BIOPHARMACEUTICS & DRUG DISPOSITION
Biopharm. Drug Dispos. **21**: 7–14 (2000)

Non-linear Regression Analysis with Errors in Both Variables: Estimation of Co-operative Binding Parameters

Georgia Valsami[a], Athanassios Iliadis[b,*] and Panos Macheras[a]
[a] *School of Pharmacy, University of Athens, Athens, Greece*
[b] *Department of Pharmacokinetics, University of Mediterranee, Marseilles, France*

ABSTRACT: Four different parameter estimation criteria, the geometric mean functional relationship (GMFR), the maximum likelihood (ML), the perpendicular least-squares (PLS) and the non-linear weighted least squares (WLS), were used to fit a model to the observed data when both regression variables were subject to error. Performances of these criteria were evaluated by fitting the co-operative drug–protein binding Hill model on simulated data containing errors in both variables. Six types of data were simulated with known variances. Comparison of the criteria was done by evaluating the bias, the relative standard deviation (S.D.) and the root-mean-squared error (RMSE), between estimated and true parameter values. Results show that (1) for data with correlated errors, all criteria perform poorly; in particular, the GMFR and ML criteria. For data with uncorrelated errors, all criteria perform equally well with regard to the RMSE. (2) Use of GMFR and ML lead to lower values for S.D. but higher biases compared with WLS and PLS. (3) WLS performs less well when equal dispersion is applied to the two observed variables. Copyright © 2000 John Wiley & Sons, Ltd.

Key words: co-operative drug–protein binding model; errors-in-variables model; estimation criteria; non-linear regression

Introduction

In various fields of pharmaceutical sciences it is not uncommon to deal with problems associated with non-linear regression analysis where both observed variables are subject to experimental error. In such situations, the regression model is known as the errors-in-variables model [1]. Typical examples are the Scatchard and Hill equations in binding studies [2,3] and the pharmacokinetic–pharmacodynamic relationships, e.g. E_{max} model, which is based on pharmacodynamic readings and plasma drug concentration measurements [4,5]. In these studies, data are analysed by regression analysis using the standard least-squares criterion, assuming that one variable is not subject to error and may be considered as the independent variable. However, when both variables are subject to experimental error, parameter estimates derived from approaches utilizing this criterion cannot be considered valid.

In statistics literature, when both observed variables are subject to error, the geometric mean functional regression criterion (GMFR) [6] has been widely applied in fitting straight lines, e.g. in fishery studies and for the estimation of growth rates [7,8]. More recently, GMFR has been applied to the analysis of drug–protein binding data, according to the Scatchard model for one class of binding sites [9]. The errors in both variables in protein binding studies usually originate from the routine, but naive use of the difference between the total and the measured bound concentration to compute the unbound concentration. Undoubtedly, the limitations of

* Correspondence to: Department of Pharmacokinetics, Faculty of Pharmacy, University of Mediterranee, 27 Blvd. Jean Moulin, 13385 Marseilles Cedex 5, France. E-mail: iliadis@pharmacie.univ-mrs.fr

Received 4 December 1999
Accepted 17 May 2000

this approach have been realised in literature and other methods that use total concentration as the independent variable have been reported [10,11].

Consideration has also been given to the errors in both variables in fitting straight lines using the perpendicular least-squares (PLS) criterion [12,13]. However, there have been sporadic applications of non-linear regression analysis with errors in both variables [1,6,14,15]. This study was undertaken in order to evaluate the performances of three different criteria, the GMFR, the PLS and the maximum likelihood (ML) criterion, to the analysis of data that are currently analysed with the least-squares criterion in spite of the fact that both variables are subject to error. To this end, the weighted forms of the three criteria were applied to co-operative drug–protein binding data using simulated data generated from the Hill model, to which error was added to both variables. For comparative purposes, the weighted least-squares (WLS) criterion was also applied.

Estimation Criteria and the Binding Model

The GMFR criterion has been recently discussed [6,9,16]. The estimates derived from the GMFR approach minimize a cost function, based on the sum of the triangular areas formed by connecting the experimental data points to the regression line, with lines parallel to the co-ordinate axes. Although the areas summed and mininized when GMFR is applied in non-linear regression are not exact triangles [6], triangular areas were considered to be formed in order to simplify the iterative process.

The PLS criterion has been also discussed recently [15]. In this criterion, the sum of squared perpendicular distances, which represent the shortest distance between the data points and the predicted curve describing the relationship between the two variables, is used as an objective function for minimization in non-linear regression.

The ML criterion is derived from the classical ML principle [17,18] when several variables are observed. From a geometrical point of view, ML minimizes the area of ellipses tangential to the predicted curve and centred at the observed data. The ML criterion is described in Appendix A.

For the WLS criterion, the sum of squared vertical distances is used as the objective function for parameter optimization [1,17].

The co-operative drug–protein interaction based on the site-oriented binding model [19,20] is described by the following equation:

$$r = \frac{n \cdot K \cdot F^h}{1 + K \cdot F^h} \tag{1}$$

where r is the ratio of the molar concentration of the bound drug divided by the total protein molar concentration, and F is the molar concentration of the free drug. In most of the protein binding studies focusing on the estimation of binding parameters, the experimenter endeavours to set the total protein concentration and the total drug concentrations at certain targeted values. This model is termed the controlled errors-in-variables model [1]. Then, r and F are observed m times with a measurement error. The unknown parameters of the Hill model are n, the number of binding sites per protein molecule; K, the binding association constant; and h, the Hill exponent. These parameters are to be estimated from the m observed $\{F_j^\bullet \quad r_j^\bullet\}$, $j = 1, m$ pairs, using the previously presented criteria.

Typical values for parameters n, K and h are 1, $5 \cdot 10^6$ M^{-1}, and 2, respectively. They lead to the observed pairs $F_j^\bullet$ and $r_j^\bullet$ ranging from $2 \cdot 10^{-6}$ to $4 \cdot 10^{-3}$ M, and from 0 to 1 or 2, respectively. Due to the problems associated with estimation of parameters of very different magnitudes, the observed variable F was rescaled by $F = a \cdot \bar{F}$. According to this scaling, the regression will be between $\bar{F}_j^\bullet$ and $r_j^\bullet$ pairs without distortion of the statistical characteristics of the measurement error. The resulting Hill model is of the same structure as Equation (1) but it embodies a new parameter to be estimated, $\bar{K} = a^h \cdot K$.

Simulation Study Design

To compare the four criteria and assess their performances in the parameter estimation

problem, a simulation study was undertaken. To this end, theoretical parameter values were set to the Hill equation, $n_0 = 1$, $K_0 = 5 \cdot 10^6$ M^{-1}, and $h_0 = 2$. Using these values and after specifying the range of F values ($2 \cdot 10^{-6}$ to $4 \cdot 10^{-3}$ M), $m = 20$ values for $\{F_j \quad r_j\}$ $j = 1, m$ pairs were computed. F_j values were logarithmically equally spaced (i.e. $\log F_j$ were linearly equally spaced) in the specified range. The scaling factor was set to $a = 10^{-3}$ and consequently, $\bar{K}_0 = 5$ M^{-1}.

These errorless data were disturbed by heteroscedastic two-dimensional measurement errors having various statistical characteristics: uniform (U) or Gaussian (G) zero-mean random errors associated with equal (E) or not equal (H) dispersion on the two components; furthermore, for the G case, we designed correlated (C) or uncorrelated (R) components. Combination of all these cases led to six types of measurement errors, UE, UH, GER, GHR, GEC and GHC. The coefficients of variation for the heteroscedastic measurement error on F and r were w/q and w, respectively. The coefficient w was set to 10% and the E and H cases were obtained with $q = 1$ and $q = 4$ (16-fold higher dispersion on the observed variable r than on F), respectively. The correlation coefficient was set to $c = -0.8$ as a compromise between the rather unusual strong correlation case and the uncorrelated case. It is known that the bound drug is calculated as the difference between total drug and free drug concentrations. Consequently, a negative value for the correlation coefficient was used, as a positive error in the measured free drug concentration will result in a negative error for the bound drug. Appendix B gives full details of the error generating processes.

Comparison of U and G models will allow analysis of sensitivity with respect to the error model distribution. Comparison of E and H will reveal the influence of the relative magnitude of dispersion in variables while comparison of R and C cases in G models will clarify the importance of correlation between variables in the drug–protein binding experiments.

A total of $N = 500$ sets of data were generated for each type of error added. Finally, the Hill model was fitted to all sets of data using successively the four criteria GMFR, PLS, ML and WLS. Optimization programmes used the Nelder–Mead simplex algorithm. Simulations and optimizations were done under the MATLAB 5.2 programming environment [21]. The number of call functions and the run time were also recorded in order to evaluate technical performances.

After parameter values, $x_i^{(k)}$ $i = 1, N$ (x may be one of n, $\bar{K}$ and h), were estimated using GMFR, ML, PLS and WLS criteria ($k = 1$, 2, 3 and 4, respectively), the difference between the theoretical x_0 and the estimated values $x_i^{(k)}$ were calculated as a measure of the validity of the approaches. The results of this simulation study were evaluated statistically. The root-mean-square relative error (RMSE) was calculated expressing the spread of the estimated values around the theoretical ones. Then, RMSE was split into its two components, namely the relative standard deviation (S.D.), which evaluates the spread of $x_i^{(k)}$ around their average value $x_\bullet^{(k)}$, and the relative bias (bias), which provides the difference between x_0 and $x_\bullet^{(k)}$. Furthermore, the ratios [S.D./RMSE] and [bias/RMSE] were computed in order to express the contribution of S.D. and bias to the total variability. The sum of these ratios equals 1 as they are complementary indexes. All these measures are presented in Appendix C and they are expressed as percentages (%).

Results

Figure 1 illustrates the fit of the Hill model to a dataset corresponding to the GER case utilizing the four criteria. WLS and PLS criteria led to estimates that predicted very similar model behaviours. Table 1 contains the estimated model parameters. Whereas h is correctly estimated by the four criteria, $\bar{K}$ is underestimated by ML and overestimated by WLS and PLS. One can also observe that estimations of n obtained by ML and GMFR are closer to the theoretical value.

For the various types of the measurement error and for each model parameter, Figures 2–4 show the associated RMSE, [S.D./RMSE] and [bias/RMSE] values, respectively.

1. According to Figure 2, the values of RMSE are strongly dependent on whether or not the

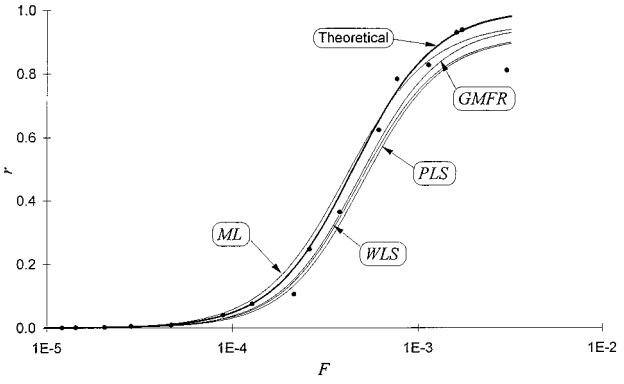

Figure 1. The fit of the Hill model to a dataset generated under the GER noise conditions using the GMFR, PLS, ML and WLS criteria (thin lines). The theoretical curve (thick line) and the observed data (●) are also presented. For the sake of clarity, the four data points with $F_j^{\bullet} < 10^{-5}$ M are not shown

Table 1. Example of model parameters estimated by the four fitting criteria

	n	$\bar{K}$	h
GMFR	0.951	5.183	2.026
ML	0.957	4.120	1.953
PLS	0.915	7.050	2.07
WLS	0.918	6.318	2.051

The dataset was generated under the GER noise conditions (see text for full details) with $n = 1$, $\bar{K} = 5\ \mathrm{M}^{-1}$ and $h = 2$.

observed variables are correlated. When correlation exists, high values for RMSE were observed and all criteria performed unsatisfactorily. Due to the extremely high RMSE values for $\bar{K}$, the ML and GMFR criteria are not recommended. Thus, subsequent comments concern only the uncorrelated types of the measurement error.

2. All criteria perform equally on RMSE for the U and G error distribution models; the RMSE scores were similar (Figure 2). The highest RMSE was found for $\bar{K}$ while the smallest for h. As expected the WLS criterion showed the poorest performance when the random error was equally dispersed on the two observed variables (Figure 2).
3. The GMFR and ML criteria performed better than WLS and PLS for both parameters $\bar{K}$ and h with regards to the ratio [S.D./RMSE] (Figure 3). However, GMFR is the best criterion for n (Figure 3).

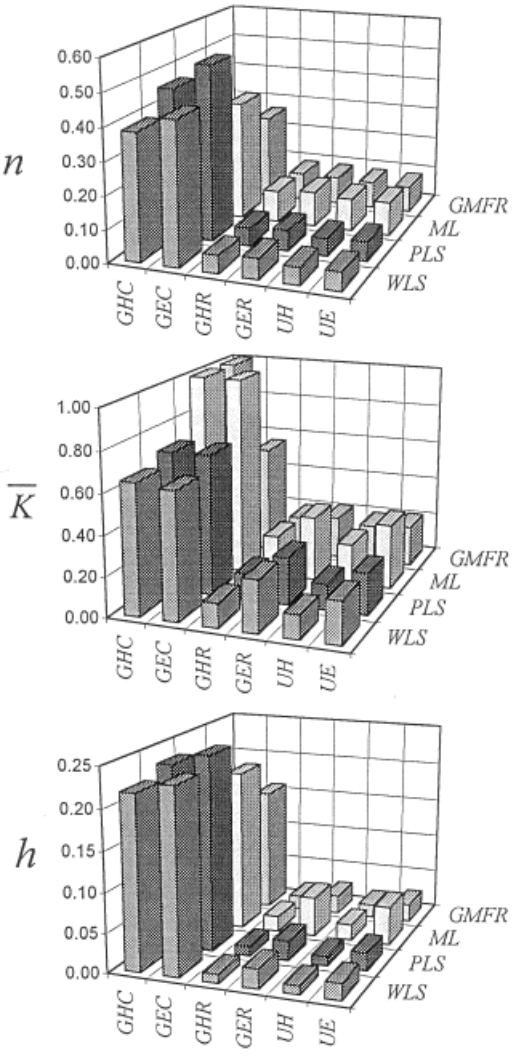

Figure 2. Bar charts for RMSE computed for the model parameters

4. On the contrary, GMFR and ML lead to more biased estimates than WLS and PLS (Figure 4). In addition, the criteria of GMFR and ML seem to be particularly influenced by the relative magnitude of the random error dispersion between the observed variables when estimating $\bar{K}$ and h.

All fits were carried out using the Nelder–Mead simplex algorithm for the non-linear regression procedure. The final parameter estimates did not depend on their initial guesses. Starting with [2 1 1] as arbitrary guess

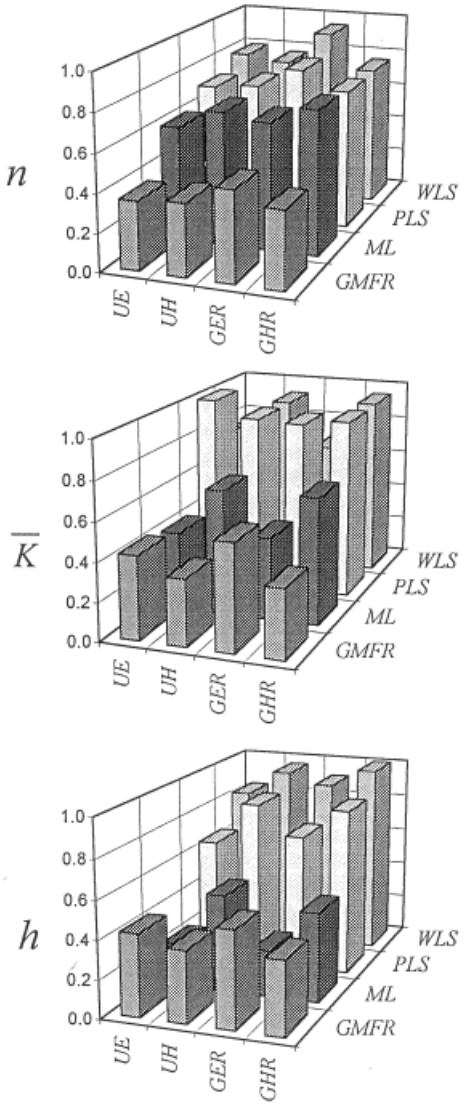

Figure 3. Bar charts for [S.D./RMSE] computed for the model parameters

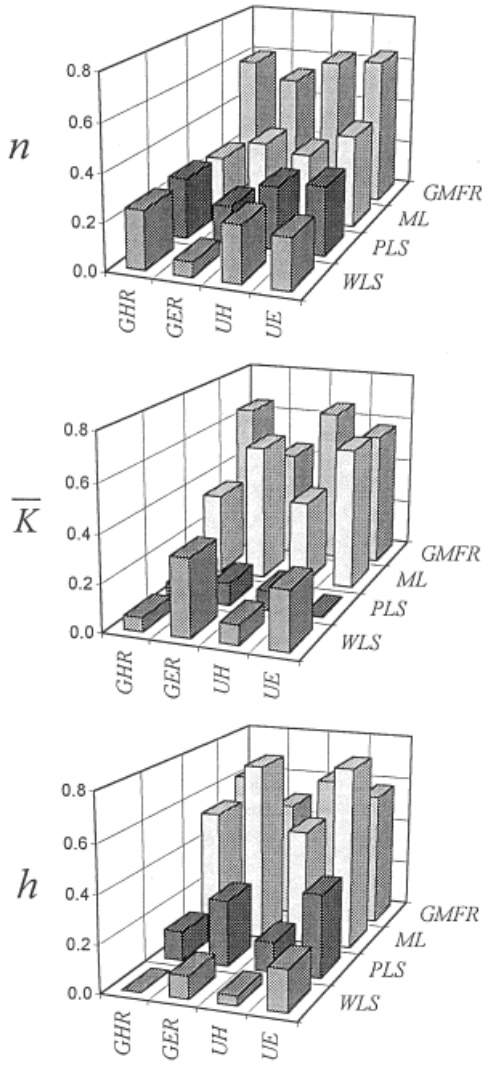

Figure 4. Bar charts for [bias/RMSE] computed for the model parameters

values, the algorithm converges quickly to the theoretical values [1 5 2] after 180–250 criterion function calls. Nevertheless, the computing time is 80-fold longer for the PLS criterion when compared with the other criteria. This is due to the fact that the PLS criterion involves a numerical bisection method to find the shortest distance between the fitted curve and the data points [15].

Discussion

Special procedures are used in the literature for fitting purposes in those situations where both variables are subject to error [1,9]. It is well known that in such errors-in-variables models, bias and lack of precision may lead to incorrect parameter estimations when the optimization criterion is not well adapted to the statistical context. Nevertheless, bias as a component of RMSE is less of a problem. Known bias can be easily removed [22], while unknown bias, as in

the case of real data, can be reduced by using a resampling procedure [23].

The present study shows that the performances of the four fitting criteria are strongly dependent on the type of measurement error. It was found that the modelling of data led to unacceptable parameter estimates when correlation existed between the errors on both variables. In such situations, it is advisable to use more sophisticated criteria, e.g. the method of scoring or others [1].

For data with uncorrelated errors, neither the PLS nor the GMFR criterion provided clearly better estimates for the model parameters when compared with the estimates derived from ML and WLS. When the errors associated with the two variables differ considerably, the GMFR criterion performs equally well with the other criteria.

A particular behaviour was noted for the estimates of n when compared with the $\bar{K}$ and h estimates; n was the less sensitive parameter in respect of the estimation criterion and the type of simulated measurement error. Indeed, n occupies a linear position in the model and it merits a special consideration in the estimation procedure [18,24].

In practice, when real data are analysed, the scaling factor a should be selected close to the maximum value of F. After parameter estimation with this value, it may be further refined by locally tuning a in order to keep the same order of magnitude for all scaled parameters.

A major problem with the errors-in-variables models is that in addition to the main parameters of interest, there are also incidental parameters relating to the distribution of the random regressors (e.g. the mean values of the regressors). This creates complications, such as inconsistent ML estimates or variance–covariance matrices that are not given by the inverse of the expected information matrix, or other relevant measures. One way of overcoming these problems is to sufficiently replicate the experiment. With replication, the expected information matrix can be evaluated and the confidence intervals computed [1]. Since the design of protein binding experiment follows what we call a controlled errors-in-variables model, the targeted F values appear as the control variables of confidence intervals. On the other hand, D-optimal designs are intended to reduce the size of confidence intervals by selecting the appropriate control variables. Using D-optimal design techniques, one may calculate an optimal distribution of F values in the experimental range (other than the logarithmically or linear spaced values) leading to more precise parameter estimates. Therefore, in order to further improve precision of parameter estimates, replication of experiments and the use of D-optimal designs are recommended [25].

Conclusions

Despite the apparent simplicity of the Hill model for co-operative binding processes, the analysis of simulated data with added noise revealed several surprises, which are due to the error components in both observed variables, r and F. To avoid unreliable parameter estimates, one must first rescale the model by a simple linear transformation on observed F data. Further, reconsider the estimation procedure by eliminating the linear parameter n in the non-linear regression [18,24].

Extreme caution should be exercised when the observed variables are highly correlated. If correlation exists, the ML and GMFR criteria are not recommended. For uncorrelated data, GMFR or ML may be used. The fitting criteria studied do not seem to be sensitive to the error distribution model. Finally, the WLS criterion is sensitive to the relative magnitude of error dispersion on the observed variables.

Appendix A

The Maximum Likelihood Criterion

The statistical model is the following:

$$F_j^{\bullet} = F_j + e_{F_j}$$
$$r_j^{\bullet} = r(F_j, \underline{x}) + e_{r_j} = r_j + e_{r_j} \quad \text{(A1)}$$

with $j = 1, m$ and where $\underline{x}$ is the vector of model parameters. Errors e_{F_j} and e_{r_j} are considered as components of the two-dimensional random measurement vector $\underline{e}_j$ i.e. $\underline{e}_j^T = [e_{F_j} \quad e_{r_j}]^T$, and

they are dependent on the values of model parameters $\underline{x}$, i.e. $\underline{e}_j(\underline{x})$

For the observed sample j, it is usually assumed that $\underline{e}_j(\underline{x})$ can be adequately described by a zero-mean Gaussian distribution with covariance matrix V_j. For a heteroscedastic model of error variance, V_j may be factored as

$$V_j = Y_j \cdot W \cdot Y_j, \quad \text{with } Y_j = \begin{bmatrix} F_j & 0 \\ 0 & r_j \end{bmatrix}$$

and W a symmetric definite-positive matrix. If the $\underline{e}_j(\underline{x})$ sequence is independent, the ML estimation principle specifies the structure of the W matrix

$$W(\underline{x}) = \frac{1}{m} \cdot \sum_{j=1}^{m} Y_j^{-1} \cdot \underline{e}_j(\underline{x}) \cdot \underline{e}_j^T(\underline{x}) \cdot Y_j^{-1}$$

and leads to the minimization with respect to $\underline{x}$ of the following criterion function:

$$J = -\sum_{j=1}^{m} \ln|Y_j| - \frac{m}{2} \cdot \ln|W(\underline{x})|$$

In the above expressions, the superscripts -1 and T denote matrix inversion and transposition, respectively, and $|\cdot|$ denotes the matrix determinant [17,18].

Appendix B

Simulation of Measurement Errors

For the simulation of various types of measurement errors, $\underline{e}_j$ was assumed to be distributed according to a uniform (U) or Gaussian (G) probability density function with zero-mean and covariance matrix V_j.

(1) For a U distribution with uncorrelated components

$$V_j = \frac{1}{12} \begin{bmatrix} a_j^2 & 0 \\ 0 & b_j^2 \end{bmatrix}$$

where a_j and b_j are the range widths of the e_{F_j} and e_{r_j} errors, respectively. For the heteroscedastic model of error variance, we have

$$\frac{a_j^2}{12} = \frac{w^2}{q^2} \cdot F_j^2 \quad \text{and} \quad \frac{b_j^2}{12} = w^2 \cdot r_j^2$$

or

$$a_j = \frac{w}{q} \cdot F_j \cdot \sqrt{12} \quad \text{and} \quad b_j = w_j \cdot r_j \cdot \sqrt{12} \tag{B1}$$

Thus, giving the theoretical $\{F_j \quad r_j\}$ $j = 1, m$ pairs, we computed first a_j and b_j from (B1) and then we generated uniform distributed random errors e_{F_j} and e_{r_j} from the ranges $[-a_j/2 \quad +a_j/2]$ and $[-b_j/2 \quad +b_j/2]$, respectively [21]. Observed pairs were obtained from (A1).

(2) For a G distribution

$$V_j = \begin{bmatrix} v_{F_j}^2 & v_{Fr_j} \\ v_{Fr_j} & v_{r_j}^2 \end{bmatrix}$$

where $v_{F_j}^2$ and $v_{r_j}^2$ are the variances, and v_{Fr_j} the covariance of e_{F_j} and e_{r_j}. For the heteroscedastic model of error variance, we have

$$v_{F_j}^2 = \frac{w^2}{q^2} \cdot F_j^2,$$

$$v_{r_j}^2 = w^2 \cdot r_j^2 \quad \text{and} \quad v_{Fr_j} = c \cdot v_{F_j} \cdot v_{r_j} \tag{B2}$$

where c is the correlation coefficient. Thus, giving the theoretical $\{F_j \quad r_j\}$ $j = 1, m$ pairs, we computed first $v_{F_j}^2$, $v_{r_j}^2$ and v_{Fr_j} from Equation (B2) and then we generated zero-mean Gaussian random errors e_{F_j} and e_{r_j} with covariance matrix V_j [21].

Appendix C

Statistical Indexes

Statistical assessment of the results was performed by evaluation of the k criterion and the parameter x

1. the RMSE

$$[\text{RMSE}^{(k)}]^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} \left[\frac{x_i^{(k)} - x_0}{x_0} \right]^2$$

2. the relative S.D.

$$[\text{S.D.}^{(k)}]^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} \left[\frac{x_i^{(k)} - x_\bullet^{(k)}}{x_0} \right]^2$$

3. the relative bias

$$\text{bias}^{(k)} = \frac{x_\bullet^{(k)} - x_0}{x_0}$$

These indexes are related by

$$[\text{RMSE}^{(k)}]^2 = [\text{S.D.}^{(k)}]^2 + [\text{bias}^{(k)}]^2$$

References

1. Seber GAF, Wild CJ. *Nonlinear Regression Analysis*. Wiley: New York, 1989; 491–527.
2. Hill AV. The combination of haemoglobin with oxygen and with carbon monoxide. *Biochem J* 1913; **7**: 471–480.
3. Scatchard G. The attraction of proteins for small molecules and ions. *Ann NY Acad Sci* 1949; **51**: 660–672.
4. Ponto LLB, Schoenwald RD. Furosemide (frusemide): a pharmacokinetic/pharmacodynamic review (part I). *Clin Pharmacokin* 1990; **18**: 381–408.
5. Ponto LLB, Schoenwald RD. Furosemide (frusemide): a pharmacokinetic/pharmacodynamic review (part II). *Clin Pharmacokin* 1990; **18**: 460–471.
6. Ebert T, Russel M. Allometry and model II non-linear regression. *J Theor Biol* 1994; **168**: 367–372.
7. Jolicoeur P. Linear regressions in fisheries research: some comments. *J Fish Res Board Can* 1975; **32**: 1491–1494.
8. Sprent P, Dolby GR. The geometric mean functional relationship. *Biometrics* 1980; **36**: 547–550.
9. Valsami G, Macheras P. The geometric mean functional relationship approach to linear regression in pharmaceutical studies: application to the estimation of binding parameters. *Pharm Sci* 1995; **1**: 551–557.
10. Feldman HA. Mathematical theory of complex ligand-binding systems at equilibrium. *Anal Biochem* 1972; **48**: 317–338.
11. Munson PJ, Rodbard D. LIGAND: a versatile computerized approach for characterization of ligand binding systems. *Anal Biochem* 1980; **107**: 220–239.
12. Brace RA. Fitting straight lines to experimental data. *Am J Physiol* 1977; **233**: R94–R96.
13. Riggs DS, Guarnieri JA, Addelman S. Fitting straight lines when both variables are subject to error. *Life Sci* 1978; **22**: 1305–1360.
14. Prior RL, Rosenthal HE. A statistical method for the estimation of binding parameters in a complex system. *Anal Biochem* 1976; **70**: 231–240.
15. Ko HQ, Jusko WJ, Ebling WF. Nonlinear perpendicular least-squares regression in pharmacodynamics. *Biopharm Drug Dispos* 1997; **18**: 711–716.
16. Barker F, Soh YC, Evans RJ. Properties of the geometric mean functional relationship. *Biometrics* 1988; **44**: 279–281.
17. Bard Y. *Nonlinear Parameter Estimation*. Academic Press: New York, 1974; 54–71.
18. Ba BB, Iliadis A, Cano JP. Pharmacokinetic modeling of ethylloflazepate (Victan) and its main active metabolites. *Ann Biomed Eng* 1989; **17**: 633–646.
19. Molyneux P, Cornarakis-Lentzos M. The interaction of polyvinylpyrrolidone with aromatic compounds in aqueous solution. Part IV. Evaluation of the co-operativity parameter, and the methylene-group contribution to the binding strength, for the alkyl-parahydroxybenzoates. *Colloid Polymer Sci* 1979; **257**: 855–873.
20. Sideris E, Valsami G, Koupparis K, Macheras P. Studies on the interaction of diflunisal with cyclodextrins using ion-selective electrode potentiometry. *Eur J Pharm Sci* 1999; **7**: 271–278.
21. MATLAB. *High-Performance Numeric Computation and Visualization Software* (5.2). The Math Works: Natick, MA, 1998.
22. Candy JV. *Signal Processing. The Model-based Approach*. McGraw-Hill: Singapore, 1987; 47–66.
23. Efron B. *The Jackknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics: Philadelphia, PA, 1982; 5–10.
24. Lawton WH, Sylvestre EA. Elimination of linear parameters in nonlinear regression. *Technometrics* 1971; **13**: 461–467.
25. Atkinson AC, Donev AN. *Optimum Experimental Designs*. Clarendon Press: Oxford, 1992; 116–133.