To Build an Optimization Model of Biomedicine Regression Analysis Decision and a Study of Its Application

Xiao Xiaonan
Xiamen University Tan kahkee College
Zhangzhou, Fujian China
xiaoxn@xujc.com

***Abstract*—Some practical uses of mathematics in biology and medical science have fully explained the modern trend of biological and medical mathematics. This article presents a discussion of how to use the pluralistic differential and the method of tropical analysis with a view to establishing mathematical model of several typical functions in biology and medical science. At the same time, it further discusses how to make biology prediction with tropical analysis. It further explores a new theory of combining biology and medical science with mathematics.**

Keywords—stochastic error; residual sum of squares; method of least square; normal equations; regression equation; stepwise regression analysis; sum of squares of partial regression

I. INTRODUCTION

Today, the science of life can be studied quantitatively with the development of biology and medical science. Biochemistry, biophysics and bio-economics, etc., are all frontier sciences which have emerged in connection with the science of life. Biomathematics as the newly emerged field develops more quickly than all the others.

The key to studying every kind of biology system of mathematics lies in the setting-up of various kinds of biomathematics models. However, many experiments in biology and medical science reveal that if one variable is affected by another one or other variables, interlaced with each other, no conclusive functional equality exists between them. Of all these we know one or more variables. We cannot work out the other variables. The most useful way to get the approximate function expression is that first, we collect the experimental data (the more, the better) from some limited groups in which variable Y varies with the change of X_i that represents one variable or more variables. Then we can get a comparatively absolute mathematics model through regression analysis. If combined with the concrete problem in practice, it is a practical experiential expression which is often of theoretical or practical value. This essay discusses how to set up the kind of important mathematics model in the scope of biology and medicine, and then goes further into the discussion about the practical application of regression expression in biology calculation[1-2].

II. THE BUILD-UP OF SEVERAL IMPORTANT BIOMATHEMATICS MODELS

In the following we mainly discuss several important mathematics models in the sphere of biology:

(1) Linear type $y = a + bx$

(2) Exponential function type $y = Ae^{B_x}$

(3) Power function type $y = Ax^B$

(4) Logarithmic function type $y = a + b\ln x$

(5) Parabolic type $y = ax^2 + bx + c$

Several accesses could be found getting near to the aforementioned mathematics models. The method of the least square is the most accurate one. The following discusses how to work out the experiential expression through the method of least square and of multipartite differential calculus. Because of the limited space, the author has omitted the correlation test of experiential formulas. Students interested in that may refer to the concerned books[3].

Specially, after the parameter estimation is born, we should have a check to it. If we find any defects of the model after passing the inspection, we must return to the stage that we have set before or the stage of the variable estimation. And there are two ways to do. First, we should reselect the dependent variable, the argument, and the function form. Second, we can have an estimation to the parameter after decorating and arranging to the data[4].

When the predicting error is born, there are four reasons in the real autoregressive model for predicting:

1) The error was caused by the error factors of the model itself.

2) The error was caused because the estimated value of the regression coefficient is different from the truth-value.

3) The error was caused because the setting value of the argument x is far away from the actual value.

4) The error was caused because the population regression coefficient in the future time has changed.

In fact, except the linear correlation, there exists nonlinear correlation frequently, while on most occasions, the nonlinear regression functions can reflect the relationship between each of the objective phenomenon more properly. For example, adopting the nonlinear regression model will be more suitable for the objective facts than the linear regression model when setting up the production function.

The nonlinear regression analysis must solve the following two questions. Firstly, how to decide the concrete form of the nonlinear function? Different from the linear regression analysis, the nonlinear regression function has various concrete forms.

But what it need is that it should make a choice by the nature of the problems and the observation value of the practical samples. Secondly, how to evaluate the variables of the function? The method used frequently in nonlinear regression analysis is still the least square. However, it should have a proper processing according to the different types of the functions[5].

III. A New Method to Statistical Decision of Digitized Information in Medical Care Based on Entropy of Information

As we know, biological medical system is a complex information system. In order to search fully for and dig from all kinds of complex information in medical decision, this paper uses the optimized entropy and modern statistical information of optimized method, and then gropes for an effective way of the optimized decision of digitized medical information by statistical analysis on a disease group.

According to information theory, in the process of disease diagnosis patients (disease groups) can be regarded as the information resources, doctors as the information destination, and all symptoms getting from every diagnosis as communicating.

Generally speaking, there are two problems that we should pay attention, when we use statistical analysis and computer to diagnose diseases: One is how to judge "the diagnostic value of symptoms and its characteristic of disease groups", and then choose the symptoms according to its value so that we can optimize decision-making. The other is how to figure out the basis of the disease diagnosis according to the larger diagnostic value of the symptoms. However, the current popular Marmum Similitude Method, Bayes Analytical Method, and Sequence Analysis Method etc., can only provide with diagnosis basis other than the diagnostic value of the symptom. So we must apply the statistical analysis optimization method of the information entropy to solve the two problems above to plan as a whole [6-7].

Supposing { $D_1, D_2, \cdots, D_m$ }is a disease group of information resource and there are m types in-compatible disease in this group. Thus, according to Shannon's "the theory of statistical information", we define the entropy of the disease group (X) as $H(X) = -\sum_{i=1}^{m} P(D_1)$, and $P(D_1)$ is the fore probability of disease D_1 in the formula, $0 < P(D_1) < 1$ and $\sum_{i=1}^{m} P(D_1) = 1$.

Let's resume that we have done r times diagnosis and exam independently. There are $S_1, S_2, \cdots, S_r$. We can get n_k incompatible symptoms from exam $S_k (1 \le K \le e)$ and S_{kj} is one of them. Now, the doctor's uncertainty of the diagnosis to his patient will reduce from $H(X)$ to

$$H(X|S_{kj}) = -\sum_{i=1}^{m} P(D_1 | S_{kj}) \log p(D_1 | D_{kj}) \cdots\cdots (1), k=1,2,\cdots,r, j=1,2,\cdots,n_k, P(D_1 | S_{kj})$$

is conditional probability of disease D_1 on the condition that symptom S_{kj} has been known. If we define

$$T(X, S_{kj}) = H(X) - H(X | S_{kj}) \cdots\cdots (2)$$

then $T(X, S_{kj})$ shows the amount of reduction of the disease's uncertainty when S_{kj} is known, namely, the amount of information which doctor can get.

If the occurrence probability of symptom S_{kj} is $P(S_{kj})$, then the average uncertainty of the disease which doctor can get from exam S_k is

$$H(X | S_k) = \sum_{j=1}^{m_k} P(S_{kj}) H(X | S_{kj}) \cdots\cdots (3),\ K = 1,2,\cdots,r$$

and amount of information is

$$T(X, S_k) = H(X) - H(X | S_k) \cdots\cdots (4), k = 1,2,\cdots,r \cdot$$

So, we can use formula (1) or (2) to assess the diagnostic value of each symptom and use formula (3) or (4) to assess the diagnostic value of each exam.

We can choose some symptoms $S_k (1 \le k \le r)$, which have higher diagnostic value after assessment of diagnostic value, then we can decide the foundation of diagnosis and assessment. For this, we rewrite formula (1) as

$H(X | S_{kj}) = \sum_{i=1}^{m} H_i(S_{kj})$ and

$H_i(S_{kj}) = -P(D_i | S_{kj}) \log(D_i | S_{kj})$, $i = 1,2 \cdots, m$

shows the partial uncertainty of disease D_i on the condition that symptom is S_{kj} and the whole uncertainty is $H(X | S_{kj})$, If the exams we did r times are independent one anther the sum of partial uncertainty of disease D_i is

$H_i = \sum_{i=1}^{m} H_i(S_{kj})$ $i = 1,2,\cdots,m$ If $\{H = \min H_i, i = 1,2,\cdots,m\}$ we can diagnose that the disease is $D_i (1 \le i \le m)$.

Because of the limitation of the research methodology and application perspective, in the current information society the traditional non-digitized decision in medical care could not meet the need of the information optimization. Therefore, how to extend the non-digitized medical information into the optimized decision of the general medical information in digitized medical care has become a significant problem that is to be urgently solved on the sustainable development of the present medical information. Hereby, this paper has deeply studied lots of the digitized complex information in medical care by making full use of the optimized entropy and modern statistical information optimization theory, and then established an optimized mathematic model with multi-indexes and multi-factors of disease diagnosis. This model can effectively overcome the defect of the traditional non-digitized information decision, exclude to the largest degree the subjective factors from doctor’s diagnosis, meanwhile, eliminate the errors of medical decision, improve the overall efficiency of medical care, and further improve the level of informative medical care. According to the analysis of the applied case and the appraisement of experts, the diagnosis preciseness of this model is up to above 99.2 percent [8-9].

IV. Prediction

In biology, the study of relationship between the two variables can not only help us expose the inner connection of various biologic characteristics, but also predict another from one variable.

For example, in high-grade breeding, we must select those seeds that have full content of protein. The standard method of determining those seeds that have full content of protein is a key method of determining nitrogen. But the determining process of this method is very complicated and not fit for the selection work of large quantity of original materials. In order to select the high-grade original materials at a high speed, we determine the content of basic amino acid in seeds by adopting the DBC method which is fit for the selection in large quantities. The regression analysis indicates there is an obvious regression relation between the content of protein in seeds and content of basic amino acid in seeds. Therefore, the regression equation of Y (the content of protein in seeds) to X (the content of basic amino acid in seeds) can be acquired. And in light of the regression equation, the content of protein in seeds can be predicted by the results which are obtained by the DBC method. But to prefigure Y from X, we must pay special attention not to exceed the research range when calculating the regression equation at random. Otherwise, the prediction value will not be true[10].

Moreover, in many problem of mathematical simulation, such as in weather forecast, the forecast of insects, the control of manufacture and hydrology, geology and earthquake, we must choose the main elements in a large number of elements and then set out regression equation so as to carry on the forecast and control. Generally speaking, there are two points which must be paid attention to in choosing the correlation elements in order to set out regression equation.

1) In order to get more reliable messages from regression equation, forecasting elements should be included as much as possible in the final regression equation, especially those that have dominant affection on the object of forecasting cannot be missed.

2) If the elements in the regression equation are too many, there would be many problems. That is, it is not only difficult for calculation but also complicated in the interaction of the elements. And the effect of the forecasting would be also influenced. Thus, the forecasting elements should be included as little as possible in the forecasting process, especially those that have no or very little effect on the object of forecasting should not be entered in the regression equation.

In order to solve the proceeding problems, usually we adopt the method of stepwise regression analysis in all the elements to be considered. We will select the most important element according to its effect on the forecasting object y and set out a regression equation in which only this element is contained, then figure out the residual regression sum of squares of the rest elements, drawing a dominant element and setting out a regression equation which contains two elements. Thereafter, every step (inputting an element or deleting an element from the regression equation is a step) of stepwise regression should be checked in a remarkable way no matter if it is the beginning or the end of the step. Through stepwise regression, the high-grade regression equation can be finally gained, and therefore the high-grade forecasting controlling results also can be achieved.

References

[1] Ren Y, Lu S P, Xia N m. Remarks on the existence and uniqueness of solutions to stochastic functional differential equations with infinite delay [J]. Comput Appl Math, 2008, 220: 364-372.

[2] Shi Kaiquan, Yao Bingxue. Function S-rough sets and las identification [J]. Science in China Series F: Information Sciences, 2008, 51(5):499-510.

[3] Yuan C G, Willian G. Approximate solutions of stochastic differential delay equations with Markovian switching [J]. Comput Appl Math, 2006, 194: 207-226.

[4] Connor J, Gross-Erdmann K G. Sequential definitions of continuity for real functions [J]. Rocky mountain J.Math. 2003, 33 (1):93-121.

[5] Richard J L, Morris L M. An introduction to Mathematical Statistics and Its Appication[M]. New Jersey: Prentice-Hall. 1986. 447-481.

[6] Kalbfleisch J G. Probability and Statistical Inference [M]. New York: Springer-Verlag, 1985, 201-241.

[7] Moricz F. Statistical convergence of multiple sequences [J]. Arch.Math, 2003, 81(1): 82-89

[8] Maritz J S. Distribution-Free Statistical Method [M]. London: Chapman and Hall, 1981, 125-144.

[9] Donald F M. Applied Linear Statistical Methods [M]. New Jersey: Prentice-Hall, 1983. 122-172.

[10] Kallianpur G, Krishnaiah P R, Ghosh J K. Statistics and Probability [M]. Amsterdam: North-Holland Publishing Company, 1982, 249-262.