International Journal of Innovative
Computing, Information and Control
Volume **8**, Number **9**, September **2012**

 ISSN 1349-4198
pp. **6259–6270**

STUDY ON QUASI-LINEAR REGRESSION METHODS

Fachao Li[1], Chenxia Jin[1], Yan Shi[2,3] and Kuo Yang[4]

[1]School of Economy and Management
[4]School of Science
Hebei University of Science and Technology
No. 70, Yuhua East Road, Shijiazhuang 050018, P. R. China
lifachao@tsinghua.org.cn; jinchenxia2005@126.com

[2]School of Industrial Engineering
Tokai University
9-1-1, Toroku, Kumamoto 862-8652, Japan
yshi@ktmail.tokai-u.jp

[3]School of Management
Dalian University of Technology
No. 2, Linggong Road, Ganjingzi District, Dalian 116024, P. R. China

Received May 2011; revised September 2011

Abstract. *Regression analysis, an important branch of statistics, is an effective tool for scientific prediction and management. In this paper, by analyzing the characteristic and weakness of the existing regression methods, using the concept of quasi-linear function with good structure and approximation properties, we establish quasi-linear regression model (denoted by QRM for short). Further we consider the parameter estimation strategy for QRM, propose parameter estimation based on genetic algorithm and the least squares method, do error testing based on residual analysis. Finally, we analyze the performance of the model by an illustrative example. The result indicates that QRM possesses generality and good operability. The regression effect can be satisfied by adjusting the freedom degree of the quasi-linear regression function. Accordingly, it can be widely used in many fields such as artificial intelligence and economic management.*
Keywords: Regression analysis, Quasi-linear function, Parameter estimation, Genetic algorithm, Residual, Hypothesis testing

1. **Introduction.** Regression, proposed by biostatistician Galton, is a useful tool dealing with the correlation between variables. Its basic idea is from mean aspect to consider the dependency relationship between random variables and controlled variables using statistical methods. The correlation exists widely in real life (for instance, height and weight, working time and achievement, input and output); therefore, the theories and methods about regression analysis are important in academia and application. There are numerous contributions focused on this aspect, such as economy, management, engineering and medicine. Regression prediction model about grain production in terms of summer rainfall was considered in [1]. A regression prediction model based on time series was proposed according to China logistics industry [2]. Regression models in animal breeding can be found in [3]. For the medicine penetration problem in clinical medicine one can refer to [4] for a regression model. You may also refer to contributions [5-9] for other applications of regression analysis in military affairs, physics, geography and biology.

Regression model is the basic element of regression analysis. Simple regression model cannot guarantee the effect of fitting, while complex regression model is difficult to estimate the parameters; therefore, how to construct a general regression model with good

structure is the key. Traditional linear regression analysis was perfect in theories and methods, but the model is too simple to describe accurately the correlation among variables. Accordingly, nonlinear regression analysis attracts more concentration both in academia and application. There are two ways to process a nonlinear regression problem, that is, transforming it into a linear one or straightly fitting. Although some interesting results have been obtained for concrete problems, it is hard to deny that there exist some defects. Linear transforming methods do not possess generality because there were not general rules for choosing the transformation function. For straightly fitting methods, there is no effective selecting mechanism for regression model, and it is hard to estimate the parameters of the regression function. Based on these defects, helpful discussions have been given in recent years. Zhou [10] pointed out that the method of least squares can do nothing to obtain satisfactory regression equation for data with the feature of nonlinearity, seasonality and strong fluctuation. Wang [11] thought that it is the nonzero mean error that results in unsatisfactory goodness of fit for regression curve. Xie [12] proposed the idea of piecewise fitting, but did not give an operable implement strategy. Zhang [13] put forward the weighted linear regression method and showed that it can approximate to the best model by an illustrative example, but he could not give an operable data grouping strategy. Chen [14] thought that almost all parameters cannot be estimated analytically through the method of least squares for nonlinear regression analysis. Although these discussions enrich the regression analysis theories and methods to a certain degree, they all did not give operable regression model and solution methods, and this is very important for real regression problems.

In the sequel, we have the following work: 1) We propose the quasi-linear regression model (denoted by QRM for short) and analyze its property from theories; 2) We give parameter estimation strategy for QRM based on genetic algorithm and the least squares method; 3) We do error testing and variance estimate based on statistics and residual analysis; 4) We analyze the performance of QRM by an illustrative example.

2. **A Summary of Regression Model.** Regression analysis is a statistical method considering the correlation among random variables and controlled variables; its general form is that,

$$y = \mu(x) + \varepsilon, \tag{1}$$

where $x = (x_1, x_2, \cdots, x_p)$ is controlled variables, $\mu(x)$ is the definite relationship on x (called regression function), ε is random error satisfying $E(\varepsilon) = 0$.

Model (1) simply describes the correlative characteristic between random variable y and controlled variable. It contains two parts: one is $\mu(x)$ with the essence which is the mathematical expectation $E(y)$; the other is random error ε. If the regression function in (1) is linear, and ε obeys normal distribution $N(0, \sigma^2)$, then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{2}$$

is called linear regression model. Here, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are regression coefficients, and can be estimated by sample $(x_{1i}, x_{2i}, \cdots, x_{pi}, y_i)$ and the Least Squares Method, that is, let

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

be an objective function, and determine the estimation $\hat{\beta}_i$ of β_i by $\partial Q / \partial \beta_i = 0$. Here, $\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$, and $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ is called empirical regression equation of $\mu(x)$.

If the regression function in (1) is nonlinear, then (1) is a nonlinear regression problem. And for the nonlinear regression analysis, the common used methods lack enough operability, and the regression effect is usually not ideal.

Because unitary regression is the basis of complex one, and has a wide application, this research mainly concentrated on unitary regression.

3. Regression Model Based on Quasi-linear Function (QRM).

3.1. Quasi-linear function and its properties.

In order to establish an operable nonlinear regression model, this section introduced the concept of quasi-linear function.

Definition 3.1. [15] *Let* $a = a_0 < a_1 < \cdots < a_n = b$, $c_0, c_1, c_2, \cdots, c_n \in (-\infty, \infty)$. *If* $f(x) = f_k(x) = c_{k-1} + (x - a_{k-1})(c_k - c_{k-1})/(a_k - a_{k-1})$, $a_{k-1} \le x \le a_k$, $k = 1, 2, \cdots, n$, *then we say* $f(x)$ *is a quasi-linear function on* $[a, b]$ *with freedom degree* n, *and written as* $f(x) = Q_L((a_0, c_0), (a_1, c_1), \cdots, (a_n, c_n))$ *for short.*

Theorem 3.1. *If* $f(x)$ *is continuous on* $[a, b]$, *then there exists a series of quasi-linear functions* $\{f^{(n)}(x)\}_{n=1}^{\infty}$ *on* $[a, b]$ *such that* $\{f^{(n)}(x)\}_{n=1}^{\infty}$ *uniformly converge to* $f(x)$.

Theorem 3.2. *If* $f(x)$ *only have finite discontinuous point, there must exist a series of quasi-linear functions* $\{f^{(n)}(x)\}_{n=1}^{\infty}$ *on* $[a, b]$ *such that* $\{f^{(n)}(x)\}_{n=1}^{\infty}$ *converge to* $f(x)$ *on the continuous point* x.

Please refer to [15] for the proof for Theorem 3.1 and Theorem 3.2.

From the above analysis, we can see that quasi-linear function can be approximated by any piecewise continuous function, and also has a good description, so we can do regression analysis by quasi-linear function.

3.2. Quasi-linear regression model (QRM).

For given samples (x_i, y_i), $i = 1, 2, \cdots, n$, if the quasi-linear function on $[a, b]$ is employed as the regression function, we say the model a quasi-linear regression model (QRM) with freedom degree n, that is

$$y = Q_L((a_0, c_0), (a_1, c_1), \cdots, (a_n, c_n)) + \varepsilon. \tag{4}$$

Obviously, (4) is the linear regression model for $n = 1$, and it is a general model with different freedom degree. We can construct a QRM as follows: 1) Determine the freedom degree of quasi-linear function according to the feature of sample data; 2) Determine the fitting interval $[a, b]$ according to the distribution of sample data; 3) Estimate the endpoints of the subintervals using the Least Squares Method.

To illustrate this method, we take a quasi-linear regression problem with freedom degree 2 as an example.

For given sample data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, we assume, without loss of generality, that $x_1 \le x_2 \le \cdots \le x_n$. Firstly, select $[a, b] = [x_1, x_n]$, including all x_i, as the fitting interval. Secondly, estimate the endpoints of the subintervals using the Least Squares Method, that is: 1) assume $A(x_A, y_A)$, $B(x_B, y_B)$ and $C(x_C, y_C)$ (here, $x_A = a, x_C = b$) denote respectively the three points on the quasi-linear curve from left to right; 2) by

$$\mu(x) = \begin{cases} \dfrac{y_A - y_B}{x_A - x_B}x + \dfrac{x_A y_B - x_B y_A}{x_A - x_B}, & x \in [x_A, x_B] \\ \dfrac{y_B - y_C}{x_B - x_C}x + \dfrac{x_B y_C - x_C y_B}{x_B - x_C}, & x \in [x_B, x_C] \end{cases} \tag{5}$$

and

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{k} (\mu(x_i) - y_i)^2 + \sum_{i=k+1}^{n} (\mu(x_i) - y_i)^2, \quad x_k < x_B < x_{k+1}, \tag{6}$$

using the Least Squares Method to determine the estimation values $\hat{y}_A$, $\hat{x}_B$, $\hat{y}_B, \hat{x}_C$ of y_A, x_B, y_B, x_C, respectively, and empirical regression equation $\hat{\mu}(x) = Q_L((x_A, \hat{y}_A), (\hat{x}_B, \hat{y}_B), (x_C, \hat{y}_C))$ of $\mu(x)$.

In fact, all the above can come down to the following programming problem:

$$\begin{cases} \min \sum_{i=1}^{n} e_i^2 \\ \text{s.t.} \quad x_A \le x_B \le x_C, \quad y_A, y_B, y_C \in (-\infty, +\infty). \end{cases} \tag{7}$$

Theorem 3.3. *If ε in (4) obeys normal distribution $N(0, \sigma^2)$, then (6) is identical with the maximum likelihood estimates of y_A, x_B, y_B, y_C.*

Proof: If (x_i, y_i) is taken as the observed values of (x_i, Y_i), then Y_i obeys normal distribution $N(\mu(x_i), \sigma^2)$, $i = 1, 2, \cdots, n$, and they are mutually independent, further we have the joint distribution density function of $Y_1, Y_2, \cdots, Y_n$:

$$L(y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\mu(x_i) - y_i)^2\right]. \tag{8}$$

Clearly, (8) is the likelihood function of $Y_1, Y_2, \cdots, Y_n$ corresponding to $y_1, y_2, \cdots, y_n$. The maximum value of $L(y_1, y_2, \cdots, y_n)$ is identical with the minimum value of (6), which shows that the conclusion is true.

This theorem means from statistics that it is feasible to determine the quasi-linear regression function using (6).

Theorem 3.4. *If ε in (4) obeys normal distribution $N(0, \sigma^2)$, and the expressions of $\hat{\mu}(x) = Q_L((x_A, \hat{y}_A), (\hat{x}_B, \hat{y}_B), (x_c, \hat{y}_c))$ on $[a, \hat{x}_B]$ and $[\hat{x}_B, b]$ are identical with the linear regression estimate functions of samples $(x_1, y_1), (x_2, y_2), \cdots, (x_k, y_k)$ and samples $(x_{k+1}, y_{k+1}), (x_{k+2}, y_{k+2}), \cdots, (x_n, y_n)$, respectively, (here, $x_k < \hat{x}_B < x_{k+1}$ and $k > 2$, $n - k > 2$), then the variance σ^2 can be estimated as*

$$\hat{\sigma}^2 = \frac{1}{n-4}\left[\sum_{i=1}^{k} (\hat{\mu}(x_i) - y_i)^2 + \sum_{i=k+1}^{n} (\hat{\mu}(x_i) - y_i)^2\right], \tag{9}$$

which is unbiased.

Theorem 3.4 can be directly proved using linear regression theories.

The above conclusion cannot be thought as the general principle to estimate the random error, but it can be an approximate estimate method under some condition.

Remark 3.1. *The above discussions hold for quasi-linear fitting problems with freedom degree n.*

Remark 3.2. *Because regression analysis is a kind of empirical inference, it is difficult to construct the internal relationship between variables. In order to obtain the random error with fixed distribution, we can improve slightly the object function of (7). Of course, the improvement should keep first the total squares sum of error attaining its minimum and also give consideration to the stability of average errors for all linear section.*

Remark 3.3. *Theorems 3.3 and 3.4 with the basis of Central Limit Theorem further enrich the parameter estimate methods of (6), and lay theory foundation for the above discussions. In real, many correlation elements interact on many random phenomena, but theses micro effect cannot bring great influence, and the sum of these independent factors with micro effect is approximate normal distribution. Accordingly, the condition in Theorems 3.3 and 3.4 is reasonable.*

4. **Solving Strategy for QRM Based on Genetic Algorithm.** Genetic algorithm [16,17] is a useful tool in intelligent computing and complex system optimization. Next we consider the solving strategy based on genetic algorithm for QRM with freedom degree 2.

1) Coding. In this paper, we use real value coding directly. Take a quasi-linear function with freedom degree 2 as an example. For the coordinates of connection points, then the real-code is (y_A, x_B, y_B, y_C).

2) Fitness function. This paper selects $G(x) = [1 + \sum_{i=1}^{n} e_i^2]^{-1}$ as fitness function.

3) Selection operator. This paper selects proportional selection operator.

4) Crossover operation. This paper introduces bit by bit arithmetic crossover operator as follows. For a given crossover probability $p_c \in [0, 1]$, crossover bit by bit the gene in populations $X = (y_A, x_B, y_B, y_C)$ and $Y = (y'_A, x'_B, y'_B, y'_C)$, for the first bit, we have:

$$X'_1 = ry_A + (1 - r)y'_A, \quad Y'_1 = (1 - r)y_A + ry'_A, \tag{10}$$

r is a random number in $[0, 1]$, and same to the other bit.

5) Mutation operation. In (7), x_B is influenced by the range of $[a, b]$. To avoid infeasible solutions, we use the following mutation methods: for any given mutation probability $p_m \in (0, 1)$, operate the individual y_A, x_B, y_B, y_C bit by bit as follows:

$$x'_B = \begin{cases} x_B + r(b - x_B), & 0 \le r \le 1 \\ x_B - r(x_B - a), & -1 \le r < 0 \end{cases}, \quad y'_i = y_i + r_i, \quad i = A, B, C, \tag{11}$$

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

For convenience, we abbreviate the above genetic algorithm as GA-QRM. Following considers the convergence of GA-QRM through Markov Chain theories.

Definition 4.1. [16] *Let $\overrightarrow{X} = (X_1(t), X_2(t), \cdots, X_N(t))$ be the t^{th} population of genetic algorithm, $f^* = \max\{f(X)|X \in S\}$ denote the global optimal value of the individuals, $Z_t = \max\{f(X_i(t)), i = 1, 2, \cdots, N\}$. If $P\{|Z_t - f^*| < \varepsilon\} \to 1$ $(t \to \infty)$ always holds for any $\varepsilon > 0$, then we say the genetic sequence $\{\overrightarrow{X}(t)\}_{t=1}^{\infty}$ converges. Here, $P(A)$ is the probability of event A happening.*

Theorem 4.1. *The GA-QRM using the elitist preserving strategy (that is, the contemporary optimal individual persevered to the next generation) is globally convergent.*

5. Implementing Steps of Quasi-linear Regression Analysis.

5.1. Implementing steps of quasi-linear regression.

This section will propose the steps of regression analysis: 1) Select fitting interval and freedom degree according to the feature of sample data; 2) Determine the regression parameter based on Genetic Algorithm, including coding, fitness function and genetic operators; 3) Test the effectiveness of regression estimate function and the random error combining with some strategy.

5.2. Testing problem of quasi-linear regression.

Because the parameter estimates of quasi-linear regression function is hard to express by analytical methods, the correlation rationality test and the estimate of variance are hard to realize by analytical methods. Considering that residual analysis does not require the exact estimates of parameters, the corresponding test of quasi-linear regression can be processed by combining residual analysis with some strategies. Residual graphics are the basic tool for residual analysis. Following will give the concrete implementing strategies of each testing.

5.2.1. *Normality test based on residual.* Because the real error in regression is unknown, using residual to approximately analyze the feature of the error is a key method. Based on statistics theories, following will give several criteria to infer the feature of the error. Firstly, we assume the random errors corresponding to controllable variables are identically and independently distributed.

Criterion 1 If they all have normal distribution $N(0, \sigma^2)$, then with a larger probability the residual coordinate should be symmetric around the line $e = 0$.

Criterion 2 If they all have common normal distribution, then with a larger probability the residual histogram should appear "Bell"-type.

Criterion 3 If they all have common normal distribution, then with a larger probability all (e_i, f_i) should be distributed around a straight line in the normal probability paper. Here, $f_i = n_i/n$, n is the sample size, n_i is the number of samples whose residual is less than or equal to e_i.

Criterion 4 If they all have common normal distribution, then with a larger probability the residual histogram of any subsample should be similar to the residual histogram of all samples.

Remark 5.1. *All the criteria above base on the statistical theories: If the random errors corresponding to the values of controllable variable are identically, independently and normally distributed, then the residuals e_i corresponding to sample points (x_i, y_i) could be considered approximately as sampling from the same population.*

Remark 5.2. *All the criteria above depend on the regression estimate function, so the selection of the regression model is the key to determine whether the test could be passed through. For QRM, the test can be passed step by step by raising gradually the level of freedom degree. But it should be noted that the raising of the level may result in the linear increasement of computing complexity. So we should select the one which has small freedom degree under the condition of passing through all tests.*

5.2.2. *Correlation test based on DW.* Correlation test of error is to judge whether the corresponding random error $\varepsilon(x)$ of controllable variables x is dependent. The common method to test the correlation between the residuals of two neighbored samples is DW-method, and the test statistic is

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2}, \tag{12}$$

and its illustrative explanation is as Figure 1.

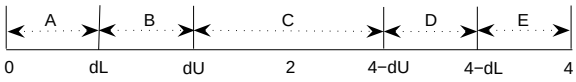


FIGURE 1. Paraphrase for DW-test

In Figure 1, 1) dL and dU are the correlation critical values that can be obtained by lookup table, and they depend on the significance level λ and the size of sample. For $n = 60$, if $\lambda = 0.01$, then dL $= 1.38$, dU $= 1.45$, if $\lambda = 0.05$, then dL $= 1.55$, dU $= 1.62$. For $\lambda = 0.05$, if $n = 100$, then dL $= 1.65$, dU $= 1.69$. 2) $A = [0, \mathrm{dL}]$ represents the positive correlation accepting region, $C = [\mathrm{dU}, 4 - \mathrm{dU}]$ the non-correlation accepting region, and $E = [4 - \mathrm{dL}, 4]$ the negative correlation accepting region, $B = (\mathrm{dL}, \mathrm{dU})$ and $D = (4 - \mathrm{dU}, 4 - \mathrm{dL})$ the region whose correlation cannot be judged. It should be

noted that only for normal variables, independence is equivalent to non-correlation. So DW-method could answer partially the problem of independence.

Remark 5.3. *Significance level λ is a decision risk parameter. The smaller (bigger) λ is, the bigger (smaller) non-correlation accepting region is, usually, we use $\lambda = 0.05$ or $\lambda = 0.01$ in real application; the smaller (bigger) n is, the bigger (smaller) non-correlation accepting region is. And in Section 6, we use $\lambda = 0.01$ and $n = 60$.*

5.2.3. *Error estimation based on subsample.* Statistical law is a decision technology based on large test, and small data cannot assure high reliability decision. That the form of regression function is unknown reminds us that the main task for regression analysis is to decide the distribution of the error and the parameters estimate. In order to make high reliable decision, we consider the feature of the random error from different aspects. We start from the feature similarity of subsamples to give a couple of methods for testing error features and estimating variance.

Criterion 5 If the random errors for controllable variables are independently distributed with mean 0, then the average of residuals of all subsamples should with a large probability fall over $(-\delta^{'}, +\delta')$. Here, $\delta^{'}$ is a given positive threshold value.

Criterion 6 If the random errors for controllable variables are independently identically distributed with mean 0, then the average of residual squares for any subsample should fall over a moderate interval $[a, a + \beta]$ with large probability, and the residual diagrams appears similar symmetry. Here, $\beta > 0$ is a given positive threshold value.

According to the discussion above, if the random errors for controllable variables are identically independently distributed with mean 0, then we can take $\hat{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma}_i^2$ as the estimate of the variance of random error. Here, k represents the size of the subsample and $\hat{\sigma}_i^2$ the square of the average residual of i^{th} subsample.

Remark 5.4. *All the above criteria base on the following statistical theories: 1) Sample mean is an unbiased estimate of population mean; 2) 2-order origin moment of sample is an unbiased estimate of population variance if the population mean is 0.*

6. **Case Study.** In this section, we will consider the feature and effectiveness of QRM through an example.

Case description. With the development of science and technology, many things including the people's living standards, working conditions, environmental consciousness, artistic appreciation changes greatly, and some enterprises gradually realize the mechanization and IT application in both production and management, which brings many electrical appliances popularization to a certain degree. In order to raise the level of service and management, Power Co. A made a series of investigation, which show that air temperature is a major factor influencing the power consumption. All the investigated data are listed in Table 1, the dependency relationship between Peak loads (KW) and temperature (°F) is expected to obtained using the data in Table 1 (here, T denotes temperature, PL denotes Peak loads).

From Table 1 we can see that, there does not exist exact numerical dependency relationship between Peak loads and temperature, but correlativity that power consumption is greater when temperature is higher or lower, and smaller when temperature is moderate. Following consider the correlativity from regression analysis.

We can see from the sample scatter diagram (See Figure 2) that the relationship between peak load and temperature appears V-type, which is not suitable for straight line fitting. We can use a quadratic function $y = ax^2 + bx + c$ or a quasi-linear function $y = Q_L((a_0, c_0), (a_1, c_1), (a_2, c_2))$ with freedom degree 2 to fit. Using Matlab and the solv-

TABLE 1. Peak loads (KW) in terms of temperature (°F)

T	55	55	56	56	57	57	58	59	60	62	65	66
PL	120	118.3	118.9	117.6	113.9	117.4	115.8	111.8	110.5	109.5	99.5	100.2
T	61	63	63	64	64	67	67	68	69	70	70	71
PL	109.5	105	106	105.2	102.8	96.5	101.6	96.3	95.6	92.5	94.5	90
T	71	72	74	74	76	79	84	85	86	87	77	78
PL	93.6	88	90.4	92.4	96.1	103.1	114.9	114.8	116.8	120	97	102.1
T	80	81	82	83	81	82	88	89	89	90	92	94
PL	104.5	108.1	111.3	112.1	106.7	108.1	124.1	125	127.8	127.8	134.1	138
T	95	96	97	98	100	100	108	106	96	94	95	86
PL	140.7	142	145.2	147.4	150.1	153.9	170.8	166.1	143	138	139	121.4

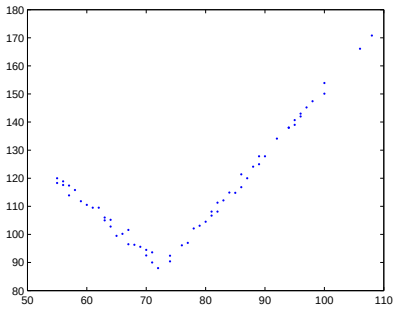


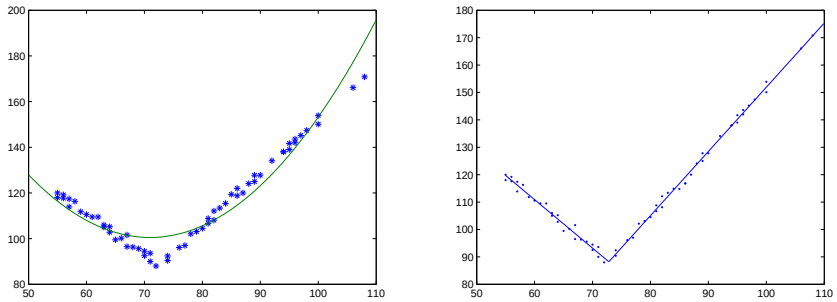
FIGURE 2. Scatter diagram for data in Table 1

ing method for QRM given in Section 4, we can obtain the quadratic estimate function and the quasi-linear estimate function with freedom degree 2 as follows:

$$\hat{y}^* = 0.0625x^2 - 8.8692x + 415.1326, \tag{13}$$

$$\hat{y}^{**} = \begin{cases} 216.9694 - 1.7680x, & x \in [50, 71] \\ -82.3854 + 2.3430x, & x \in [71, 110], \end{cases} \tag{14}$$

the corresponding residuals list in Table 2 (where $\hat{y}^* = \hat{y}^*(x_i)$, $\hat{y}^{**} = \hat{y}^{**}(x_i)$, $e_i^* = y_i - \hat{y}_i{}^*$, $e_i^{**} = y_i - \hat{y}_i^{**}$). And one can find the fitting curves, regression residual diagrams, regression residual histograms and normal probability plots in Figures 3-6, (a) for quadratic regression and (b) for QRM.

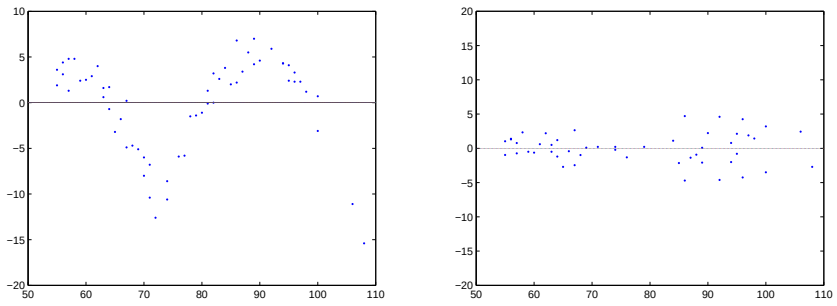


(a) (b)

FIGURE 3. Fitting curves

TABLE 2. Regression residual estimates

x_i	55	55	56	56	57	57	58	59	60	62	65	66
y_i	120	118.3	118.9	117.6	113.9	117.4	115.8	111.8	110.5	109.5	99.5	100.2
$\hat{y}_i^*$	116.4	116.4	114.5	114.5	112.6	112.6	111	109.4	108	105.5	102.7	102
$\hat{y}_i^{**}$	119.7	119.7	117.9	117.9	116.2	116.2	114.4	112.6	110.9	107.4	102.0	100.3
e_i^*	3.6	1.9	4.4	3.1	1.3	4.8	4.8	2.4	2.5	4	–3.2	–1.8
e_i^{**}	0.3	–1.4	1	–0.3	–2.3	1.2	1.4	–0.8	–0.4	2.1	–2.5	–0.2
x_i	61	63	63	64	64	67	67	68	69	70	70	71
y_i	109.5	105	106	105.2	102.8	96.5	101.6	96.3	95.6	92.5	94.5	90
$\hat{y}_i^*$	106.6	104.4	104.4	103.5	103.5	101.4	101.4	101	100.7	100.5	100.5	100.4
$\hat{y}_i^{**}$	109.1	105.6	105.6	103.8	103.8	98.5	98.5	96.7	94.9	93.2	93.2	91.4
e_i^*	2.9	0.6	1.6	1.7	–0.7	–4.9	0.2	–4.7	–5.1	–8	–6	–10.4
e_i^{**}	0.4	–0.6	0.4	1.4	–1	–2	3.1	–0.4	0.7	–0.7	1.3	–1.4
x_i	71	72	74	74	76	79	84	85	86	87	77	78
y_i	93.6	88	90.4	92.4	96.1	103.1	114.9	114.8	116.8	120	97.0	102.1
$\hat{y}_i^*$	100.4	100.6	101.0	101.0	102	104.5	111.1	112.8	114.6	116.6	102.8	103.6
$\hat{y}_i^{**}$	91.4	89.6	91.0	91.0	95.7	102.7	114.4	116.7	119.1	121.4	98	100.3
e_i^*	–6.8	–12.6	–10.6	–8.6	–5.9	–1.4	3.8	2	2.2	3.4	–5.8	–1.5
e_i^{**}	2.2	–1.6	–1.4	1.4	0.4	0.4	0.5	–1.9	–2.3	–1.4	–1	1.8
x_i	80	81	82	83	81	82	88	89	89	90	92	94
y_i	104.5	108.1	111.3	112.1	106.7	108.1	124.1	125	127.8	127.8	134.1	138
$\hat{y}_i^*$	105.6	106.8	108.1	109.5	106.8	108.1	118.6	120.8	120.8	123.2	128.2	133.7
$\hat{y}_i^{**}$	105.0	107.4	109.7	112.1	107.4	109.7	123.8	126.1	126.1	128.5	133.1	137.8
e_i^*	–1.1	1.3	3.2	2.6	–0.1	0	5.5	4.2	7	4.6	5.9	4.3
e_i^{**}	–0.5	0.7	1.6	0	–0.7	–1.6	0.3	–1.1	1.7	–0.7	1	0.2
x_i	95	96	97	98	100	100	108	106	96	94	95	86
y_i	140.7	142.0	145.2	147.4	150.1	153.9	170.8	166.1	143.0	138	139	121.4
$\hat{y}_i^*$	136.6	139.7	142.9	146.2	153.2	153.2	186.2	177.2	139.7	133.7	136.6	114.6
$\hat{y}_i^{**}$	140.2	142.5	144.9	147.3	151.9	151.9	170.7	166.0	142.5	137.8	140.2	119.1
e_i^*	4.1	2.3	2.3	1.2	–3.1	0.7	–15.4	–11.1	3.3	4.3	2.4	6.8
e_i^{**}	0.5	–0.5	0.3	0.1	–1.8	2	0.1	0.1	0.5	0.2	–1.2	2.3

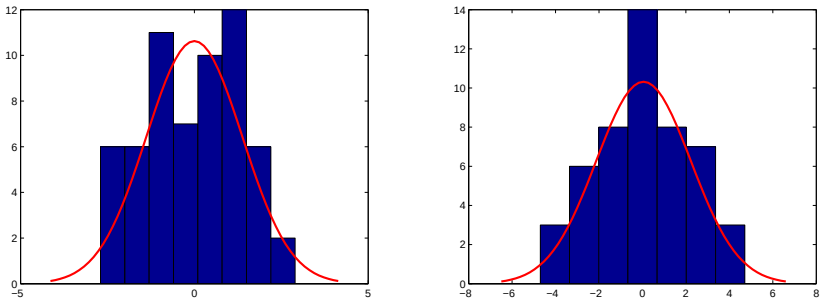


(a) (b)

FIGURE 4. Residual diagrams

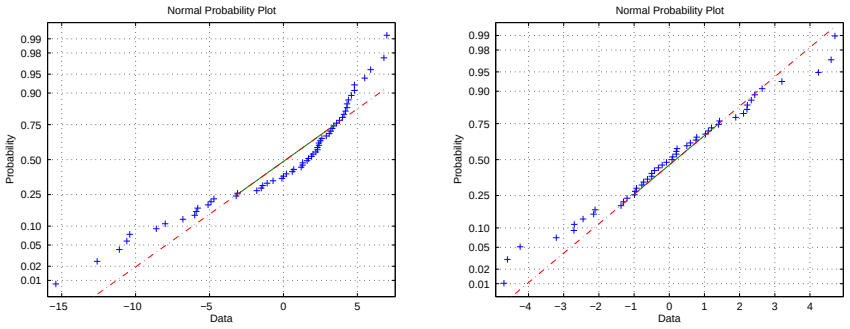
It can be deduced from the above figures and results in Table 2 that the two regression results have the following points:

1) The fitting effect of QRM is better than that of quadratic regression. i) In reflecting the feature of sample data, please see Figure 3; ii) In comparing the sizes of sum of squared residuals: $\sum(e_i^*)^2 = 1600.58$, $\sum(e_i^{**})^2 = 99.38$.



(a) (b)

FIGURE 5. Residual histograms



(a) (b)

FIGURE 6. Normal probability plots

2) Random error of QRM is normally distributed with mean 0 but quadratic regression is not. Please see Figures 4-6. i) All the residuals of QRM are uniformly distributed around $e = 0$ but quadratic regression is not, please see Figure 4; ii) Residual histograms of QRM has normal distribution but quadratic regression does not, please see Figure 5; iii) From normal probability plots, the corresponding residual of QRM almost lies in a straight line but quadratic regression does not, please see Figure 6.

3) Random errors of QRM are approximately independent, but quadratic regression is not. In fact, by Table 2 and (12) we can obtain the value of DW is 2.513, 0.6043 respectively for QRM and quadratic regression. Under $\lambda = 0.01$, by $\mathrm{dL} = 1.38, \mathrm{dU} = 1.45$, we know 2.513 for QRM is in the correlation rejection region $[1.45, 2.55]$, and 0.6043 for quadratic regression is in the positive correlation acceptance region is $[0, 1.38]$.

4) Random errors of QRM are approximately independent and the expectation value is always zero, but quadratic regression is not. i) All the residuals of QRM are uniformly distributed around $e = 0$ but quadratic regression is not, please see Figure 4; ii) According to the variation range of controllable variables $[55, 64]$, $[64, 81]$, $[81, 100]$, the total sample can be divided into three subsamples, the corresponding residual histograms of QRM is similar in Symmetry on $e = 0$, while the corresponding residual histograms of quadratic regression is obviously different, please see Figure 4.

All the above analysis indicates that the QRM possesses good interpretability and structure. It is a general regression model. The satisfactory fitting effect and error estimate can be obtained through adjusting the freedom degree of the model.

The above analysis results can provide practical value for power department to forecast load and to formulate electric power dispatch program, so that realize shift peak load, gradually relieve the tight situation of electricity supply.

7. **Conclusion.** In this paper, by analyzing the characteristic and weakness of the existing regression methods, using the concept of quasi-linear function and approximation properties, we establish quasi-linear regression model (denoted by QRM for short). Further we consider the parameter estimation and give the operating strategy based on genetic algorithm and the least squares method. For the feature testing of random errors, we give the basic criteria based on residual analysis and statistical principle, propose the variance estimation based on subsample. Finally, we discuss the performance of QRM by an illustrative example. All these indicate that QRM not only possesses generality and operability, but also results in satisfactory results through adjusting the freedom degree of quasi-linear regression function, for example, making random errors of QRM $\varepsilon(x)$ be approximately independent and obey normal distribution $N(0, \sigma^2)$. Accordingly, the results in this paper enrich the current regression theories and methods, and can be widely used in many fields such as artificial intelligence and economic management.

Acknowledgment. This research is supported by the National Natural Science Foundation of China (71071049) and the Natural Science Foundation of Hebei Province (F2011208 056).

REFERENCES

[1] B. Parthasarathy, A. A. Munot and D. R. Kothawale, Regression model for estimation of indian food grain production from summer monsoon rainfall, *Agricultural and Forest Meteorology*, vol.42, no.2-3, pp.167-182, 1988.

[2] L. Yang and Z. Liu, Application of linear regression model in predicting the demand in logistics, *Culture of Business*, vol.10, pp.173-175, 2007.

[3] L. R. Schaeffer, Application of random regression models in animal breeding, *Livestock Production Science*, vol.86, no.1-3, pp.35-45, 2004.

[4] Y. Liu, Y. Shou, J. Xu et al., Estimation for drug penetration parameters using a nonlinear regression model, *Journal of Biomedical Engineering of China*, vol.22, no.1, pp.37-42, 2003.

[5] W. Bich, G. D'agostino et al., 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*, pp.16-18, 2007.

[6] K. V. Kumar, K. Porkodi and F. Rocha, Isotherms and thermodynamics 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*, vol.151, no.2-3, pp.794-804, 2008.

[7] B. S. Soumya, M. Sekhar, J. Riotte et al., Non-linear regression model for spatial variation in precipitation chemistry for South India, *Atmospheric Environment*, vol.43, no.5, pp.1147-1152, 2009.

[8] K. V. 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, no.5, pp.1198-1202, 2006.

[9] Y. Nishihara, J. Irie, T. Yamaguchi, T. Yamazaki and K. Inoue, The statistical estimation method of muscle activity based on surface EEG, *ICIC Express Letters, Part B: Applications*, vol.2, no.3, pp.603-608, 2011.

[10] Z. Zhou, Analysis and research about the model of the rational approximation of regression function, *Mathematics in Practice and Theory*, vol.34, no.7, pp.113-117, 2004.

[11] Z. Wang and Z. Feng, On the problem and improved measure in some quasi-linearization regression with one argument, *The Journal of Northeast Normal University*, vol.38, no.4, pp.45-52, 2006.

[12] Y. Xie, Sectional line fitting by means of least sequence method, *Journal of Nanchang Institute of Aeroautical Technology*, vol.1, pp.19-24, 1992.

[13] Z. Zhang and H. Wang, On the weighted quasi-linear regression, *Journal of Central University of Financial and Economics*, vol.3, pp.110-112, 1992.

[14] X. Chen, Least absolute regression, *Application of Statistics and Management*, vol.1, pp.48-55, 1989.

[15] F. Li, C. Jin and L. Liu, Quasi-linear fuzzy number and its application in fuzzy programming, *Systems Engineering – Theory and Practice*, vol.29, no.4, pp.119-127, 2009.
[16] W. Zhang and Y. Liang, *Mathematical Foundation for Genetic Algorithm*, 2nd Edition, Xi'an Jiaotong University Press, Xi'an, 2003.
[17] D. Y. Ge, X. F. Yao, S. S. Jiang and Y. Z. Gu, On hand-eye calibration of manipulator with orthogonal neural network and genetic algorithm, *ICIC Express Letters, Part B: Applications*, vol.2, no.4, pp.899-904, 2011.