A Genetic Algorithm-Based Quasi-Linear Regression Method and Application

Fachao Li and Kena Zhang
School of Economy and Management
Hebei University of Science and Technology
Shijiazhuang, China
e-mail: lifachao@tsinghua.org.cn , panpan_summer@yeah.net

***Abstract*—Regression analysis, as an important branch of statistics, is an effective tool for scientific prediction. Genetic algorithm is an optimization search algorithm in computational mathematics. In this paper, a new regression model named quasi-linear regression model is established. Further, its implementation method is introduced in detail. Then by taking the population development of Hebei province as an example, we conduct the fitting problem and short-term prediction. Moreover, we compare the fitting effect and the prediction results of two models.**

Keywords-quasi-linear function; regression analysis; genetic algorithm; prediction

I. INTRODUCTION

As a large population country, the development of our country is restricted by its population. In the past half century, China's population has increased sharply. And up to now, it has been a serious problem. In recent years, some new characteristics of our country's demographic structure are presented: accelerated aging population, elevated sex ratio at birth, rural population urbanization, etc. Population change will have an impact on social economy, education, climate and many other aspects, especially it will affect our country's economic development. The prediction is very important to the research of population. Regression analysis is a very common method. But it still has many limitations and deficiencies. So the prediction methods also have attracted the attention of many scholars.

Over the years, there have been many studies on the population prediction and regression analysis. And also it has obtained some achievements. [1] presented the "endogenous efficiency-augmenting mechanism". It suggested that sustained economic growth and a declining population can coexist in the long-run. [2] analyzed the relationship among economy transformation, population growth and the long-run world income distribution. There was evidence that contradicts the conclusion that population growth is adverse to economic growth in [3]. [4] made a research and prediction on the demographic trend of Shandong province since 1952 with ARIMA model and residual autoregressive model. This study provided a basis for the population policy adjustment. An empirical research was done in [5] and showed that the impact of population growth on economic development is a complicated nonlinear relationship. [6] comprehensively investigated the impact of population change on economy with a life cycle hypothesis economic dynamic model. There are also many researches on regression methods. Regression prediction model about the relationship between summer rainfall and grain production was established in [7]. A time series based regression prediction model about China logistics industry was established in [8]. In [9] a regression model in animal breeding can be found. One can refer to [10] for a regression model for the medicine penetration problem in clinical medicine. You may also refer to contributions [11-14] for other applications of regression analysis in physics, biology, military affairs and geography. The Quasi-linear regression model was put forward and explained by one case in [15]. In this paper, we mainly study and predict the population growth rate in Hebei province through the quasi-linear regression and genetic algorithm. And we further compare the results of quasi-linear regression with quadratic regression.

II. QUASI-LINEAR REGRESSION MODEL

Quasi-linear function has piecewise linear feature. And it has good structure characteristics and computation performance. The quasi-linear function with freedom degree 2 is,

$$f(x)=\begin{cases} c_0+\dfrac{(x-a_0)(c_1-c_0)}{a_1-a_0}, & a_0\le x\le a_1, \\ c_1+\dfrac{(x-a_1)(c_2-c_1)}{a_2-a_1}, & a_1\le x\le a_2. \end{cases} \tag{1}$$

Given sample data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, and we assume that $x_1 \le x_2 \le \cdots x_n$. If we use the quasi-linear function on $[a, b]$ as the regression function, then we can get a quasi-linear regression model (QRM for short) with freedom degree *n*,

$$y = Q_L((x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)) + \varepsilon \tag{2}$$

We can construct QRM as follows:

1) determine scatter diagram,

2) determine the freedom degree of the quasi-linear function,

3) solve and get the regression function.

Normally, we can take $a = \min(x_i)$, $b = \max(x_i)$. Also, we can get the following theorem:

Theorem 1: If $[a, b] \supseteq [\min(x_i), \max(x_i)]$, then the results of the final regression curve will not be influenced.

The object of study will be not the whole sample, but a subsample when the range of $[a, b]$ is included in

Published by Atlantis Press, Paris, France.

$[\min(x_i), \max(x_i)]$, so the final regression results will be influenced possibly. And the degree of influence depends on the degree of $[a, b]$ included in $[\min(x_i), \max(x_i)]$. If the range of $[a, b]$ includes $[\min(x_i), \max(x_i)]$, obviously, the final regression results will not be influenced.

We take a quasi-linear regression with freedom degree 2. First, select $[a, b] = [x_1, x_n]$. Second, let $A(x_A, y_A)$, $C(x_C, y_C)$ and $B(x_B, y_B)$ denote the three points respectively on the quasi-linear curve from left to right ($x_A = a$, $x_C = b$). Then the regression function is,

$$\mu(x) = \begin{cases} \dfrac{y_A - y_B}{x_A - x_B} x + \dfrac{x_A y_B - x_B y_A}{x_A - x_B}, & x \in [x_A, x_B], \\ \dfrac{y_B - y_C}{x_B - x_C} x + \dfrac{x_B y_C - x_C y_B}{x_B - x_C}, & x \in [x_B, x_C]. \end{cases} \tag{3}$$

Using the Least Squares Method to determine the estimation value $\hat{y}_A, \hat{x}_B, \hat{y}_B, \hat{y}_C$,

$$\begin{cases} \min \sum_{i=1}^{n} e_i^2 \\ \text{s.t. } x_A \le x_B \le x_C, \\ \quad y_A, y_B, y_C \in (-\infty, +\infty). \end{cases} \tag{4}$$

If we solve the function with Least Square Method, it's very difficult to get the solution due to the too much variables and the complexity of the function. And with the increase of the freedom degree, the unknown variables increase in pairs. That makes it more difficult. And we almost can not get the results with usual numerical method. In the following, we use genetic algorithm (GA) to get the solution.

III. A Genetic Algorithm Based Method For QRM

Genetic algorithm (GA) is an evolutionary algorithm. It is a useful tool for optimization problem. Usually the numerical method mainly depends on iterative operation. General iterative method is subject to fall into local minimum trap and appear "endless loop". Genetic algorithm, as a global optimization algorithm, overcomes this weakness.

Compared with traditional optimization method, genetic algorithm has good convergence, strong currency, etc. It has solved the limitation of current analytical methods. And it has been widely used in combinatorial optimization, signal processing, adaptive control, machine learning, etc. Also it has been one of the key technologies in intelligent and complex system optimization.

The basic operation of the genetic algorithm has three parts: selection, crossover, mutation.

A. *Characteristics of genetic algorithm*

- Compared with traditional optimization algorithm, GA can cover large area, conducive to global merit.
- GA use fitness value to search. It almost can handle any problems as it only demands general information fitness value and coding, etc.
- GA has high fault-tolerant ability.
- The selection, crossover and mutation are random operation.
- GA has scalability, so it is easy to mix with other technology.

B. *The solving strategy based on genetic algorithm for QRM with freedom degree 2*

- Coding.
- Fitness function. Just as model (4), we can select $G(x) = \left[1 + \sum_{i=1}^{n} e_i^2\right]^{-1}$ as fitness function.
- Selection operator. We selects proportional selection operator.
- Crossover operation. We select bit by bit arithmetic crossover operator.
- Mutation operation. We use the mutation method as follows to avoid infeasible solutions: given mutation probability $p_c \in (0, 1]$, bit by bit operate individual (y_A, x_B, y_B, y_C) as follows:

$$x_B' = \begin{cases} x_B + r(b - x_B), & 0 \le r \le 1 \\ x_B - r(x_B - a), & -1 \le r < 0 \end{cases}, \quad y_i' = y_i + r_i, \; i = A, B, C \tag{5}$$

r is a random number in [-1,1], r_i is a random number with normal distribution $N(0, \sigma^2)$.

IV. Hebei Province Population Growth Research and Prediction

Both the natural change and migration change of population are main factors that affect the change of total population of province. After the founding of new China, there was a baby boom in the 50's or 60's. At that time, the natural population growth rate is relatively high, and showed ascendant trends. In the rest of 70's, as the family planning work thoroughly development, the birth rate is greatly reduced, and the natural population growth rate appeared downward trend. In summary, the population of our country still grow over fast. Especially, in the recent years, the population growth rate of our province appeared upward trend again. The population grows much faster. The economic fluctuation and the rising prices are larger than before. The multifaceted impact results the intensification of many social problems of our province such as the employment, education, housing and medical. In order to guide the development of our province better, the research of population growth rate is essential.

Next, we further analyze the characteristics and performance of QRM according to the population data (Table 1) of Hebei province from 1986 to 2007.

TABLE I. Hebei Province Population Data

Year	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996

Total population (ten thousand)	5627	5710	5795	5881	6159	6220	6275	6334	6388	6437	6484
Natural population growth rate (%)	14.3	16.5	14.85	14.75	13.64	9.86	8.9	9.32	8.43	7.61	7.3
Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Total population (ten thousand)	6525	6569	6614	6674	6699	6735	6769	6809	6851	6898	6943
Natural population growth rate (%)	6.29	6.83	6.73	5.09	4.98	5.28	5.16	5.79	6.09	6.23	6.55

A. *Modelling*

First, do the scatter diagram. From the scatter diagram (figure 1) of the data in table 1, we can see that the population growth rate is not obviously scalar trend. But it presents the feature on the whole that first decline then rising. It is roughly U-shaped.

So we can take the quasi-linear function with freedom degree 2 as the regression model.

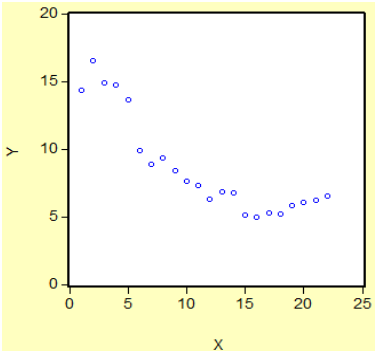

Figure 1. Scatter diagram of Table 1

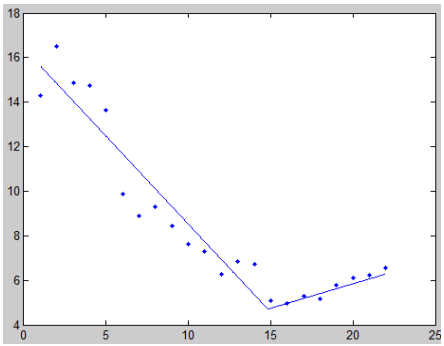

Figure 2. Fitted curve of Quasi-linear model

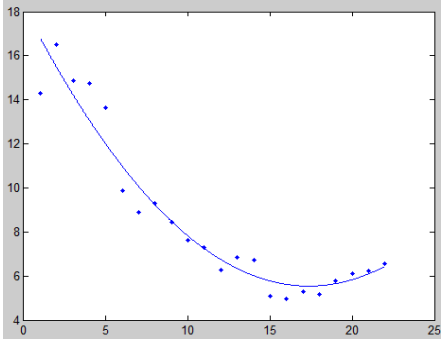

Figure 3. Fitted curve of Quadratic regression model

Then, we can construct the regression model by quasi-linear function.

Finally, using the genetic algorithm and formulas (3) and (4), we can get the quasi-linear function by MATLAB.

Get the result $y_A = 15.64$, $y_C = 6.25$, $x_B = 14.85$, $y_B = 4.69$. Then the quasi-linear function can be got,

$$\hat{y} = \begin{cases} -0.79x + 16.43, & x \in [1, 14.85], \\ 0.22x + 1.44 \quad , & x \in (14.85, 22]. \end{cases} \tag{6}$$

B. *The analysis of results*

Through the above results, we can make the fitted curve of quasi-linear model as Figure 2.

Then we can get the sums of square of deviation of quasi-linear,

$$RSS = \sum e_i^2 = 21.6 \tag{7}$$

The total dispersion square sum is that,

$$TSS = \sum_{i=1}^{22} (y_i - \bar{y})^2 = 286.2691 \ , \tag{8}$$

here $\bar{y} = \left(\sum_{i=1}^{22} y_i \right) / 22$.

The test of goodness of fit can be determined by coefficient of determination,

$$R^2 = 1 - \frac{RSS}{TSS} . \tag{9}$$

The goodness of fit of quasi-linear is 0.9245. And it can be deduced that it meets the test of normality from its residuals.

Secondly, DW (Durbin-Watson) test is a very important test for the regression test.

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2} \tag{10}$$

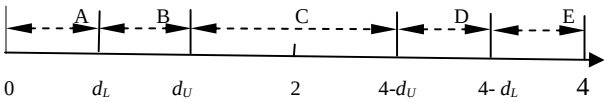

Figure 4. paraphrase for DW correlation test

By this, we can get the DW of quasi-linear is 1.25. It is in the negative correlation accepting region [d_u, 4-d_u]= [1.17, 2.83]. It does not require any further amendments.

In summary, the result of quasi-linear regression is effective. Below we compare it with quadratic regression.

We can get the quadratic regression function by EVIEWS,

$$\hat{y}^* = 0.042x^2 - 1.46x + 18.23 . \tag{11}$$

From Figure 3, we can know that the goodness of fit is also good by quadratic. And R^2 is 0.9363. But the best goodness of fit does not mean that its prediction result is the best.

Published by Atlantis Press, Paris, France.

The natural population growth rate of 2008-2010 predicted with quasi-linear are 6.5, 6.72, 6.94, and with quadratic regression is 6.87, 7.38, 7.98. Actual population growth rate are 6.55, 6.5, 6.81. Compared with the three groups, it can be concluded that the results by quasi-linear regression are more accurate than the results by quadratic regression.

In the example, the difference of quasi-linear and quadratic regression results is not very much, but we can still recognize the advantages of quasi-linear regression. Quasi-linear regression model has better generality and operability. And we can improve the freedom degree to satisfy the tests gradually, and make its fitting effect better.

V. CONCLUSION

By analyzing the above results, a conclusion can be drawn that the population growth rate in our province will continue to rise slowly in the next few years. But it will not rise so quickly in consideration of the policy implementation about population and the change of concept about the fertility problems. Due to the rapid growth of the population, the education, housing and others will still receive attention widely.

For the model, quasi-linear regression method is not suitable for all the situations. It is necessary to determine the model according to the scatter diagram of the research data. Under suitable conditions, the prediction results of quasi-linear are very effective. And it also makes up for the deficiencies of existing regression methods. When the scatter diagram of the data turns sharply, quasi-linear regression is superior to conventional regression methods either in fitting effect or prediction results, especially when the data close to the prediction interval has the linear trend.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (71071049, 71210107002) and the Natural Science Foundation of Hebei Province (F2011208056).

REFERENCES

[1] Ceyhun Elgin and Semih Tumen, “Can sustained economic growth and declining population coexist?” Economic Modelling, vol. 29, 2012, pp: 1899-1908.

[2] Marcos Chamon and Michael Kremer, “Economic transformation, population growth and the long-run world income distribution,” Journal of International Economics, vol. 79, 2009, pp: 20-30.

[3] D. Gale Johnson and Ph. D, “Population and economic development,” China Economic Review, vol. 10, 1999, pp:1-16.

[4] Bo Li, Shuai Wang and Shuang Zhang, “The time series analysis of population growth rate,” Science and Technology Review, 2010, pp:244.

[5] Xuejin Zuo, “The affect population growth on economic growth,” International Economic Review. Vol. 6, 2010, pp: 127-135.

[6] Juhuang He, “Influence of Population Change on Economy,” Quantitative and technology economics, vol. 12, July 2003, pp: 41-46.

[7] B. Parthasarathy, A. A. Munot and D. R. Kothawale, “Regression model for estimation of indian foodgrain production from summer monsoon rainfall,” Agricultural and Forest Meteorology, vol. 42, March 1988, pp. 167-182.

[8] Lin Yang and Zhongbo Liu, “Application of linear regression model in predicting the demand in logistics,” Culture of Business, vol. 10, Oct. 2007, pp. 173-175.

[9] L. R. Schaeffer, “Applilcation of random regression models in animal breeding,” Livestock Production Science, vol. 86, March 2004, pp: 35-45.

[10] Yuhong Liu, Yanjun Shou, Jingfeng Xu and Mei Zhang, “Estimation for drug penetration parameters using a nonlinear regression model,” Journal of Biomedical Engineering of China, vol. 22, 2003, pp: 37-42.

[11] Walter Bich, Giancarlo and D Agostino, “Pennecchi Uncertainty Propagation in A Non-linear Regression Analysis: Application to Ballistic Absolute Gravimenter (IMGC-02),” International Workshop on Advanced Methods for Uncertainty Estimation in Measurement, 2007, pp:16-18.

[12] K. Vasanth Kumar, K. Porkodi and F. Rocha, “Isotherms and Thermo Dynamics by Linear and Non-linear Regression Analysis for the Sorption of Methylene Blue onto Activated Carbon: Comparison of Various Error Functions,” Journal of Hazardous Materials, 2008, pp:794–804.

[13] B. Siva Soumya, M. Sekhar, J. Riotte and JJ Braun, “Non-linear Regression Model for Spatial Variation in Precipitation Chemistry for South India,” Atmosheric Environment, vol.43, 2009, pp: 1147-1152.

[14] K. Vasanth Kumer and S. Sivanesan, “Pseudo second order kinetic models for safranin onto rice husk: Comparison of linear and non-linear regression analysis,” Process Biochemistry, vol.41, 2006, pp: 1198–1202A.

[15] Fachao Li, Chenxia Jin, Yan Shi and Kuo Yang, “Study on Quasi-linear Regression Methods,” ICIC International, 2012 ISSN 1349-4198, pp: 6259-6270