Two-Stage Genetic Algorithm in the Analysis of Biochemical Relations

Biyokimyasal İlişkilerin Analizinde İki Aşamalı Genetik Algoritma Yöntemi

Barış AŞIKGİL,[a]
Aydın ERAR[a]

[a]Department of Statistics,
Mimar Sinan Fine Arts University,
Faculty of Science and Letters,
İstanbul

Geliş Tarihi/*Received:* 16.11.2011
Kabul Tarihi/*Accepted:* 01.03.2012

This study is the re-reviewed form of the presentation titled "Nonlinear Parameter Estimation By Using Two-Stage Genetic Algorithm" which was presented in the "Fifth Conference of the Eastern Mediterranean Region of the International Biometric Society, 10-14 May 2009, İstanbul, Turkey".

Yazışma Adresi/*Correspondence:*
Barış AŞIKGİL
Mimar Sinan Fine Arts University,
Faculty of Science and Letters,
Department of Statistics, İstanbul,
TÜRKİYE/TURKEY
basikgil@msgsu.edu.tr

ABSTRACT Objective: Nonlinear regression analysis is usually used in medical or biochemical areas in order to estimate unknown parameters. Ordinary least squares method can be considered for parameter estimation. However, efficient parameter estimates can not be obtained with the help of this method when errors are autocorrelated. In this paper, a method called two-stage genetic algorithm (TSGA) is proposed in order to obtain efficient parameter estimates. **Material and Methods:** Two-stage least squares (TSLS) method is preferred in order to overcome the autocorrelation problem. However, the method has some problems in the parameter estimation process if initial values of parameters are not known. Therefore, TSGA can be used in the presence of unknown initial values for nonlinear parameters. In this paper, two data sets taken from Faculty of Pharmacy have been examined on a preferred nonlinear model and it has been noticed that residuals are autocorrelated. Because of both unknown initial values of parameters and the autocorrelation problem among errors, TSGA method has been applied by using MATLAB 7.1. **Results:** The results obtained by using genetic algorithm (GA) and TSGA have been compared in view of the correlograms. p-values in the correlograms obtained from GA are <0.001 for all lags. This means that the autocorrelation problem exists among errors. However, p-values in the correlograms obtained from TSGA are greater than α=0.05 for all lags. This means that the autocorrelation problem is removed. Moreover, mean squared error values for TSGA (4.86 and 0.03) are smaller than mean squared error values for GA (11.28 and 0.04). **Conclusion:** It can be concluded that TSGA can give efficient parameter estimates in the presence of autocorrelated errors.

Key Words: Nonlinear dynamics; regression analysis; least-squares analysis

ÖZET Amaç: Doğrusal olmayan regresyon analizi, genellikle medikal ya da biyokimyasal alanlarda bilinmeyen parametreleri tahmin etmek için kullanılır. Parametre tahmini için alışılagelen en küçük kareler yöntemi kullanılabilir. Ancak, hataların otokorelasyonlu olduğu durumlarda bu yöntem ile etkin parametre tahminleri elde edilemeyebilir. Bu çalışmada, etkin parametre tahminleri elde etmek için iki aşamalı genetik algoritma (İAGA) olarak adlandırılan bir yöntem önerilmiştir. **Gereç ve Yöntemler:** Otokorelasyon sorununun üstesinden gelmek için iki aşamalı en küçük kareler (İAEKK) yöntemi önerilir. Ancak, eğer bu analiz sürecinde parametrelerin başlangıç değerleri önceden bilinmiyorsa bu yöntemin kullanımında bazı sorunlarla karşılaşılır. Bu nedenle, doğrusal olmayan parametrelerin başlangıç değerleri bilinmediğinde İAGA kullanılabilir. Bu çalışmada, Eczacılık Fakültesi'nden alınan iki veri kümesi, önerilen doğrusal olmayan bir model üzerinde incelenmiş ve parametre kestirimi sonucu ortaya çıkan artıkların otokorelasyonlu olduğu görülmüştür. Bu nedenle, hem parametreler için başlangıç değerlerinin bilinmemesi hem de otokorelasyon sorunu nedeniyle MATLAB 7,1'de bir program oluşturularak İAGA yöntemi uygulanmıştır. **Bulgular:** Genetik algoritma (GA) ve İAGA kullanılarak elde edilen sonuçlar, artıkların korelogramları göz önünde bulundurularak karşılaştırılmıştır. GA ile elde edilen korelogramlarda tüm gecikmeler için p-değerlerinin <0,001 olduğu belirlenmiştir. Bu durum, otokorelasyon sorununun varlığını belirtmektedir. Ancak, İAGA kullanılarak elde edilen korelogramlarda tüm gecikmeler için p-değerlerinin α=0,05'den oldukça büyük çıkması, otokorelasyonun ortadan kalktığını göstermektedir. Bunun yanında, İAGA'dan elde edilen hata kareler ortalaması değerleri (4,86 ve 0,03), GA'dan elde edilen hata kareler ortalaması değerlerinden (11,28 ve 0,04) küçüktür. **Sonuç:** İAGA'nın otokorelasyonlu hataların varlığında etkin parametre tahminlerini verebildiği kararına varılabilir.

Anahtar Kelimeler: Doğrusal olmayan dinamikler; regresyon çözümlemesi; en küçük kareler analizi

Turkiye Klinikleri J Biostat 2012;4(2):49-54

Nonlinear regression analysis is usually used in medical or biochemical areas in order to estimate unknown parameters. A nonlinear regression model has unknown parameters (θ_j) at least one of which exists in nonlinear form in the model

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \ . \tag{1}$$

For nonlinear parameter estimation it is necessary to minimize the objective function $S(\boldsymbol{\theta})$ given as

$$S(\boldsymbol{\theta}) = \left(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})\right)' \left(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})\right). \tag{2}$$

Gauss-Newton, which is one of the most widely used algorithm for nonlinear parameter estimation, is not easily controllable by practioners. It requires much auxiliary information to work properly.[1,2] Moreover, it may not be convergent or it may only converge to a local optimum. Because of all the limitations, genetic algorithm (GA) can be used for nonlinear parameter estimation.

GA, first introduced by John Holland in the early seventies, is a powerful stochastic algorithm based on the principles of natural selection and natural genetics which has been quite successfully applied in machine learning and optimization problems.[3] GA begins with a collection of parameter estimates (called a chromosome) and each is evaluated for its fitness in solving the given minimization task. At each generation (algorithm timestep), the most fit chromosomes are allowed to mate and bear offspring. The children (new parameter estimates) then form the basis for the next generation. The biological analogy suggests that such a procedure will be likely to lead to workable solutions for complex nonlinear problems.[4]

In GA for nonlinear problems the parameters are encoded into genes and chromosomes as a fixed-length binary string. The length of the string depends on the domain of the parameters and the required precision. The initial values of the parameters are randomly assigned. Therefore, N chromosomes are generated as random binary strings at the beginning of the estimation process. If it is necessary to decode a binary string into a real number, then the binary string can be converted from the base two to the base ten and the real value of the corresponding parameter can be calculated. When the number of parameters is greater than one, the string of each parameter can be united into a single string and this can be solved as parameter vector on its own.[3,4]

GA includes three types of genetic operators given as follows:

- *Selection* chooses chromosomes in the population for reproduction. The selection strategy is chiefly based on the fitness level of the chromosomes actually presented in the population. There are many different selection strategies based on fitness. These are the rank selection, the tournament selection, the fitness-proportionate selection etc. The rank selection strategy used in this paper is based on the rank of the fitness values of the objective function $S(\boldsymbol{\theta})$ of chromosomes in the population. If $S(\theta_i,t)$ is the value of the objective function $S(\boldsymbol{\theta})$ of the chromosome θ_i in the population at generation t and $n_i(t)$ is the rank of $S(\theta_i,t)$ by sorting $\{S(\theta_i,t);\ i = 1, 2, ..., n\}$ descendingly, then the probability that chromosome θ_i will be selected into the next generation of the population as a parent to be operated on by genetic operators is

$$r(\theta_i,t) = \frac{2n_i(t)}{N(N+1)} \ , \tag{3}$$

where N is the population size, i.e., the number of chromosomes in the population.[3,5]

- *Crossover* randomly chooses a locus and exchanges the subsequences before and after the locus between two chromosomes to create two offspring. If two chromosomes given as

$$\theta_1 = (10000100) \text{ and } \theta_2 = (11111111)$$

are crossed over after the third locus in each, then two resulting offspring will be

$$\theta_1' = (10011111) \text{ and } \theta_2' = (11100100).$$

The operator roughly mimics the biological recombination between two single-chromosome (haploid) organisms.[5]

- *Mutation* randomly flips some of the bits in a chromosome. If a chromosome defined as

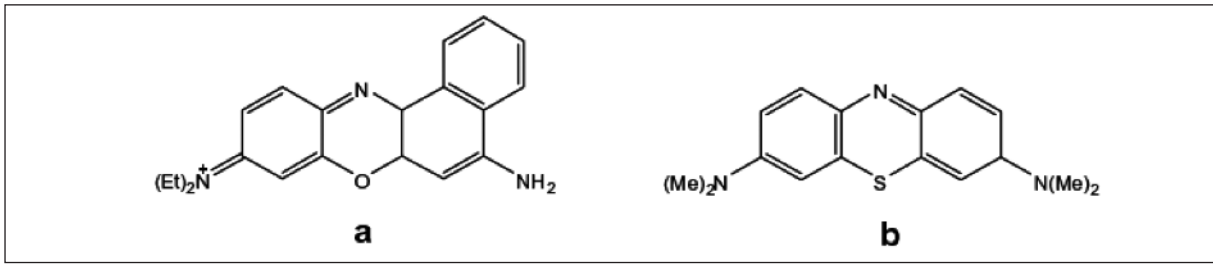

FIGURE 1: The structures of the ligands. **(a)** Nile blue, **(b)** Methylene blue.[6]

(00000100) is mutated in its second position, then it will be (01000100). The operator can occur at each bit position in a chromosome with some probability, usually very small.[5]

MATERIAL AND METHODS

Butyrylcholinesterase (BChE) is a serine hydrolase abundant in liver and plasma. It is expressed in a variety of other tissues including the central nervous system. Although it is originally presumed to serve mainly as a scavenger of xenobiotics, it has more recently been implicated in neurogenesis and overall brain function including the progression of Alzheimer's disease.[6] In this paper, a biochemical nonlinear regression model is examined by evaluating two dyes which act as complex inhibitors of human plasma cholinesterase (BChE). These two dyes are called nile blue (NB), which is a cationic phenoxazine dye, and methylene blue (MB), which is the structurally related phenothiazine.[6] Their chemical structures are given in Figure 1.

Two data sets related to NB and MB have been got from Faculty of Pharmacy. The data sets are based on a system of chemical reaction given in Figure 2.

In Figure 2, (E) is the catalyst and it produces (P) from the substance (S). The velocity of the existence of (P) depends on the concentration of (E) and (S). However, the substance (I) prevents the existence of (P) by interacting one or more ways with (E). The velocity of transformation from (S) to (P) can be modelled as

$$v = \frac{K_1K_2K_3V_{max}S}{K_1K_2K_3(K_s+S)+(K_2K_3K_s+K_1K_3S)I+K_2K_sI^2}, \qquad (4)$$

where $V_{max} \approx 6$, $K_S \approx 40$, S and I are independent variables. Since the initial values of parameters (K_1, K_2, K_3) are not known it is considered that GA can be used to estimate the parameters.

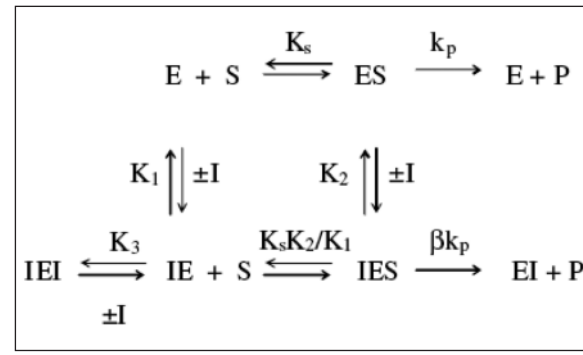

FIGURE 2: System for 2-site mixed inhibiton.[6]

In this paper, two-stage genetic algorithm (TSGA) is proposed as a new approach for nonlinear parameter estimation in the presence of autocorrelated errors in pharmacology. TSGA method which is explained below can be applied by using MATLAB 7.1.

In the model (1) errors can be autocorrelated. Specifically, errors have the autoregressive form AR(q)

$$\varepsilon_t + a_1\varepsilon_{t-1} + a_2\varepsilon_{t-2} + \cdots + a_q\varepsilon_{t-q} = \varepsilon_t^*, \; t = 0, \pm 1, \pm 2, \ldots, \qquad (5)$$

where ε_t^* is independently and identically distributed random variable with mean zero and variance σ^2.

The parameter estimators can not be efficient when errors are autocorrelated.[7,8] In order to overcome this problem TSGA can be used. This method is similar to two-stage least squares (TSLS) proposed by Gallant and Goebel.[9] TSGA is given as follows:

- Residuals (e_i) are obtained by using GA for the model (1).

■ Estimated autocovariances are calculated by using

$$\hat{\gamma}(h)=\frac{1}{n}\sum_{i=1}^{n-h}e_i e_{i+h} \quad , \quad h=0,\ 1,\ ...,\ q\ , \tag{6}$$

where n is the sample size, q is the lag value and h denotes the lag of residuals.

■ (q x q) dimensional estimated variance-covariance matrix and (q x 1) dimensional estimated autocovariance vector of errors are obtained as

$$\hat{\boldsymbol{\Gamma}}_q=\begin{bmatrix}\hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(q-1)\\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(q-2)\\ \vdots & \vdots & \cdots & \vdots\\ \hat{\gamma}(q-1) & \hat{\gamma}(q-2) & \cdots & \hat{\gamma}(0)\end{bmatrix}, \hat{\boldsymbol{\gamma}}_q=\left[\hat{\gamma}(1)\ \ \hat{\gamma}(2)\ \cdots\ \hat{\gamma}(q)\right]'. \tag{7}$$

■ By using Yule-Walker equations estimated parameters of equation (5) are obtained as

$$\hat{\boldsymbol{a}}=-\hat{\boldsymbol{\Gamma}}_q^{-1}\hat{\boldsymbol{\gamma}}_q \quad , \quad \hat{\sigma}^2=\hat{\gamma}(0)+\hat{\boldsymbol{a}}'\hat{\boldsymbol{\gamma}}_q\ . \tag{8}$$

■ By using the Cholesky method $\hat{\boldsymbol{\Gamma}}_q^{-1}=\hat{\boldsymbol{P}}_q'\hat{\boldsymbol{P}}_q$ is obtained and (n x n) dimensional weight matrix is defined as

$$\hat{\boldsymbol{P}}=\begin{bmatrix}\sqrt{\hat{\sigma}^2}\,\hat{\boldsymbol{P}}_q & & & & 0\\ \hat{a}_q & \hat{a}_{q-1} & \cdots & \hat{a}_1 & 1\\ & \hat{a}_q & \hat{a}_{q-1} & \cdots & \hat{a}_1 & 1\\ & & \ddots & \ddots & & \ddots & \ddots\\ & & & \hat{a}_q & \hat{a}_{q-1} & \cdots & \hat{a}_1 & 1\end{bmatrix}, \tag{9}$$

where the first row contains q rows.

Then, the method continues with the second stage:

■ By using the weight matrix some transformations are made

$$\boldsymbol{w}=\hat{\boldsymbol{P}}\boldsymbol{y} \quad , \quad \boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta})=\hat{\boldsymbol{P}}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta}) \quad , \quad \boldsymbol{\upsilon}=\hat{\boldsymbol{P}}\boldsymbol{\varepsilon}\ . \tag{10}$$

■ A new model is obtained as

$$\boldsymbol{w}=\boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta})+\boldsymbol{\upsilon}\ . \tag{11}$$

The two-stage estimator of parameters can be obtained by using GA for the new model.

RESULTS

The results obtained by using GA and TSGA are given in Table 1. According to Table 1, there are two different data sets which are related to NB including 72 observations and MB including 113 observations. The estimated values of parameters K_1, K_2, K_3 for NB by using GA are given as 0.69, 3.99, 0.01, respectively. The mean squared error (MSE) value and the mean absolute error (MAE) value are obtained as 11.28 and 2.21 by using the parameter estimates. On the other hand, GA gives 0.90, 3.54 and 0.29 as the parameter estimates for MB. Also, it is clear that MSE and MAE values are obtained as 0.04, 0.16, respectively. It can be seen by comparing the MSE and MAE values for NB and MB that the two data sets have different behaviours about fitting the model given in (4).

In order to check the assumptions some graphs are examined and the correlograms of residuals are given in Figure 3 for both NB and MB. It can be seen from Figure 3 that the residuals for NB and MB are autocorrelated as AR(1) since the first lags are out of the confidence limits and also, the p-values are smaller than 0.001 both for NB and MB. Therefore, TSGA can be used as a new approach to estimate the parameters in order to overcome the autocorrelation problem.

After using TSGA, the estimated values of parameters K_1, K_2, K_3 for NB are 2.02, 5.25, 0.003, respectively. TSGA causes a decrease of MSE and MAE values which are obtained as 4.86 and 1.43. On the other hand, TSGA gives 1.40, 5.61 and 0.19 as the parameter estimates for MB. Again, a decrease of MSE and MAE which are obtained as 0.03 and 0.14 can be seen by examining Table 1.

TABLE 1: Parameter estimation results by using GA and TSGA.

		n	K1	K2	K3	MSE	MAE
GA	NB	72	0.69	3.99	0.01	11.28	2.21
	MB	113	0.90	3.54	0.29	0.04	0.16
TSGA	NB	72	2.02	5.25	0.003	4.86	1.43
	MB	113	1.40	5.61	0.19	0.03	0.14

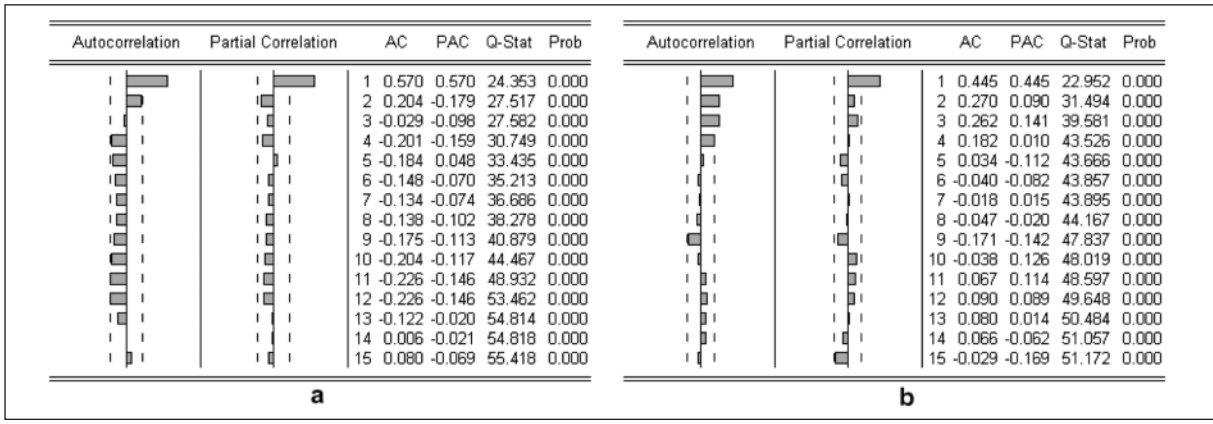

FIGURE 3: Correlograms of residuals obtained from GA. **(a)** Nile blue, **(b)** Methylene blue.

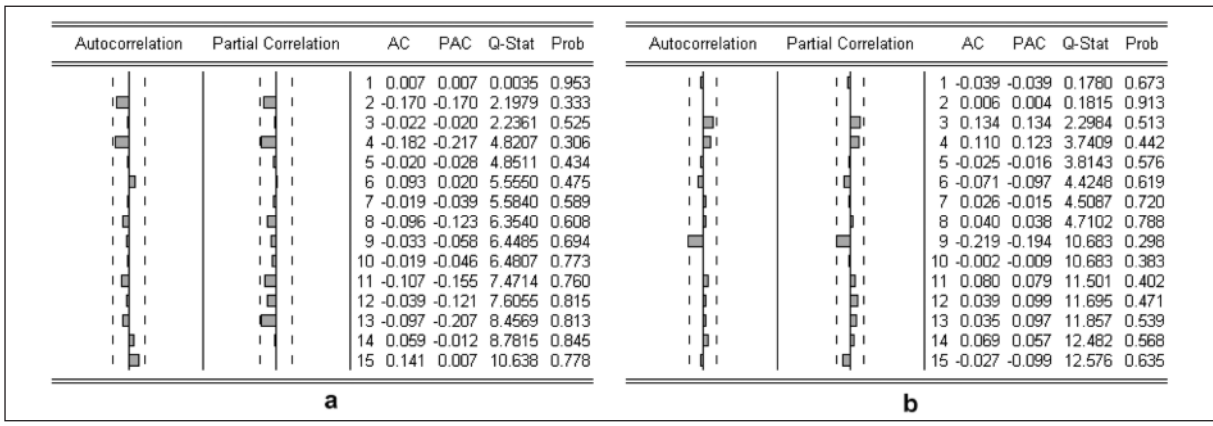

FIGURE 4: Correlograms of residuals obtained from TSGA. **(a)** Nile blue, **(b)** Methylene blue.

Apart from the decrease of MSE and MAE values, it has to be observed for efficient parameter estimates that the autocorrelation problem is removed. After using TSGA, some graphs are examined and the correlograms of residuals are given in Figure 4 in order to check the situation. It can be seen from Figure 4 that the p-values of the first lags for NB and MB are obtained as 0.953 and 0.673, respectively. These values are greater than α=0.05. Also, both the p-values of all lags given in Figure 4 are greater than α=0.05 and no lags are out of confidence limits for NB and MB. These results specify that there is no autocorrelation among errors obtained from TSGA.

DISCUSSION AND CONCLUSION

The violations of assumptions in regression analysis are usually seen in the examination of real data sets. Especially, the autocorrelation problem among errors can be met in medical areas. A big problem can emerge when parameter estimation is considered in the presence of autocorrelated errors. The model can be estimated inaccurately. Also, the comments and the decision about an experiment can guide the researchers to a different way. Therefore, in these cases the autocorrelation problem should be removed in order to obtain efficient parameter estimates.

Estimation of nonlinear parameters is a process that should be examined carefully. There are several numerical methods that can be used in nonlinear parameter estimation, but they are not efficient in the presence of autocorrelated errors. TSLS method is proposed for this case. However, if the initial values of the parameters in nonlinear regression are not known TSLS can not give efficient

results. Therefore, GA is evaluated for the problem of unknown initial values and TSLS method is modified as TSGA in this paper.

According to the numerical example and the results, TSGA overcomes both the problems of autocorrelation and unknown initial values for nonlinear parameters. It can be seen from the last correlograms that there is no more autocorrelation problem. Also, both MSE and MAE values decrease by using TSGA. This is a sign for efficient parameter estimation. Moreover, if TSLS is used by taking randomly chosen initial values instead of TSGA for the complex nonlinear model, then the algorithm may be reached to a local optimum near the initial values instead of the global optimum. This causes different and inaccurate parameter estimates.

To sum up, it can be said that TSGA can be used for better parameter estimation when the initial values of parameters are not known and there is an autocorrelation problem among errors.

Acknowledgements

The authors are grateful to Prof.Dr. İnci ÖZER from Faculty of Pharmacy for the data sets. Also, the helpful comments of two anonymous referees are gratefully acknowledged.

REFERENCES

1. Gallant AR. Univariate nonlinear regression. Nonlinear Statistical Models. 1st ed. New York: John Wiley and Sons; 1987. p.26-30.
2. Seber GAF, Wild CJ. Estimation methods. Nonlinear Regression. 1st ed. New York: John Wiley and Sons; 1989. p.25-8.
3. Pan Z, Chen Y, Kang L, Zhang Y. Parameter Estimation by Genetic Algorithms for Nonlinear Regression. Proceedings. of International Conference on Optimization Techniques and Applications. Hackensack, NJ: World Scientific; 1995. p. 946-53.
4. Yao L, Sethares WA. Nonlinear parameter estimation via the genetic algorithm. IEEE Transactions on Signal Processing 1994;42(4): 927-35.
5. Mitchell M. Genetic algorithms: an overview. An Introduction to Genetic Algorithms. 1st ed. Cambridge, Mass: MIT Press; 1996. p.8-12.
6. Küçükkilinç T, Ozer I. Multi-site inhibition of human plasma cholinesterase by cationic phenoxazine and phenothiazine dyes. Arch Biochem Biophys 2007;461(2):294-8.
7. Glasbey CA. Nonlinear regression with autoregressive time series errors. Biometrics 1980;36(1):135-40.
8. Glasbey CA. Examples of regression with serially correlated errors. The Statistician 1988;37(3):277-91.
9. Gallant AR, Goebel JJ. Nonlinear regression with autocorrelated errors. J Am Stat Assoc 1976;71(356):961-7.