Доклади на Българската академия на науките
Comptes rendus de l'Académie bulgare des Sciences

Tome 66, No 1, 2013

MATHEMATIQUES

Modèles mathématiques

STATISTICAL PROPERTIES OF THE LEAST SQUARE ESTIMATORS FOR DRUG-PROTEIN BINDING

Krassimira Prodanova

(*Submitted by Academician P. Kenderov on July 5, 2012*)

Abstract

In the present paper the statistical properties of the estimated parameters for drug protein binding are discussed. An approximation of the covariance matrix and confidence intervals for these parameters are obtained.

Key words: nonlinear regression, drug protein binding, statistical properties of the estimators

2000 Mathematics Subject Classification: 62H15, 62P10

1. Introduction. In [1] a new mathematical procedure for analysis of data for studying drug protein binding obtained by Circular Dichroism Titration Method (CDTM) was reported. This minimization algorithm for estimating the number of classes of binding sites and the association constants is based on regression with constrains over the parameters. The procedure estimates also the concentration of an unbound drug, which is unobservable. In addition, the procedure allows an easy and natural way to choose the number of classes of binding sites based on F-criteria and residual sums of squares.

CDTM depends on the phenomenon that the binding of a protein (serum albumin) can induce optical activity in the molecule. When a drug binds to a protein molecule the molar ellipticity change $\Delta\theta$ is obtained at wavelengths where the drug has absorption bands and $\Delta\theta$ is proportional to the concentration of drug-protein complexes ($D_t \cup P_T$), where P_T is the total protein concentration and D_t is the total drug concentration [2–5].

The proposed method in [1] is used for analyzing data for drug protein binding in several publications [2,3,6–10].

In the present paper the statistical properties of some of the estimated parameters given in [1] are discussed. An approximation of the covariance matrix and confidence intervals for these parameters are obtained.

The paper is organized as follows. In Section 2 the mathematical model and the results obtained in [1] are presented. Section 3 contains the new results about statistical errors of estimated parameters. In Section 4 these results are applied to experimental data given in [1] and [5].

2. Mathematical model. Molar ellipticity change is proportional to the concentration of drug – protein complexes. If there are m classes of binding sites of protein molecules then

$$\Delta\theta = \sum_{i=1}^{m} \Delta\theta_i \tag{1}$$

where $\Delta\theta_i$ is the concentration of the i-th class of binding sites in the total molar ellipticity change. For the i-th class we have the presentation

$$\Delta\theta_i = e_i D_i, \tag{2}$$

where D_i is the complex concentration of the i-th class of binding and e_i is the corresponding proportionality coefficient. For the concentration D_i in the literature [1–5], the following representation is cited:

$$D_i = \frac{k_i N_i D_f}{1 + k_i D_f} P_T. \tag{3}$$

Here D_f is the concentration of unbound drug, N_i is the number of binding sites for the i-th class of binding sites, P_T is the total protein concentration, k_i is the association constant for the i-th class. So, each class is characterized by the parameters $\{e_i, k_i, N_i\}$.

If D_t is the total drug concentration then

$$D_t = \sum_{i=1}^{m} D_i + D_f. \tag{4}$$

The obtained data $\Delta\theta$ and D_t by CDTM are observable, but values of D_f are not observable. In [1], the following approach for analysis is proposed:

Eq. (3) is presented in the form

$$D_i = N_i P_T \frac{k_i P_T \dfrac{D_f}{P_T}}{1 + k_i P_T \dfrac{D_f}{P_T}} = N_i P_T \varphi\left(\tilde{k}_i u\right), \tag{5}$$

where $\tilde{k}_i = k_i P_T$, $u = \dfrac{D_f}{P_T}$ and $\varphi(z) = \dfrac{z}{1+z}$. Substituting (5) in (2) and (4) model equations, which described $\Delta\theta$ and D_t as a function of D_f are presented by the system

$$\left| \begin{array}{l} x = \dfrac{\Delta\theta}{P_T} = \displaystyle\sum_{i=1}^{m} e_i N_i \varphi\left(\tilde{k}_i u\right), \\ y = \dfrac{D_t}{P_T} = \displaystyle\sum_{i=1}^{m} N_i \varphi\left(\tilde{k}_i u\right) + u. \end{array} \right. \tag{6}$$

The assumptions for model (6) are: both measuring points $(\Delta\theta)_j$ and $(D_t)_j$ $(j = 1, \ldots, n)$ are not exact, because of some objective reasons of CDTM and the measurements have equal relative errors, i.e. the errors of x_j and y_j will be equal (P_T is a known constant for CDTM). In (6) x and y are expressed by the function φ.

In order to estimate the errors of e_i, k_i, N_i $(i = 1, 2, \ldots, m)$, u_j $(j = 1, 2, \ldots, n)$, the model in (6) is reformulated as a stochastic one and thus makes possible the use of the a priori information about the error either caused by measurements or by different other factors. Let the observable data x_j, y_j, $(j = 1, 2, \ldots, n)$ satisfy the following statistical model:

$$\left| \begin{array}{l} x = \displaystyle\sum_{i=1}^{m} e_i N_i \varphi\left(\tilde{k}_i u\right) + \varepsilon_j = x\left(u_j\right) + \varepsilon_j, \\ y = \displaystyle\sum_{i=1}^{m} N_i \varphi\left(\tilde{k}_i u\right) + u_j + \delta_j = y\left(u_j\right) + \delta_j, \quad j = 1, 2, \ldots, n, \end{array} \right. \tag{7}$$

where ε_j, δ_j are independent, identically distributed normal random variables. According to the theory of maximum likelihood the least square method was chosen to estimate:

- the parameters e_i, k_i, N_i $(i = 1, 2, \ldots, m)$ and
- the concentration of unbound drug u_j $(j = 1, 2, \ldots, n)$, corresponding to the point (j), which is unobservable,

i.e. the function

$$\Phi\left(e_i, k_i, N_i, u_j\right) = \sum_{j=1}^{n} \left(x_j - x(u_j)\right)^2 + \left(y_j - y(u_j)\right)^2 \tag{8}$$

is minimized.

In [1], for minimizing (8), the programme CONSTR (Simplex method with constraints) from the package MATLAB [13] was used. As N_i are integers, the minimization of (8) was made for all possible combination of N_i. The procedure of minimization of (8) was repeated with different values of N_i in order to estimate

the number of binding classes of binding sites. After finding the m numbers, we calculate the minimum

$$S_i = \min_{e_i,k_i,N_i,u_j} \Phi\left(e_i, k_i, N_i, u_j\right), \quad i = 1,2,\ldots,m;\ j = 1,2,\ldots,n. \tag{9}$$

The statistical F-criteria for answering the question: "How many classes of binding sites i,$(i = 1,2,\ldots,m)$ are best fitted by the experimental data" is applied. For this purpose the following sequences of statistical hypothesis are tested:

$(H_0)_r$: The model with (r) classes of binding is not essentially better than the model with $(r-1)$ classes of binding $(r = m, m-1, m-2, \ldots, 2)$.

The statistics in [11,12]

$$F_r = \frac{(S_{r-1} - S_r)(n-2r)}{S_r\,(2r - 2(r-1))} \tag{10}$$

is compared with the corresponding value $F_{1-\alpha}(\nu_1,\nu_2)$ of F-distribution with $\nu_1 = (2r-2(r-1))$ and $\nu_2 = (n-2r)$ degree of freedom and probability $(1-\alpha) = 0.95$. The $(H_0)_r$ is rejected if $F_r > F_{1-\alpha}(2,\ n-2r)$. The testing is stopped at the first rejected hypothesis $(H_0)_r$ and the model with (r) classes is chosen. Now, let us consider the statistical properties of the triple $\left\{\hat{e}_i,\ \hat{k}_i,\ \hat{u}_j\right\}$.

3. Statistical properties of the estimators and approximate confidence intervals for the model parameters. Let the model with $\hat{m} = r$ classes of binding and numbers $\hat{N}_i = N_i$, $(i = 1,2,\ldots,r)$ of binding cites in i-th class are fixed.

Let us denote by Θ the vector which components are e_i, k_i, u_j

$$\Theta = (e_1,\ldots,e_r,\ k_1,\ldots,k_r,\ u_1,\ldots,u_n)^T, \tag{11}$$

where $(^T)$ means transpose of the matrix. Vector Θ has $p = 2r + n$ components. Employing the least square method a maximum likelihood estimator $\hat{\Theta} = \hat{\Theta}\,(x_1,\ldots,x_n,y_1,\ldots,y_n)$ of the parameter Θ is obtained. As it is well known [11,12], the maximum likelihood estimator has asymptotic normal distribution, i.e. $\hat{\Theta} \in N\,(\Theta,\Omega)$. Now we have to find an estimator $\hat{\Omega}$ of the covariance matrix Ω. For this purpose, let us present the stochastic model from (7) in the form

$$\left|\begin{array}{l} x_j = f_j(\Theta) + \varepsilon_j, \\ y_j = g_j(\Theta) + \delta_j,\ j = 1,2,\ldots,n, \end{array}\right. \tag{12}$$

where

$$\begin{aligned} f_j(\Theta) &= \sum_{i=1}^{r} e_j N_j \varphi\left(\tilde{k}_i u_j\right), \\ g_j(\Theta) &= \sum_{i=1}^{r} \left[N_i \varphi\left(\tilde{k}_i u_j\right)\right] + u_j. \end{aligned} \tag{13}$$

Let Θ_0 is some value of Θ. To examine the statistical properties of the estimated parameters we shall use an appropriate linearization of the functions $f(\Theta)$, $g(\Theta)$. For this purpose, the model functions (13) are expanded by Taylor series in the vicinity of Θ_0, neglecting the terms of second and higher order

$$f_j(\Theta) \approx f_j(\Theta_0) + [\underset{\Theta}{\text{grad}}\, f_j(\Theta_0)]^T (\Theta - \Theta_0)$$
$$g_j(\Theta) \approx g_j(\Theta_0) + [\underset{\Theta}{\text{grad}}\, g_j(\Theta_0)]^T (\Theta - \Theta_0), j = 1, 2, \ldots, n.$$

The first r components of $[\underset{\Theta}{\text{grad}}\, f_j(\Theta_0)]$ are $\dfrac{\partial f_j}{\partial e_i} = N_i \varphi\left(\tilde{k}_i u_j\right)$, $i = 1, 2, \ldots, r$, the second r are $\dfrac{\partial f_j}{\partial k_i} = N_i \dfrac{e_i u_j P_T}{\left(1 + \tilde{k}_i u_j\right)^2}$, $i = 1, 2, \ldots, r$ and, the last n components are $\dfrac{\partial f_j}{\partial u_l} = \sum\limits_{i=1}^{r} N_i \dfrac{e_i \tilde{k}_i}{\left(1 + \tilde{k}_i u_j\right)^2}$ if $l = j$ and $\dfrac{\partial f_j}{\partial u_l} = 0$, if $l \neq j$.

The first r components of $[\underset{\Theta}{\text{grad}}\, g_j(\Theta_0)]$ are $\dfrac{\partial g_j}{\partial e_i} = 0$, $i = 1, 2, \ldots, r$, the second r are $\dfrac{\partial g_j}{\partial k_i} = N_i \dfrac{u_j P_T}{\left(1 + \tilde{k}_i u_j\right)^2}$, $i = 1, 2, \ldots, r$ and, the last n components are $\dfrac{\partial g_j}{\partial u_l} = \sum\limits_{i=1}^{r} \left[N_i \dfrac{\tilde{k}_i}{\left(1 + \tilde{k}_i u_j\right)^2} \right] + 1$, if $l = j$ and $\dfrac{\partial g_j}{\partial u_l} = 0$, if $l \neq j$.

Now, the stochastic model from (12) takes the form

$$\begin{aligned} f_j(\Theta) &\approx f_j(\Theta_0) + [\underset{\Theta}{\text{grad}}\, f_j(\Theta_0)]^T (\Theta - \Theta_0) + \varepsilon_j, \\ g_j(\Theta) &\approx g_j(\Theta_0) + [\underset{\Theta}{\text{grad}}\, g_j(\Theta_0)]^T (\Theta - \Theta_0) + \delta_j, \quad j = 1, 2, \ldots, n. \end{aligned} \tag{14}$$

Now the function given in (8), which we minimize, takes the form

$$\Phi(\Theta) = \sum_{j=1}^{n} (x_j - f_j(\Theta))^2 + (y_j - g_j(\Theta))^2. \tag{15}$$

Let us denote the vectors

$$\vec{f}(\Theta) = (f_1(\Theta), \ldots, f_n(\Theta))^T,\ \vec{g}(\Theta) = (g_1(\Theta), \ldots, g_n(\Theta))^T,\ \vec{u} = (u_1, \ldots, u_n)^T.$$

Let us write the function $\Phi(\Theta)$ in the form

$$\Phi(\Theta) = \left\| \vec{f}(\Theta) - X \right\|^2 + \|\vec{g}(\Theta) - Y\|^2.$$

Now, from (14) we get

$$\Phi(\Theta) = \left\| \vec{f}(\Theta) + [\underset{\Theta}{\operatorname{grad}}\, \vec{f}(\Theta_0)]^T (\Theta - \Theta_0) - X \right\|^2$$
$$= \left\| \vec{g}(\Theta) + [\underset{\Theta}{\operatorname{grad}}\, \vec{g}(\Theta_0)]^T (\Theta - \Theta_0) - Y \right\|^2$$
$$= \left\| \begin{pmatrix} \vec{f}(\Theta_0) \\ \vec{g}(\Theta_0) \end{pmatrix} + (FG)^T (\Theta - \Theta_0) - \begin{pmatrix} X \\ Y \end{pmatrix} \right\|^2,$$

where $\begin{pmatrix} \vec{f}(\Theta_0) \\ \vec{g}(\Theta_0) \end{pmatrix}$ and $\begin{pmatrix} X \\ Y \end{pmatrix}$ are vectors with $2n$ components, $(FG) = [\underset{\Theta}{\operatorname{grad}}\, \vec{f}(\Theta_0)][\underset{\Theta}{\operatorname{grad}}\, \vec{g}(\Theta_0)]$ is $(p \times 2n)$ – matrix and $(\Theta - \Theta_0)$ is a vector with p components. After the minimization of $\Phi(\Theta)$ we get the estimator $\hat{\Theta}$ of Θ. Let us denote its $p = 2r+n$ components by $\hat{\Theta}_1 = \hat{e}_1, \ldots, \hat{\Theta}_{r+1} = \hat{k}_1, \ldots, \hat{\Theta}_p = \hat{u}_n$. According to the theory of multiple regression models [11, 12] the random vector $\left(\hat{\Theta} - \Theta\right)$ has approximately normal distribution

$$\left(\hat{\Theta} - \Theta\right) \in N\left(0, \operatorname{cov}\left(\hat{\Theta} - \Theta\right)\right),$$

with the covariance matrix $\operatorname{cov}\left(\hat{\Theta} - \Theta\right) = \hat{\sigma}^2 \left[(FG)(FG)^T\right]^{-1}$, $([\ldots]^{-1}$ – inverse matrix), where $\hat{\sigma}^2 = \dfrac{\text{REE}}{2n-p}$ and

$$\text{REE} = \Phi\left(\hat{\Theta}\right) = \sum_{j=1}^{n} \left(\left(x_j - f_j(\hat{\Theta})\right)^2 + \left(y_j - g_j(\hat{\Theta})\right)^2 \right). \tag{16}$$

REE is the minimum of the function given in (15).

Let us denote the elements (i,j) of the matrix $\left[(FG)(FG)^T\right]^{-1}$ by c_{ij}, $(i, j = 1, 2, \ldots, p)$. According to the theory [11, 12]:

- the statistic $\dfrac{\hat{\Theta}_i - \Theta}{\hat{\sigma}\sqrt{c_{ii}}}$, $(i = 1, 2, \ldots, p)$ has approximately t-distribution with $\nu = 2n - p$ degrees of freedom.
- approximate $(1-\alpha)100\%$ confidence interval for the component Θ_i, $(i = 1, 2, \ldots, p)$ of Θ is

$$\left(\hat{\Theta}_i - \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{ii}}; \hat{\Theta}_i + \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{ii}}\right), \tag{17}$$

where $t_{1-\frac{\alpha}{2}}(\nu)$ is the value of the $\left(1 - \dfrac{\alpha}{2}\right)$-th quintile of t-distribution with $\nu = 2n - p$ degree of freedom. From (11) and (17) it follows that the corresponding

approximate $(1-\alpha)100\%$ confidence intervals for the model parameters are

$$\left(\hat{e}_i - \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{ii}};\quad \hat{e}_i + \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{ii}}\right),\quad i = 1, 2, \ldots, r;$$

$$\left(\hat{k}_i - \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{r+i,r+i}};\quad \hat{k}_i - \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{r+i,r+i}}\right),\quad i = 1, 2, \ldots, r;$$

$$\left(\hat{u}_i - \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{2r+i,2r+i}};\quad \hat{u}_i - \hat{\sigma} t_{1-\frac{\alpha}{2}}(\nu)\sqrt{c_{2r+i,2r+i}}\right),\quad i = 1, 2, \ldots, n.$$

4. Application to the experimental data. In [1], the proposed procedure for estimation of the model parameters was applied to the experimental data for binding of the drug BROMDIAZEPOXIDE [5] to human serum albumin. The experimental points $x_j = (\Delta\theta/P_T)_j$, $y_j = (D_t/P_T)_j$, $j = 1, 2, \ldots, 15$ and $P_T = 9.5 \times 10^{-6}$ were fitted according to the equations (6) for $m = 1, 2, 3$, i.e. the one-, two- and three-model classes are compared by F-criteria. The most appropriate model was found to be one with two classes of binding. The following values for the estimators of the parameters $\hat{e}_i$, $\hat{k}_i$, $\hat{N}_i$, $(i = 1, 2)$ were reported in [1]: $\hat{e}_1 = 1.107273$, $\hat{e}_2 = 0.157495$, $\hat{k}_1 = 2.269809$, $\hat{k}_2 = 0.1415498$ and $\hat{N}_1 = 1$, $\hat{N}_2 = 3$. The value of the sum of residuals (see (16)) was reported to be REE = 0.00436. From [1] we shall give only the first and the last estimates of unbound drug concentration: $\hat{u}_1 = 0.0764$, $\hat{u}_{15} = 3.3688$.

According to the theory given in the previous section, following approximately 95% confidence intervals for the model parameters $\hat{e}_i$, $\hat{k}_i$, $\hat{u}_j$ are obtained:

- $\hat{e}_1 : (0.868492; 1.346053)$;
- $\hat{e}_2 : (0.078367; 0.393357)$;
- $\hat{k}_1 : (1.640533, 2.897646)$;
- $\hat{k}_2 : (0.130775; 0.160221)$;
- $\hat{u}_1 : (0.030315; 2.455956)$ and $\hat{u}_{15} : (0.728405; 5.021654)$.

It is obvious that the confidence intervals of the estimated parameters are large, that is to be expected with such a small number of observations ($n = 15$). The calculations of the covariance matrix and confidence limits are made by MATLAB [13].

5. Conclusions. As a result of this investigation we obtain the approximately confidence intervals of estimated parameters $\hat{e}_i$, $\hat{k}_i$ $(i = 1, 2, \ldots, r)$, where r is an estimator of the number of binding classes. It is of great importance to estimate the confidence limits of the concentration of unbound drug $\hat{u}_j$ (which is unobservable) because only unbound drug cures. The estimation of the parameters of these models can be used to considerably improve the drug efficiency.

REFERENCES

[1] Vandev D., K. Prodanova, V. Russeva. Arzneim.-Forsch./Drug Res., **48**, 1998, No 12, 1190–1193.

[2] Zsila F. In: Electronic Circular Dichroism Spectroscopy, New York, J. & Wiley, 2010, 1–61.

[3] Burton M., M. Leslie, J. Shentag, W. Evans. In: Applied pharmacokinetics & pharmacodynamics, New York, Lipp. Will. & Wilk., 2006, 1–114.

[4] Chignell C. F. Molec. Pharmac., **5**, 1969, No 2, 244–252.

[5] Roosdorp N., I. Sjoholm. Biochem. Pharmacol., **25**, 1976, No 19, 2141–2145.

[6] Russeva V., Z. Zhivkova, K. Prodanova, R. Rakovska. J. of Pharmacy and Pharmacol., **51**, 1999, No 1, 49–52.

[7] Prodanova K. Proc. XXIV Wokshop "Applications of Math. in Eng. and Economics", Sofia, Heron Press, 1999, 10–12.

[8] Liu F. K. J. of Chin. Chemical Society, **49**, 2002, No 4, 611–618.

[9] Hage D. S. J. of Cromat., **768**, 2002, No 1, 3–30.

[10] Clarke W., D. S. Hage. Separat. and Purficat. Rewiews, **32**, 2003, No 1, 19–60.

[11] Jobson J. D. Applied Multivariate Data Analysis, Springer Verlag, 1991, 1–382.

[12] Anderson F. B. In: Statistical methods for comparative studies, New York, J & Wiley, 1982, 184–203.

[13] `www.mathworks.com/patents`

Technical University of Sofia
Faculty of Applied Mathematics and Informatics
8, Kl. Ohridski Blvd
1000 Sofia, Bulgaria
e-mail: `kprod@tu-sofia.bg`