          | 0.1                 | 99.90   | 99.90   | 99.90   |
| 2        | 10                  | 89.05   | 75.22   | 90.12   |
|          | 5                   | 94.21   | 84.16   | 94.98   |
|          | 1                   | 98.63   | 94.23   | 99.03   |
|          | 0.1                 | 99.81   | 98.73   | 99.90   |
| 3        | 10                  | 0.13    | 0.14    | 0.00    |
|          | 5                   | 0.40    | 0.44    | 0.01    |
|          | 1                   | 3.45    | 3.73    | 0.14    |
|          | 0.1                 | 22.85   | 22.21   | 2.71    |

Table 1. Percentage of simulated data sets for which a statistical difference has not been found between the two methods compared at four different levels of significance. 100,000 data sets were obtained for each original data set.

**Data set 2.** The results of applying BLS, OLS and WLS joint confidence interval tests for the intercept and the slope to the 100,000 data sets derived from data set 2 are summarized in Table 1. The percentage of data sets for which the theoretical point of zero intercept and unity slope falls within the region defined by the joint confidence intervals based on the BLS method agrees quite closely with the corresponding theoretical values given by the pre-established levels of significance. Similar results are obtained when using WLS conditions. However, as is to be expected from the heteroscedasticity in the  $y$ -axis, the percentage of data sets for which the theoretical point of zero intercept and unity slope fall within the region defined by the joint confidence interval based on the OLS method differ quite considerably (up to 10% with the present data set) from the theoretical values.

It should be pointed out that although there are heteroscedastic uncertainties on the abscissa axis, WLS gives good results (even though it does not take these uncertainties into account). This may be due to the fact that, in this data set, there is a constant relation between the uncertainties present in the two axes, so the heteroscedasticity considered only in the ordinate axis leads to results which are very similar to the ones obtained with the BLS method, the application of which is rigorously more correct.

Those users with data sets which have a similar structure to the ones analysed in this section are therefore compelled to use at least the joint confidence test based on WLS if they want to prevent systematic errors. The absence of readily available information and software to carry out such a test enhances the usefulness of the BLS test developed here.

**Data set 3.** The results of applying the BLS, OLS and WLS joint confidence interval tests for the intercept and the slope to the data sets derived from data set 3 are summarized in Table 1. Because 100,000 data sets are analysed all derived

from data with a slope of 1.1 and because the points are perfectly fitted to a line, the percentage of data sets obtained using the Monte Carlo method in which the theoretical point of zero intercept and unity slope falls within the joint confidence interval is low in all three methods. This serves to confirm that if there is a significant difference between the results of the method represented on the abscissa and ordinate axis, the theoretical point of zero intercept and unity slope will also fall outside the region defined by the joint confidence interval when BLS is used.

**Data set 4.** The results of applying BLS, OLS and WLS methods to data set 4 are summarized in Table 2. According to the joint confidence test based on OLS conditions, AAS and SIA with multivariate detection techniques give results which are not statistically different at a significance level  $\alpha=5\%$  but this is not true for the test based on BLS and WLS regression techniques.

| Data set | BLS | OLS | WLS |
|----------|-----|-----|-----|
| 4        | no  | yes | no  |
| 5        | yes | no  | yes |
| 6        | yes | no  | no  |
| 7        | no  | yes | no  |
| 8        | no  | yes | yes |

Table 2. Verification of whether the theoretical point of zero intercept and unity slope falls within the region defined by the joint confidence test using different calibration methods.

Figure 3a shows the joint confidence intervals obtained by applying the three regression techniques studied. It can be seen that the joint confidence interval obtained by BLS is lower than the one obtained by OLS. One factor that may have an influence is that the slopes obtained with both methods are very similar (OLS:  $\hat{b}=0.73\pm 0.11$ ; BLS:  $\hat{b}=0.74\pm 0.10$ ) but that the intercept obtained with OLS ( $\hat{a}=43\pm 8$ ) is slightly larger than the one obtained with BLS ( $\hat{a}=39\pm 16$ ). The negative correlation between the intercept and the slope means that if the data set has high or low intercepts and slopes, the theoretical point of zero intercept and unity slope is more likely to fall within the region defined by their joint confidence interval. So, OLS detects differences between the compared data sets while BLS does not.

What is more, in accordance with what could be expected of a descriptive analysis, the presence of considerable uncertainties in the  $x$ -axis makes the BLS regression line oscillate in a broader interval with the consequent broadening of the joint confidence interval. So, the ellipse obtained with BLS is larger than the one obtained with WLS which only takes into account the errors in the ordinate axis. Furthermore, in this data set there are no great differences in the values of the regression coefficients obtained by the three methods, and it is quite clear that the WLS and BLS methods give less importance to the points with most uncertainty, giving smaller ellipses in general. So, it is usually easier to find statistically significant differences between the two data sets tested.

**Data set 5.** The results of applying the joint confidence interval test obtained using BLS, OLS and WLS regression techniques to data set 5 are summarized in Table 2 and Figure 3b. The joint confidence interval test obtained with OLS indicates the lack of similarity between the two methods of measuring thickness but the theoretical point of zero intercept and unity slope falls into the joint confidence region for the BLS and WLS methods for a level of significance  $\alpha=5\%$ .

Although it is difficult to visually observe differences in the regression coefficients of the three regression lines in Figure 2b, the different conclusions reached using OLS and BLS/WLS techniques might be due to the different correlation between slopes and intercepts obtained using these three regression techniques. Although these differences are not considerable in absolute terms, the different values obtained for the regression coefficients are enough for ellipses to be constructed with significant differences in tilt, so giving conflicting conclusions about the similarity between measurement techniques, as can be observed in Figure 3b.

It is interesting to note the role of the correlation between the variables analysed (analytical methods compared) and their incidence in the statistical test which is clear from the last two sets of analysed data. There is a low correlation between variables in data set 4, just the opposite of data set 5. In contrast to OLS and the Mandel and Linnig test, a low correlation between variables, together with heteroscedasticity, makes it easier to find statistically significant differences between the methods studied.

**Data set 6.** The results of applying BLS, OLS and WLS methods to the data sets derived from data set 6 are summarized in Table 2 and Figure 3c. The effect of considerable heteroscedasticity in both axes is quite clear in this example. Three points of high uncertainty can be seen to have a strong effect on the calculations of the regression coefficients obtained with the OLS, WLS and BLS techniques. In this case, the result (Figure 3c) is that a statistically significant difference is found between the two methodologies tested with the joint confidence interval test derived from OLS and WLS but not with BLS for a level of significance  $\alpha=5\%$ . BLS clearly gives more importance to the four points with limited analytical results than to the three points which have a high degree of uncertainty. This tendency does not exist for OLS and is greatly exaggerated for WLS.

**Data set 7.** The conclusions reached with data set 6 are clearly confirmed in this data set. The presence of two points with high leverage but with a different degree of associated uncertainty (Figure 2d) is the reason for the clearly different behaviour of the tests based on BLS, OLS and WLS regression techniques i.e., while the joint confidence interval test based on OLS does not find statistically significant differences between the two methods compared, BLS and WLS do find these differences for a level of significance  $\alpha=5\%$ . So the presence of the point with high uncertainty means that the slope obtained by OLS is higher than unity and the intercept much smaller than zero and that the confidence intervals are much broader. These conditions make it more likely that the theoretical point of zero intercept and unity slope will fall within the region defined by the joint confidence interval. In contrast, the fact that BLS and WLS regression techniques take into account the uncertainties minimizes the effect of the aforementioned point, so slopes and intercepts are obtained which are smaller than unity and zero, respectively. These are not favourable conditions for the theoretical point of zero intercept and unity slope to fall within the region defined by the joint confidence interval (Figure 3d).

Despite the similarity in the conclusions obtained, the differences between this data set and data set 4 should be noted. Although the correlation between the variables on each axis is low in both cases, the origin of the discrepancy between the behaviour of the test based on OLS and the test based on BLS and WLS is quite different.

**Data set 8.** This real data set exemplifies the case in which the joint confidence interval test based on OLS or WLS regression analysis leads to results which are in conflict with the results given by the BLS regression technique (Figure 3e, Table 2). According to the latter, there is a significant statistical difference between the two methods compared whereas there is not according to the parameters obtained with OLS/WLS for a level of significance  $\alpha=5\%$ .

Although there is a slight difference in the regression coefficients obtained by the three methods which displaces the centre of the ellipse obtained by BLS and gives values which are different from the ones obtained by OLS and WLS (Figure 2e), the smaller joint confidence interval using BLS is much more important for obtaining these differing results between the three methods. It is smaller because there is a series of points with greater uncertainties than the others which means that for the WLS and BLS methods, these points are not so important when carrying out the linear calibration and so the uncertainty associated to the intercept and the slope is reduced.

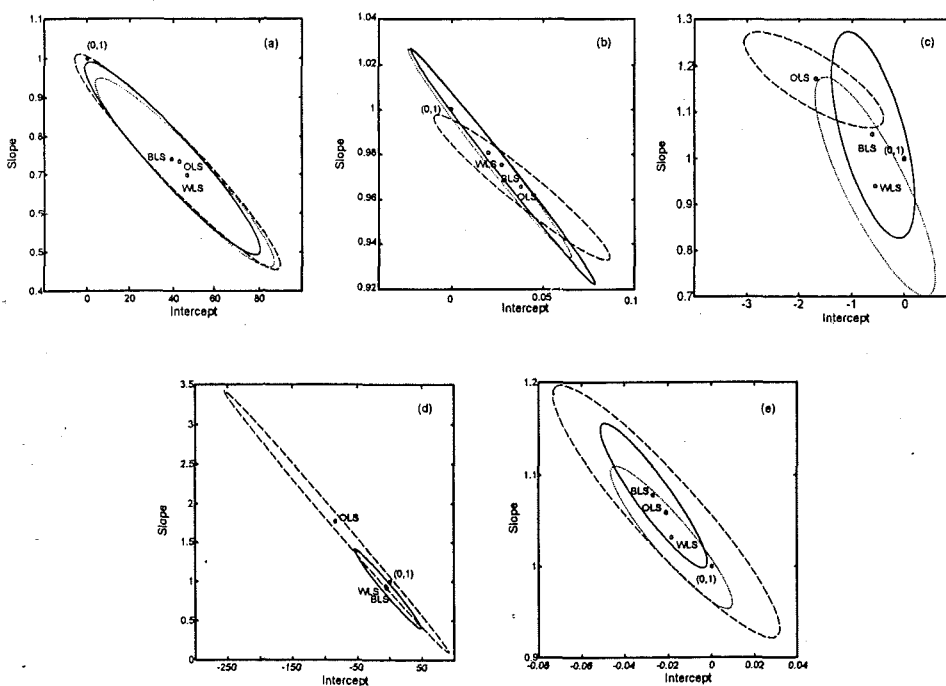


Figure 3. Joint confidence intervals based on BLS, OLS and WLS methods for the five real data sets studied in the text. BLS, solid lines; OLS, dashed lines; WLS, dotted lines.

## CONCLUSIONS

The new joint confidence test for regression coefficients based on the BLS calibration method was applied to three simulated data sets and proved to give correct results. The new test is not always necessary for drawing accurate conclusions about method equivalency. However, as is shown by its application to five real data sets, if it is not used two overall conflictive situations may be caused:

i) the theoretical point defined by zero intercept and unity slope is located within the joint confidence region derived using the BLS method, but not using the OLS or WLS methods. This is important for those users who have developed a new method which is accurate but which it has not been recognised as such until now. The result of this is that good, new methods are being rejected if the joint confidence test based on OLS or WLS methods are used. The old method will probably still be used and the figures of merit of a new method will be unnecessarily wasted.

ii) the theoretical point defined by zero intercept and unity slope is located within the joint confidence region derived using the OLS/WLS methods, but not using BLS, i.e, there are new analytical methods that give results that do not statistically differ from the reference method using the OLS or WLS methods but they do differ if the more reliable BLS regression technique is used. In these cases, the new method is wrongly considered to be accurate when using the joint confidence test based on OLS or WLS methods, and may easily give biased results whenever it is used to analyse new samples.

The new method is for general use, with no restrictive statistical constraints. It has the additional advantage of being invariant upon switching axes for the regression coefficients. The presence of possible outliers containing considerable uncertainties has the effect of reducing the joint confidence intervals in BLS and WLS with

respect to OLS. However, the user should be aware of two weaknesses; the first, general to most BLS techniques is the lack of robustness in the presence of outliers with low individual uncertainty; and the second is that the individual variances associated to each data point are needed, which may mean a longer analysis time. If these variances are omitted, the application of the BLS technique gives identical results to the ones obtained with classical OLS or WLS methods.

## ACKNOWLEDGMENTS

The authors would like to thank the DGICYT (project num. BP93-0366) for financial support and the Spanish Ministry of Education and Science for providing Jordi Riu's doctoral fellowship.

## LITERATURE CITED

- (1) Mandel, J.; Linnig, F. J.; *Anal. Chem.* **1957**, *29*, 743-749
- (2) Draper, N.; Smith, H. *Applied Regression Analysis, Second Edition*; John Wiley & Sons: New York, 1981; pp 5-128
- (3) De la Guardia, M.; Salvador, A.; Berenguer, V. *Ann. Quim.* **1980**, *77*, 129-132
- (4) Myers, R. H. *Classical and Modern Regression with Applications (2nd Edition)*; Duxbury Press: Belmont, California, 1990; pp 18-21
- (5) Wentworth, W. E. *J. Chem. Educ.* **1965**, *42*, 96-103
- (6) Wentworth, W. E. *J. Chem. Educ.* **1965**, *42*, 162-167
- (7) Riu, J.; Rius, F. X. *J. Chemom.* **1995**, *9*, 343-362
- (8) Lisý, J. M.; Cholvadová, A.; Kutej, J. *Computers Chem.* **1990**, *14*, 189-192

ISBN: 978-84-691-1897-9/D.L.: T-353-2008

- (9) Mood, A. McF. *Introduction to the Theory of Statistics (2nd Edition)*; McGraw-Hill: New York, 1963; pp 112-301
- (10) Scheff J, H. *The Analysis of Variance*, John Wiley & Sons, New York (1959); pp 3-41
- (11) Meloun, M.; Militk, J.; Forina, M. *Chemometrics for Analytical Chemistry Vol II*; Ellis Horwood: London, 1994; pp 20-24
- (12) Meier, P. C.; Zünd, R. E. *Statistical Methods in Analytical Chemistry*; John Wiley & Sons: New York, 1993; pp 145-150
- (13) Güell, O.; Holcombe, J. A. *Anal. Chem.* **1990**, *60*, 529A-542A
- (14) Ruisánchez, I.; Rius, A.; Larrechi, M. S.; Callao, M. P.; Rius, F. X. *Chemom. Int. Lab. Sys.* **1994**, *24*, 55-63
- (15) Boqué, R.; Rius F.X.; Massart D.L. *J. Chem. Educ. (Computer Series)* **1994**, *71*, 230-232
- (16) Gamsky, C. J.; Howes, G. R.; Taylor, J. W. *Anal. Chem.* **1994**, *66*, 1015-1020
- (17) Langenfeld, J. J.; Hawthorne, S. B.; Miller, D. J.; Pawliszyn, J. *Anal. Chem.* **1994**, *66*, 909-916
- (18) López-Avila, V.; Young, R.; Beckert, W. *Anal. Chem.* **1994**, *66*, 1097-1106
- (19) Riu, J.; Rius, F. X. In preparation.

## Method comparison using regression with uncertainties in both axes

Jordi Riu\* and F. Xavier Rius

*Departament de Química. Universitat Rovira i Virgili.*

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

### ABSTRACT

Method comparison is one of the most important activities within the domain of method validation. The results of newly developed analytical methods are regressed against those obtained using reference methods for a series of samples with different concentrations of the analyte of interest. The linear model, based on the bivariate least squares (BLS) calibration method, developed taking into account the comparable errors in both axes, should fit a straight line where the intercept is not significantly different from zero and the slope not significantly different from one. To check these premises, the joint confidence interval for the intercept and the slope is usually applied. This paper compares the results of this method with the joint confidence test for the intercept and the slope based on the ordinary least squares (OLS) and weighted least squares (WLS) methods, using two real data sets.

## The problem

The comparison of analytical methods using regression analysis has been used since the fifties when Linnig and Mandel [1] first applied the joint confidence interval test for the intercept and the slope to chemical problems. However, this test applied to the regression parameters derived from the least squares method assumes that the results in the  $x$ -axis (often the reference method) are error-free, or that the errors assigned to the reference method are negligible with respect to the ones given by the new method ( $y$ -axis). This is not always true since the precision of both methods must be often taken into account. These precisions can be considered using the different existing approaches for calculating regression coefficients and related statistical parameters considering errors in both axes. [2]

A new procedure for comparing analytical methods that takes into account the errors in both methods has recently been developed. [3] This approach, based on the bivariate least squares (BLS) calibration method, [4] computes the joint confidence interval for the intercept and the slope taking into account the errors in both methods. This paper compares the results of this test with the joint confidence interval test for the intercept and the slope based on the ordinary least squares (OLS) and weighted least squares (WLS) methods, applied to the recovering of polycyclic aromatic hydrocarbons (PAHs), and to the determination of As(V) in water.

## The computer program

All computations were made using home-made Matlab subroutines (Matlab 4.0 for Microsoft Windows 3.1 or higher, The MathWorks, Inc., Massachusetts, USA). The program, structured in menus, calculates the joint confidence intervals for the slope and the intercept using the data input by the user, either manually or from a file in ASCII code. The procedure is divided in two steps:

- calculation of the regression coefficients using the bivariate least squares method.
- development of the joint confidence interval of the regression coefficients.

```

% Calculation of the regression parameters by means of BLS

xmin=min(x);
xmax=max(x);
b=0;
a=0;
slope=0.1;
cvxy=0;
while abs(slope-b)>10^(-12);
    slope=b;
    w=var(y+b^2*varx);
    r=(y-a-b*x).*(y-a-b*x);
    r1=1./w;
    r2=x./w;
    r4=(x.*x)./w;
    g1=(y./w);
    g2=(x.*y)./w+.5.*(r./(w.*w)).*(2.*b.*varx-2.*cvxy);
    R=[sum(r1) sum(r2);sum(r2) sum(r4)];
    g=[sum(g1);sum(g2)];
    det=sum(r1)*sum(r4)-(sum(r2))^2;
    T=[sum(r4)/det -sum(r2)/det; -sum(r2)/det sum(r1)/det];
    sol=T*g;
    a=sol(1,1);
    b=sol(2,1);
end;

% T*s^2 is variance-covariance matrix
% R*(1/s^2):coefficient matrix in quadratic form of the distribution (a-ahat)(b-bhat)

vara=T(1,1);
varb=T(2,2);
covar=T(2,1);
aa=sol(1,1);
bb=sol(2,1);
su=(1./w).*r;
suma=sum(su);

% ss is s^2, estimation of the experimental error
ss=suma/(n-2);

% desa i desb are the respective standard deviations of a and b
desa=sqrt((vara)*suma/(n-2));
desb=sqrt((varb)*suma/(n-2));

```

Fig. 1. Matlab code for the calculation of the regression coefficients.

ISBN:978-84-691-1897-9/D.L.: T-353-2008

As an example, the code for the calculation of the regression coefficients is shown in Figure 1.

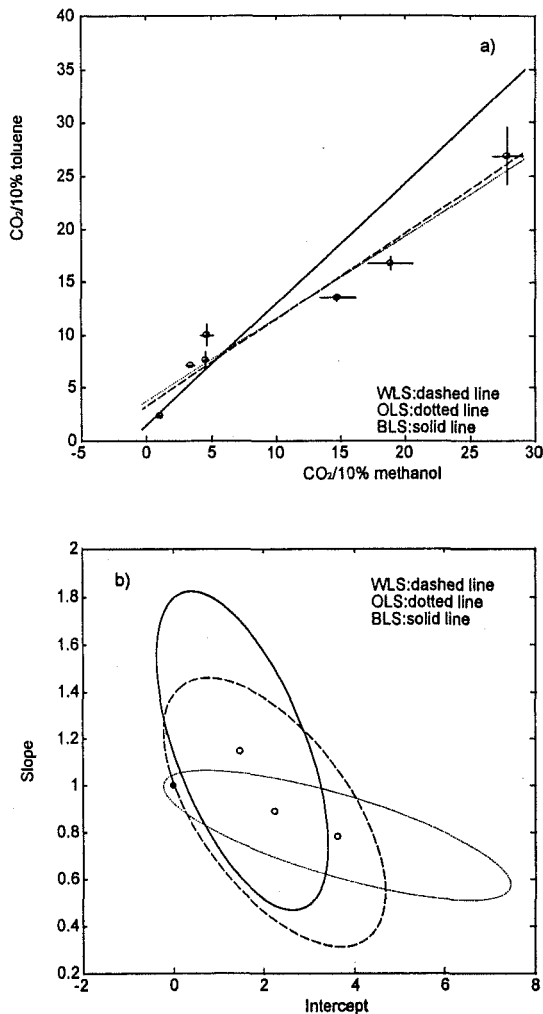


Figure 2.a) BLS, OLS and WLS calibration lines for the methods for recovering polycyclic aromatic hydrocarbons (PAHs) from railroad bed soil together with the individual points and their uncertainties. b) Joint confidence intervals based on BLS, OLS and WLS techniques. Empty points indicate the ellipse centroid while the solid one indicates the theoretical point of zero intercept and unity slope.

For a slope precision of eight decimal places, and for several data sets tested, the time taken to do the calculations and draw the figures ranges between 3-5 seconds

with a 486 processor at 66Mhz. The program is available from the authors on request.

It is of interest to note that the joint confidence interval test is not affected if the values are shifted by the same amount along the coordinate axes, i.e. the same amount is added or subtracted from each coordinate of each point. By so doing the value of the scale is reduced and the possibility of introducing numerical instability into the calculation process decreases. The slope and its standard deviation are not affected by this shift, but both the value of the intercept and its standard deviation are. When the scale is increased, the intercept value increases in an amount proportional to its standard deviation, so the results given by the joint confidence interval remain unchanged. Problems related to differences in errors limits of the two measurements that are compared are being studied. [7]

### Two examples of application

*SFE analysis of soils.* Two methods for recovering polycyclic aromatic hydrocarbons (PAHs) from railroad bed soil after supercritical fluid extraction (SFE) are compared. [5] The modifiers used are CO<sub>2</sub>/10% methanol (represented on the *x*-axis) and CO<sub>2</sub>/10% toluene (represented in the *y*-axis). The uncertainties are based on triplicate supercritical fluid extraction in each condition. Figure 2a shows the regression lines obtained by applying ordinary least squares (OLS), weighted least squares (WLS) and bivariate least squares (BLS). The main statistical parameters obtained by applying the different regression techniques are listed in Table 1. In this particular data set, for some points the uncertainty in the *x*-variable is more important and for others it is the uncertainty in the *y*-variable. The fact that the regression parameters are calculated by means of OLS means that these differences in the uncertainty are ignored because the *x*-axis is considered to be error free and the *y*-axis homoscedastic. Using WLS brings us closer to the real situation, since it considers the uncertainties in the *y*-axis, but it still considers the

$x$ -axis to be error free. The use of these methods is reflected in the regression parameters shown in Table 1.

|     | $a$  | $s_a$ | $b$  | $s_b$ |
|-----|------|-------|------|-------|
| OLS | 3.63 | 1.14  | 0.78 | 0.08  |
| WLS | 2.25 | 0.70  | 0.89 | 0.17  |
| BLS | 1.50 | 0.56  | 1.15 | 0.20  |

Table 1. Main statistical parameters obtained applying the different regression techniques to the method comparison study for the SFE analysis of soils.

Figure 2b shows the joint confidence intervals for the intercept and the slope with different regression techniques: OLS, WLS and BLS. The level of significance chosen for the joint confidence interval for the intercept and the slope was  $\alpha=5\%$  in all three cases.

According to Figure 2b, applying the BLS technique different results are obtained by both methods at the level of significance chosen. It is interesting to note that if the joint confidence test for the intercept and the slope with OLS conditions were to be applied to this data set, it would be concluded that the two recovery methods studied provide comparable results since the theoretical point of zero intercept and unity slope lies within the limits of the OLS ellipsoid. Therefore, the erroneous conclusion that the two methods are comparable when in fact they are not would be drawn. As a consequence, unless BLS is applied, a procedure is accepted as good when in fact it does not produce accurate results. The fact that WLS is used produces results that are closer to the correct ones than using OLS.

*Determination of As(V) in water.* Two methods for determining arsenic (V) in natural waters are compared: [6] continuous selective reduction and atomic

absorption spectrometry (represented on the  $x$ -axis), and non-selective reduction, cold trapping, and atomic emission spectrometry (represented on the  $y$ -axis). Figure 3a shows the regression lines obtained when applying ordinary least squares (OLS), weighted least squares (WLS) and bivariate least squares (BLS). The main statistical parameters obtained by applying the different regression techniques are listed in Table 2. Figure 3a has strong heteroscedasticity in both axes which is not taken into account if the regression parameters are calculated by the OLS method. The WLS method only takes into account the heteroscedasticity in the  $y$ -axis and ignores what is present in the  $x$ -axis, which in this particular data set tends to be as important for nearly all the points as the uncertainty in the  $y$ -axis.

|     | $a$  | $s_a$ | $b$  | $s_b$ |
|-----|------|-------|------|-------|
| OLS | 0.54 | 0.26  | 0.85 | 0.05  |
| WLS | 0.01 | 0.01  | 0.89 | 0.07  |
| BLS | 0.00 | 0.01  | 1.01 | 0.09  |

Table 2. Main statistical parameters obtained applying the different regression techniques to the method comparison study for the determination of As(V) in water.

Figure 3b shows the joint confidence intervals for the intercept and the slope found taking into account OLS, WLS and BLS conditions. The level of significance chosen by the joint confidence interval for the intercept and the slope was  $\alpha=5\%$  in all three cases.

From Figure 3b it can be concluded that the two methods for analysing arsenic(V) in natural waters give comparable results when the BLS technique is applied at the level of significance chosen. If the joint confidence test for the intercept and the slope with OLS conditions had been applied to this data set, the theoretical point of

ISBN:978-84-691-1897-9/D.L: T-353-2008

zero intercept and unity slope would have been placed outside the region defined by the joint confidence interval, and so the erroneous conclusion would have been drawn that the two methods are not comparable when in fact they are. In this particular case, a procedure would be rejected that in fact gives accurate results. Applying WLS gives results that are closer to the correct ones than by applying OLS.

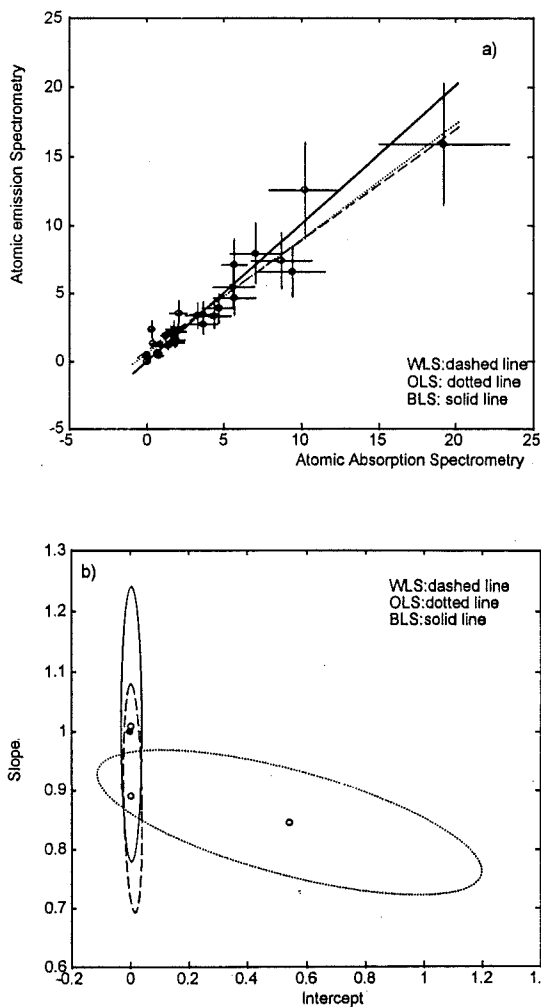


Figure 3. a) BLS, OLS and WLS calibration lines for the methods for determining As(V) in water together with the individual points and their uncertainties. b) Joint confidence intervals based on BLS, OLS and WLS. Empty points indicate the ellipse centroid while the solid one indicates the theoretical point of zero intercept and unity slope.

## Conclusions

The achievement of traceability is a permanent concern of regulatory bodies. Method comparison is a way of demonstrating traceability but, as shown in the present work, comparison must be carried out using not only the appropriate chemical procedures but also the suitable statistical techniques as well. Chemical literature might be full of erroneously considered traceable or accurate methods when in fact they are not, since the commonly used joint confidence interval for the slope and the intercept based on OLS regression has been used whenever the statistical hypotheses are not accomplished. Using the joint confidence interval for the slope and the intercept with WLS conditions produces, in some cases, better results than the ones obtained with OLS conditions, but it can also lead to any or the two aforementioned incorrect situations [3]. The test developed for the BLS conditions has a broad application and overcomes this type of problems but requires the knowledge of the individual uncertainties of the points that are involved in the regression.

## Acknowledgements

The authors would like to thank the DGICyT (project num. BP93-0366) for financial support and the Spanish Ministry of Education and Science for providing Jordi Riu's doctoral fellowship.

## References

- [1] J. Mandel and F.J. Linnig, *Anal. Chem.*, 29 (1957) 743.
- [2] J. Riu and F.X. Rius, *J. Chemom.*, 9 (1995) 343.
- [3] J. Riu and F.X. Rius, *Anal. Chem.*, 68 (1996) 1851.

ISBN:978-84-691-1897-9/D.L: T-353-2008

[4] J.M. Lisý, A. Cholvadová and J. Kutej, *Computers Chem.*, 14 (1990) 189.

[5] J.J. Langenfeld, S.B. Hawthorne, D.J. Miller and J. Pawliszyn, *Anal. Chem.*, 66 (1994) 909.

[6] B.D. Ripley and M. Thompson, *Analyst*, 112 (1987) 377.

[7] R. Boqué, Personal communication

## Detection of bias in method-comparison studies

Jordi Riu\* and F. Xavier Rius

*Departament de Química. Universitat Rovira i Virgili de Tarragona.*

*Pl. Imperial Tàrraco, 1 43005 Tarragona. Catalonia, Spain*

### SUMMARY

In this study we compare a newly developed chemometric technique for detecting the presence of bias in analytical results in comparison to previously used techniques. The new technique is based on the joint confidence interval test for the slope and intercept of the straight line which were obtained by regressing the results for different concentration values of the analyte of interest which had previously been determined by two different analytical methodologies. The new test takes into account both heteroscedasticity and errors in both regression axes and is compared to the joint confidence interval for the slope and intercept developed by Mandel and Linnig, the individual tests method of De la Guardia et al. and the simultaneous test of a composite hypothesis. All the tests were applied to the results obtained when two methods for determining ethylene content in poly(propylene-ethylene) copolymers were compared. The Mandel and Linnig and the De la Guardia et al. methods and the simultaneous test of a composite hypothesis lead to conclusions that are different from the new method.

## INTRODUCTION

The validation step is fundamental in the development of a new analytical methodology. One procedure for verifying the trueness of a new method, i.e., the absence of systematic errors or bias, [1] is to compare its results with the ones obtained using an already validated analytical method, usually a reference method. So, the coefficients of the least squares regression line, calculated by regressing the results for different concentration values of the analyte of interest are compared to the hypothetical values of a straight line obtained without bias, that is to say, a slope of one and an intercept of zero.

Following these premises, Mandel and Linnig [2] first applied in 1957 the joint confidence interval test for the intercept and the slope to analytical chemistry, concluding that the methods under study are comparable if the theoretical point corresponding to zero intercept and unity slope falls into the ellipse defined by the joint confidence interval. This procedure has been extensively used in the bibliography but may lead to incorrect results since it is based on statistical hypotheses that very often are not accomplished. One of these hypotheses states that the uncertainties of the experimental values (normally measured as their standard deviation) are constant throughout the calibration range (homoscedasticity). In chemical analysis, the uncertainty of the measurements is normally related to the absolute value of the measurand. The second hypothesis states that the error associated to the  $x$ -values is negligible compared to the  $y$ -values. This premise is normally accepted in the calibration step but it might be unacceptable when comparing analytical methods.

30 years later, De la Guardia et al. [3] pointed out these limitations and suggested using individual confidence intervals for the intercept and the slope to assess the absence of bias of an analytical method. To find the coefficients of the regression line and other statistical parameters to derive the individual confidence intervals,

De la Guardia et al. applied a weighted regression. This kind of regression considers heteroscedasticity in the data set and the presence of uncertainties in the values of the  $x$ -variable. The coefficients of the calibration line obtained with this approach may lead to better results than those obtained by applying the ordinary least squares method, OLS, but they do not consider the covariance between the two straight line coefficients and, therefore, give rise to possible erroneous conclusions.

Another test which has been proposed to verify the trueness of analytical methods is the simultaneous test of a composite hypothesis. [4] It compares the squared sum of residuals of the regression line with the squared sum of residuals of the theoretical line with intercept zero and slope one.

In this paper, the previous approaches are compared to a procedure based on a calibration method that takes into account the errors in both axes to derive the joint confidence interval for the intercept and the slope. [5] Of all the different existing approaches for calculating regression coefficients considering errors in both axes, Lisý's method [6] was selected because of its speed in estimating the correct results for the regression coefficients and because the simplicity of programming its algorithm. The joint confidence interval test for the intercept and the slope taking into account the errors in both axes has been shown to produce reliable results in different experimental conditions. [5,7] The procedure considers the presence of heteroscedastic data, usually obtained in analytical methods, and errors in both the reference and the newly tested methods. When the Mandel and Linnig and De la Guardia et al. procedures and the simultaneous test of a composite hypothesis are applied to data from a comparison study of the application of Near Infrared Spectrometry (NIRS) and two multivariate calibration techniques (PCR selecting the factors for their ability to model the concentration values and PLS1) for determining ethylene content in poly(propylene-ethylene) copolymers, the

conclusions are not the same as when the joint confidence interval for the intercept and the slope taking into account the errors in both axes is applied.

## LINEAR REGRESSION IN METHOD COMPARISON

The method developed by Mandel and Linnig [2] calculates the joint confidence interval for the intercept and the slope of the regression line using the results obtained from the reference method ( $x$ -variable) and the method to be validated ( $y$ -variable). The relationship between  $x$  and  $y$  for the  $n$  points in the data set is defined by:

$$y_i = a + bx_i + \varepsilon_i \quad (1)$$

$\varepsilon_i$  being the  $i$ -th residual for a random error with a normal distribution, and the coefficients  $a$  and  $b$  of the calibration line are found using the OLS method. If no systematic errors are present, the calibration line must have an intercept which is not significantly different from zero and a slope which is not significantly different from one, and the theoretical point (intercept = 0, slope = 1) is found in the region defined by the joint confidence interval for the intercept and the slope. The joint confidence region is an ellipse, and its tilt is the consequence of the correlation between the intercept and the slope. The method has the drawback that it does not take into account either the errors in the  $x$ -variable (i.e., the reference method) or the heteroscedasticity in both  $x$  and  $y$  variables.

De la Guardia et al. [3] reported a statistical test which is based on the individual confidence intervals for the intercept and the slope obtained by a weighted linear calibration. They considered that each data pair incorporates a known variance ( $s_{x_i}^2$  and  $s_{y_i}^2$ ), and depending on to these variances, a weighting factor is applied to each experimental point before carrying out the regression:

$$g_i = \frac{1}{s_{y_i}^2 + b_0^2 s_{x_i}^2} \quad (2)$$

where  $b_0$  is the first estimation of the slope, made by OLS. However, this procedure does not take into account that the term for the slope that appears in the weighting factor, Eq.(2), also appears in the regression model, Eq.(1), and an iterative procedure should be carried out until the convergence of the slope. This results in biased results. This method allows for errors in both axes but there are considerable discrepancies between the regression coefficients obtained and the ones calculated with any of the correct methods that consider errors in both axes (e.g. Lisý's method et al. [6,8]).

Once the regression coefficients have been found, the De la Guardia et al. statistical test for checking the trueness of an analytical method compares each of the regression coefficients ( $M$ , with standard deviation  $s_M$ ) obtained by their weighted lineal calibration, with its theoretical value if the two methods would be comparable,  $M_0$  (0 for the intercept, 1 for the slope), employing the discriminant function:

$$t_{\text{exp}} = \frac{M - M_0}{s_M} \quad (3)$$

The  $t_{\text{exp}}$  value obtained from Eq.(3) is compared with the  $t_{1-\alpha}$  value given by the Student's  $t$ -distribution with  $n-2$  degrees of freedom and an  $\alpha$  significance level. However, the use of individual confidence intervals does not take into account the correlation between intercept and slope. The use of the individual  $t$ -distributions with an  $\alpha$  significance level defines a rectangular confidential region. Mandel and Linnig pointed out that if the correlation between the distributions is considered, the confidence region becomes an ellipse. [2] So the correlation between the

intercept and the slope may cause a method to be considered free from bias if the individual confidence intervals are taken into account but to be rejected if the joint confidence interval is considered. [9]

Another method which assesses the trueness of analytical methods using linear regression is the simultaneous test of a composite hypothesis. [4] This method tests the composite hypothesis  $H_0: a=0$  and  $b=1$ , against the alternative hypothesis  $H_A: a \neq 0$  and  $b \neq 1$ , for the regression coefficients found using ordinary least squares. The method finds the residual sum of squares RSS, assuming that the composite hypothesis is valid:

$$RSS = \sum_{i=1}^n (y_i - 0 - 1x_i)^2 \quad (4)$$

and the corresponding residual sum of squares  $RSS_1$  if the composite hypothesis is false:

$$RSS_1 = \sum_{i=1}^n (y_i - a - bx_i)^2 \quad (5)$$

The difference ( $RSS_1 - RSS$ ) corresponds to an increase in the residual sum of squares caused by the validity of the null hypothesis  $H_0$ . The test criterion takes the form:

$$F_1 = \frac{(RSS_1 - RSS) \times (n-2)}{RSS \times 2} \quad (6)$$

which has the Fisher-Snedecor distribution with 2 and  $n-2$  degrees of freedom. If the experimental value  $F_1$  found in Eq.(6) for the regression line is higher than the value of the Fisher-Snedecor distribution with two and  $n-2$  degrees of freedom for

an  $\alpha$  significance level, the null hypothesis  $H_0$  must be rejected, thus showing that the two methods are not comparable.

Recently, a new statistical test for method validation that finds the joint confidence interval for the intercept and the slope, allowing for both heteroscedasticity and errors in both axes has been developed. [5] This method finds the regression coefficients in Eq.(1) using Bivariate Least Squares (BLS) Lisý's procedure, [6] an iterative process that takes into account the errors in both axes by minimizing the expression:

$$S = \sum_{i=1}^n \frac{(y_i - a - bx_i)^2}{w_i} \quad (7)$$

The weighting factor of the method is the variance of the  $y$ -residuals ( $s_{e_i}^2$ ), which residuals can be expressed using the Taylor series, even when the covariance between each data pair is not zero:

$$w_i = s_{e_i}^2 = s_{y_i}^2 + b^2 s_{x_i}^2 - 2b \text{cov}(x_i, y_i) \quad (8)$$

Due to the presence of the slope in the weighting factor and in the expression to be minimized, an iterative process is applied until the slope converges. This method allows the user to easily find the variance-covariance matrix of the coefficients of the regression line, which is used in the development of the joint confidence interval. As in Mandel and Linnig's method, the tested analytical method and the reference method are comparable if the theoretical point (intercept zero, slope one) falls in the joint confidence region, which is ellipse-shaped. This procedure overcomes the limitations of the above mentioned approaches (i.e., homoscedasticity and errors in only one axis in Mandel and Linnig's method), and

the individual confidence intervals derived from incorrect regression coefficients in De la Guardia's et al. method.

## EXPERIMENTAL SECTION

The percentage of ethylene in 86 samples of poly(propylene-ethylene) copolymers was studied with Near Infrared Spectroscopy (NIRS) and two multivariate calibration techniques (PCR selecting the factors for their ability to model the concentration values and PLS1) [10] and the data obtained was used to compare the different procedures. Absorbances at 178 wavelengths were measured, and finally four factors for PLS1 and five factors for PCR were selected to built the model. Validation was carried out using cross-validation both for PCR and PLS1. The uncertainties in each axis are derived from the root mean squared error of prediction (RMSEP) calculated by using the two multivariate regression models.

| method               | $a$    | $s_a$  | $b$    | $s_b$  | $s^2_{y/x}$ |
|----------------------|--------|--------|--------|--------|-------------|
| OLS                  | 0.5625 | 0.2211 | 0.9249 | 0.0288 | 0.2179      |
| De la Guardia et al. | 0.5990 | 0.2083 | 0.9198 | 0.0281 | 0.1984      |
| BLS                  | 0.3066 | 0.2106 | 0.9605 | 0.0284 | 0.1948      |

Table 1. Coefficients of the calibration line and estimation of pure error found using the OLS, De la Guardia et al. and BLS methods.

## RESULTS

The individual points corresponding to the concentration values together with their associated uncertainties and derived regression lines are plotted in Figure 1. The regression coefficients and their respective standard deviations calculated using

OLS, De la Guardia at al. and Lisy's calibration methods are listed in Table 1. It should be pointed out that the differences among these values are mainly attributed to uncertainties in the  $x$ -axis.

Table 2 shows the  $t$ -values for the individual confidence intervals for the slope and intercept. They were calculated using Eq.(3) for each of the three calibration methods which were applied to the data set in Fig. 1.

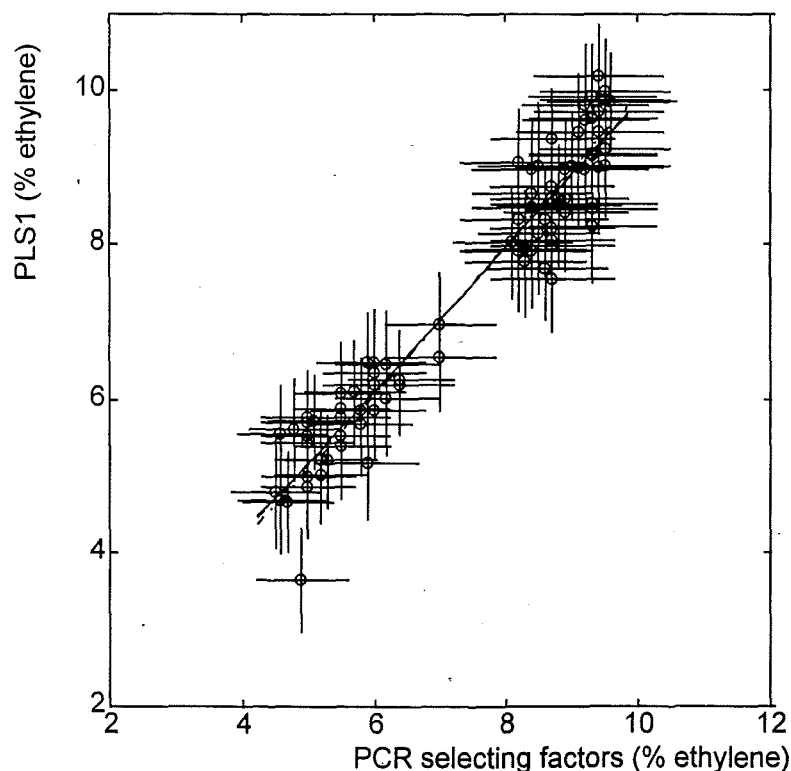


Figure 1. Graphical representation of the 86 samples used in the method comparison study together with their individual uncertainties. De la Guardia, OLS and BLS regression lines are nearly coincident.

ISBN:978-84-691-1897-9/D.L: T-353-2008

When these values are compared to the theoretical  $t$ -value for a significance level of 5% and 84 degrees of freedom ( $t=1.960$ ), it can be concluded that PCR selecting the factors and PLS1 produce comparable results when they are used to determine the percentage of ethylene in poly(propylene-ethylene) with the BLS calibration method, while the least squares and the De la Guardia et al. methods conclude that the two methods are not statistically comparable.

The  $F$ -values for the simultaneous composite hypothesis test, Eq.(6), were also calculated using the regression coefficients which had been calculated with the OLS, De la Guardia et al. and BLS calibration methods. The results are listed in Table 3. When these values are compared with the theoretical  $F$ -value for a significance level of 5% and 84 and 2 degrees of freedom ( $F=3.105$ ), it can be concluded that PCR selecting factors and PLS1 give comparable results when they are used to determine the ethylene content in poly(propylene-ethylene) with the BLS method, while the least squares and the De la Guardia et al. methods conclude that the two multivariate methods PCR selecting factors and PLS1 are statistically different at the level of significance chosen of  $\alpha=0.05$ .

| method               | $t_{\text{exp } a}$ | $t_{\text{exp } b}$ | $t_{\text{exp}} < t_{\text{theoretical}}$ | Bias detected |
|----------------------|---------------------|---------------------|---|---------------|
| OLS                  | 2.544               | 2.613               | Neither $a$ nor $b$                       | Yes           |
| De la Guardia et al. | 2.876               | 2.855               | Neither $a$ nor $b$                       | Yes           |
| BLS                  | 1.455               | 1.390               | Yes, $a$ and $b$                          | No            |

Table 2.  $t$ -values for the intercept and slope calculate from Eq.(3) and using the OLS, De la Guardia et al. and BLS calibration methods ( $t_{\text{theor}} = 1.960$ ).

| method        | $F_1$ | $F_{\text{exp}} < F_{\text{theoretical}}$ | Bias<br>detected |
|---------------|-------|---|------------------|
| OLS           | 3.156 | No  | Yes              |
| De la Guardia | 3.141 | No  | Yes              |
| BLS           | 2.427 | Yes                                       | No               |

Table 3.  $F$ -values for the simultaneous composite hypothesis test with the regression coefficients found using the OLS, De la Guardia et al. and BLS calibration methods ( $F_{\text{theor}} = 3.105$ ).

Similarly, Figure 2 shows both the joint confidence interval for the intercept and the slope based on OLS and BLS conditions, and the individual confidence intervals based on the De la Guardia et al. method. The position of the theoretical point of zero intercept and unity slope, indicates that the two methods are not comparable for the Mandel and Linnig and the De la Guardia procedures (apart from the fact that the De la Guardia et al. method does not take into account the correlation between intercept and slope), while the joint test based in BLS conditions concludes that there is no reasons to presume that the two methods PCR selecting factors and PLS1 produce different results at the level of significance chosen of 0.05.

## CONCLUSIONS

The fact that there are no statistical tests appropriate to the needs of analytical problems has meant that analysts have been using erroneous validation procedures for a long time, despite being fully aware of their limitations. This is the case of the

well known Mandel and Linnig test for assessing the presence of bias by comparing two procedures with linear regression. In this study we have shown, with a set of real data, the limitations of three procedures in comparison to the new BLS methodology which is capable of overcoming the previous drawbacks.

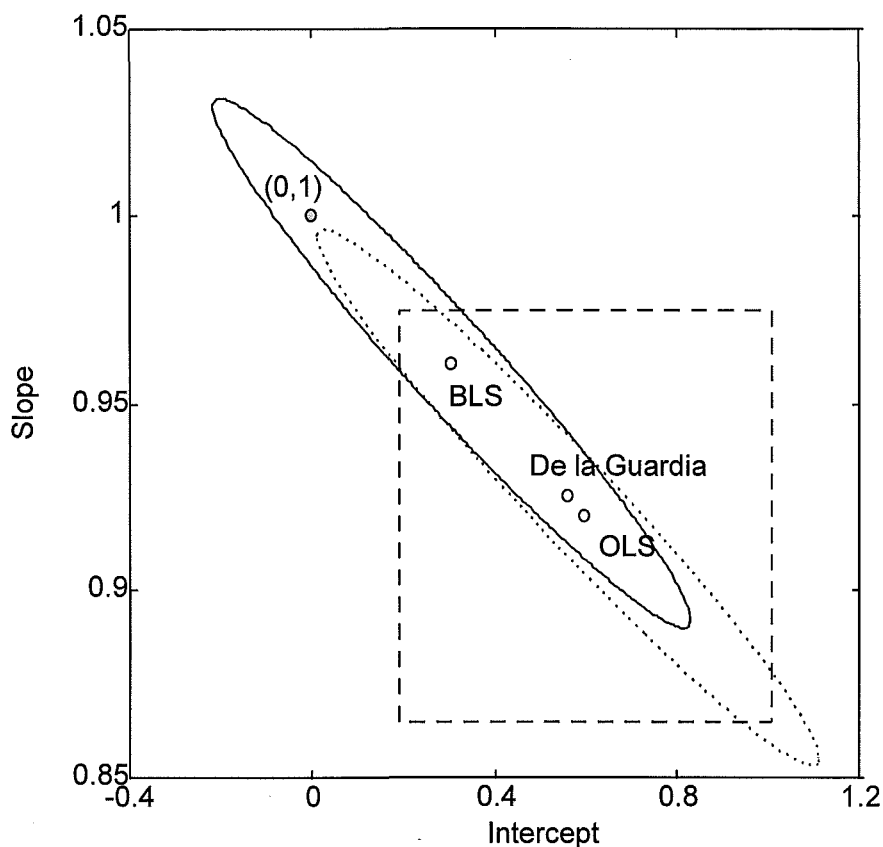


Figure 2. The joint confidence interval for the intercept and the slope based on the ordinary least squares and bivariate least squares calibration method and the individual confidence interval based on the De la Guardia calibration method. De la Guardia: dashed line, OLS: dotted line, BLS: solid line.

## ACKNOWLEDGEMENTS

The authors would like to thank the Spanish Ministry of Education and Science (DGICyT project num. BP93-0366) for their financial support. Jordi Riu also thanks the Spanish Ministry of Education and Science for providing his doctoral fellowship.

## REFERENCES

- [1] BAM Federal Institute for Material Research and Testing Statistical Techniques for Interlaboratory Studies and Related Topics, Technical Report No. 193/ No. 13, Berlin 1993
- [2] J. Mandel and F.J. Linnig, *Anal. Chem.*, 29, 743, 1957
- [3] M. De la Guardia, A. Salvador and V. Berenguer, *Ann. Quim.*, 77, 129, 1980
- [4] M. Meloun, J. Militký and M. Forina, *Chemometrics for Analytical Chemistry*. Volume 2, Ellis Horwood London, 1994
- [5] J. Riu and F.X. Rius, *Anal. Chem.*, 68, 1851, 1996
- [6] J.M. Lisý, A. Cholvadová and J. Kutek, *Computers Chem.*, 14, 189, 1990
- [7] J. Riu and F.X. Rius, *Trends Anal. Chem.*, In Press
- [8] J. Riu and F.X. Rius, *J. Chemom.*, 9, 343, 1995
- [9] D. L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a text book*, Elsevier Amsterdam, 1988
- [10] C. Villagrasa, J. Ferré, M.S. Larrechi, F.X. Rius and C. García, Submitted for publication.

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 5

---

**Regressió lineal multivariant considerant errors en tots els eixos. Aplicació a la comparació de múltiples mètodes**

Després del desenvolupament al capítol 4 del test conjunt per a la ordenada a l'origen i el pendent basat en el mètode BLS, una de les seves extensions lògiques era l'ampliació d'aquest test per tal de poder comparar els resultats de més de dos mètodes analítics. La comparació dels resultats obtinguts amb múltiples mètodes analítics es pot dur a terme a un nivell de concentració mitjançant l'anàlisi de la variància (ANOVA),<sup>1</sup> però per comparar a diversos nivells de concentració, aquesta tècnica és tediosa i requereix moltes rèpliques.

La complexitat de la comparació de múltiples mètodes analítics a diversos nivells de concentració és superior respecte a la comparació de només dos mètodes. Per exemple, en la comparació de dos mètodes es disposa de la representació gràfica de la recta de regressió entre els resultats, cosa que facilita l'anàlisi visual de les dades. En la comparació de múltiples mètodes no es disposa d'aquesta representació (excepte quan es comparen tres metodologies, en què es pot disposar d'una gràfica tridimensional de la representació dels resultats obtinguts amb els tres mètodes), i lògicament el procés de càlcul també es complica, i encara més com més mètodes estiguin involucrats en la comparació.

La regressió dels resultats obtinguts analitzant una sèrie de mostres mitjançant  $j$  mètodes analítics dona un hiperplà  $j$ -dimensional. En aquest capítol es descriu un test estadístic basat en la comparació de l'ordenada i els pendents de l'hiperplà de regressió (test conjunt per a l'ordenada i la suma de pendents) amb els seus valors teòrics, per tal d'establir si els  $j$  mètodes analítics en comparació proporcionen resultats que no difereixen estadísticament entre si per al nivell de significança escollit, tot i tenir en compte els errors individuals de cada mètode analític. Aquest test no és només vàlid en processos de comparació de múltiples metodologies analítiques, sinó que també pot ser aplicable, per exemple, a exercicis interlaboratoris, on es volen comparar els resultats produïts pels diversos laboratoris que participen en l'exercici.

El gruix de la revisió crítica es troba al final del capítol, en l'article titulat *Multiple Analytical Method Comparison Using Regression with Uncertainties in all Axes*, que s'ha enviat per a la seva publicació a la revista *Analytical Chemistry*.

### 5.1 Regressió multivariant amb errors heteroscedàstics individuals en tots els eixos

La regressió multivariant (secció 1.3.2) és àmpliament emprada en l'actualitat. Però quasi totes les tècniques de regressió multivariant només consideren els errors presents en la variable resposta, i deixen normalment les variables predictores lliures d'error, com per exemple, la tècnica de regressió lineal múltiple (*multiple linear regression*, MLR). Una de les justificacions principals per negligir els errors en tots els eixos era la dificultat computacional per tal de desenvolupar i programar algorismes útils per ser utilitzats per la comunitat científica. Recentment, i dins de les tècniques de regressió esbiaixades, s'han desenvolupat les tècniques de regressió per components principals de màxima versemblança (*maximum likelihood principal components regression*, MLPCR) i regressió per arrel latent de màxima versemblança (*maximum likelihood latent root regression*, MLLRR),<sup>2</sup> totes dues basades en l'anàlisi per components principals de màxima versemblança (*maximum likelihood principal component analysis*, MLPCA),<sup>3</sup> que considera els errors en totes les variables, la qual cosa dóna idea que un cop s'han disposat de les eines de càlcul adequades, s'ha començat a tenir en compte les incerteses en tots els eixos.

Dins de les tècniques no esbiaixades, els antecedents que tenen en compte els errors en tots els eixos són escassos.<sup>4,5</sup> Per desenvolupar el test per comparar els resultats de múltiples mètodes analítics ens hem basat en la tècnica de regressió de mínims quadrats multivariants (*multivariate least squares*, MLS),<sup>6</sup> que té en compte els errors individuals en tots els eixos. A més a més que la programació del seu algorisme és ràpida i els resultats dels coeficients de regressió se solen trobar

ISBN: 978-84-691-1897-9/D.L.: T-353-2008

en un curt període de temps, la tècnica proporciona també la matriu variància-covariància dels coeficients de regressió, de gran utilitat per tal de calcular posteriors tests estadístics útils en processos de comparació de mètodes analítics. Aquesta tècnica no és res més que l'extensió al camp multivariant del mètode de regressió amb errors individuals en dos eixos BLS, revisat al capítol 2. La regressió multivariant entre  $j$  variables (on en processos de comparació de mètodes correspondria a la comparació entre els resultats de  $j$  mètodes analítics  $m_1, m_2, \dots, m_j$ ) dóna un hiperplà de regressió, que correspon a l'expressió següent:

$$m_1 = \hat{b}_1 + \hat{b}_2 m_2 + \hat{b}_3 m_3 + \dots + \hat{b}_j m_j \quad (5.1)$$

El procediment per trobar els coeficients de l'hiperplà de regressió amb el mètode MLS consisteix en la minimització de la suma de residuals ponderats ( $S$ ), expressats en l'equació següent:

$$S = \sum_{i=1}^n \frac{1}{w_i} (m_{1i} - \hat{m}_{1i})^2 = \sum_{i=1}^n \frac{1}{w_i} (m_{1i} - \hat{b}_1 + \hat{b}_2 m_{2i} + \hat{b}_3 m_{3i} + \dots + \hat{b}_j m_{ji})^2 \quad (5.2)$$

on el factor de ponderació  $w_i$  és funció de les variàncies del resultat en el punt  $i$  obtingut amb cada un dels mètodes analítics i de les covariàncies entre els resultats esmentats, les quals se solen considerar negligibles. És important puntualitzar que, com que amb el mètode de regressió MLS es tenen en compte totes les incerteses individuals, no importa quin dels  $j$  mètodes analítics en comparació s'escull per ser  $m_1$  a l'equació 5.1. S'obtidrien els mateixos resultats dels coeficients de regressió si qualsevol altre mètode s'escollís per ser  $m_1$ .

Per tal de trobar els coeficients de l'hiperplà de regressió (on  $\hat{b}_1$  correspon a la ordenada a l'origen i  $\hat{b}_2 \dots \hat{b}_j$  corresponen als pendents), cal dur a terme un procediment iteratiu en els pendents. Aquest procediment iteratiu sol ser bastant

ràpid. Per exemple, la programació de l'algorisme MLS en Matlab 4.0 per a Windows 3.1 o superior<sup>7</sup> i la seva aplicació sobre un ordinador HP Vectra VE Pentium 75 a diversos conjunts de dades reals i simulades de la comparació d'entre 4 i 8 mètodes analítics, dona coeficients de regressió en temps inferiors als 3 segons per a quasi tots els conjunts; només arriba el temps de càlcul fins als 29 segons en un sol dels conjunt de dades utilitzats. Però un dels inconvenients que presenta la tècnica de regressió multivariant MLS està íntimament lligat al procés iteratiu de cerca dels coeficients de l'hiperplà de regressió. Com que és un procés iteratiu, l'algorisme requereix una estimació inicial dels coeficients de l'hiperplà de regressió (que pot ser per exemple considerar-los tots zero) per tal de poder iniciar el procés fins a arribar a la convergència d'aquests mateixos coeficients, moment en què s'acaba el procés i es presenten els resultats. S'ha observat que en alguns conjunts de dades, segons quina sigui l'estimació inicial dels coeficients de regressió (per exemple fixant-los tots a zero, o introduint-hi les estimacions obtingudes amb el mètode MLR), el procés iteratiu pot arribar a la convergència en diferents resultats dels coeficients de regressió. Això és degut al fet que la funció a minimitzar ( $S$ , equació 5.2), no presenta un sol mínim, sinó que a més del mínim absolut (la solució real que hauria de trobar-se idealment sempre) hi ha diversos mínims relatius. N'hi ha més a mesura que la dimensionalitat és més gran (com més augmenta el nombre de mètodes en comparació), i segons quina sigui l'estimació inicial dels coeficients de regressió es pot arribar a trobar un mínim relatiu. Cal dir, però, que aquesta situació, tot i ser possible, no és freqüent i normalment se solen arribar sempre als mateixos resultats dels coeficients de regressió, sigui quina sigui la seva estimació inicial. Per tant, com a treball posterior caldria una revisió de l'algorisme de càlcul per fer-lo més robust (per exemple, trobar els coeficients de regressió amb diverses estimacions inicials i quedar-se amb l'òptim o introduir-hi un procediment de reiteració per tal d'assegurar-se que s'arriba al mínim absolut).

## 5.2 Test conjunt per a l'ordenada a l'origen i la suma de pendents de l'hiperplà de regressió

Per tal de comprovar si els resultats produïts per múltiples mètodes analítics (o múltiples instruments o múltiples laboratoris, etc.) a diferents nivells de concentració no difereixen estadísticament entre si, ens basarem en el test conjunt per a l'ordenada a l'origen i el pendent desenvolupat en el capítol 4 i l'ampliarem per al supòsit multivariant. Com que normalment els mètodes analítics en comparació presenten errors del mateix ordre de magnitud, els coeficients de la recta de regressió seran trobats amb el mètode de regressió MLS.

| $m_1$ | $m_2$ | $m_3$ | $m_4$ | $m_5$ | $m_6$ | $m_7$ |
|-------|-------|-------|-------|-------|-------|-------|
| 3.4   | 3.4   | 3.4   | 3.4   | 3.4   | 3.4   | 3.4   |
| 5.6   | 5.6   | 5.6   | 5.6   | 5.6   | 5.6   | 5.6   |
| 7.8   | 7.8   | 7.8   | 7.8   | 7.8   | 7.8   | 7.8   |
| 10.2  | 10.2  | 10.2  | 10.2  | 10.2  | 10.2  | 10.2  |
| 15.6  | 15.6  | 15.6  | 15.6  | 15.6  | 15.6  | 15.6  |
| 23.4  | 23.4  | 23.4  | 23.4  | 23.4  | 23.4  | 23.4  |
| 28.7  | 28.7  | 28.7  | 28.7  | 28.7  | 28.7  | 28.7  |
| 36.4  | 36.4  | 36.4  | 36.4  | 36.4  | 36.4  | 36.4  |
| 40.0  | 40.0  | 40.0  | 40.0  | 40.0  | 40.0  | 40.0  |
| 48.6  | 48.6  | 48.6  | 48.6  | 48.6  | 48.6  | 48.6  |

Taula 4.1. Resultats de l'anàlisi de 10 mostres mitjançant 7 mètodes analítics que donen exactament els mateixos resultats.

La primera idea que hom pot tenir per dur a terme la comparació de  $j$  mètodes analítics a diferents nivells de concentració és analitzar diverses mostres per tots els mètodes analítics en comparació, i trobar l'hiperplà de regressió dels resultats de tots els mètodes (equació 5.1). Llavors, per analogia amb el test conjunt per a l'ordenada a l'origen i el pendent (capítol 4), es podria dir que els mètodes en

comparació produeixen resultats que no difereixen estadísticament entre ells si l'ordenada a l'origen ( $\hat{b}_1$ ) no difereix significativament del valor teòric 0, i si cada un dels pendents ( $\hat{b}_2 \dots \hat{b}_j$ ) no difereix significativament del valor teòric 1, tot i tenir en compte la correlació entre tots els coeficients de l'hiperplà de regressió (és a dir, considerant les covariàncies entre els coeficients). Examinem, però, les dades representades a la taula 4.1, on es representa la comparació de set mètodes analítics en l'anàlisi de 10 mostres en el supòsit teòric que tots els mètodes produïssin exactament els mateixos resultats. Seguint el raonament exposat anteriorment, com que els set mètodes en comparació proporcionen resultats exactament iguals, l'hiperplà de regressió hauria de ser:

$$m_1=0+1m_2+1m_3+1m_4+1m_5+1m_6+1m_7$$

però en realitat el mètode MLR ens dona l'expressió següent:

$$m_1=0+1m_2+0m_3+0m_4+0m_5+0m_6+0m_7$$

(si es consideressin les incerteses individuals de cada resultat i es busquessin els coeficients de regressió mitjançant la tècnica MLS, els resultats obtinguts en l'hiperplà de regressió haurien estat anàlegs; en les dades de la taula 4.1 no es consideren les incerteses individuals per simplificar el procés). De fet, si es canvien les dades de la taula 4.1, s'agafa un nombre diferent de punts i es canvia el nombre de mètodes analítics en comparació -mantenint la hipòtesi que tots produeixen resultats exactament iguals-, s'obtindrà sempre un hiperplà de regressió on l'ordenada a l'origen serà exactament igual a 0, un dels pendents serà exactament igual a 1, i tots els altres igual a 0. Però què passa quan les dades experimentals no s'ajusten exactament a l'hiperplà de regressió, que és el que normalment passa en els casos reals? El conjunt de dades de la taula 4.2 està basat en la taula 4.1. Les dades estan extretes aleatòriament (emprant Matlab 4.0 per a Windows 3.1 o

superior) d'una distribució normal amb mitjana de cada un dels valors individuals de la taula 4.1 i amb desviació estàndard 1/10000 de cada un dels valors de la taula 4.1. Així, doncs, els resultats entre els set mètodes teòrics en comparació seran pràcticament iguals, però no exactament idèntics. Per al conjunt de dades de la taula 4.2, el mètode MLR ens dona l'expressió següent de l'hiperplà de regressió entre els resultats dels diferents mètodes analítics en comparació:

$$m_1 = -0.0002 + 0.6956m_2 - 0.7333m_3 - 0.5490m_4 + 0.4655m_5 + 0.7832m_6 + 0.3380m_7$$

Es pot observar que en aquest cas, tot i que els resultats dels set mètodes en comparació són molt semblants, ara no hi ha un pendent que sigui igual a 1 i tots els altres igual a 0 (el valor de l'ordenada a l'origen de l'hiperplà sí que continua trobant-se molt proper a 0). Però per altra banda, si se sumen els valors dels sis pendents, s'obté el resultat de 1.0000. Per tant, es pot arribar a la conclusió que els mètodes analítics en comparació donen resultats comparables si l'ordenada a l'origen de l'hiperplà de regressió no difereix significativament de 0 i simultàniament si la suma de pendents no difereix significativament d'1. Els resultats obtinguts amb les dades de la taula 4.2 han estat confirmats amb diferents conjunts de dades, variant el nombre de punts i el nombre de mètodes analítics teòrics en comparació, i tant considerant errors en tots els eixos (és a dir, buscant l'hiperplà de regressió amb el mètode MLS) o bé sense considerar-los (emprant el mètode MLR).

Per tal de comparar els resultats produïts per  $j$  mètodes analítics ens haurem de fixar simultàniament en l'ordenada de l'hiperplà de regressió (que haurà de ser no significativament diferent de 0 per a un nivell de significança  $\alpha$  escollit) i simultàniament en la suma dels pendents (que haurà de ser no significativament diferent d'1 per al mateix nivell de significança  $\alpha$  escollit). Per tal de comprovar això s'ha desenvolupat el test conjunt per a l'ordenada a l'origen i la suma de pendents considerant els errors individuals en tots els eixos, que està descrit en

ISBN:978-84-691-1897-9/D.L: T-353-2008

l'article al final del capítol. Per poder aplicar aquest test es necessiten la variància de l'ordenada a l'origen de l'hiperplà de regressió, la variància de la suma de pendents de l'hiperplà i la covariància entre l'ordenada a l'origen i la suma de pendents. La variància de l'ordenada a l'origen és fàcilment assolible, ja que el procés iteratiu del mètode MLS proporciona la matriu variància-covariància dels coeficients de regressió. Per a la variància de la suma de pendents i la covariància entre l'ordenada a l'origen i la suma de pendents s'empraran les variàncies individuals dels pendents i les covariàncies entre pendents (també fàcilment assolibles a través de la matriu variància-covariància proporcionada pel mètode MLS), tal com es veurà en de l'article que hi ha al final d'aquest capítol.

| $m_1$   | $m_2$   | $m_3$   | $m_4$   | $m_5$   | $m_6$   | $m_7$   |
|---------|---------|---------|---------|---------|---------|---------|
| 3.4001  | 3.4005  | 3.3994  | 3.4001  | 3.3999  | 3.3996  | 3.3999  |
| 5.5998  | 5.5999  | 5.6010  | 5.6001  | 5.6015  | 5.6000  | 5.6005  |
| 7.7993  | 7.8000  | 7.7988  | 7.8005  | 7.7998  | 7.7997  | 7.7999  |
| 10.1996 | 10.2003 | 10.1989 | 10.2006 | 10.1999 | 10.2014 | 10.2019 |
| 15.6018 | 15.5987 | 15.6001 | 15.5985 | 15.6006 | 15.6021 | 15.6003 |
| 23.4020 | 23.4027 | 23.3966 | 23.4041 | 23.3972 | 23.4015 | 23.4030 |
| 28.6970 | 28.7005 | 28.7013 | 28.6989 | 28.6992 | 28.7014 | 28.6999 |
| 36.3993 | 36.4057 | 36.4020 | 36.3967 | 36.3942 | 36.4029 | 36.3945 |
| 39.9948 | 39.9992 | 40.0053 | 39.9960 | 40.0042 | 39.9969 | 40.0015 |
| 48.6015 | 48.6080 | 48.5980 | 48.6058 | 48.6015 | 48.6018 | 48.5990 |

Taula 4.2. Resultats de l'anàlisi de 10 mostres mitjançant set mètodes analítics que produeixen resultats molt semblants.

El test conjunt per a l'ordenada i la suma de pendents assumeix la normalitat en la distribució dels coeficients de l'hiperplà de regressió. Si en els coeficients de la recta de regressió considerant errors individuals en dos eixos ja no es podia considerar que la distribució era sempre normal (apartat 3.1), probablement els

ISBN: 978-84-9014-197-9

coeficients de regressió de l'hiperplà trobat considerant errors en dos eixos tampoc segueixin la distribució normal. Però de la mateixa manera que l'error comès en considerar la normalitat en els coeficients de la recta de regressió tenint en compte els errors individuals en dos eixos era poc important, es pot pensar que els coeficients de l'hiperplà de regressió seguiran aquesta tendència. De fet, com es veurà en l'apartat de la validació del test conjunt per a l'ordenada a l'origen i la suma de pendents de l'hiperplà de regressió, en l'article que hi ha al final d'aquest capítol, els resultats trobats coincideixen bastant aproximadament amb els valors teòrics. Això sembla indicar que es pot treballar amb la hipòtesi de normalitat dels coeficients de l'hiperplà de regressió tenint en compte els errors en tots els eixos.

### 5.3 Referències

1. H. Scheffé, *The Analysis of Variance*, John Wiley & Sons, New York (1959)
2. P.D. Wentzell, D.T. Andrews, B.R. Kowalski, *Analytical Chemistry*, **69** (1997) 2299
3. P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, *Journal of Chemometrics*, **11** (1997) 339
4. J. Schnute, *Journal of American Statistical Association*, **79** (1984) 1
5. C. Chatfield, A.J. Collins, *Introduction to Multivariate Analysis*, Chapman & Hall, New York (1980)
6. J.M. Lisý, A. Cholvadová, J. Kutej, *Computers and Chemistry*, **14** (1990) 189
7. Matlab, The Mathworks Inc., Natick, MA

# Multiple analytical method comparison using regression with uncertainties in all axes

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

This paper discusses a new approach for comparing multiple analytical methods which takes into account all errors. The new test uses the multivariate least squares (MLS) regression method to calculate the coefficients of the regression hyperplane and takes into account the uncertainties in all the methods. A joint confidence interval test is applied to check if the intercept and the sum of the slopes of the regression hyperplane do not differ significantly from the theoretical values zero and one respectively. The new technique is validated using three simulated and four real data sets. Statistical analysis using the simulated reference data sets show the correctness of the new approach which is then applied to four real data sets to detect the differences between the results obtained for the joint confidence interval using MLS and multivariate linear regression (MLR), which does not take into account the errors in all the methods.

## INTRODUCTION

Comparing the performance of several analytical methods is a frequent task in the chemical laboratory. However, until now, it has only been possible to compare two methodologies simultaneously when studying a range of analyte concentrations. Recently, it has been shown<sup>1</sup> that the joint confidence interval test for the slope and the intercept of the regression line calculated considering the errors in the two methods to be compared (represented in the coordinated axes) provided promising results for detecting biased methods or accepting correct new ones. This approach is based on the bivariate least squares (BLS) calibration method,<sup>2</sup> a regression technique that calculates the regression coefficients taking into account the presence of non-constant uncertainties in both axes.

Analysis of variance (ANOVA)<sup>3</sup> is commonly used to compare the concentration results obtained by several analytical methods. However, a number of replicates need to be analyzed for each concentration level in order to obtain sufficient degrees of freedom for statistical significance. Moreover the higher the number of concentration levels, the larger the number of replicates to be analyzed.

This paper discusses a new procedure for assessing the presence of bias in at least one of the several analytical methods that are compared for a range of analyte concentrations. The analytical results obtained for each method are used to build a multivariate model that takes into account all of their uncertainties. The calculated regression parameters are compared to the theoretical ones by means of a new joint confidence interval test that considers the different errors associated to the individual results for each method and to the level of analyte concentration. Therefore, the new method not only considers errors in all coordinate axes but can also cope with highly heteroscedastic data.

To validate the new approach for method comparison studies, three simulated and four real data sets were used. The Monte Carlo technique was used to generate simulated data sets that show that the results obtained with the new approach are in agreement with the theoretical results expected. The results of using the joint confidence test based on multivariate linear regression, MLR, and multivariate least squares considering uncertainties in all axes, MLS for four real data sets may lead to different conclusions about the validity of the methods compared, which reinforces the usefulness of the regression methods that take into account the errors in all the axes.

The procedure is general in nature, that is, it can be applied to any experimental problem in which several analytical methodologies, techniques, analysts or instruments are to be compared using samples with different levels of analyte concentrations. If one or more methods are biased, a significant difference between the theoretical and calculated regression parameters will be detected. Further analyses would then be necessary to determine the specific method(s) which may be the origin of the detected systematic error.

## BACKGROUND AND THEORY

The multivariate least squares model that relates the analytical methods to be compared is built taking into the account the errors in each individual result. Therefore the uncertainties resulting from analyzing  $n$  samples with each method are needed. For one single sample  $i$  analyzed by all the methods in comparison ( $m_k$ ,  $k=1..j$ ), an experimental point will be described by:  $(m_{2_i}, m_{3_i}, \dots, m_{j_i}, m_{1_i})$  where  $m_{k_i}$  ( $k=1..j$ ) is the concentration result obtained by the  $k$ th method analyzing the  $i$ th sample. Replications of each sample analyzed by each method are required in order to estimate the standard deviation for each  $m_{k_i}$  value ( $s_{m_{k_i}}$ ).

Once all the data have been collected, the multivariate model in eq 1 can be built:

$$m_1 = \hat{b}_1 + \hat{b}_2 m_2 + \hat{b}_3 m_3 + \dots + \hat{b}_j m_j \quad (1)$$

where  $\hat{b}_k$  ( $k=1..j$ ) are the estimated regression coefficients of the regression hyperplane expressed in eq 1 ( $\hat{b}_1$  corresponds to the intercept and  $\hat{b}_l$ ,  $l=2..j$ , to the slopes). Normally all the analytical methods to be compared (or the laboratories in an interlaboratory study) include random errors of about the same order of magnitude. Therefore the regression coefficients must be found using regression methods that take into account the errors in all the axes. Here, we used the multivariate least squares (MLS) method<sup>2</sup>. We also used the multiple linear regression method (MLR)<sup>4</sup> for comparison purposes but in this case only constant errors are considered for the results given by  $m_1$  method while all the other methods are considered error-free. It is important to note that since all the individual errors in all the methods are taken into account using the MLS regression method, it is not of importance which one of the  $j$  methods in comparison is chosen as to be  $m_1$  in eq 1. The same results for the regression coefficients are obtained if another method is chosen as to be  $m_1$ .

If  $j$  analytical methods that give rise to exactly identical analytical results are regressed in order to find the regression coefficients of the hyperplane expressed in eq 1, the intercept ( $\hat{b}_1$ ) will be 0, one of the  $j-1$  slopes will be 1 and all the other slopes will be 0, giving rise to a simple regression line. If  $j$  analytical methods that give rise to nearly identical analytical results are regressed in order to find the regression coefficients of the hyperplane, the intercept ( $\hat{b}_1$ ) will be almost 0 (i.e. it will not statistically differ from 0), and in this case there will not be one single slope being 1 and all the other ones being 0, but the sum of all the slopes

$(\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j)$  will be almost 1 (the sum will not statistically differ from 1).

Regarding this, it will be said that the  $j$  analytical methods in comparison give rise to results that do not statistically differ at an  $\alpha$  level of significance if the intercept of the regression hyperplane is not statistically different from 0 and simultaneously if the sum of the slopes of the regression hyperplane are not statistically different from 1. This is checked with the joint confidence test for the intercept and the sum of the slopes of the regression hyperplane, which is presented in the next sections.

### Multivariate Least-Squares (MLS)

Multivariate least-squares is the generic name for a set of techniques which regress multivariate data with errors in all the axes. Of all the approaches for calculating regression coefficients and related statistical parameters, we selected Lisý's method<sup>2</sup> because of its speed in estimating the correct results for the regression coefficients and the simplicity of programming its algorithm. The method consists of minimizing the sum of weighted residuals,  $S$ , expressed in eq 2:

$$S = \sum_{i=1}^n \frac{1}{w_i} (m_{1i} - \hat{m}_{1i})^2 = \sum_{i=1}^n \frac{1}{w_i} (m_{1i} - \hat{b}_1 - \hat{b}_2 m_{2i} - \hat{b}_3 m_{3i} - \dots - \hat{b}_j m_{ji})^2 = (n-j) \cdot \hat{s}^2 \quad (2)$$

where  $w_i$  is the weighting factor corresponding to the  $i$ th point,  $\hat{s}^2$  is the estimation of the experimental error,  $s^2$ , and it is measured in terms of variance for the set of  $n$  experimental data points according to eq 3:

$$\hat{s}^2 = \frac{1}{(n-j)} \sum_{i=1}^n \frac{1}{w_i} (m_{1i} - \hat{m}_{1i})^2 = \frac{1}{(n-j)} \sum_{i=1}^n \frac{1}{w_i} (m_{1i} - \hat{b}_1 - \hat{b}_2 m_{2i} - \hat{b}_3 m_{3i} - \dots - \hat{b}_j m_{ji})^2 \quad (3)$$

The residual corresponding to the  $i$ th point ( $\varepsilon_i$ ) is expressed in eq 4:

$$\varepsilon_i = m_{1i} - \hat{b}_1 - \hat{b}_2 m_{2i} - \hat{b}_3 m_{3i} - \dots - \hat{b}_j m_{ji} \quad (4)$$

This method uses the variance of the residuals ( $s_{\varepsilon_i}^2$ ), which can be expressed using the Taylor series, as a weighting factor even when the covariances between experimental data are not zero:

$$s_{\varepsilon_i}^2 = w_i = s_{m_{1i}}^2 + \sum_{k=2}^j \hat{b}_k^2 s_{m_{ki}}^2 - 2 \sum_{k=2}^j \hat{b}_k \text{cov}(m_{1i}, m_{ki}) + 2 \sum_{k=2}^j \sum_{l=k+1}^j \hat{b}_k \hat{b}_l \text{cov}(m_{ki}, m_{li}) \quad (5)$$

By minimizing the sum of the weighted residuals,  $S$  (eq 2), in relation to the regression coefficients, ( $\hat{b}_k$ ,  $k=1\dots j$ ),  $j$  non-linear equations are obtained. By including the partial derivatives of the squared residuals, eq 6 and the equivalent eq 7 can be written in matrix form:

$$\mathbf{R}\hat{\mathbf{b}} = \mathbf{g} \quad (6)$$

$$\begin{pmatrix} \sum_{i=1}^n \frac{1}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{2i}}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{3i}}{s_{\varepsilon_i}^2} & \dots & \sum_{i=1}^n \frac{m_{ji}}{s_{\varepsilon_i}^2} \\ \sum_{i=1}^n \frac{m_{2i}}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{2i}^2}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{2i} m_{3i}}{s_{\varepsilon_i}^2} & \dots & \sum_{i=1}^n \frac{m_{2i} m_{ji}}{s_{\varepsilon_i}^2} \\ \sum_{i=1}^n \frac{m_{3i}}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{2i} m_{3i}}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{3i}^2}{s_{\varepsilon_i}^2} & \dots & \sum_{i=1}^n \frac{m_{3i} m_{ji}}{s_{\varepsilon_i}^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \frac{m_{ji}}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{2i} m_{ji}}{s_{\varepsilon_i}^2} & \sum_{i=1}^n \frac{m_{3i} m_{ji}}{s_{\varepsilon_i}^2} & \dots & \sum_{i=1}^n \frac{m_{ji}^2}{s_{\varepsilon_i}^2} \end{pmatrix} \times \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \vdots \\ \hat{b}_j \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \left[ \frac{m_{1i}}{s_{\varepsilon_i}^2} + \frac{1}{2} \left[ \frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right]^2 \frac{\partial s_{\varepsilon_i}^2}{\partial \hat{b}_1} \right] \\ \sum_{i=1}^n \left[ \frac{m_{1i} m_{2i}}{s_{\varepsilon_i}^2} + \frac{1}{2} \left[ \frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right]^2 \frac{\partial s_{\varepsilon_i}^2}{\partial \hat{b}_2} \right] \\ \sum_{i=1}^n \left[ \frac{m_{1i} m_{3i}}{s_{\varepsilon_i}^2} + \frac{1}{2} \left[ \frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right]^2 \frac{\partial s_{\varepsilon_i}^2}{\partial \hat{b}_3} \right] \\ \vdots \\ \sum_{i=1}^n \left[ \frac{m_{1i} m_{ji}}{s_{\varepsilon_i}^2} + \frac{1}{2} \left[ \frac{\varepsilon_i}{s_{\varepsilon_i}^2} \right]^2 \frac{\partial s_{\varepsilon_i}^2}{\partial \hat{b}_j} \right] \end{pmatrix} \quad (7)$$

where the column matrix  $\hat{\mathbf{b}}$  is defined by the estimated regression coefficients. The regression coefficients, which are components of vector  $\hat{\mathbf{b}}$  in eq 6 and 7, can be determined by carrying out an iterative process on the following matrix form:

$$\hat{\mathbf{b}} = \mathbf{R}^{-1}\mathbf{g} \quad (8)$$

With this method, the variance-covariance matrix of the regression coefficients,  $\mathbf{B}$ , is obtained by multiplying the final matrix  $\mathbf{R}^{-1}$ , by the experimental error,  $s^2$ .<sup>5,6</sup> As the experimental error is unknown, the estimated value,  $\hat{s}^2$ , expressed in eq 3 should be used.<sup>7</sup> It should be pointed out that, if the situation were to be  $s_{e_i}^2 = w_i = 1$  (i.e., there are only homoscedastic errors in the  $m_1$  method, and all the other methods are error-free), then the expressions obtained would be the same as if the MLR regression method were applied.

The model expressed in eq 1 corresponds to a hyperplane of  $j$ -dimensions. For instance, a hyperplane for a three method comparative study with 13 experimental data points is shown in Figure 1 together with the individual points and their uncertainties. The equation of the hyperplane is  $m_1 = 0.0003 + 0.4228m_2 + 0.6183m_3$ . It should be pointed out that the regression coefficients found using the MLS regression method are invariant upon switching axes

### **Dimensionality reduction. Joint confidence interval for the intercept and the sum of slopes**

Once the regression coefficients have been found, the differences between the results given by the  $j$  analytical methods to be compared are not statistically significant at the significance level  $\alpha$  if the intercept of the hyperplane,  $\hat{b}_1$ , is not

significantly different from 0, while the sum of the slopes  $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j$  is not significantly different from 1. This can be checked with the joint confidence interval test for the intercept and the slope where the intercept is the intercept of the hyperplane ( $\hat{b}_1$ ) and the slope is the sum of the slopes of the hyperplane ( $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j$ ).

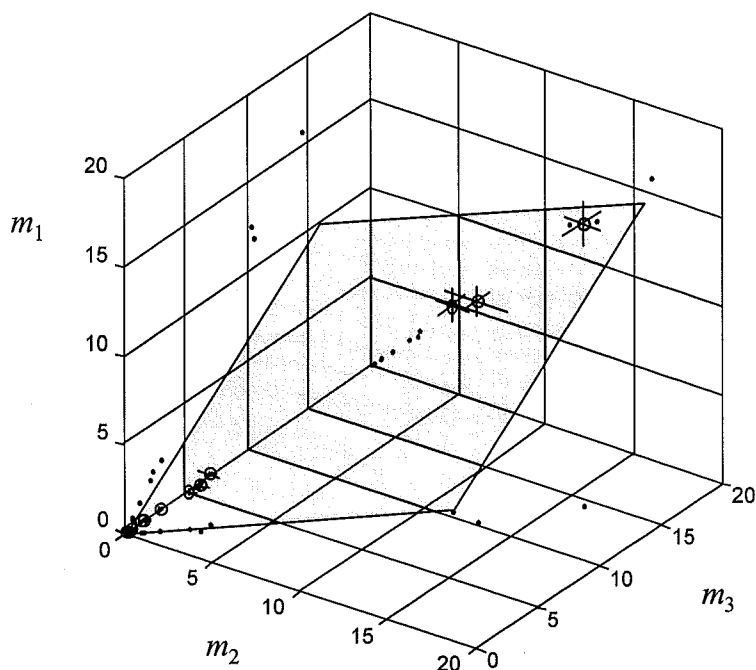


Figure 1. Regression hyperplane calculated using the MLS regression technique corresponding to 13 experimental data points obtained in a three-methods comparison study. The projections of each point in the three-dimensional space are represented as points on the three planes defined by each pair of methods.

In order to apply this joint confidence interval test, one must transform the variance-covariance matrix of the regression coefficients ( $\mathbf{B} = \mathbf{R}^{-1} \cdot \hat{s}^2$ ) of the initial hyperplane, given by eq 1, into the variance-covariance matrix

corresponding to intercept  $\hat{b}_1$  and the sum of slopes  $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j$ . The first step transforms the hyperplane expressed in eq 1 into the following model:

$$m_1 = \hat{b}_1 + \hat{b}'m' \quad (9)$$

where the sum of the slopes of the original hyperplane ( $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j$ ) must be equal to the  $\hat{b}'$  coefficient in eq 9. So the combination of results given by  $j-1$  methods ( $m'$  in eq 9) so that eq 9 is equivalent to eq 1 must be found: i.e.  $\hat{b}_2m_2 + \hat{b}_3m_3 + \dots + \hat{b}_jm_j = \hat{b}'m'$ . Hence,  $m'$  corresponds to:

$$m' = \frac{\hat{b}_2m_2 + \hat{b}_3m_3 + \dots + \hat{b}_jm_j}{\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j} \quad (10)$$

It can be seen that the term  $m'$  corresponds to a combination of the results obtained with  $j-1$  analytical methods being compared and  $j-1$  regression coefficients of the hyperplane (eq 1) found using MLS regression. The term  $m_1$  in eq 9 coincides with the term  $m_1$  in eq 1 so no change need be made in the original results, nor in their uncertainties, for the method  $m_1$ . As regression with errors in all axes is being considered, the uncertainty due to  $m'$  must be found, which will be a combination of the uncertainties of  $j-1$  methods. Following the error propagation theory,<sup>8</sup> the variance associated to the  $i$ th point given by the method combination  $m'$  is:

$$s_{m_i}^2 = \left[ \frac{\hat{b}_2}{\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j} \right]^2 s_{m_{2_i}}^2 + \left[ \frac{\hat{b}_3}{\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j} \right]^2 s_{m_{3_i}}^2 + \dots + \left[ \frac{\hat{b}_j}{\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j} \right]^2 s_{m_{j_i}}^2 \quad (11)$$

The experimental error associated to eq 9, which is needed to find the variance-covariance matrix corresponding to intercept  $\hat{b}_1$  and the sum of the slopes  $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j$ , can be found according to eq 12:

$$\hat{s}'^2 = \frac{\sum_{i=1}^n (m_{1i} - \hat{b}_1 - \hat{b}'m'_i)^2 / w_i}{n-2} \quad (12)$$

where  $w_i$  corresponds to the weighting factor for the  $i$ th point, i.e. the variance of the  $i$ th residual of eq 9. To obtain the variance-covariance matrix of intercept  $\hat{b}_1$  and of the sum of the slopes  $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j = \hat{b}'$ , the expression for the variance of  $\hat{b}_1$  ( $s_{\hat{b}_1}^2$ ) corresponds to the first term of the variance-covariance **B** matrix of the regression coefficients of the original hyperplane ( $\mathbf{B} = \mathbf{R}^{-1} \cdot \hat{s}^2$ ) and the expressions for the variance of the sum of slopes ( $s_{\hat{b}'}^2$ ) and the covariance between the intercept  $\hat{b}_1$  and the sum of slopes ( $\text{cov}(\hat{b}_1, \hat{b}')$ ) correspond to:

$$s_{\hat{b}'}^2 = \left[ \sum_{k=2}^j s_{\hat{b}_k}^2 + 2 \sum_{k=2}^{j-1} \sum_{l=k+1}^j \text{cov}(\hat{b}_l, \hat{b}_k) \right] \times \hat{s}^2 \quad (13)$$

$$\text{cov}(\hat{b}_1, \hat{b}') = \left[ \sum_{k=2}^j \text{cov}(\hat{b}_1, \hat{b}_k) \right] \times \hat{s}^2 \quad (14)$$

where the variances of the regression coefficients ( $s_{\hat{b}_k}^2$ ) stand for the  $(k,k)$  value of the variance-covariance **B** matrix of the regression coefficients of the original hyperplane, and the covariances of the regression coefficients ( $\text{cov}(\hat{b}_l, \hat{b}_k)$ ) stand for the  $(l,k)$  value of the variance-covariance **B** matrix. In short, the variance-

covariance matrix corresponding to the regression coefficients expressed in eq 9 can be re-expressed as eq 15:

$$\begin{vmatrix} s_{\hat{b}_1}^2 & \text{cov}(\hat{b}_1, \hat{b}') \\ \text{cov}(\hat{b}_1, \hat{b}') & s_{\hat{b}'}^2 \end{vmatrix} = \mathbf{T}' \cdot \hat{s}'^2 \quad (15)$$

where the  $\mathbf{T}'$  matrix is used in the process of finding the joint confidence interval for intercept  $\hat{b}_1$  and the sum of the slopes  $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j = \hat{b}'$ .

Once the variance-covariance matrix of the regression coefficients in eq 9 was found the joint confidence test for the slope and the intercept taking into account errors in two axes was applied.<sup>1</sup> We concluded that the differences between the results provided by the  $j$  analytical methods being compared are not statistically significant if the theoretical point zero intercept and unity slope falls into the ellipsoid defined by the joint confidence region for the intercept and the slope shown in eq 16:

$$\mathbf{T}'_{(1,1)}^{-1}(\hat{b}_1 - b_1)^2 + 2\mathbf{T}'_{(1,2)}^{-1}(\hat{b}_1 - b_1)(\hat{b}' - b') + \mathbf{T}'_{(2,2)}^{-1}(\hat{b}' - b')^2 = 2\hat{s}'^2 F_{1-\alpha(2,n-2)} \quad (16)$$

where  $\hat{b}_1$  and  $\hat{b}'$  are the estimated regression coefficients expressed in eq 9, and  $b_1$  and  $b'$  are their assigned values. Eq 16 defines an ellipse, with their limits depending on the experimental errors and on the significance level chosen. Its tilt is consequence of the correlation between the intercept ( $\hat{b}_1$ ) and the sum of slopes ( $\hat{b}_2 + \hat{b}_3 + \dots + \hat{b}_j = \hat{b}'$ ).

All computations were done with home-made Matlab subroutines (Matlab for Microsoft Windows ver. 4.0, The MathWorks, Inc., Natick, MA). Its application with

a HP Vectra VE Pentium 75 computer to several real and simulated data sets coming from method comparison studies of between 4 and 8 analytical methods give rise to the several joint confidence intervals for the intercept and the sum of slopes in less than 3 seconds for all the studied data sets, only reaching the computation time 29 seconds in one of the data sets.

### Validation process

The objective of the validation process is to assess whether the procedure proposed for multiple method comparison taking into account the errors in all the methods provides correct results; i.e. that it does not detect significant differences between methodologies or laboratories that have no statistical discrepancies at the level of significance chosen, but that it does detect significant differences between methodologies or laboratories that do have statistical discrepancies at the level of significance chosen.

To assess the correctness of the procedure for multiple method comparison taking into account the errors in all the methods, three simulated data sets were produced for which uncertainties were considered in all the methods. Some of these sets were homoscedastic while others were not and they were differently distributed throughout the regression range. The first two simulated data sets gave rise to a regression line characterized by a zero intercept and unity sum of slopes. The third data set gave rise to a line in which there was a significant difference between the results of the three methods being compared, the intercept being zero and the sum of slopes 1.1. By adding a random error to each value of the data points, the Monte Carlo method<sup>9</sup> generated 100,000 different data sets for the original regression data sets 1 and 3. Because of the time involved in the calculation process, only 10,000 data sets were generated for data set 2. The validation involves checking whether the method for obtaining the joint confidence interval with errors in both axes, once the dimensionality has been reduced, for an  $\alpha$  significance level gives results such

that the theoretical point zero intercept and unity slope falls into the joint confidence region in  $(1-\alpha)\%$  of the cases.

Finally, the procedure for multiple method comparison taking into account the individual errors in all the methods was applied to four different real data sets to show that it can lead to conclusions that are different from those of previous method comparison studies.

## EXPERIMENTAL SECTION

### Data sets and software

The following three simulated data sets and four real data sets were used to validate the methodology.

*Data set 1.* Homoscedastic data set containing 20 data points for a three-method comparison study. The 20 data points are equally distributed within the range 2-40 on all the axes. Standard deviations of 1 were considered for all the results given by the three methods. The regression coefficients for the initial data set are  $b_1=0$ ,  $b_2=1$ ,  $b_3=0$ . The data set is shown in Figure 2.

*Data set 2.* Twenty data points corresponding to a five-method comparison study distributed throughout the range between 25 and 425 units. Heteroscedasticity with random standard deviation is present, in such a way that the standard deviation of any point is higher than the standard deviation of the previous point. The standard deviation is, at most, 9% of the value of the individual point. The regression coefficients for the initial data set are  $b_1=0$ ,  $b_2=1$ ,  $b_3=0$ ,  $b_4=0$ ,  $b_5=0$ . The data set is shown in Table 1.

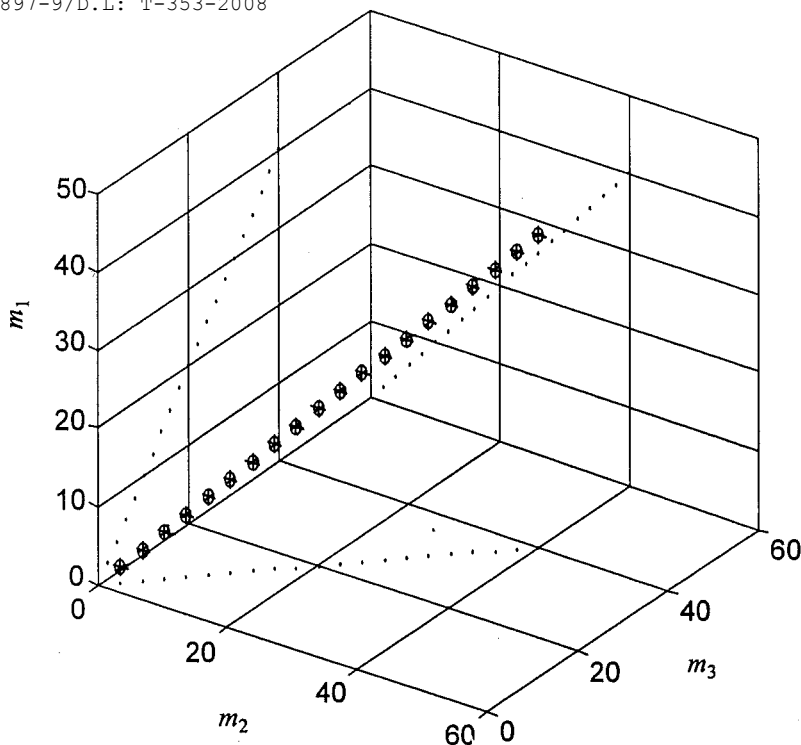


Figure 2. Graphical representation of the individual points in data set 1 and their individual uncertainties. The projections of each three-dimensional point on the planes defined by each pair of methods indicate that there is no difference between the results obtained with the three chemical analytical procedures.

*Data set 3.* Twenty data points corresponding to a three-method comparison study between the values given by the methods  $m_2$  and  $m_3$  distributed in the range between 19 and 360 and the values given by the method  $m_1$  which are 10% higher than the other two methods. The theoretical values of the regression coefficients are  $b_1=0$ ,  $b_2=0$  and  $b_3=1.1$ . Heteroscedasticity with random standard deviation is present, in such a way that each standard deviation is, at most, 9% of each individual value. The data set is shown in Figure 3.

*Data set 4.* Effectiveness of various dissolution procedures for determining nickel in twelve different materials: the nitric-perchloric hot plate procedure, the microwave-assisted procedure, potassium hydrogen-sulfate fusion and sodium peroxide fusion. These results are then compared with the nominal nickel

concentration.<sup>10</sup> Values are expressed in % m/m. Heteroscedasticity is present in the data set in such a way that each standard deviation ranges between 0.2-9.1% of each individual value. The data set is shown in Table 2.

| point number | mean value | method 1 | method 2 | method 3 | method 4 | method 5 |
|--------------|------------|----------|----------|----------|----------|----------|
| 1            | 19.02      | 0.90     | 0.12     | 0.93     | 0.64     | 0.66     |
| 2            | 31.34      | 0.26     | 1.76     | 0.82     | 1.25     | 1.37     |
| 3            | 72.12      | 4.23     | 5.73     | 0.65     | 5.04     | 4.07     |
| 4            | 73.56      | 2.73     | 1.79     | 6.85     | 2.31     | 5.00     |
| 5            | 110.89     | 7.00     | 4.36     | 0.81     | 5.83     | 3.13     |
| 6            | 135.23     | 11.06    | 9.31     | 6.69     | 10.22    | 8.10     |
| 7            | 140.45     | 9.60     | 6.02     | 5.32     | 8.01     | 5.68     |
| 8            | 146.67     | 3.47     | 3.15     | 4.03     | 3.31     | 3.63     |
| 9            | 164.90     | 0.70     | 4.08     | 14.93    | 2.92     | 10.94    |
| 10           | 188.02     | 12.45    | 6.08     | 9.86     | 9.79     | 8.19     |
| 11           | 212.32     | 6.26     | 3.18     | 9.75     | 4.96     | 7.24     |
| 12           | 221.26     | 12.58    | 9.68     | 20.59    | 11.22    | 16.08    |
| 13           | 281.76     | 19.13    | 22.71    | 1.39     | 20.99    | 16.08    |
| 14           | 305.54     | 27.20    | 24.96    | 22.99    | 26.10    | 23.99    |
| 15           | 314.04     | 10.32    | 1.72     | 23.94    | 7.40     | 16.97    |
| 16           | 326.68     | 7.25     | 26.54    | 26.72    | 19.45    | 26.62    |
| 17           | 334.83     | 29.62    | 15.21    | 4.16     | 23.54    | 11.15    |
| 18           | 341.33     | 22.18    | 15.85    | 0.54     | 19.27    | 11.21    |
| 19           | 348.19     | 23.60    | 9.99     | 23.72    | 18.11    | 18.19    |
| 20           | 359.34     | 21.05    | 31.87    | 30.85    | 27.01    | 31.37    |

Table 1. Values for data set 2. The mean values assigned to all methods are listed in column 2. The standard deviation for each method assigned to each mean value are listed in columns 3 to 7.

*Data set 5.* The total Cr content (mg/kg) is determined in six soil samples using four different separation/extraction methods: HNO<sub>3</sub>, deionized water, KCl and acetate buffer solution.<sup>11</sup> Heteroscedasticity is present in the data set in such a way that each standard deviation ranges between 2-14% of each individual value. Only one individual point exceeds this range with a standard deviation of 59% of the individual value. The data set is listed in Table 3.

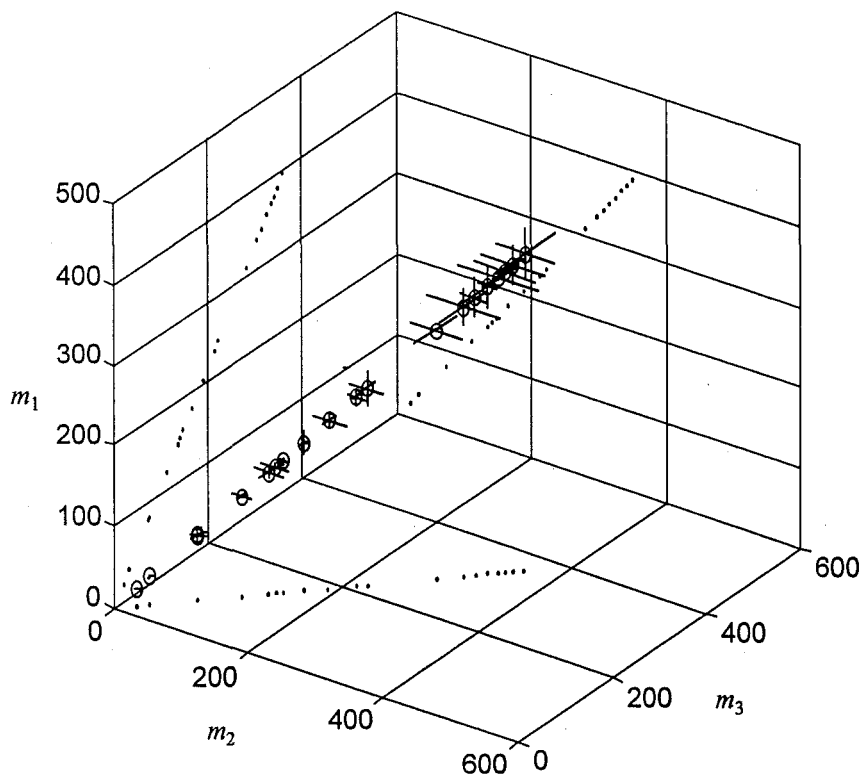


Figure 3. Graphical representation of the individual points in data set 3 and their individual uncertainties. The projections of each three-dimensional point on the  $m_1$ - $m_3$  and  $m_1$ - $m_2$  planes show differences between the results obtained by means of  $m_1$  analytical method and the other two.

*Data set 6.* Six method comparison study to determine polycyclic aromatic hydrocarbons (ng/g) in twelve samples of fly ash: Soxhlet method with toluene, SFE with CO<sub>2</sub> and toluene, SFE with CO<sub>2</sub> and 1% TEA in toluene, SFE with CO<sub>2</sub> and 1% TEA in toluene and additional spike, SFE with CO<sub>2</sub> and 1% TFA in toluene, SFE with CO<sub>2</sub> and 1% TFA in toluene and additional spike.<sup>12</sup> High heteroscedasticity is present in the data set and the standard deviations range between 0.3-56% of the individual value. The data set is shown in Table 4.

*Data set 7.* Degree of recovery of guanine in ten compounds by differential-pulse adsorptive stripping voltammetry<sup>13</sup>. Four different concentrations of some purine

derivatives were used to determine how these purines interfere in the degree of recovery. Slight heteroscedasticity is present in the data set with a maximum standard deviation of 1.6% of the individual value. The data set is listed in Table 5.

| Nominal nickel concentration | Nitric-perchloric hot-plate procedure | Microwave-assisted procedure | Potassium hydrogen-sulfate fusion | Sodium peroxide fusion |
|------------------------------|---------------------------------------|------------------------------|-----------------------------------|------------------------|
| 75.640 ± 0.020               | 72.280 ± 2.250                        | 73.990 ± 1.638               | 73.170 ± 2.465                    | 74.180 ± 8.468         |
| 26.510 ± 0.144               | 25.930 ± 0.084                        | 25.910 ± 0.281               | 26.900 ± 0.608                    | 26.120 ± 0.656         |
| 3.240 ± 0.006                | 2.990 ± 0.003                         | 3.160 ± 0.005                | 3.270 ± 0.008                     | 3.240 ± 0.003          |
| 1.510 ± 0.002                | 1.440 ± 0.001                         | 1.470 ± 0.000                | 1.540 ± 0.020                     | 1.490 ± 0.002          |
| 1.480 ± 0.005                | 1.430 ± 0.001                         | 1.450 ± 0.001                | 1.410 ± 0.001                     | 1.480 ± 0.001          |
| 72.900 ± 0.250               | 72.900 ± 1.000                        | 72.700 ± 1.000               | 71.000 ± 2.250                    | 71.700 ± 1.690         |
| 45.900 ± 5.290               | 47.400 ± 3.610                        | 44.000 ± 5.760               | 47.800 ± 1.210                    | 46.300 ± 0.810         |
| 38.300 ± 1.210               | 37.300 ± 0.160                        | 37.700 ± 1.690               | 39.900 ± 1.210                    | 37.200 ± 1.690         |
| 27.800 ± 1.210               | 25.700 ± 4.000                        | 27.800 ± 1.210               | 28.200 ± 1.000                    | 27.500 ± 0.490         |
| 28.400 ± 1.000               | 26.900 ± 0.090                        | 28.300 ± 1.000               | 27.900 ± 1.210                    | 27.000 ± 0.640         |
| 11.000 ± 0.040               | 10.400 ± 0.040                        | 11.000 ± 0.040               | 10.700 ± 0.040                    | 10.500 ± 0.040         |
| 2.800 ± 0.010                | 2.700 ± 0.010                         | 2.700 ± 0.010                | 2.700 ± 0.010                     | 2.600 ± 0.010          |

Table 2. Nominal nickel concentration (% m/m) in twelve materials and its determination using four dissolution procedures and Atomic Absorption Spectrometry. Results are expressed as mean ± standard deviation.<sup>10</sup>

All computations were done with home-made Matlab subroutines (Matlab for Microsoft Windows ver. 4.0, The Mathworks, Inc., Natick, MA).

## RESULTS AND DISCUSSION

Once the joint confidence interval for the intercept and the sum of the slopes had been validated using the simulated data sets (data sets 1-3), for the real data sets (data sets 4-7) it is of interest to find the differences between the conclusions reached by the authors, whenever they are available (mainly based in the examination of the data), and the conclusions reached using the joint confidence

interval for the intercept and the sum of slopes.

| HNO <sub>3</sub> | deionized water | KCl           | acetate buffer solution |
|------------------|-----------------|---------------|-------------------------|
| 4.5 ± 0.2        | 2.5 ± 0.1       | 2.8 ± 0.3     | 1.9 ± 0.1               |
| 27.0 ± 1.4       | 1.5 ± 0.1       | 1.3 ± 0.1     | 1.4 ± 0.2               |
| 48.0 ± 3.1       | 14.0 ± 1.0      | 14.0 ± 1.6    | 22.0 ± 0.5              |
| 63.0 ± 3.8       | 3.7 ± 2.2       | 4.2 ± 0.4     | 3.6 ± 0.3               |
| 923.0 ± 91.0     | 563.0 ± 34.0    | 563.0 ± 18.0  | 673.0 ± 33.0            |
| 1420.0 ± 120.0   | 1380.0 ± 110.0  | 1380.0 ± 52.0 | 1290.0 ± 44.0           |

Table 3. Determination of total chromium content (mg/kg) using four different separation/extraction methods. Results are expressed as mean ± standard deviation.<sup>11</sup>

| Soxhlet method with toluene | SFE with CO <sub>2</sub> and toluene | SFE with CO <sub>2</sub> and 1% TEA in toluene | SFE with CO <sub>2</sub> and 1% TEA in toluene and additional spike | SFE with CO <sub>2</sub> and 1% TFA in toluene | SFE with CO <sub>2</sub> and 1% TFA in toluene and additional spike |
|-----------------------------|--------------------------------------|--|---|--|---|
| 8 ± 1                       | 9 ± 1                                | 10 ± 1   | 24 ± 8  | 9 ± 1  | 11 ± 1  |
| 1109 ± 33                   | 1398 ± 30                            | 777 ± 266                                      | 1163 ± 4  | 848 ± 47                                       | 863 ± 104   |
| 15 ± 17                     | 41 ± 10                              | 68 ± 14  | 108 ± 3   | 23 ± 3   | 77 ± 43   |
| 857 ± 38                    | 776 ± 15                             | 408 ± 102                                      | 816 ± 23  | 530 ± 14                                       | 637 ± 161   |
| 583 ± 174                   | 693 ± 23                             | 265 ± 48                                       | 617 ± 14  | 213 ± 27                                       | 413 ± 228   |
| 78 ± 63                     | 45 ± 5                               | 27 ± 6   | 100 ± 1   | 24 ± 2   | 66 ± 33   |
| 185 ± 13                    | 65 ± 13                              | 30 ± 9   | 123 ± 1   | 33 ± 1   | 79 ± 32   |
| 220 ± 12                    | 31 ± 3                               | 26 ± 5   | 109 ± 2   | 26 ± 1   | 86 ± 39   |
| 88 ± 12                     | 14 ± 1                               | 10 ± 3   | 50 ± 1  | 18 ± 4   | 42 ± 23   |
| 27 ± 23                     | 29 ± 2                               | 28 ± 2   | 69 ± 2  | 28 ± 1   | 56 ± 12   |
| 158 ± 88                    | 32 ± 2                               | 36 ± 3   | 46 ± 5  | 36 ± 2   | 48 ± 8  |
| 172 ± 24                    | 42 ± 1                               | 42 ± 2   | 56 ± 7  | 43 ± 1   | 55 ± 9  |

Table 4. Determination of polycyclic aromatic hydrocarbons (ng/g) with six different procedures. Results are expressed as mean ± standard deviation.<sup>12</sup>

**Data set 1.** Table 6 summarizes the results of applying the joint confidence interval test (after reducing the dimensionality of the MLS and MLR regression techniques) to the 100,000 data sets obtained by applying the Monte Carlo simulation to data set 1. It can be seen that the percentage of data sets for which the theoretical point

of zero intercept and unity slope falls into the region defined by the joint confidence intervals based on the MLS method and dimensionality reduction approximately agrees with the theoretical values given by the level of significance. Because the uncertainties in all methods are homoscedastic and of the same size ( $s_{k_i} = 1$  in this case), the data structure of the  $m_1$  method meets the theoretical MLR conditions. However, even in this case, worse results are given by the MLR method than by the MLS approach.

| $5 \times 10^{-8}$ mol/l purines | $1 \times 10^{-7}$ mol/l purines | $5 \times 10^{-7}$ mol/l purines | $1 \times 10^{-6}$ mol/l purines |
|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| 102.0 ± 0.6                      | 125.0 ± 0.6                      | 150.0 ± 0.9                      | 155.0 ± 1.2                      |
| 100.0 ± 0.5                      | 108.0 ± 0.6                      | 130.0 ± 0.4                      | 135.0 ± 0.7                      |
| 100.0 ± 0.7                      | 99.0 ± 0.7                       | 125.0 ± 0.7                      | 135.0 ± 0.9                      |
| 100.0 ± 0.9                      | 100.0 ± 0.9                      | 100.0 ± 0.2                      | 100.0 ± 0.6                      |
| 100.0 ± 0.3                      | 101.0 ± 0.3                      | 110.0 ± 0.5                      | 120.0 ± 0.4                      |
| 100.0 ± 0.4                      | 100.0 ± 0.4                      | 100.0 ± 0.3                      | 100.0 ± 0.6                      |
| 101.0 ± 0.3                      | 100.0 ± 1.0                      | 75.0 ± 0.4                       | 50.0 ± 0.8                       |
| 110.0 ± 0.9                      | 120.0 ± 0.9                      | 115.0 ± 0.6                      | 70.0 ± 0.9                       |
| 100.0 ± 1.0                      | 100.0 ± 0.4                      | 101.0 ± 0.8                      | 102.0 ± 1.0                      |
| 101.0 ± 0.6                      | 101.0 ± 0.7                      | 102.0 ± 0.7                      | 102.0 ± 1.1                      |

Table 5. Degree of recovery of guanine under four different concentrations of purine derivatives. Results are expressed as mean ± standard deviation.<sup>13</sup>

**Data set 2.** The results of applying the MLS and MLR joint confidence interval test for the intercept and the slope after dimensionality reduction to the 10,000 data sets derived from data set 2 are also summarized in Table 6. The percentage of data sets for which the theoretical point of zero intercept and unity slope falls into the region defined by the joint confidence intervals based on the MLS method is quite similar to the corresponding theoretical values given by the preestablished levels of significance. As is to be expected from the heteroscedasticity in all the methods, the percentage of data sets for which the theoretical point of zero intercept and unity slope falls into the region defined by the joint confidence interval based on

the MLR method is quite different (up to 75% for the present data set) from the theoretical values.

| data set | $\alpha$ value (%) | MLS (%) | MLR (%) |
|----------|--------------------|---------|---------|
| 1        | 10                 | 88.73   | 70.66   |
|          | 5                  | 93.99   | 78.80   |
|          | 1                  | 98.92   | 88.45   |
|          | 0.1                | 99.90   | 93.39   |
| 2        | 10                 | 88.52   | 70.13   |
|          | 5                  | 93.94   | 70.88   |
|          | 1                  | 99.00   | 71.89   |
|          | 0.1                | 99.88   | 73.27   |
| 3        | 10                 | 13.14   | 8.21    |
|          | 5                  | 16.31   | 10.34   |
|          | 1                  | 20.79   | 16.63   |
|          | 0.1                | 27.97   | 24.53   |

Table 6. Percentage of simulated data sets for which a statistical difference is not significant for all the methods being compared at four different levels of significance. 100,000 data sets were obtained for each of the original data sets 1 and 3. 10,000 data sets were obtained for data set 2.

**Data set 3.** The results of applying the MLS and MLR joint confidence interval test for the intercept and the slope after dimensionality reduction to the 100,000 data sets derived from data set 3 are summarized in Table 6. Since 100,000 data sets are analyzed, all derived from initial data with regression coefficients:  $b_1=0$ ,  $b_2=0$ ,  $b_3=1.1$  (which indicates that the three methods being compared do not produce comparable results), the percentage of data sets obtained using the Monte Carlo method in which the theoretical point of zero intercept and unity slope falls into the joint confidence interval is low for both MLS and MLR with dimensionality reduction. This percentage (data sets for which the differences between the results given by the analytical methods being compared are falsely considered as not statistically significant) is linked to the type II error ( $\beta$  probability of error).<sup>14</sup>

**Data set 4.** The results of applying the joint confidence interval test for the intercept and the sum of slopes based on MLR and MLS to data set 4 are summarized in Figure 4a. The joint confidence interval test obtained with MLR indicates the lack of similarity between the four methods of measuring nickel and the reference values, but the theoretical point of zero intercept and unity slope falls into the joint confidence region for the MLS method at a level of significance of  $\alpha=5\%$ . This agrees with the author's comments, who states that the four methods are suitable for determining nickel.<sup>10</sup> Despite being closer the MLR than the MLS experimental regression point to the theoretical point, the covariance between the intercept and the sum of slopes in MLR makes the ellipse to be tilted and so the theoretical point falls outside of its confidence limits for this  $\alpha$  significance level.

**Data set 5.** Figure 4b summarizes the results of applying the joint confidence interval test based on MLR and MLS to data set 5. The joint confidence interval test obtained with MLR indicates the lack similarity between the four methods of measuring the total chromium content in soils. On the other hand, the theoretical point of zero intercept and unity slope falls into the joint confidence region for MLS for a level of significance of  $\alpha=5\%$ . These results do not agree with the author's comments, who states that one separation/extraction method is significantly better than the others.<sup>11</sup> This data set exemplifies the need for a statistical test which can compare several methods taking into account the errors associated to each method. At first sight, the data set may suggest that the nitric extraction method is more efficient than the other three. Statistically, however, this difference is not significant, so the extraction levels achieved with the four media are comparable at the level of significance of  $\alpha=5\%$ . The large differences in the uncertainties of the individual points give the joint confidence region based on MLS a large confidence ellipse and allow the theoretical point to fall inside. These differences are also enhanced by the fact that only six points are measured in the

ISBN:978-84-691-1897-9/D.L: T-353-2008

data set, which gives larger  $F$ -values (eq 16) than, for instance, in data set 4, where there are also large differences in the uncertainties of the individual points.

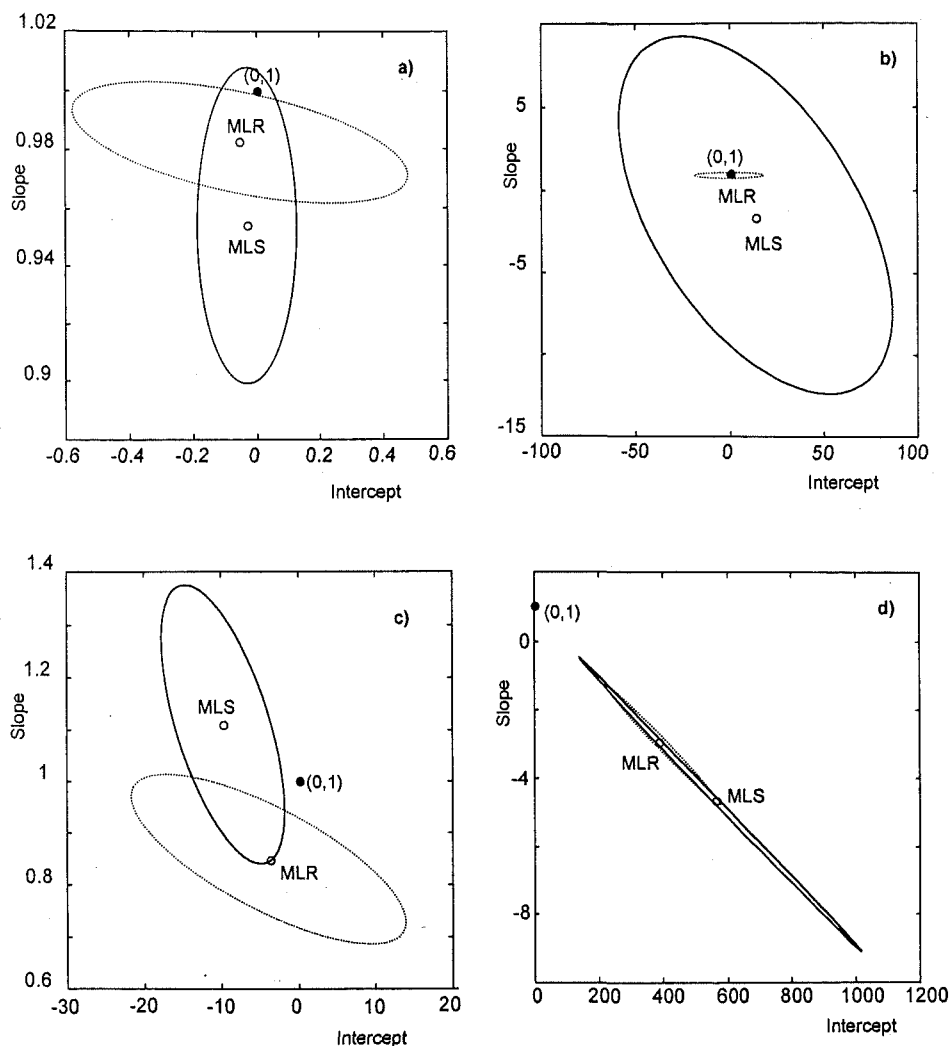


Figure 4. Joint confidence intervals based on the MLS and MLR methods for the four real data sets studied in the text. MLS, solid lines; MLR, dotted lines.

Larger confidence intervals make the theoretical point to have more probabilities to fall inside them (and hence to say that the methods in comparison give rise to comparable results), but it may also indicate that the experimental error (eqs 3 or

12) is high enough to make one suspect about a possible lack of fit of the experimental points to the model.

**Data set 6.** The results of applying the joint confidence interval test based on MLR and MLS to data set 6 are summarized in Figure 4c. The joint confidence intervals obtained with MLS and MLR indicate the lack of similarity between the six methods for determining polycyclic aromatic hydrocarbons in fly ash for a level of significance of  $\alpha=5\%$ . The authors state that the differences between the results provided by all the methods being compared are not statistically significant.<sup>12</sup> Like data sets 4 and 5, the large heteroscedasticity in the data sets means that MLS and MLR have different joint confidence intervals for the intercept and the slope. As well as the size of the confidence ellipse, in this data set the position of the center of the ellipse (the regression coefficients) and the use of one regression method upon the other (MLR or MLS) are also important.

**Data set 7.** Figure 4d summarizes the results of applying the joint confidence interval test based on MLR and MLS to data set 7. The joint confidence interval test obtained with MLR and MLS shows that the presence of four different concentrations of purine derivatives gives different results when determining guanine by differential pulse adsorptive cathodic stripping voltammetry while the authors state that they do not interfere.<sup>13</sup> The data set has low heteroscedasticity in all the axes which means that the joint confidence intervals for the slope and the intercept calculated using MLS and MLR are quite similar. This is the data set which most resembles the MLR conditions and therefore gives joint confidence regions which are similar in shape. The joint confidence region for MLS is larger because it takes into account the uncertainties in all the axes, and in this particular data set the experimental error is enlarged.

## CONCLUSIONS

The new joint confidence interval for multiple method comparison taking into account the errors in all the methods was applied to three reference simulated data sets and proved to give correct results. The application of the new test to four real data sets showed that, as expected, visual examination of the results is not a reliable method for checking the similarity between different methods. However, this new approach, which is based on statistical analysis, is a reliable tool for comparing multiple analytical methodologies.

The new method is of a general nature, with no restrictive statistical constraints. It has the additional advantage of being invariant upon switching axes for the regression coefficients. The main drawback is that the individual variances associated with each data point are needed, which may mean a longer analysis time. If these variances are omitted, the application of the MLS technique gives results which are identical to the ones obtained with MLR method.

Additional work is being carried out in our laboratory to further improve the multiple methods comparison analysis. For instance, a test to detect the possible lack of fit of the experimental points to the regression hyperplane or the detection of outliers in multivariate linear regression when there are heteroscedastic errors in all the axes is under development. We are also working on tests for detecting which of the individual methods causes the rejection of the null hypothesis for the joint confidence interval test, or the incorporation of the second type-error in the new joint confidence interval for multiple method comparison taking into account the errors in all the methods.

## Acknowledgements

The authors would like to thank the DGICYT (project num. BP96-1008) for financial support and the Spanish Ministry of Education and Science for providing Jordi Riu's doctoral fellowship.

## REFERENCES

- (1) Riu, J.; Rius, F.X. *Anal. Chem.* 1996, 68, 1851-1857
- (2) Lisý, J.M.; Cholvadová, A.; Kutej, J. *Comput. Chem.* 1990, 14, 189-192
- (3) Scheffé, H. *The Analysis of Variance*; John Wiley & Sons: New York, 1959; pp 3-41
- (4) Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y.; Kaufman, L. *Chemometrics: a textbook*; Elsevier: Amsterdam, 1988; pp 181-189
- (5) Wentworth, W.E. *J. Chem. Educ.* 1965, 42, 96-103
- (6) Wentworth, W.E. *J. Chem. Educ.* 1965, 42, 162-167
- (7) Draper, N.; Smith, H. *Applied Regression Analysis*, 2nd ed.; John Wiley & Sons: New York, 1981; pp 5-128
- (8) Meloun, M.; Militký, J.; Forina, M. *Chemometrics for Analytical Chemistry, Volume 1: PC-aided statistical data analysis*; Ellis Horwood: London, 1992
- (9) Meier, P.C.; Zünd, R. E. *Statistical Methods in Analytical Chemistry*; John Wiley and Sons: New York, 1993; pp 145-150
- (10) Butler, O.T.; Bradley, S.D.; Howe, A.M. *Analyst*, 1995, 120, 2089-2095
- (11) Fodor, P. and Fischer L. *Fresenius J. Anal. Chem.*, 1995, 351, 454-455
- (12) Friedrich, C., Cammann, K. And Kleiböhmer, W. *Fresenius J. Anal. Chem.*, 1995, 352, 730-734
- (13) Shubietah, R. M.; Abu Zhuri, A. Z. And Fogg, A.G. *Fresenius J. Anal. Chem.*, 1994, 348, 754-757



## Capítol 6

---

### **Predicció de les variables resposta i predictora en regressió lineal considerant errors en dos eixos**

En el camp de comparació de mètodes, un procés útil en determinades ocasions és la predicció de la variable predictor o de la variable resposta donat un valor determinat de l'altra variable. Per exemple, pot ser convenient saber, a partir de dades actuals o històriques del valor de la concentració i l'interval de confiança d'una determinada mostra proporcionats per un mètode, el valor de la concentració i l'interval de confiança associats que s'obtindrien amb l'aplicació d'un altre mètode. Aquests nous valors de la concentració i l'interval de confiança es poden aconseguir mitjançant la regressió lineal dels resultats obtinguts per ambdós mètodes. Com ja s'ha comentat a bastament al llarg d'aquesta tesi, els dos mètodes analítics normalment solen presentar errors heteroscedàstics del mateix ordre de magnitud, per la qual cosa en aquest capítol es desenvoluparan les expressions per calcular els intervals de confiança tant de la variable predictor com de la variable resposta a partir del mètode de regressió BLS.

Malgrat que el treball del present capítol se centrarà en els intervals de confiança associats a la recta de la regressió obtinguda en processos de comparació de dos mètodes analítics, aquests intervals de confiança poden tenir altres aplicacions. Per exemple, es poden utilitzar en l'assignació de l'origen de mostres arqueològiques desconegudes, on a cada un dels dos eixos s'hi situaria una sèrie de paràmetres (com poden ser les concentracions de diversos metalls) de cada un dels orígens coneguts on es volen intentar classificar les mostres arqueològiques d'origen desconegut.<sup>1</sup> Un altre exemple pot ser l'establiment d'una relació entre dues tècniques de datació, normalment emprant patrons radioquímics, on la incertesa associada a aquests no es pot negligir.

El gruix del present capítol es troba a l'article del final del capítol *Confidence Intervals in Linear Regression Taking into Account Uncertainties in Both Axes*, que ha estat enviat per a la seva publicació a la revista *Analytical Chemistry*.

### 6.1 Interval de confiança considerant només errors en la variable resposta

La predicció tant de la variable predictor com de la variable resposta donat un determinat valor de l'altra variable, així com el càlcul dels seus intervals de confiança associats, es troben àmpliament descrits en la literatura pel mètode de regressió OLS.<sup>2-5</sup> Les remarcables propietats matemàtiques del mètode OLS, així com algunes característiques pràctiques com poden ser la senzillesa de les seves expressions o la rapidesa en l'obtenció dels resultats, fan que aquest mètode sigui àmpliament emprat a l'hora d'obtenir els intervals de confiança de la variable predictor o resposta en regressió lineal univariant. La representació gràfica dels intervals de confiança per a la variable resposta donen les clàssiques hipèrboles de confiança, com es pot veure representat a la figura 6.1 per a un conjunt de dades determinat.

Els intervals de confiança associats a la predicció de la variable resposta segons el mètode OLS partint d'un valor de la variable predictor  $x_0$  segueixen l'expressió següent:

$$y = y_0 \pm t_{\alpha/2, n-2} \cdot \sqrt{\left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \cdot \hat{s} \quad (6.1)$$

L'equació 6.1 és l'expressió general de l'interval de confiança de la variable resposta al punt  $y_0$  d'una mostra futura, sortida de fer  $q$  rèpliques sobre el valor de la variable predictor  $x_0$ . Cal tenir present que si es vol trobar l'interval de confiança per al valor veritable de la variable resposta al punt  $y_0$ , el terme  $1/q$  de l'equació 6.1 ha d'eliminar-se,<sup>6</sup> ja que en aquest cas s'assimila que  $q = \infty$ <sup>i</sup>. La

<sup>i</sup> Les expressions que es presenten en aquest capítol per calcular els intervals de confiança associats a la predicció tant de la variable predictor com de la variable resposta són per a una mostra futura,

ISBN:978-84-691-1897-9/D.L: T-353-2008

mateixa expressió també es troba descrita en forma matricial.<sup>2</sup> De l'equació 6.1 es poden deduir els aspectes en què cal incidir per tal de disminuir el valor dels intervals de confiança: augmentar el nombre de rèpliques  $q$ , incrementar el nombre de punts de la recta de regressió  $n$  i situar els valors de la variable predictor a ambdós extrems de l'interval de regressió (per tal de maximitzar el terme  $\sum_{i=1}^n (x_i - \bar{x})^2$  de l'equació 6.1), minimitzar l'error experimental i mesurar al centre de l'interval, on  $x_0$  és semblant a  $\bar{x}$ .

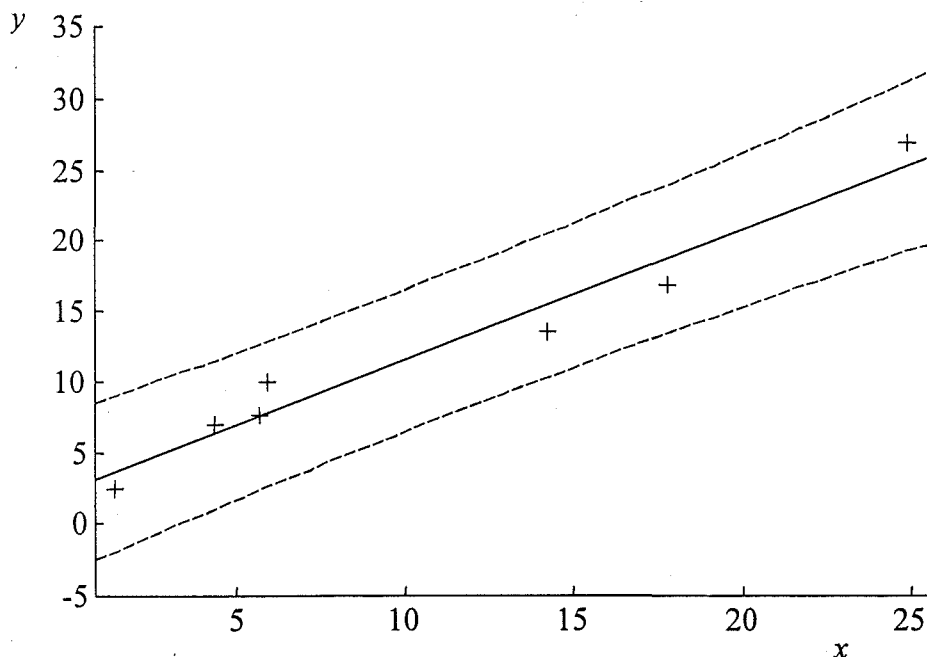


Figura 6.1 Representació dels intervals de confiança per a la variable resposta trobats amb el mètode OLS.

Per calcular múltiples intervals de confiança de la variable resposta partint de diferents valors de la variable predictor, el valor  $t_{\alpha/2, n-2}$  de l'equació 6.1 (i de les

---

sortida de fer  $q$  rèpliques al valor de la variable que no s'està predient. Per calcular l'interval de confiança per al valor veritable de la variable que s'està predient, caldrà agafar en totes les expressions  $q = \infty$ .

altres equacions que es veuran posteriorment per calcular els intervals de confiança de les variables predictor i resposta basades en els mètodes de OLS o WLS) ha de ser substituït pel valor  $\sqrt{2F_{2,n-1,\alpha}}$ . D'aquesta forma es té en compte el fet que la recta de regressió pot presentar tota la sèrie de combinacions entre ordenada a l'origen i pendent que pertanyen al seu interval conjunt. Aquests intervals de confiança s'anomenen de Working-Hotelling.<sup>7,8</sup>

La predicció d'un valor de la variable predictor  $x_0$  a partir d'un valor específic de la variable resposta  $y_0$ , així com el càlcul de l'interval de confiança associat a  $x_0$  solen anomenar-se regressió inversa. Els intervals de confiança associats a la predicció de variable predictor calculats amb el mètode OLS no sempre són simètrics respecte al valor estimat de la variable predictor, a diferència dels intervals de confiança trobats per a la variable resposta. Aquests intervals es troben resolent l'equació 6.1 com una equació quadràtica en  $x$  per a una  $y$  donada. D'aquesta manera, l'interval de confiança per al valor de la variable predictor al punt  $x_0$ , sortit de fer  $q$  rèpliques al valor de la variable resposta  $y_0$  presenta l'expressió següent:

$$x = x_0 + \frac{(x_0 - \bar{x}) \cdot g \pm \frac{t_{\alpha/2, n-2} \cdot \hat{s}}{\hat{b}} \cdot \sqrt{(1-g) \cdot \frac{n/q + 1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}{1-g} \quad (6.2)$$

on el paràmetre  $g$  correspon a:

$$g = \frac{t_{\alpha/2, n-2}^2}{\hat{b}^2 / \left[ \hat{s}^2 / \sum_{i=1}^n (x_i - \bar{x})^2 \right]} = \frac{t_{\alpha/2, n-2}^2}{\hat{b}^2 / \hat{s}_b^2} \quad (6.3)$$

Es pot observar que els intervals de confiança obtinguts amb l'equació 6.2 no sempre seran simètrics al voltant de  $y_0$ . El terme  $\sqrt{g}$  està relacionat amb el test de significança pel pendent de la recta de regressió. Si aquest test és significat al nivell de significança  $\alpha$  escollit (és a dir, si  $\hat{b}/\hat{s}_b > t_{\alpha/2, n-2}$ ), es dirà que el pendent és significat i el paràmetre  $g$  serà menor que 1. Si el test és altament significat (és a dir,  $g$  és prou petit), es pot negligir el paràmetre  $g$  en l'equació 6.2 i aquesta queda reduïda a:<sup>2,9,10</sup>

$$x = x_0 \pm \frac{t_{\alpha/2, n-2} \cdot \hat{s}}{\hat{b}} \cdot \sqrt{\frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (6.4)$$

que dona uns intervals de confiança simètrics al voltant del punt  $x_0$ . Aquesta és l'expressió que normalment sol utilitzar-se per calcular els intervals de confiança de la variable predictor emprant el mètode OLS.

El paràmetre  $g$  de l'equació 6.3 tendirà a ser gran si el valor del pendent és baix o l'error experimental alt. Per aquesta causa, la regressió inversa no té molta utilitat pràctica si no és que el test de significança del pendent sigui elevat, cosa que implica que el paràmetre  $g$  ha de ser petit (normalment menor de 0.2). Això sol comportar que la recta de regressió estigui ben determinada. Quan la recta de regressió no està ben determinada, pot ser que no s'obtinguin extrems de l'interval de confiança o que només se n'obtingui un (és a dir, que en resoldre l'equació quadràtica abans esmentada per trobar els intervals de confiança per a la variable predictor s'obtinguin arrels complexes o arrels reals, però al mateix costat de la recta de regressió). Aquests dos casos es poden observar gràficament a la figura 6.2.

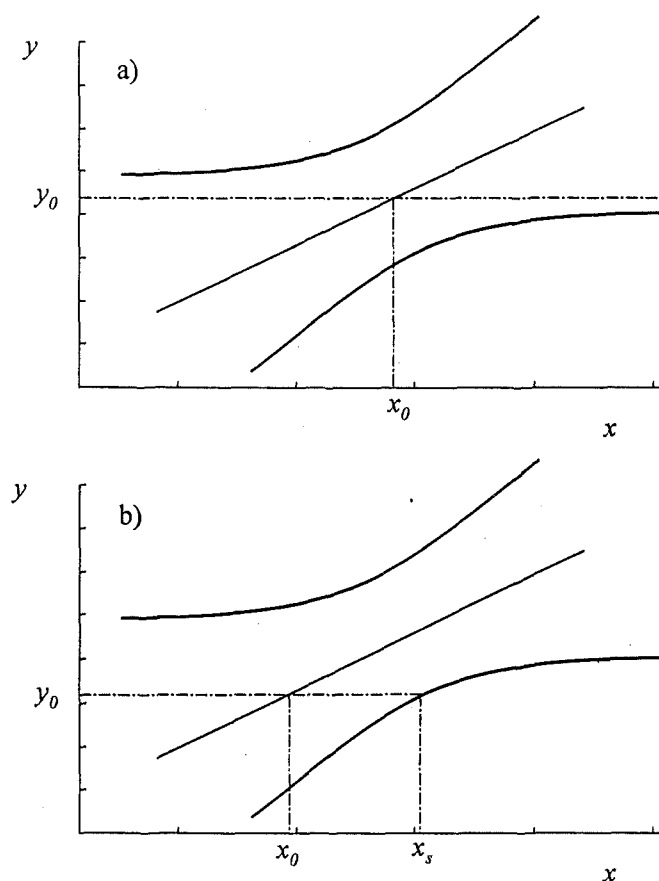


Figura 6.2. Particularitats de la regressió inversa: a) arrels complexes; b) arrels reals però al mateix costat de la recta de regressió. Els respectius intervals de confiança són a)  $(-\infty, \infty)$ , b)  $(-\infty, x_s)$ .

És important assenyalar que els intervals de confiança trobats amb el mètode OLS no són reversibles: si es calcula l'interval de confiança per a la variable resposta donat un valor de la variable predictor i llavors s'efectua un intercanvi d'eixos (la variable resposta anterior passa a ser la variable predictor i viceversa), el nou interval de confiança per a la variable predictor trobat al mateix valor anterior de la variable resposta no coincideix amb l'interval de confiança per a la variable resposta calculat anteriorment. Això es pot veure emprant les dades de la taula 6.1, que pertanyen a un procés de comparació de dos mètodes per a l'extracció d'hidrocarburs aromàtics policíclics (PAH) en sòls mitjançant extracció per fluids supercrítics emprant dos modificadors diferents:  $\text{CO}_2/10\%$  metanol i  $\text{CO}_2/10\%$

toluè.<sup>11</sup> Els resultats estan expressats en  $\mu\text{g}$  de PAH per grams de sòl.

| $\text{CO}_2/10\%$ metanol | $\text{CO}_2/10\%$ toluè |
|----------------------------|--------------------------|
| 27.9                       | 26.9                     |
| 18.8                       | 16.8                     |
| 14.7                       | 13.5                     |
| 4.6                        | 7.6                      |
| 4.7                        | 10.0                     |
| 3.4                        | 7.1                      |
| 1.0                        | 2.4                      |

Taula 6.1 Determinació de PAH en sòl mitjançant extracció per fluids supercrítics emprant dos modificadors diferents.

A la figura 6.3a es pot veure representat el procés de predicció de la variable predictor i el seu interval de confiança associat per a una sola repetició ( $q=1$ ) d'un valor de la variable resposta de  $15 \mu\text{g/g}$ . A continuació, s'intercanvien les variables d'eix i es representa el procés de predicció de la variable resposta i el seu interval de confiança associat per a una sola repetició del valor de la variable predictor de  $15 \mu\text{g/g}$  (figura 6.3b). És important recordar que amb el mètode OLS no només els intervals de confiança no són reversibles, sinó que la regressió de  $y$  sobre  $x$  dona una recta diferent de la regressió de  $x$  sobre  $y$ .

Això és important, per exemple, en processos de comparació de mètodes, en què el valor de l'interval de confiança dependrà també de l'eix on se situï cada mètode en comparació, quan no hauria d'importar quin dels dos mètodes se situés en un eix o en l'altre.

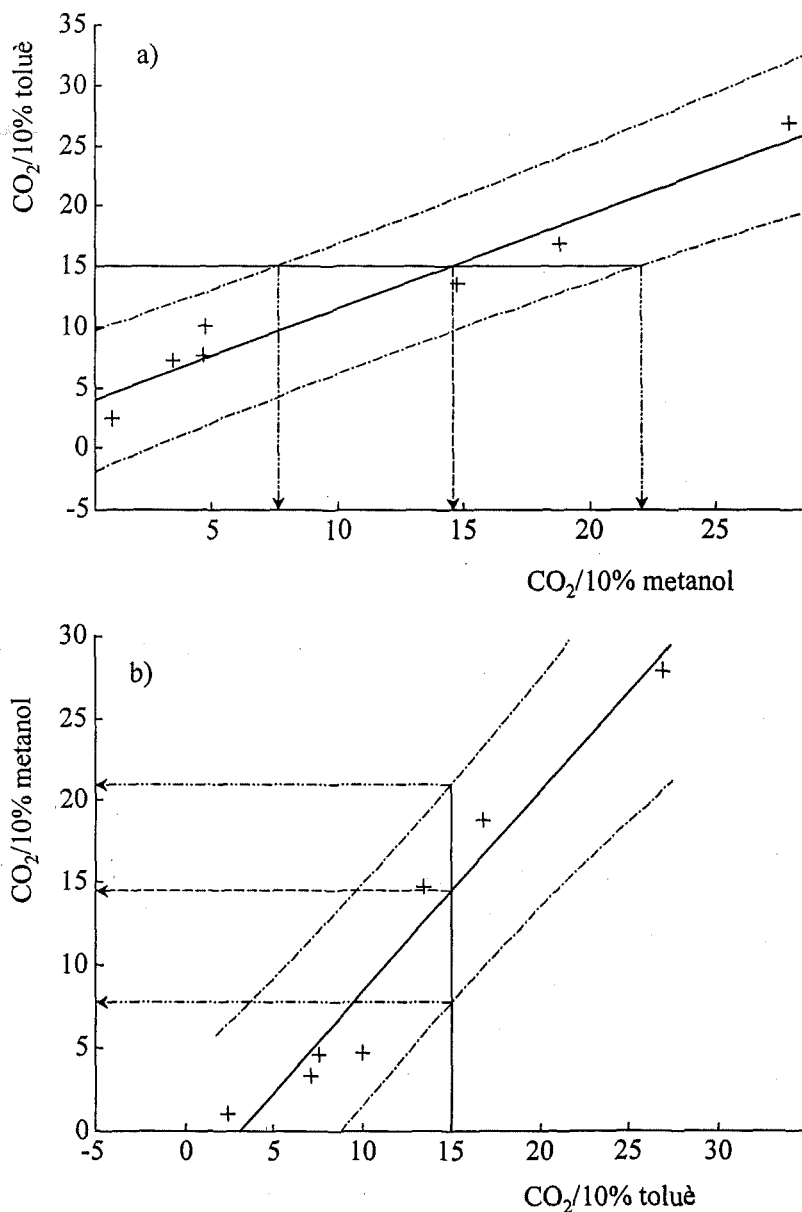


Figura 6.3. Interval de confiança quan es du a terme un intercanvi d'eixos amb les dades de la taula 6.1. a) per a un valor de  $y_0=15 \mu\text{g/g}$  s'obté un valor de  $x_0=14.50\pm 6.93 \mu\text{g/g}$ ; b) per a un valor de  $x_0=15 \mu\text{g/g}$  un cop fet un intercanvi d'eixos s'obté un valor de  $y_0=14.31\pm 6.76 \mu\text{g/g}$ .

En situacions en què el model escollit per a la recta de regressió no presenti ordenada a l'origen (el que es coneix com regressió per l'origen), és a dir, que el model de la recta de regressió sigui  $y = bx$ , els intervals de confiança presenten una

ISBN: 978-84-691-1897-9/D.L.: T-353-2008

altra forma, ja que el seu valor disminueix a mesura que s'acosten a l'origen.<sup>12</sup>  
Això és consistent amb l'assumpció feta en construir aquest model, en què se sap que el valor de la variable resposta és zero quan la variable predictor és zero. Una representació d'aquests intervals de confiança per a la variable resposta emprant un conjunt de dades determinat es pot veure en la figura 6.4.

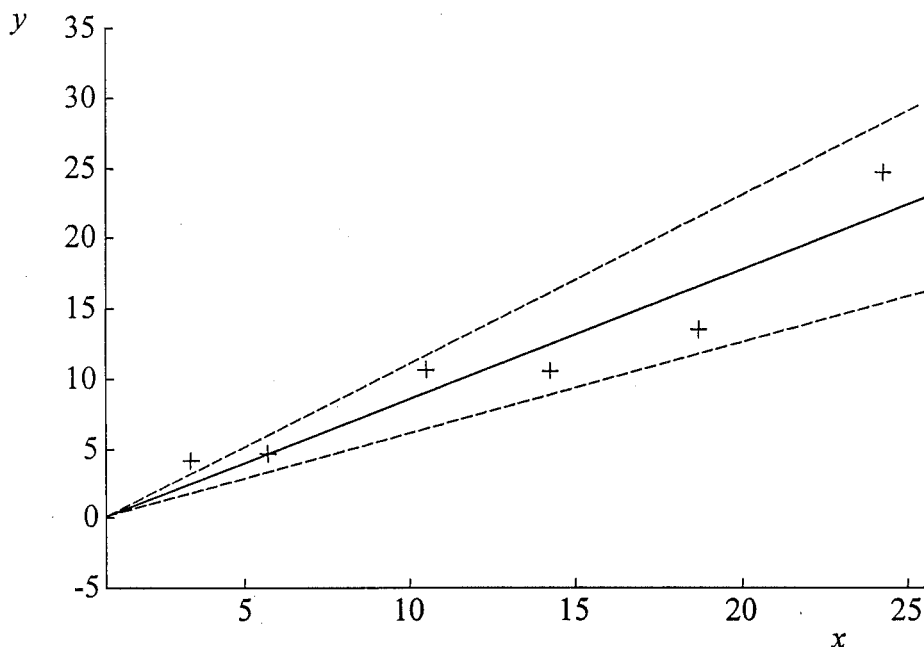


Figura 6.4 Intervals de confiança per a la variable resposta trobats amb el mètode OLS. Regressió per l'origen.

Considerant heteroscedasticitat en la variable resposta, el càlcul dels intervals de confiança emprant el mètode WLS aplicat al camp químic es du a terme principalment sobre la variable predictor. D'aquesta manera, l'interval de confiança per al valor de la variable predictor al punt  $x_0$ , sortit de fer  $q$  rèpliques al valor de la variable resposta  $y_0$  calculat amb el mètode WLS correspon a l'expressió següent:

$$x_0 \pm \frac{t_{\alpha/2, n-2} \cdot \hat{s}}{\hat{b}} \cdot \sqrt{\frac{w_0}{q} + \frac{1}{\sum_{i=1}^n \frac{1}{w_i}} + \frac{(y_0 - \bar{y}_p)^2 \sum_{i=1}^n \frac{1}{w_i}}{\hat{b}^2 \left( \sum_{i=1}^n \frac{1}{w_i} \sum_{i=1}^n \frac{x_i^2}{w_i} - \left( \sum_{i=1}^n \frac{x_i}{w_i} \right)^2 \right)}} \quad (6.5)$$

on els diversos paràmetres de l'equació 6.5 estan definits a la secció 1.3.1, i fan que els intervals de confiança siguin simètrics al voltant del punt  $x_0$ .

També tenint en compte l'heteroscedasticitat en la variable resposta, alguns autors<sup>13,14</sup> proposen uns altres intervals de confiança simètrics per a un valor  $x_0$  de la variable predictor, derivats de la recta de regressió trobada mitjançant un altre tipus de regressió ponderada diferent a WLS:

$$x_0 \pm \frac{t_{\alpha/2, n+q-3} \cdot \hat{s}}{\hat{b}} \cdot \sqrt{\frac{x_0^p}{q} + \frac{1}{\sum_{i=1}^n \frac{1}{w_i}} + \frac{(x_0 - \bar{x}_w)^2}{\sum_{i=1}^n \frac{1}{w_i} \cdot (x_i - \bar{x}_w)^2}} \quad (6.6)$$

$p$  és la potència del factor de ponderació emprat (per exemple, si el factor de ponderació emprat per trobar els coeficients de la recta de regressió és  $x^2$ , llavors  $p=2$ ). S'ha de puntualitzar que el factor de ponderació escollit en aquest cas no cal que sigui igual que l'emprat en el mètode WLS ( $w_i = s_{y_i}^2$ ), i que a la bibliografia es descriuen diversos factors de ponderació<sup>15</sup>);  $\bar{x}_w$  és el factor de ponderació mitjà dels  $n$  punts emprats per construir la recta de regressió:

$$\bar{x}_w = \frac{\sum_{i=1}^n \frac{x_i}{w_i}}{\sum_{i=1}^n \frac{1}{w_i}} \quad (6.7)$$

Aquest mètode per calcular els intervals de confiança es coneix com mètode paramètric aproximat (*approximate parametric method*, APM). També s'ha proposat una solució paramètrica asimètrica (coneguda com mètode exacte paramètric, *exact parametric method*, EPM) per als intervals de confiança de la variable predictor donat un valor de la variable resposta. Aquestes expressions es troben mitjançant un procediment iteratiu en els extrems inferior i superior de l'interval de confiança ( $x_l$ ,  $x_s$ ) fins a minimitzar les funcions de diferències relatives següents:

$$\min \left\{ \left( \left[ y_0 - \left[ \hat{a} + \hat{b}x_l + t_{\alpha, n+q-3} \cdot \hat{s} \cdot \left( x_l^p / q + 1 / \sum_{i=1}^n \frac{1}{w_i} + (x_l - \bar{x}_w)^2 / \sum_{i=1}^n \frac{1}{w_i} (x_i - \bar{x}_w)^2 \right)^{1/2} \right] \right] / y_0 \cdot 100\% \right) \right\} \quad (6.8)$$

$$\min \left\{ \left( \left[ y_0 - \left[ \hat{a} + \hat{b}x_s - t_{\alpha, n+q-3} \cdot \hat{s} \cdot \left( x_s^p / q + 1 / \sum_{i=1}^n \frac{1}{w_i} + (x_s - \bar{x}_w)^2 / \sum_{i=1}^n \frac{1}{w_i} (x_i - \bar{x}_w)^2 \right)^{1/2} \right] \right] / y_0 \cdot 100\% \right) \right\} \quad (6.9)$$

Totes les variables en les equacions 6.8 i 6.9 són trobades mitjançant el mètode de regressió ponderada escollit excepte les dues variables sobre les quals es du a terme el procediment iteratiu ( $x_l$  i  $x_s$ ). Els intervals de confiança obtinguts amb els mètodes APM i EPM s'han comparat amb els intervals de confiança proporcionats pel mètode no paramètric de *bootstrap*.<sup>16</sup> El *bootstrap* és una tècnica de computació intensiva que proporciona una estimació no paramètrica de l'error estadístic d'un model en termes de biaix i variància. El procediment imita el procés de seleccionar un nombre elevat de mostres de la mateixa mida a partir d'una

població donada per tal de calcular un interval de confiança. El conjunt de dades compost per  $n$  observacions no es considera una mostra de la població, sinó la població mateixa, a partir de la qual mostres de mida  $n$ , anomenades mostres *bootstrap*, són seleccionades amb substitució. Això s'aconsegueix assignant un número a cada observació del conjunt de dades i després generant mostres aleatòries aparellant les dades als números que corresponen a les observacions.<sup>17,18</sup> Els autors arriben a la conclusió que els intervals de confiança obtinguts amb la tècnica *bootstrap* assoleixen el millor balanç entre precisió i exactitud quan el nombre de rèpliques sobre la variable resposta ( $q$ ) està entre 2 i 4.

## 6.2 Intervals de confiança considerant errors en les variables predictor i resposta

Els errors en la variable predictor han estat considerats en molt poques ocasions a l'hora de calcular els intervals de confiança de la recta de regressió. Spiegelman i col·laboradors<sup>19</sup> van desenvolupar les expressions per calcular els intervals de confiança en processos de calibració per a la determinació de potassi mitjançant espectrometria d'emissió de flama (FES), on la resposta instrumental (variable resposta) sol ser tan estable que difícilment es poden negligir les incerteses en els patrons de calibració (variable predictor). Aquests intervals de confiança consisteixen en un eixamplament dels intervals de confiança calculats considerant només els errors en la resposta instrumental. Les expressions trobades estan basades en un procediment desenvolupat per Lwin i Spiegelman<sup>20</sup> aplicable quan els errors en la variable predictor són molt baixos. Assumint un error màxim ( $\phi_i$ ) associat als patrons de calibració del 0.5% del valor individual de la concentració de cada patró ( $\phi_i = 0.5\% \cdot x_i$ ), en aquest cas concret els intervals de confiança es veuen augmentats en un valor  $\Delta(x)$ :

$$\Delta(x_i) = \varphi_i \sum_{i=1}^n |c_i| m(x_i) \quad (6.10)$$

on  $m(x_i)$  és una funció positiva donada i les constants  $c_i$  depenen dels valors vertaders de la concentració dels estàndards (es compleix que  $\sum_{i=1}^n c_i = 1$ ). Tenint en compte aquest procediment, per exemple els intervals de confiança per a una mostra desconeguda de concentració 4  $\mu\text{g/ml}$  es veuen augmentats aproximadament en un 0.5%.

L'aproximació CVR (apartat 2.2) també ha estat emprada per calcular la variància associada a la predicció de les variables resposta i predictor. La variància per el valor de la variable resposta al punt  $y_0$  corresponent al punt  $x_0$  de la variable predictor ve donada per la següent expressió:<sup>21</sup>

$$s_{y_0}^2 = \hat{b}^2 s_{\delta}^2 + \left[ \frac{1}{n} + (1 + k\hat{b})^2 \cdot \frac{(x_0 - \bar{x})^2}{S_{uu}} \right] \cdot s_e^2 \quad (6.11)$$

on

$$S_{uu} = \sum_{i=1}^n (x_i - \bar{x})^2 + 2k \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + k^2 \sum_{i=1}^n (y_i - \bar{y})^2 \quad (6.12)$$

$$s_e = \sqrt{\left( \hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2\hat{b} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2 \right) / (n-2)} \quad (6.13)$$

$$s_{\delta} = s_e / \sqrt{\hat{b}^2 + \lambda - 2\hat{b}\theta} \quad (6.14)$$

El paràmetre  $k$  de les equacions 6.11 i 6.12 s'introdueix en el procés de càlcul dels coeficients de regressió i correspon a:<sup>21</sup>

$$k = \frac{\hat{b} - \theta}{\lambda - \hat{b}\theta} \quad (6.15)$$

on  $\lambda$  correspon a la relació entre les variàncies de la variable resposta i la variable predictor i  $\theta$  correspon a la covariància entre les variàncies resposta i predictor multiplicada per  $\sqrt{\lambda}$ .<sup>21</sup> El valor  $s_8$  està relacionat amb l'error associat a la variable predictor i  $s_e$  està associat a l'estimació de l'error experimental. L'expressió pel càlcul de la variància per el valor de la variable predictor al punt  $x_0$  corresponent al punt  $y_0$  de la variable resposta ve donada per la següent expressió:

$$s_{x_0}^2 = \frac{s_e^2}{\hat{b}^2} + \frac{1}{\hat{b}^2} \left[ \frac{1}{n} + (1 + k\hat{b})^2 \cdot \frac{(y_0 - \bar{y})^2}{\hat{b}^2 S_{uu}} \right] \cdot s_e^2 \quad (6.16)$$

on els paràmetres  $S_{uu}$ ,  $s_e$  i  $k$  es troben definits respectivament a les equacions 6.12, 6.13 i 6.15, i  $s_8$ , que està relacionat amb l'error associat a la variable resposta, ve donat segons la següent expressió:

$$s_e = s_8 \sqrt{\lambda} \quad (6.17)$$

Com es podrà comprovar a l'article que es troba al final del capítol, el fet de calcular la variància associada a la predicció de les variables resposta i predictor emprant l'aproximació CVR (o en el seu cas particular de OR) dóna lloc a resultats que es troben lluny dels valors teòrics obtinguts mitjançant el mètode de simulació de Monte Carlo,<sup>22</sup> i molt pitjors que els obtinguts mitjançant les expressions derivades amb el mètode BLS.

Per tal de desenvolupar els intervals de confiança associats a la predicció tant de la variable resposta com de la predictor considerant els errors heteroscedàstics individuals en totes dues sense limitació en el tipus d'incertesa associat a les

variables, s'utilitzarà el mètode de regressió BLS que té en compte heteroscedasticitat tant en la variable predictor com en la variable resposta. Amb aquest objectiu se seguiran dues estratègies: intentar adaptar les expressions dels mètodes OLS i WLS, i aplicar la teoria de la propagació dels errors sobre la variable de la qual es vulgui calcular el seu interval de confiança. D'aquesta manera s'obindrà la variància per a la variable predictor o resposta donat un valor de l'altra variable. El fet de trobar la variància de la variable predictor o resposta per dos camins independents constitueix en si mateix una validació, ja que, com s'exposarà més detalladament a l'article del final del capítol, els resultats obtinguts amb les dues estratègies són exactament idèntics. A més, les simulacions dutes a terme amb el mètode de Monte Carlo donen una bona concordança entre els valors obtinguts amb les expressions per calcular els intervals de confiança basats en el mètode BLS i els valors teòrics sortits de les simulacions, i molt millors que els resultats obtinguts amb els mètodes OLS, WLS, CVR i OR quan es troben presents errors en els dos eixos. A partir de la variància es construirà l'interval de confiança per a les variables predictor i resposta assumint la hipòtesi que la seva distribució és normal, tot i tenir en compte que malgrat que les seves distribucions no són normals poden acceptar-se aquestes hipòtesis sense cometre un error elevat (l'aplicació dels mètodes de comprovació de la normalitat de Cetama, Kolmogorov i gràfiques de probabilitat normal a les variables resposta i predictor donen uns resultats completament anàlegs als obtinguts amb la comprovació de la normalitat de l'ordenada a l'origen i el pendent de la recta de regressió trobades mitjançant BLS, apartat 3.1).

### 6.3 Referències

1. J. Yellin, *Analyst*, 14 (1995) 37
2. N. Draper, H. Smith, *Applied Regression Analysis*, 2nd ed., Wiley, New York (1981)

3. S.J. Haswell (ed.), *Practical Guide to Chemometrics*, Marcel Dekker Inc., New York (1992)
4. M. Meloun, J. Militký, M. Forina, *Chemometrics for Analytical Chemistry. Volume 2. PC-aided Regression and Related Methods*, Ellis Horwood, London (1994)
5. 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)
6. S. Weisberg, *Applied Linear Regression*, 2nd Ed., John Wiley & Sons, New York (1985)
7. H. Working, H. Hotelling, *Journal of American Statistical Association*, **24** (1929) 73
8. R.G. Miller, *Simultaneous Statistical Inference*, McGraw-Hill, New York (1966)
9. G.W. Snedecor, W.G. Cochran, *Statistical Methods*, 8th ed., Iowa State University Press, Ames (1989)
10. P.D. Lark, B.R. Crowen, R.L.L. Bosworth, *The Handling of Chemical Data*, Pergamon Press, Oxford (1968)
11. J.J. Langenfeld, S.B. Hawthorne, D.J. Miller, J. Pawliszyn, *Analytical Chemistry*, **66** (1994) 909
12. J.O. Rawlings, *Applied Regression Analysis: A Research Tool*, Wadsworth & Brooks/Cole Advanced Books & Software, Belmont (1988)
13. J.C. Miller, J.N. Miller, *Statistics for Analytical Chemists*, Ellis Horwood, Chichester (1984)
14. J.N. Miller, *Analyst*, **116** (1991) 3
15. A.G. Asuero, A.G. González, *Microchemical Journal*, **40** (1989) 216
16. P.L. Bonate, *Analytical Chemistry*, **65** (1993) 1367
17. I.E. Frank, R. Todeschini, *The Data Analysis Handbook*, Elsevier, Amsterdam (1994)
18. P. Hall, *The Annals of Statistics*, **14** (1986) 1431

3. S.J. Haswell (ed.), *Practical Guide to Chemometrics*, Marcel Dekker Inc., New York (1992)
4. M. Meloun, J. Militký, M. Forina, *Chemometrics for Analytical Chemistry. Volume 2. PC-aided Regression and Related Methods*, Ellis Horwood, London (1994)
5. 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)
6. S. Weisberg, *Applied Linear Regression*, 2nd Ed., John Wiley & Sons, New York (1985)
7. H. Working, H. Hotelling, *Journal of American Statistical Association*, **24** (1929) 73
8. R.G. Miller, *Simultaneous Statistical Inference*, McGraw-Hill, New York (1966)
9. G.W. Snedecor, W.G. Cochran, *Statistical Methods*, 8th ed., Iowa State University Press, Ames (1989)
10. P.D. Lark, B.R. Crowen, R.L.L. Bosworth, *The Handling of Chemical Data*, Pergamon Press, Oxford (1968)
11. J.J. Langenfeld, S.B. Hawthorne, D.J. Miller, J. Pawliszyn, *Analytical Chemistry*, **66** (1994) 909
12. J.O. Rawlings, *Applied Regression Analysis: A Research Tool*, Wadsworth & Brooks/Cole Advanced Books & Software, Belmont (1988)
13. J.C. Miller, J.N. Miller, *Statistics for Analytical Chemists*, Ellis Horwood, Chichester (1984)
14. J.N. Miller, *Analyst*, **116** (1991) 3
15. A.G. Asuero, A.G. González, *Microchemical Journal*, **40** (1989) 216
16. P.L. Bonate, *Analytical Chemistry*, **65** (1993) 1367
17. I.E. Frank, R. Todeschini, *The Data Analysis Handbook*, Elsevier, Amsterdam (1994)
18. P. Hall, *The Annals of Statistics*, **14** (1986) 1431

19. C.H. Spiegelman, R.L. Watters, L. Hungwu, *Chemometrics and Intelligent Laboratory Systems*, **11** (1991) 121
20. T. Lwin, C.H. Spiegelman, *Journal of the Royal Statistical Society Series C*, **35** (1986) 256
21. J. Mandel, *Journal of Quality Technology*, **16** (1984) 1
22. P.C. Meier, R.E. Zünd, *Statistical Methods in Analytical Chemistry*, John Wiley & Sons, New York (1993)

## Confidence intervals in linear regression taking into account uncertainties in both axes

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

This study reports the expressions for the variances associated to the response and predictor variables which are calculated with the bivariate least squares (BLS) regression technique, which takes into account the errors in both axes. The calculated results are compared to those obtained from a simulation process applied to six different real data sets. The mean error of the results found with the new expressions is between 4 and 5% whereas mean errors can be as high as 85%, 277%, 637% and 1697% when weighted least squares, ordinary least squares, constant variance ratio approach and orthogonal regression are used respectively. An important property of the confidence intervals calculated using the BLS regression technique is the invariance of the results when axes are switched.

## INTRODUCTION

The remarkable mathematical properties of ordinary least squares, OLS, together with its practical performance characteristics are the main reasons why it is the regression technique which is most commonly used by the analytical chemistry community. However, this technique is based on a set of mathematical hypotheses such as the homoscedasticity on the  $y$  axis or the absence of errors on the  $x$  axis, that are sometimes not fulfilled. This leads to biased regression coefficients of the straight line and, consequently, to erroneous predicted results.<sup>1,2</sup> Method comparison studies, where the errors associated to both methods are usually of the same order of magnitude, or calibration lines, where the errors in the instrumental responses are comparable to the errors associated to the concentration values,<sup>3</sup> are situations in which the application of OLS often provides biased results.

An improvement on the OLS technique is the weighted least squares (WLS) technique<sup>1,4</sup> that takes into account heteroscedasticity in the  $y$ -axis. However, WLS still considers the  $x$  axis as being error free.

The errors-in-variables regression<sup>5</sup>, also called constant variance ratio (CVR) approach,<sup>6-8</sup> 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)<sup>9</sup>, in which the errors are of the same order of magnitude in the response and predictor variable (i.e.  $\lambda = 1$ ). In the literature, this latter case is also called orthogonal distance regression (ODR)<sup>2</sup> or total least squares regression (TLS).<sup>10</sup>

The bivariate least squares (BLS) method<sup>11,12</sup> is a linear regression technique capable of overcoming the limitations of the previous methods i.e. the fact that the

individual uncertainties in both variables are not considered. This technique calculates the straight line regression coefficients by taking into account the heteroscedastic uncertainties in both axes. BLS has been applied in method validation studies to detect bias in newly developed analytical methodologies.<sup>13</sup>

The calculation of predicted values in regression analysis considering individual heteroscedastic errors in both axes is an important issue in practical instances that has merited little attention up to date. The calculation of the measurement results and the uncertainty of a newly developed method from the historical values recorded by using a previously established methodology, or the establishment of relationships between two dating methodologies, both incorporating uncertainty, so as to assign the chronological origin of archaeological samples, are two examples where confidence intervals from linear regression taking into account uncertainties in both axes should be considered.

This paper develops and validates new expressions for calculating the confidence intervals for predicted values of the response variable given a value of the predictor variable, and vice versa, using the BLS regression technique, i.e. by considering the individual uncertainties of every experimental point. The expressions for the predictor intervals considering errors in both axes have been derived from a generalisation of the existing OLS and WLS expressions. The same results have also been found using the error propagation theory.<sup>14</sup> To validate the appropriateness of the new confidence intervals, six real data sets were used. Random errors based on the individual uncertainties of each real point were added to the data sets using the Monte Carlo method. The values obtained with the new expressions based on BLS do agree with the theoretical values more than the results obtained using the expressions based on OLS, WLS, OR or CVR. One of the most important properties of the BLS confidence intervals is their invariance when axes are switched.

## BACKGROUND AND THEORY

**Bivariate least squares technique.** Of the various regression techniques that consider errors in both axes, bivariate least squares (BLS) was chosen because it can readily provide the regression coefficients and their associated variances and covariances, and because of the simplicity of programming its algorithm. The prediction step using the straight line model is expressed in eq. 1:

$$\hat{y}_i = \hat{a} + \hat{b}x_i \quad (1)$$

where  $\hat{a}$  represents the intercept,  $\hat{b}$  the slope, and  $\hat{y}_i$  is the predicted value for the observed value  $x_i$ . The method consists of minimising the sum of the weighted residuals of the regression straight line:

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

where  $n$  is the number of experimental data points and  $w_i$  is the weighting factor that corresponds to the variance of the  $i$ th-residual:

$$w_i = s_{y_i}^2 + \hat{b}^2 s_{x_i}^2 - 2\hat{b} \text{cov}(x_i, y_i) \quad (3)$$

where  $s_{x_i}^2$  and  $s_{y_i}^2$  are, respectively, the variances of the  $i$ th-point for the predictor and response variables of the straight line expressed in the eq. 1, and  $\text{cov}(x_i, y_i)$  is the covariance between the predictor and the response variable, which is normally set to zero.

It is interesting to note that whenever the variances of the predictor variable values are zero and all the variances on the response variable are the same (i.e., all errors are constant and only due to the experimental measurement in the  $y$ -axis), the results obtained are identical to those obtained with the OLS method. Since in the BLS regression model the unobserved  $\hat{x}$  and  $\hat{y}$  values are affected by a random error, and it is only possible to observe the  $x$  and  $y$  values, BLS can be considered a structural regression model, in contrast to the functional models in which the  $x$ -variable is fixed and known to be without error.<sup>5</sup>

**Variance of the response variable.** In the OLS method, the well known expression for the variance of the predicted observation of the response variable  $y_0$  obtained as the mean of  $q$  observations performed at  $x_0$  is given by eq. 4:

$$s_{y_0}^2 = \left[ \frac{1}{q} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \cdot \hat{s}^2 \quad (4)$$

where  $\bar{x}$  is the mean value of the predictor variable and  $\hat{s}^2$  is the estimation of the experimental error ( $s^2$ ):

$$\hat{s}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} \quad (5)$$

Eq. 4 can also be expressed in matrix form:

$$s_{y_0}^2 = \left( \frac{1}{q} + \mathbf{X}'_0 \cdot (\mathbf{X}' \cdot \mathbf{X})^{-1} \cdot \mathbf{X}_0 \right) \cdot \hat{s}^2 \quad (6)$$

where  $\mathbf{X}_0$  is a two-element column vector formed by a 1 in the first row and the predictor variable ( $x_0$ ) in the second row, and  $\mathbf{X}$  is an  $n \times 2$  matrix in which the first column is a column of ones and the second is formed by the  $n$  values of the predictor variable corresponding to the experimental points.

For the WLS technique, which takes into account heteroscedastic errors in the response variable, the variance for the predicted observation  $y_0$ , calculated as the mean of  $q$  observations performed at a selected value of  $x_0$  is given by eq. 7:

$$s_{y_0}^2 = \left[ \frac{1}{q} + \mathbf{X}'_0 \cdot (\mathbf{X}' \cdot \mathbf{V}^{-1} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}_0 \right] \cdot \hat{s}^2 \quad (7)$$

where  $\mathbf{V}$  is an  $n \times n$  diagonal matrix the  $i$ th element of which corresponds to the variance of  $y_i$  ( $s_{y_i}^2$ ), and where  $\hat{s}^2$ , the estimation of the experimental error, now takes into account the variances of the response variable as the weighting factor:

$$\hat{s}^2 = \frac{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{s_{y_i}^2}}{n - 2} \quad (8)$$

In the CVR approach, the expression for the variance of the prediction given a value of the predictor variable, is given by eq. 9:

$$s_{y_0}^2 = \hat{b}^2 s_b^2 + \left[ \frac{1}{n} + \frac{(1 + k\hat{b})^2 \cdot (x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2 + 2k \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + k^2 \sum_{i=1}^n (y_i - \bar{y})^2} \right] \cdot s_e^2 \quad (9)$$

where  $k$ ,  $s_{\hat{y}_0}^2$  and  $s_e^2$  are defined in the process of finding the regression coefficients ( $s_{\hat{y}_0}^2$  refers to the error associated to the predictor variable and  $s_e^2$  is associated to the estimate of the experimental error).<sup>8</sup> If  $\lambda$ , which appears in the coefficients  $k$  and  $\hat{b}$  in eq. 9 is chosen to be unity, then results for the OR method are obtained.

The expression for the variance of the prediction given a value of the predictor variable for the BLS regression technique, which takes into account heteroscedastic errors in both axes, is given by eq. 10:

$$s_{y_0}^2 = \mathbf{X}'_0 \cdot (\mathbf{X}' \cdot \mathbf{W}^{-1} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}_0 \cdot \hat{s}^2 \quad (10)$$

where the matrix  $\mathbf{W}$  is a  $n \times n$  diagonal matrix the  $i$ th-diagonal element of which is the weighting factor  $w_i$  defined in eq. 3. This weighting factor takes into account the errors in both axes. The estimation of the experimental error is now:

$$\hat{s}^2 = \frac{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{w_i}}{n-2} \quad (11)$$

However, in this expression there is still a term to consider when the errors of the predictor variable ( $x_0$ ) are also taken into account. In order to correct the difference in ranges between the two axes, the factor corresponding to the square of the slope has to be introduced. The expression for the variance of the response true mean value at a given observation  $x_0$  is shown in eq. 12:

$$s_{y_0}^2 = \left[ \mathbf{X}'_0 \cdot (\mathbf{X}' \cdot \mathbf{W}^{-1} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}_0 + s_{x_0}^2 \cdot \hat{b}^2 \right] \cdot \hat{s}^2 \quad (12)$$

On the other hand, an independent expression for the variance associated to the predicted response variable can be found by applying the error propagation theory<sup>14</sup>

to eq. 1. This expression is presented in eq. 13. The covariances between the regression coefficients and  $x_0$  are assumed to be negligible.

$$s_{y_0}^2 = \hat{s}_a^2 + x_0^2 \hat{s}_b^2 + \hat{b}^2 s_{x_0}^2 + 2x_0 \text{cov}(\hat{a}, \hat{b}) \quad (13)$$

where  $\hat{s}_a^2$  and  $\hat{s}_b^2$  are the estimates of the variances of the intercept and the slope respectively, and  $\text{cov}(\hat{a}, \hat{b})$  is the covariance between the two regression coefficients. The variances and covariances of the regression coefficients are easily obtained during the iterative process to find the regression coefficients provided by BLS.<sup>12</sup> The coincidence of the results when expressions 12 and 13 are used to calculate the variance of the response variable given a value of the predictor variable is an internal validation of the derived expressions.

The uncertainty of the predicted observation of the response variable using the BLS technique must take into account the variance of the regression line (eqs. 12 or 13) and the variance of the new observation. Eq. 14 gives the final matrix expression for the calculation of the variance of the response variable  $y_0$  obtained as the mean of  $q$  observations performed at  $x_0$ .

$$s_{y_0}^2 = \left[ \frac{1}{q} + \mathbf{X}'_0 \cdot (\mathbf{X}' \cdot \mathbf{W}^{-1} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}_0 + s_{x_0}^2 \cdot \hat{b}^2 \right] \cdot \hat{s}^2 \quad (14)$$

**Variance of the predictor variable.** The study of the variance associated to the predicted predictor variable given a value of the response variable is similar to the study of the prediction of the response variable. Only the new expressions developed for the BLS method are presented here. Taking into account the errors in both axes, the resulting expression is eq. 15:

$$s_{x_0}^2 = \left[ \mathbf{Y}'_0 \cdot (\mathbf{Y}' \cdot \mathbf{W}^{-1} \cdot \mathbf{Y})^{-1} \cdot \mathbf{Y}_0 + s_{y_0}^2 \cdot \frac{1}{\hat{b}^2} \right] \cdot \frac{\hat{s}'^2}{\hat{b}^2} \quad (15)$$

where  $\mathbf{Y}_0$  is a two-element column vector with a 1 in the first row and the response variable  $y_0$  in the second row,  $\mathbf{Y}$  is an  $n \times 2$  matrix in which the first column is a column of ones and the second is made up of the  $n$  values corresponding to the response variables of the experimental points,  $\mathbf{W}$  is an  $n \times n$  diagonal matrix whose  $i$ th-diagonal element is the weighting factor  $w'_i$ , and  $\hat{s}'^2$  is the experimental error associated to the predictions on the  $x$  axis, which corresponds to:

$$\hat{s}'^2 = \frac{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{w'_i}}{n-2} \quad (16)$$

The weighting factor  $w'_i$  is now defined in eq. 17.

$$w'_i = s_{x_i}^2 + \frac{1}{\hat{b}^2} s_{y_i}^2 - 2 \frac{1}{\hat{b}} \text{cov}(x_i, y_i) \quad (17)$$

where normally the covariance between the predictor and response variables is neglected. The variance of the predictor variable  $x_0$  is the mean of  $q$  observations at  $y_0$  and is found according to eq. 18:

$$s_{x_0}^2 = \left[ \frac{1}{q} + \mathbf{Y}'_0 \cdot (\mathbf{Y}' \cdot \mathbf{W}^{-1} \cdot \mathbf{Y})^{-1} \cdot \mathbf{Y}_0 + s_{y_0}^2 \cdot \frac{1}{\hat{b}^2} \right] \cdot \frac{\hat{s}'^2}{\hat{b}^2} \quad (18)$$

**Predictor intervals.** The hypothesis of normality can be assumed in the distributions of both the intercept and the slope,<sup>15</sup> and the results are not appreciably biased when linear regression with errors in both axes is used. Furthermore, three methods for testing the normality (Kolmogorov test,<sup>16</sup> normal

probability plots<sup>14</sup> and Cetama method<sup>17</sup>) were applied to the Monte Carlo simulation data of the response and predictor variables and the results (not shown) indicate that the response and predictor variables, despite being non-normally distributed in most cases, are very close to normality. The hypothesis that their distribution is normal is, therefore, acceptable.

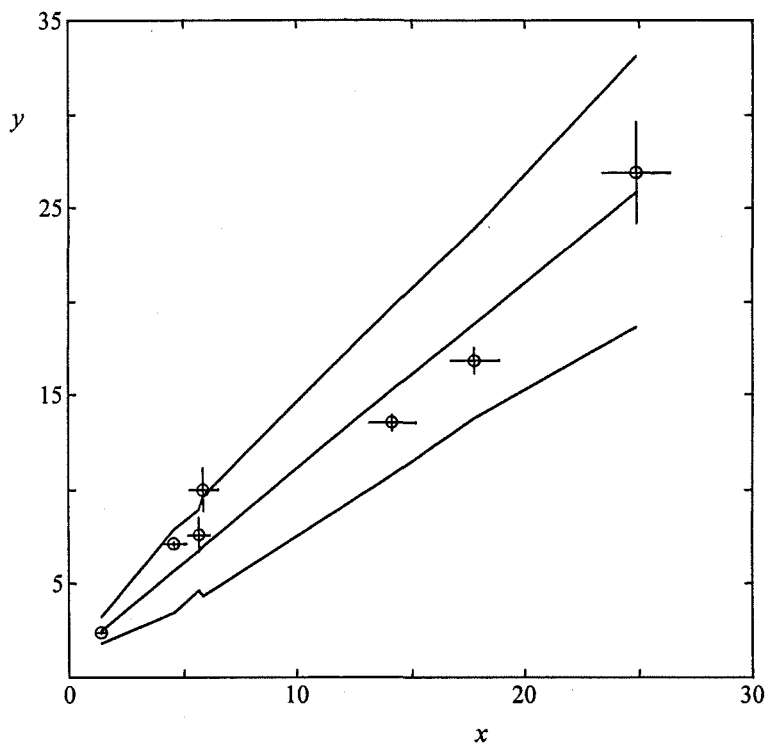


Figure 1° Experimental points for data set 3, calculated BLS regression line and confidence intervals associated to the response variable. A significance level of  $\alpha = 0.1$  was selected.

The expressions of the confidence intervals for the response and predictor variables are then defined by the following equations:

$$y_0 \pm t_{\alpha, n-2} s_{y_0} \quad (19)$$

$$x_0 \pm t_{\alpha, n-2} s_{x_0} \quad (20)$$

where  $t_{\alpha, n-2}$  is the  $t$ -value for a given level of significance  $\alpha$  and  $n-2$  degrees of freedom. As an example, the confidence interval associated to the prediction of the response variable which takes the uncertainties in both axes into account for data set 3 in the Experimental Section is shown in Figure 1. In linear regression taking into account errors in both axes, the patterns of the confidence intervals are very irregular. This is due to the variance of the predictor variable (i.e. the last term within brackets in eqs. 12 and 14, or in eqs. 15 and 18). If these terms were constant throughout the regression interval, then the confidence interval would have the shape of the classical hyperbola that is found in OLS. However, since heteroscedasticity is usually present, these terms are not constant and the pattern for the confidence intervals which take into account errors in both axes can be calculated at a given point provided that the individual uncertainty at this point is known. The continuous line for the confidence interval along the regression line is drawn by interpolating between contiguous points, since only the confidence interval at the points used for predicting can be strictly calculated.

## EXPERIMENTAL SECTION

**Data sets and software.** Six real data sets were used to validate the expressions for calculating the variance of the response variable given a value of the predictor variable and vice versa. In the data sets studied, mainly about method comparison studies, the established method is normally placed on the  $x$  axis and the new method on the  $y$  axis. Data sets 3 and 6, which are not related to method comparison studies, were introduced to show the usefulness of the new expressions in other fields. These six data sets are plotted in Figure 2. For the sake of clarity, only the BLS, OLS and WLS regression lines have been drawn in Figure 2.

*Data Set 1:* Concentrations of polycyclic aromatic hydrocarbons (PAHs) recovered from railroad bed soil after supercritical fluid extraction (SFE) with  $\text{CO}_2$  as the modifier on the  $x$  axis, and  $\text{CO}_2/10\%$  toluene as the modifier on the  $y$  axis.<sup>18</sup> The

standard deviations are the averages of three determinations at each of the 7 experimental points. The data set ranges between 1.4 and 26.9  $\mu\text{g/g}$  of soil. The standard deviations associated to all experimental point are similar in both methods.

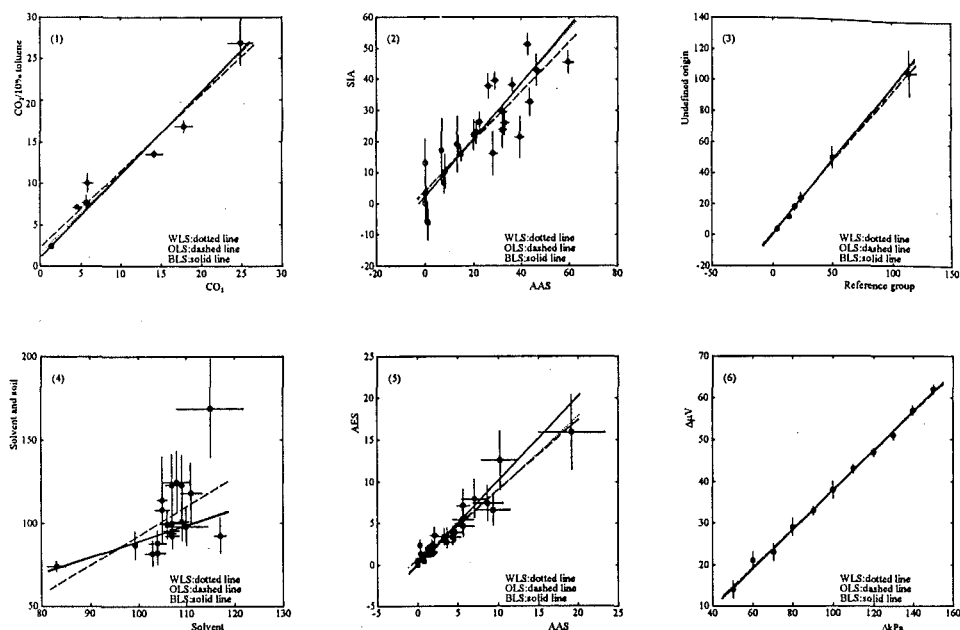


Figure 2 BLS (solid line), OLS (dashed line) and WLS (dotted line) regression lines for the six real data sets assayed. The experimental points are represented with their associated uncertainties.

*Data Set 2:* A method comparison study for analyzing  $\text{Mg}^{2+}$  in natural waters with atomic absorption spectrometry (AAS) on the  $x$  axis, and sequential injection analysis (SIA) on the  $y$  axis.<sup>19</sup> The uncertainties in AAS are derived from the uncertainties associated to the univariate calibration line. The uncertainties in the SIA method are calculated from the multivariate regression model developed using the partial least squares (PLS) technique. The comparison consists of 26 data pairs within the range 0.4 and 46.3  $\text{mg/l}$ . In all cases, the uncertainties associated to the SIA method are larger than the ones provided by AAS.

*Data Set 3:* The composition of a set of archaeological samples of unknown origin (on the  $y$ -axis) is compared to a reference set of known origin (on the  $x$ -axis) with neutron activation analysis (NAA). Concentrations of six metal ions (Ce, Co, Cr, Fe, La and Sc) expressed in ppm, except for Fe which is in percent, are determined for a number of pottery jar handles found in Tell en-Nasbeh.<sup>20</sup>

*Data Set 4:* The percentage of recovery for several organochlorine pesticides after microwave-assisted extraction (MAE) with solvent (hexane/acetone 1:1) on the  $x$  axis, and solvent/soil suspensions spiked with the target compounds on the  $y$  axis.<sup>21</sup> The standard deviations are the average of three determinations at each point. The experiment consists of 20 points with recoveries ranging between 83 and 169%. The variances on both axes are quite large, and there is a possible outlier at high recovery values.

*Data Set 5:* A method comparison study for the determination of arsenic in natural water using continuous selective reduction and atomic absorption spectrometry (AAS) on the  $x$  axis, and reduction, cold trapping and atomic emission spectrometry (AES) on the  $y$  axis.<sup>22</sup> The study consists of 30 points ranging between 0 and 19.3 mg/l. The uncertainties are proportional to the concentration determined by both methods.

*Data Set 6:* Data from the measurement of the CO<sub>2</sub> Joule-Thompson coefficient.<sup>23</sup> The data correspond to thermocouple measured voltage differences ( $\Delta \mu\text{V}$ ) on the  $y$  axis, as a function of pressure increments ( $\Delta \text{kPa}$ ) on the  $x$  axis. There were 11 equally distributed data pairs with estimated unity  $x$  axis uncertainties. The  $y$  axis uncertainties were estimated to range between one and two units.

All calculations were performed using customized software using MATLAB.<sup>24</sup>

**Validation process.** The Monte Carlo simulation technique was applied<sup>25</sup> to validate the expressions derived to calculate the variances of the response and the predictor variables using regression considering errors in both axes (eqs. 12, 13 and 15). The Monte Carlo method generated 10,000 different data sets for each of the six initial real data sets using the individual uncertainties of each experimental point. For each of the 10,000 new data sets, the regression line was calculated, and used to predict a value of the response or the predictor variable. Finally, the variance of the 10,000 values for each original data set was calculated and compared to the predicted variance given by the theoretical expressions. This study was performed for two random values of each data set. The reversibility of the axes was also checked using the expressions for calculating both predictor and response variables. In the CVR approach, the  $\lambda$  parameter was chosen to be the ratio between the average of the variances of the response variable and the average of the variances of the predictor variable for each data set.

## RESULTS AND DISCUSSION

**Variance associated to the response variable.** Table 1 shows the results of the variance of the response variable calculated using the BLS expressions at two randomly selected values for the six data sets described in the Experimental Section. It can be observed that all the results obtained from eq. 12 and eq. 13 coincide up to the eighth decimal place. Therefore, the two expressions, which were found independently, have to be considered equivalent.

**Reversibility of axes. Variance associated to the predictor and the response variables.** An interesting feature of the BLS regression technique is that it is invariant upon switching axes. OLS or WLS regression techniques do not have this feature, since only homoscedastic or heteroscedastic errors are taken into account on the  $y$  axis, and two different regression lines with different confidence intervals are obtained depending on which variable is placed on each axis. The CVR

approach, and its particular case of OR, are also invariant when the axes are switched.

| Data Set | $y_0$<br>Predicted | $s_{y_0}^2$<br>Equation 12 | $s_{y_0}^2$<br>Equation 13 | Differences<br>(%) |
|----------|--------------------|----------------------------|----------------------------|--------------------|
| 1        | 18.81              | 6.40655210                 | 6.40655210                 | 0.00               |
| 1        | 66.97              | 1.71674460                 | 1.71674460                 | 0.00               |
| 2        | 31.69              | 8.01060550                 | 8.01060550                 | 0.00               |
| 2        | 2.74               | 5.69862276                 | 5.69862276                 | 0.00               |
| 3        | 13.20              | 0.57148232                 | 0.57148232                 | 0.00               |
| 3        | 3.59               | 0.03850903                 | 0.03850903                 | 0.00               |
| 4        | 92.03              | 7.63379592                 | 7.63379592                 | 0.00               |
| 4        | 93.93              | 7.06714730                 | 7.06714730                 | 0.00               |
| 5        | 7.07               | 4.17691152                 | 4.17691152                 | 0.00               |
| 5        | 4.68               | 1.88624149                 | 1.88624149                 | 0.00               |
| 6        | 55.72              | 0.14971572                 | 0.14971572                 | 0.00               |
| 6        | 23.84              | 0.21554948                 | 0.21554948                 | 0.00               |

Table 1.- Comparison of expressions 12 and 13, for the calculation of the variance associated to the prediction of the response variable.

To check the reversibility of the axes, the variance corresponding to the response variable on the  $y$  axis (e.g., for a namely new method) was calculated for a fixed value of the predictor variable on the  $x$  axis (corresponding to a namely established method) using eqs. 12 or 13. Then, both methods were switched upon the axes and the variance of the variable on the  $x$  axis (the former so called new method) was calculated at the same value of the predictor variable (the established method on the  $y$  axis) using eq. 15. The process can be seen in Figure 3. The reversibility of the axes was tested for two random points in each of the six data sets studied. It can be observed in Table 2 that placing the methods on either of the two axes does not change the results for the variances of the predicted value. Table 2 also shows the agreement between the expressions for calculating the variance of the predictor and response variables, since the results are identical.

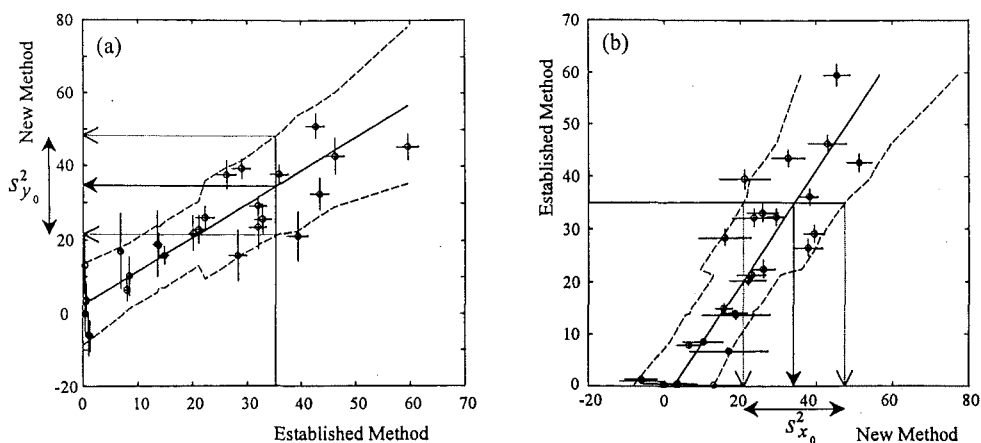


Figure 3 a) Variance associated to the predicted value of a new method (response variable) at a given value of an established method (predictor variable) b) Predicted values upon switching axes. In this case the calculated variance is associated to the predicted value of the new method (predictor variable), at a given value of the established method (response variable).

**Prediction of the variances taking into account errors in both axes.** The expressions which were derived to calculate the variance of the true mean associated to the predicted variables were validated by comparing the calculated variance values and those found by means of Monte Carlo simulations. The values obtained by means of the simulation process were also compared to the values obtained using the expressions for OLS, WLS, CVR and OR, and it was shown that errors may be significant if these techniques are used in situations in which there are heteroscedastic errors in both axes. The results obtained for the variance of the true mean of the response variable are shown in Table 3a, and the results for the variance of the true mean of the predictor variable are shown in Table 3b. In all the individual cases assayed (except two for the response variable and one for the predictor variable), the agreement between the simulated and predicted variances associated to the response and predictor variables obtained with BLS is significantly better than the agreement obtained with the other four methods. The variance for the response variable in data set 2 using the BLS expressions is

overestimated by up to 11% whereas the error obtained using the WLS expression was unusually low, 2.3%. The mean errors for the response and predictor variables found with BLS, WLS, OLS, CVR and OR are between 4-5%, 57-85%, 277-205%, 444-637% and 1697-462% respectively.

| Data Set | $x_0 / y_0$ | $s_{x_0}^2$         | $s_{y_0}^2$  | Differences (%) |
|----------|-------------|---------------------|--------------|-----------------|
|          | Predicted   | Equations 12 and 13 | Equation 15  |                 |
| 1        | 15.78       | 3.45781347          | 3.45781347   | 0.00            |
| 1        | 8.94        | 4.95883603          | 4.95883603   | 0.00            |
| 2        | 29.48       | 20.99017650         | 20.99017650  | 0.00            |
| 2        | 0.91        | 100.63858697        | 100.63858697 | 0.00            |
| 3        | 12.44       | 0.53514345          | 0.53514345   | 0.00            |
| 3        | 4.01        | 0.07533935          | 0.07533935   | 0.00            |
| 4        | 92.36       | 79.88020685         | 79.88020685  | 0.00            |
| 4        | 126.10      | 881.92789365        | 881.92789365 | 0.00            |
| 5        | 7.85        | 7.90043654          | 7.90043654   | 0.00            |
| 5        | 3.86        | 1.92173104          | 1.92173104   | 0.00            |
| 6        | 140.60      | 1.67904956          | 1.67904956   | 0.00            |
| 6        | 68.22       | 5.84275834          | 5.84275834   | 0.00            |

Table 2.- Comparison between the expressions for the variances of the predictor and response variables when these variables are switched upon the axes.

| $y_0$           | $s_{y_0}^2$ | $s_{y_0}^2$ | Error (%) | $s_{y_0}^2$ | Error (%) | $s_{y_0}^2$ | Error (%) | $s_{y_0}^2$ | Error (%) | $s_{y_0}^2$ | Error (%) |          |
|-----------------|-------------|-------------|-----------|-------------|-----------|-------------|-----------|-------------|-----------|-------------|-----------|----------|
| Pred.           | Simul.      | BLS         |           | OLS         |           | WLS         |           | CVR         |           | OR          |           |          |
| 1               | 18.81       | 6.9337      | 6.4066    | 7.60        | 0.8715    | 87.43       | 3.1740    | 54.22       | 1.9556    | 71.80       | 2.4708    | 64.37    |
| 1               | 6.97        | 1.7601      | 1.7167    | 2.47        | 0.6510    | 63.01       | 0.2547    | 85.53       | 1.7330    | 1.54        | 2.2459    | 27.60    |
| 2               | 31.69       | 8.0788      | 8.0106    | 0.84        | 2.9733    | 63.20       | 2.0940    | 74.08       | 5.0474    | 37.52       | 30.1071   | 272.67   |
| 2               | 2.74        | 6.4005      | 5.6986    | 10.97       | 5.5663    | 13.03       | 6.2511    | 2.33        | 7.6424    | 19.40       | 32.9836   | 415.33   |
| 3               | 13.20       | 0.5568      | 0.5715    | 2.64        | 0.7833    | 40.68       | 0.0366    | 93.43       | 1.1714    | 110.38      | 2.3364    | 319.61   |
| 3               | 3.59        | 0.0387      | 0.0385    | 0.52        | 1.0131    | 2517.83     | 0.0143    | 63.05       | 1.4007    | 3519.38     | 2.5659    | 6530.23  |
| 4               | 92.03       | 7.7837      | 7.6338    | 1.93        | 20.3506   | 161.45      | 5.0557    | 35.05       | 59.8349   | 668.72      | 976.4771  | 12445.15 |
| 4               | 93.93       | 7.3417      | 7.0671    | 3.74        | 17.0338   | 132.01      | 5.8768    | 19.95       | 56.3773   | 667.91      | 9.5010    | 29.41    |
| 5               | 7.07        | 4.4771      | 4.1769    | 6.71        | 0.0610    | 98.64       | 0.0289    | 99.35       | 0.4397    | 90.18       | 0.5412    | 87.91    |
| 5               | 4.68        | 2.0327      | 1.8862    | 7.21        | 0.0383    | 98.12       | 0.1176    | 94.21       | 0.4165    | 79.51       | 0.5178    | 74.53    |
| 6               | 55.72       | 0.1486      | 0.1497    | 0.74        | 0.1633    | 9.89        | 0.0832    | 44.01       | 0.2218    | 49.26       | 0.2873    | 93.34    |
| 6               | 23.84       | 0.2153      | 0.2186    | 1.53        | 0.1193    | 44.59       | 0.1622    | 24.66       | 0.1778    | 17.42       | 0.2433    | 13.01    |
| Mean error (%): |             |             | 3.91      |             | 277.49    |             | 57.49     |             | 444.42    |             | 1697.76   |          |

Table 3a.- Comparison of the variance values of the new method (response variable) calculated using equations 12 or 13 with the experimental values from the simulation process on the six real data sets and the results obtained using OLS, WLS, CVR and OR.

|                        | $x_0$  | $S_{x_0}^2$ | $S_{x_0}^2$ | Error       | $S_{x_0}^2$ | Error         | $S_{x_0}^2$ | Error        | $S_{x_0}^2$ | Error         | $S_{x_0}^2$ | Error         |
|------------------------|--------|-------------|-------------|-------------|-------------|---------------|-------------|--------------|-------------|---------------|-------------|---------------|
|                        | Pred.  | Simul.      | BLS         | (%)         | OLS         | (%)           | WLS         | (%)          | CVR         | (%)           | OR          | (%)           |
| 1                      | 15.78  | 3.6885      | 3.4578      | 6.25        | 0.7942      | 78.47         | 2.4092      | 34.68        | 3.6207      | 1.84          | 2.7991      | 24.11         |
| 1                      | 8.94   | 5.0595      | 4.9588      | 1.99        | 0.6060      | 88.02         | 0.6077      | 87.99        | 3.3658      | 33.48         | 2.6145      | 48.32         |
| 2                      | 29.48  | 21.3897     | 20.9902     | 1.87        | 4.2472      | 80.14         | 2.3623      | 88.96        | 87.7860     | 310.41        | 43.8907     | 105.20        |
| 2                      | 0.91   | 103.0027    | 100.6386    | 2.30        | 7.8508      | 92.38         | 7.1846      | 93.02        | 91.7295     | 10.94         | 47.1192     | 54.25         |
| 3                      | 12.44  | 0.5340      | 0.5351      | 0.21        | 0.9533      | 78.52         | 0.0289      | 94.59        | 4.3668      | 717.75        | 3.0540      | 471.91        |
| 3                      | 4.01   | 0.0740      | 0.0753      | 1.76        | 1.1956      | 1515.68       | 0.0148      | 80.00        | 4.6337      | 6161.76       | 3.2965      | 4354.73       |
| 4                      | 92.36  | 91.1230     | 77.2692     | 15.20       | 11.0476     | 87.88         | 5.2274      | 94.26        | 69.8071     | 23.39         | 4.5703      | 94.98         |
| 4                      | 126.10 | 994.3954    | 870.0950    | 12.50       | 6.7616      | 99.32         | 23.7078     | 97.62        | 69.0507     | 93.06         | 3.2380      | 99.67         |
| 5                      | 7.85   | 8.4438      | 7.9004      | 6.44        | 0.1242      | 98.53         | 0.4586      | 94.57        | 1.0325      | 87.77         | 0.9325      | 88.96         |
| 5                      | 3.86   | 2.0086      | 1.9217      | 4.33        | 0.0510      | 97.46         | 0.1108      | 94.48        | 1.0018      | 50.12         | 0.8615      | 57.11         |
| 6                      | 140.60 | 1.7145      | 1.6790      | 2.07        | 0.7621      | 55.55         | 0.3849      | 77.55        | 3.6430      | 112.48        | 3.3568      | 95.79         |
| 6                      | 68.22  | 5.8198      | 5.8428      | 0.40        | 0.5833      | 89.98         | 0.7755      | 86.67        | 3.4417      | 40.86         | 3.1780      | 45.39         |
| <b>Mean error (%):</b> |        |             |             | <b>4.61</b> |             | <b>205.16</b> |             | <b>85.37</b> |             | <b>636.99</b> |             | <b>461.70</b> |

Table 3b.- Comparison of the variance values of the reference method (predictor variable) calculated using equation 15 with the experimental values from the simulation process on the six real data sets and the results obtained using OLS, WLS, CVR and OR.

The lowest errors using the BLS expressions are obtained with data sets 3 and 6 that seem to present the best goodness of fit of the regression line to the experimental points, which confirms the assumption that the closeness of the experimental points to the regression line is an important factor for predicting the correct variances of the response and predictor variables. On the other hand, the errors are highest for data sets 2 and 4, the maximum being around 15% for the predictor variable. Data set 4 enables the behaviour of the BLS technique to be examined in presence of data sets with a low correlation between the variables, and with two possible outliers which have a very different degree of uncertainty at the extremes of the regression range. Since the BLS technique negatively weights the influence of points with high uncertainties, the point at the furthest extreme of the range relatively affects the value of the calculated regression coefficients. This feature is partially present in WLS but absent in the other methods because they do not take into account the individual uncertainties. Therefore, the resulting regression coefficients and associated variances of the five techniques are quite different, and again, the variances corresponding to the variables predicted using BLS are closer to the simulated results than the ones calculated using the other methods. Data sets 1 and 5 give errors ranging from 1% to 7% for the response and

predictor variables, although data set 1 is made up of seven experimental points and data set 5 of 30, which suggests that the number of experimental points does not significantly influence the correct estimation of the variances of the response and predictor variables.

## CONCLUSIONS

The new expressions for calculating the variance of the predicted values in the  $x$  and  $y$  axes taking into account heteroscedastic individual errors in both axes have been developed and validated by means of simulation studies on six real data sets. The possibility of deriving two predicting expressions that give identical results for the calculation of the variances of the  $y$  axis using BLS shows the reliability of the results obtained and the validity of the mathematical hypotheses which were assumed to find these expressions. The expressions developed are of a general nature and can be applied to predict values and associated uncertainties of any type, such as measurement results when using two different methods, analytical techniques, observers or laboratories.

BLS-based calculations can be made rapidly with an iterative process. The main limitation of this technique is that the uncertainties in both axes of each experimental point used in the regression analysis need to be known. However, this will probably not be unusual in the future since the international standards recommend that uncertainties be stated for every measurement result. Nevertheless, it is important to note that, in those cases where only the errors in one variable are considered, BLS gives results which are identical to those obtained using OLS or WLS regression techniques.

CVR and OR appear to produce acceptable results when the data structure meet their requirements, but as the individual errors are not taken into account, their results may be far from the real ones. It should be pointed out that despite the high

mean errors shown by CVR and OR methods in Tables 3a and 3b, these are mainly due to their application to data set 3. If this data set had not been taken into account, the results obtained using CVR and OR methods would have been more similar to those obtained using WLS and OLS expressions.

A feature of the BLS method is that it provides results that are invariant upon switching axes. This property is of practical significance since, in method comparison studies, for instance, it should be of no importance which of the axes is used to represent the method to be compared, as long as all the uncertainties are considered in both axes. Further studies are in progress based on the present results. Of particular interest may be the development of estimators for detection and quantitation limits.

## ACKNOWLEDGMENTS

The authors thank the Spanish Ministry of Education and Science (DGICYT project no. BP96-1008) for financial support.

## REFERENCES

- (1) Draper, N.; Smith, H. *Applied regression analysis*, 2nd ed.; John Wiley: New York, 1981; pp 8-70, 108-17.
- (2) Massart, D.L.; Vandeginste, B.G.M.; Buydens, L.M.C., de Jong, S., Lewis, P.J., Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997; pp 75-8.
- (3) Watters, R.L.; Carroll, R.J.; Spiegelman, C.H. *Anal. Chem.*, 1987, 59, 1639-43.
- (4) Rawlings, J.O. *Applied Regression Analysis*; Wadsworth & Brooks/Cole: Belmont, 1988; pp 315-8.

- (5) Fuller W.A. *Measurement Error Models*; John Wiley & Sons: New York, 1987; pp 1-5, 30-6, 74-9.
- (6) Anderson R.L. *Practical Statistics for Analytical Chemists*, Van Nostrand Reinhold: New York, 1987.
- (7) Creasy M.A. *J. Roy. Stat. Soc. B*, **1956**, *18*, 65-9.
- (8) Mandel J. *J. Qual. Tech.*, **1984**, *16*, 1-14.
- (9) Hartmann C.; Smeyers-Verbeke J.; Penninckx W.; Massart D.L. *Anal. Chim. Acta*, **1997**, *338*, 19-40.
- (10) Van Huffel S.; Vandewalle J. *The Total Least Squares Problems. Computational Aspects and Analysis*, Siam: Philadelphia, 1991; pp 1-18.
- (11) Lisý, J.M.; Cholvadová, A.; Kutej, J. *Computers Chem.*, **1990**, *14*, 189-92.
- (12) Riu, J.; Rius, F.X. *J. Chemom.*, **1995**, *9*, 343-62.
- (13) Riu, J.; Rius, F.X. *Anal. Chem.*, **1996**, *68*, 1851-7.
- (14) Meloun M.; Militký J.; Forina M. *Chemometrics for Analytical Chemistry. Volume 1: PC-aided statistical data analysis*, Ellis Horwood: Chichester, 1992; pp 61-8.
- (15) Martínez, A.; del Río, F. J.; Riu, J.; Rius, F.X. In preparation.
- (16) Kateman G.; Pijpers F.W. *Quality Control in Analytical Chemistry*, John Wiley & Sons: New York, 1981; pp 135-41.
- (17) Commission d'Établissement des Méthodes d'Analyses du Commissariat à l'Énergie Atomique (Cetama) *Statistique Appliquée a l'exploitation des Mesures*, Masson: Paris, 1986; pp 55-63.
- (18) Langenfeld, J.J.; Hawthorne, S.B.; Miller, D.J.; Pawliszyn, J. *Anal. Chem.*, **1994**, *66*, 909-16.
- (19) Ruisánchez, I.; Rius, A.; Larrechi, M.S.; Callao, M.P.; Rius, F.X. *Chemom. Intell. Lab. Syst.*, **1994**, *24*, 55-63.
- (20) Yellin, J. *Trends Anal. Chem.*, **1995**, *14*, 37-44.
- (21) López-Ávila, V.; Young, R.; Beckert, W.F. *Anal. Chem.*, **1994**, *66*, 1097-106.
- (22) Ripley, B.D.; Thompson, M. *Analyst*, **1987**, *112*, 377-83.

- (23) Ogren, P.J.; Norton, J.R. *J. Chem. Edu.*, **1992**, *69*, A130-1.
- (24) Mathworks Inc., Natick, Massachussets, USA.
- (25) Meier, P.C.; Zünd, R.E. *Statistical Methods in Analytical Chemistry*; John Wiley & Sons: New York, 1993; pp 145-50.

## Capítol 7

---

### Conclusions

## 7.1 Conclusions generals

El principal objectiu de la comparació de mètodes analítics emprant regressió lineal és comprovar si la sèrie de mètodes que es comparen produeixen resultats que no difereixen estadísticament entre ells a diversos nivells de concentració de l'analít que es vol determinar, considerant sempre que sigui possible les probabilitats d'error  $\alpha$  i  $\beta$  associades. Altres objectius són trobar el valor de la concentració i la incertesa associats a una mostra qualsevol analitzada per un nou mètode analític a partir dels resultats obtinguts amb un altre mètode de rutina, o detectar la presència d'errors sistemàtics proporcionals o constants.

És important tenir present que en la comparació de mètodes, el model de regressió construït a partir dels resultats obtinguts amb els mètodes analítics en comparació (línia recta si es comparen dos mètodes, hiperplà si es comparen més de dos mètodes) ha de elaborar-se tenint en compte els errors associats a tots els mètodes, ja que normalment aquests seran del mateix ordre de magnitud. En calibració lineal, on normalment es vol relacionar la concentració de l'analít que es vol determinar (variable predictora) amb la resposta instrumental proporcionada per una determinada tècnica analítica (variable resposta), tal com s'ha exposat a l'apartat 1.4 hi ha una sèrie de casos on no es poden negligir els errors associats a les dues variables.

En el model de regressió de línia recta, el mètode tradicionalment més emprat és el mètode OLS degut a les seves bones propietats matemàtiques i a la rapidesa en la obtenció dels coeficients de regressió i altres paràmetres relacionats. Però sota certes condicions (per exemple l'existència d'errors associats al mètode situat a l'eix de les  $x$  o d'errors no constants en el mètode situat a l'eix de les  $y$  en processos de comparació de mètodes analítics) OLS condueix a estimacions incorrectes dels coeficients de regressió i paràmetres relacionats com poden ser les

variàncies dels coeficients. El mètode WLS constitueix una millora respecte al mètode OLS, doncs ja té en compte la possible heteroscedasticitat al mètode situat a l'eix de les  $y$ . No obstant, en cas d'existència d'errors associats a la variable predictora també pot donar lloc a estimacions incorrectes dels coeficients de regressió donat que continua considerant que aquesta variable no té error. No obstant això, OLS o WLS poden ser aplicats perfectament en processos de comparació de mètodes on es tingui la certesa de que els errors proporcionats per un dels dos mètodes en comparació són molt més petits que els errors proporcionats per l'altre mètode. En aquest cas el mètode que dona menors errors hauria de situar-se a l'eix de les  $x$  i l'altre a l'eix de les  $y$ . Si el mètode situat a l'eix de les  $y$  proporciona errors homoscedàstics al llarg de tot l'interval de comparació, es podrà emprar el mètode OLS, mentre que si els errors són heteroscedàstics, s'haurà d'utilitzar el mètode WLS. En cas de que hi hagi errors presents en els dos mètodes en comparació, s'haurà d'utilitzar algun mètode que consideri els errors en els dos eixos. Si es disposa de la informació de que la relació entre els errors dels dos mètodes es constant al llarg de l'interval de comparació, es pot emprar l'aproximació CVR, on cal determinar el paràmetre  $\lambda$  (apartat 2.2), corresponent a la relació entre els errors dels dos mètodes. Un cas particular de l'aproximació CVR el constitueix el mètode OR, el qual es pot emprar quan els errors dels dos mètodes analítics són iguals per a cada punt experimental ( $\lambda=1$ ). Per últim, si hi ha present errors en els dos mètodes i no es pot assegurar que hi hagi una relació constant entre els errors dels dos mètodes al llarg de l'interval de comparació, s'hauria d'utilitzar un mètode que considerés els errors individuals en els dos mètodes analítics. Dins de tots els mètodes de regressió que consideren els errors individuals en els dos eixos, hem triat el mètode de Lisí i col·laboradors, anomenat també mètode BLS, ja que proporciona les estimacions correctes dels coeficients de la recta de regressió. Altres mètodes que consideren els errors en els dos eixos també arriben a les estimacions correctes dels coeficients de regressió (Taula 2, pàgina 86), però l'algorisme de càlcul emprat en el mètode BLS proporciona la

matriu variància-covariància dels coeficients de regressió, que és de gran utilitat pel posterior desenvolupament de tests estadístics associats.

Cal tenir present que en aquells casos on es cregui que l'estructura dels errors present a les dades és una quan en realitat n'és un altre (és a dir, que per exemple s'apliqui el mètode de regressió OR creient que la relació entre els errors dels dos mètodes és constant i igual a 1 en tots els punt, quan en realitat aquesta relació no és constant al llarg de l'interval de comparació, i per tant es tindria que haver aplicat per exemple el mètode BLS), la recta de regressió i les conclusions obtingudes amb l'aplicació d'algun test estadístic basat en algun dels seus coeficients poden estar lluny de les conclusions o resultats reals, tal com s'ha demostrat a bastament al llarg d'aquesta tesi doctoral.

En processos de calibració, la discussió seria completament anàleg, havent-se d'utilitzar el mètode BLS en aquells casos que presenten errors heteroscedàstics no constants tant a la variable predictora com a la variable resposta. Aquest procés es veu reflectit a l'esquema 7.1.

En la comparació dels resultats proporcionats per més de dos mètodes analítics, els coeficients de l'hiperplà de regressió s'han trobat mitjançant el mètode MLS, que és l'extensió del mètode BLS al camp multivariant, i que per tant considera els errors individuals associats a cada punt experimental de tots els mètodes analítics.

Una de les propietats importants dels mètodes BLS i MLS és que la seva aplicació a un conjunt de dades que no presenti errors en la variable o variables predictoros i els errors en la variable resposta siguin homoscedàstics, condueix als mateixos resultats que mitjançant l'aplicació dels mètodes OLS i MLR respectivament.

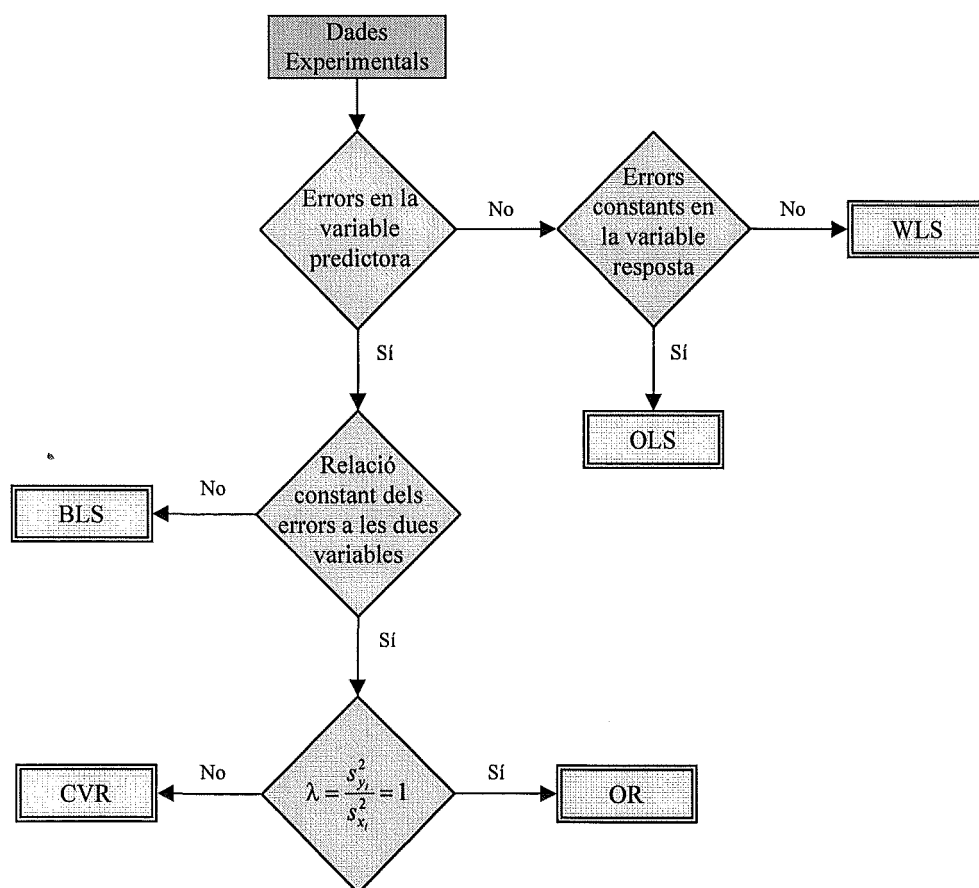
Tots els tests estadístics desenvolupats en aquesta tesi doctoral es basen en

l'assumpció de que la distribució dels coeficients de la recta i l'hiperplà de regressió segueixen la distribució normal. En la recta de regressió trobada amb el mètode BLS, els coeficients de regressió no solen seguir la distribució normal (Taula 1, pàgina 132), però els errors comesos acceptant aquesta hipòtesi són petits i en tot cas menors que emprant els mètodes OLS o WLS quan es tenen errors en els dos eixos. No s'ha estudiat encara la normalitat dels coeficients de l'hiperplà de regressió trobats amb el mètode MLS, però en analogia amb el mètode BLS es pot dir que els coeficients de regressió probablement tampoc seguiran la distribució normal. Però malgrat aquest fet, probablement la seva desviació no serà significativa, ja que els resultats obtinguts en l'etapa de validació dels processos de comparació de múltiples mètodes analítics assumint la normalitat en els coeficients de l'hiperplà de regressió concorden bastant bé amb els valors teòrics (Taula 6, pàgina 237).

Malgrat la complexitat matemàtica que pot suposar passar dels mètodes OLS o WLS a BLS en regressió univariant, o de MLR a MLS en regressió multivariant, el temps de càlcul necessari per tal de trobar els coeficients de la recta o hiperplà de regressió considerant els errors individuals en tots els eixos no és gaire elevat. Per exemple, el procés de càlcul per tal d'obtenir la representació gràfica del test conjunt per a l'ordenada a l'origen i el pendent basat en el mètode BLS per a diverses sèries de dades sol tardar entre 1 i 2 segons en un ordinador HP Vectra VE Pentium 75. En regressió multivariant, l'algorisme MLS programat sobre el mateix ordinador sol trobar els coeficients de l'hiperplà de regressió de diverses sèries de dades aplicades a la comparació d'entre 4 i 8 mètodes analítics en un temps inferior als 3 segons per a la majoria de conjunts assajats.

Una de les principals dificultats a l'hora d'aplicar els tests estadístics basats en les mètodes de regressió BLS o MLS és la necessitat de conèixer el valor de la variància individual de cada variable en cada un dels punts experimentals, ja que

les variàncies normalment impliquen repeticions i per tant més temps d'anàlisi i un cost econòmic superior. Cal recordar, però, que la sèrie de normes EN 45000/UNE 66500 especifiquen que cada resultat analític ha d'anar acompanyat del valor de la seva incertesa. De totes maneres, els analistes són, a la llarga, els qui poden escollir si volen ser més o menys restrictius a l'hora d'estimar o calcular les incerteses associades a cada resultat analític, tot i tenir clar que el fet de considerar adequadament les incerteses o no, pot arribar a fer que s'acceptin com a bons mètodes analítics esbiaixats o que es rebutgin mètodes analítics no esbiaixats, tal com s'ha pogut comprovar al capítol 4.



Esquema 7.1. Procés de trobada del mètode de regressió lineal adequat segons les característiques del conjunt de dades experimental.

## 7.2 Conclusions del capítol 3

En aquest capítol s'ha procedit a determinar la distribució dels coeficients de la recta de regressió calculada segons el mètode BLS. S'ha demostrat, per a una sèrie de conjunts de dades que abasten diverses situacions que es poden donar en processos de comparació de mètodes, que tot i obtenir-se distribucions no normals, es pot acceptar la hipòtesi de normalitat en els coeficients de la recta de regressió trobada segons el mètode BLS sense cometre un error significatiu. D'altra banda, s'han desenvolupat i validat les expressions per a la detecció d'un error sistemàtic constant o proporcional en processos de comparació de dos mètodes analítics (o per exemple per a la detecció d'efectes de matriu o correccions del blanc) considerant les probabilitats d'error  $\alpha$  i  $\beta$  associades i tenint en compte el biaix màxim que no es considera significatiu en el procés de comparació. També s'han desenvolupat les expressions (en forma de procediment iteratiu) pel càlcul *a priori* del número de punts que ha de tenir la recta de regressió per tal d'obtenir els coeficients de regressió amb probabilitats prefixades d'errors  $\alpha$  i  $\beta$ .

Un aspecte posterior a considerar seria l'avaluació del biaix màxim que no es considera significatiu, ja que en la detecció d'errors sistemàtics o constants aquest biaix ve donat en termes d'ordenada a l'origen o de pendent, i pot arribar a ser difícil la seva translació a unitats de concentració.

## 7.3 Conclusions del capítol 4

En aquest capítol s'ha desenvolupat el test conjunt per a l'ordenada a l'origen i el pendent de la recta de regressió trobada segons el mètode BLS (equació 11, pàgina 169) per tal de comparar els resultats de dos mètodes analítics. S'ha comparat aquest test conjunt basat en el mètode BLS amb els tests conjunts basats en els

mètodes OLS i WLS, així com amb altres tests estadístics, i s'ha demostrat que el fet d'ignorar els errors associats als dos mètodes analítics en comparació a l'hora d'establir la recta de regressió pot donar lloc a conclusions incorrectes: es poden acceptar com a bons, mètodes incorrectes, i es poden rebutjar bons mètodes. La seva validació mitjançant el procés de simulació de Monte Carlo ha donat lloc a resultats que concorden molt aproximadament amb els teòrics.

Pel que fa al programa desenvolupat pel càlcul i representació del test conjunt per a l'ordenada a l'origen i el pendent en Matlab 4.0 per a Windows 3.1 o superior, permet el càlcul i la visualització del test conjunt en pocs segons (normalment entre 1 i 2) escollint diverses probabilitats d'error  $\alpha$  i pels tests conjunts basats en els mètodes OLS, WLS i BLS.

#### 7.4 Conclusions del capítol 5

En aquest capítol s'ha desenvolupat un test per a la comparació simultània dels resultats de múltiples mètodes analítics a diversos nivells de concentració: el test conjunt per a l'ordenada i la suma de pendents de l'hiperplà de regressió. Aquest test està basat en el mètode de regressió MLS, que és l'ampliació del mètode BLS al camp multivariant. Aquest test també ha estat validat emprant el mètode de simulació de Monte Carlo, i els resultats també coincideixen en un grau elevat amb els valors teòrics.

Els resultats del procés de comparació de múltiples mètodes analítics per a la majoria de la sèrie de conjunts de dades comprovats s'obtenen en un període curt de temps.

## 7.5 Conclusions del capítol 6

En aquest capítol s'han desenvolupat els intervals de confiança associats a la predicció de la variable predictora donat un valor de la variable resposta o viceversa en regressió lineal considerant els errors individuals en dos eixos. Aquests intervals de confiança, apart de ser útils en processos de comparació de mètodes, poden aplicar-se a altres camps com per exemple la datació per radiocarboni o l'assignació d'òrgens a peces arqueològiques desconegudes. Les variàncies de la variable predictora o resposta, peça clau en la construcció dels seus intervals de confiança, han estat trobats seguint dos camins independents que condueixen a resultats idèntics, el que constitueix una validació interna. A més, aquestes expressions han estat també validades mitjançant el mètode de simulació de Monte Carlo, assolint-se resultats que concorden aproximadament amb els teòrics, i molt millors que els intervals de confiança obtinguts amb l'aplicació dels mètodes de regressió OLS, WLS, OR o CVR quan hi ha presents errors heteroscedàstics individuals en dos eixos.

## 7.6 Perspectives de futur

Com a perspectiva de futur, i dins d'un objectiu global que consistiria en millorar la qualitat de la informació proporcionada pels mètodes que consideren errors en ambdós eixos, la recerca es pot enfocar cap al desenvolupament de tècniques de regressió robusta que consideressin els errors heteroscedàstics individuals en tots els eixos. Una altra alternativa consistiria en el desenvolupament de tècniques de detecció de punts discrepants per ambdós mètodes, així com de tests estadístics per tal de comprovar el bon ajust dels punts experimentals a la recta o hiperplà de regressió. Aquest punt és important perquè una recta o hiperplà de regressió que presenti algun punt discrepant o un mal ajust dels seus punts experimentals, donarà

com a resultat un increment en el valor de l'error experimental i en conseqüència de tots els intervals de confiança dels coeficients de regressió o paràmetres derivats, essent aquest increment només degut a la falta d'ajust.

Fins a la data, el test conjunt per a l'ordenada a l'origen i el pendent considerant errors en dos eixos s'ha desenvolupat i validat per a la comparació dels resultats de la determinació d'un sol analit per dos mètodes diferents. Però hi ha tot un seguit de mètodes analítics que poden proporcionar informació simultània de diversos analits en una sola mostra (per exemple les tècniques cromatogràfiques). Per tant, una extensió del test conjunt per a l'ordenada a l'origen i el pendent considerant errors en dos eixos seria l'estudi de la seva aplicació per tal de comparar els resultats proporcionats per dos mètodes de simultàniament més d'un analit.

De la mateixa manera que s'ha introduït el càlcul de les probabilitats d'error  $\beta$  en els tests individuals per a l'ordenada a l'origen i el pendent, un següent pas consistiria en el càlcul de les probabilitats d'error  $\beta$  associades al test conjunt per a l'ordenada a l'origen i el pendent basat en el mètode BLS i al test conjunt per a l'ordenada a l'origen i la suma de pendents basat en el mètode MLS, és a dir, a la quantificació de les probabilitats d'acceptar com a bons, mètodes esbiaixats en un estudi de comparació de mètodes analítics. Dins del mètode de regressió MLS també caldria estudiar els coeficients de l'hiperplà de regressió per tal de comprovar si la seva distribució és normal, o en cas que no ho sigui, si s'allunya significativament de la normalitat, així com també fer l'algorisme de trobada dels coeficients de l'hiperplà regressió menys sensible al punt inicial escollit per tal de poder executar amb més fiabilitat el procediment iteratiu que condueix a la seva obtenció.

Altres actuacions dins de la calibració univariant consistirien en el desenvolupament de límits de decisió, detecció i quantificació considerant els

errors en ambdós eixos, i en el desenvolupament de tècniques de regressió no lineal, útil en casos com la datació per radiocarboni de materials arqueològics mitjançant mesura per centelleig líquid, on la relació entre concentració i resposta es sol ajustar a un polinomi, normalment de tercer grau.

Per últim, i dins de la calibració multivariant, una futura actuació consistiria en intentar desenvolupar un nou mètode de calibració multivariant basat en la descomposició per components principals que tingués en compte els errors individuals en tots els eixos: primer dur a terme una descomposició de les dades inicials segons la tècnica MLPCA i llavors aplicar el mètode MLS a les dades descompostes.



Expressions emprades en la comprovació de la normalitat de les distribucions de l'ordenada a l'origen i el pendent mitjançant el mètode de Cetama

### Coefficients generals

$$k_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = s_x^2 \quad (\text{A.1})$$

$$k_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n (x_i - \bar{x})^3 \quad (\text{A.2})$$

$$k_4 = \frac{n(n+1) \sum_{i=1}^n (x_i - \bar{x})^4 - 3(n-1) \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}{(n-1)(n-2)(n-3)} \quad (\text{A.3})$$

### *Corba simètrica limitada pel domini $\bar{x} \pm d_1$*

La distribució segueix una corba simètrica limitada pel domini  $(\bar{x} - d_1, \bar{x} + d_1)$ , amb el coeficient  $d_1$  calculat segons:

$$d_1 = \sqrt{\frac{2b_2 s_x^2}{3 - b_2}} \quad (\text{A.4})$$

El paràmetre  $m_1$  de l'equació 3.12 ve definit per:

$$m_1 = \frac{5b_2 - 9}{2(3 - b_2)} \quad (\text{A.5})$$

i  $f_0$  de la mateixa equació 3.12 és una constant tal que:

---

 Apèndix
 

---

$$\int_{\bar{x}-d_1}^{\bar{x}+d_1} f(x) dx = 1 \quad (\text{A.6})$$

***Corba simètrica il·limitada pels dos costats ( $-\infty, +\infty$ )***

Els paràmetres  $m_2$  i  $m_3$  de l'equació 3.13 venen definits segons:

$$m_2 = \sqrt{\frac{2b_2 s_x^2}{b_2 - 3}} \quad (\text{A.7})$$

$$m_3 = \frac{5b_2 - 9}{2(b_2 - 3)} \quad (\text{A.8})$$

i  $f_0$  és una constant tal que:

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (\text{A.9})$$

***Corba no simètrica limitada als dos costats ( $d_2 \dots d_3$ )***

Els paràmetres  $d_2$  i  $d_3$  es calculen segons les següents expressions:

$$d_2 = \bar{x} - \frac{w \cdot q_1}{r_1} \quad (\text{A10})$$

$$d_3 = w + d_2 \quad (\text{A.11})$$

amb els paràmetres  $r_1$ ,  $w$  i  $q_1$  que venen donats per:

$$r_1 = \frac{6(b_2 - b_1 - 1)}{3b_1 - 2b_2 + 6} \quad (\text{A.12})$$

$$w = 2\sqrt{s_x^2(1-k)(1+r_1)} \quad (\text{A.13})$$

$$q_1 = \frac{r_1}{2} \left( 1 - \sqrt{\frac{-k}{1-k}} \right) \quad (\text{A.14})$$

El paràmetre  $q_2$  de l'equació 3.15 ve expressat segons:

$$q_2 = r_1 - q_1 \quad (\text{A.15})$$

i  $f_0$  de l'equació 3.15 és una constant tal que:

$$\int_{a_2}^{a_3} f(x) dx = 1 \quad (\text{A.16})$$

El valor màxim (moda) de la distribució, correspon a:

$$m = \bar{x} + \frac{w \cdot (q_2 - q_1)}{r_1(r_1 - 2)} \quad (\text{A.17})$$

*Corba no simètrica il·limitada als dos costats  $(-\infty, +\infty)$*

Els diferents paràmetres de l'equació 3.16 es troben definits segons:

$$r_2 = \frac{6(b_2 - b_1 - 1)}{2b_2 - 3b_1 - 6} \quad (\text{A.18})$$

$$q = 1 + \frac{r_2}{2} \quad (\text{A.19})$$

## Apèndix

$$p = r_2 \sqrt{\frac{k}{1-k}} \quad (\text{A.20})$$

$$v = \sqrt{s_x^2 (r_2 - 1)(1 - k)} \quad (\text{A.21})$$

$$c_1 = \bar{x} - \frac{p \cdot v}{r_2} \quad (\text{A.22})$$

i  $f_0$  és una constant tal que:

$$f_0 = \int_{-\infty}^{\infty} f(x) dx = 1 \quad (\text{A.23})$$

La moda d'aquest tipus de distribucions correspon a:

$$m = \bar{x} + \frac{2p \cdot v}{r_2(r_2 + 2)} \quad (\text{A.24})$$

*Corba no simètrica limitada a un costat ( $d_4 \dots + \infty$ )*

La corba es troba limitada per l'esquerra segons el següent paràmetre  $d_4$ :

$$d_4 = \bar{x} - c_2 \frac{q_3}{r_2} \quad (\text{A.25})$$

i la distribució definida a l'equació 3.17 presenta la següent moda:

$$m = \bar{x} - \frac{c_2(q_3 + q_4)}{r_2(r_2 + 2)} \quad (\text{A.26})$$

Els nous coeficients de les equacions A.25–A.26 es defineixen segons:

$$q_3 = \frac{r_2}{2} \left[ \sqrt{\frac{k}{k-1}} - 1 \right] \quad (\text{A.27})$$

$$q_4 = q_3 + r_2 \quad (\text{A.28})$$

$$c_2 = 2\sqrt{s_x^2(r_2 - 1)(k - 1)} \quad (\text{A.29})$$

i  $f_0$  de l'equació 3.17 és una constant tal que:

$$\int_{d_4}^{\infty} f(x) dx = 1 \quad (\text{A.30})$$

*Corba no simètrica limitada a un costat (-∞ ...  $d_5$ )*

La corba es troba limitada per la dreta segons el següent paràmetre  $d_5$ :

$$d_5 = \bar{x} + c_2 \frac{q_3}{r_2} \quad (\text{A.31})$$

El paràmetre  $f_0$  de l'equació 3.18 és una constant tal que:

$$\int_{-\infty}^{d_5} f(x) dx = 1 \quad (\text{A.32})$$

i la distribució presenta la següent moda:

$$m = \bar{x} + \frac{c_2(q_3 + q_4)}{r_2(r_2 + 2)} \quad (\text{A.33})$$

## GLOSSARI

|        |   |
|--------|---|
| APM    | Mètode paramètric aproximat<br><i>Approximate parametric method</i>   |
| BLS    | Mínims quadrats bivariants<br><i>Bivariate least squares</i>  |
| CVR    | Relació constant de variàncies<br><i>Constant variance ratio</i>  |
| EPM    | Mètode paramètric exacte<br><i>Exact parametric method</i>  |
| GLS    | Mínims quadrats generalitzats<br><i>Generalized least squares</i>   |
| IANOVA | Anàlisi d'informació de la variància<br><i>Informational analysis of variance</i>                                       |
| ILF    | Funció lineal implícita<br><i>Implicit linear function</i>  |
| IRWLS  | Mínims quadrats iterativament ponderats<br><i>Iteratively reweighted least squares</i>                                  |
| LSM    | Mínims quadrats de la mediana<br><i>Least median squares</i>  |
| MLLRR  | Regressió per arrel latent de màxima versemblança<br><i>Maximum likelihood latent root regression</i>                   |
| MLPCA  | Anàlisi per components principals de màxima versemblança<br><i>Maximum likelihood principal component analysis</i>      |
| MLPCR  | Regressió per components principals de màxima versemblança<br><i>Maximum likelihood principal components regression</i> |
| MLR    | Regressió lineal múltiple<br><i>Multiple linear regression</i>  |
| MLS    | Mínims quadrats multivariants<br><i>Multivariate least squares</i>  |
| ODR    | Regressió de la distància ortogonal<br><i>Orthogonal distance regression</i>  |
| OLS    | Mínims quadrats<br><i>Ordinary least squares</i>  |
| OR     | Regressió ortogonal<br><i>Orthogonal regression</i>   |
| PCA    | Anàlisi per components principals<br><i>Principal components analysis</i>   |
| PCR    | Regressió per components principals<br><i>Principal components regression</i>   |
| PLS    | Regressió per mínims quadrats parcials<br><i>Partial least squares</i>  |

Glossari

---

|     |  |
|-----|--|
| TLS | Mínims quadrats totals<br><i>Total least squares</i>       |
| WLS | Mínims quadrats ponderats<br><i>Weighted least squares</i> |