ICIC Express Letters
Volume **10**, Number **4**, April **2016**

ICIC International ©2016 ISSN 1881-803X
pp. **985–993**

RESEARCH ON SYNTHESIZING EFFECT-BASED REGRESSION METHOD

Fachao Li, Xiaoxiao Su and Chenxia Jin

School of Economics and Management
Hebei University of Science and Technology
No. 26, Yuxiang Street, Shijiazhuang 050018, P. R. China
lifachao@tsinghua.org.cn; 1301141201@qq.com; jinchenxia2005@126.com

Received September 2015; accepted December 2015

Abstract. *Regression analysis is a common prediction method. Determining the regression equation with high precision is a core of regression analysis, which has been the hot research content in the application and academia. However, the existing regression methods did not systematically consider the reliability of sample. In the paper, we firstly analyze the characteristics of reliability, put forward the sample description system by regarding reliability as the auxiliary index, and give the concept of basic effect function which reflects the reliability. Then we establish a regression model based on synthesizing effect (denoted by BSE-RM), and further analyze the characteristics of BSE-RM from theory and application. The results show that BSE-RM not only has good structure characteristics and interpretability, but also extends and perfects the existing regression analysis methods.*

Keywords: Regression analysis, Basic effect function, Regression function, Reliability, Prediction

1. Introduction.

Regression analysis is a statistical tool used for finding the relationship between one variable and another variables. It quantizes the relationship which exists between these variables. It has been successfully used to solve many management and prediction problems. For example, [1] started from the factors which may influence the airport passenger throughput, and made an analysis on the correlation between each influence factor and the throughput, and then applied this method to the airport in southwest; [2] builded a sensor-based forecasting model using support vector regression and applied it to an empirical data set from a multi-family residential building in New York City. For the problem that support vector regression has deficiency to solve the problem that highly nonlinear characteristics appear in the electric load forecasting, [3] proposed a chaotic particle swarm optimization algorithm and provided a theoretical exploration of the electric load forecasting support system; [4] firstly applied support vector machines to regression prediction of stock index futures; [5] presented a robust hourly cooling-load forecasting method based on time-indexed autoregressive with exogenous inputs models, in which the coefficients are estimated through a two-stage weighted least squares regression; [6] applied Gaussian process regression (GPR) to probabilistic stream flow forecasting; [7] gave an analysis for the influence to pedestrian crossing delay made by right-turn cars and models. The relationship between independent variables and dependent variables was fitting to linear regression model and a multiple linear regression model based on the observed data. People in the study found new forecasting methods and models constantly to perfect the existing ones; [8] presented the least squares regression based on some basic assumptions, and any deviation hypothetical situation will affect the regression results. The article discussed the main problems in the regression analysis deviating from the basic assumption which may affect the results of regression analysis and gives the corresponding ways to find and remedy problems; [9] proposed a new criterion based on weighted error to evaluate the prediction methods and verified its superiority; [10] established the dynamic exponential smoothing method to predict the processor running time in grid environment; [11] builded a regression model based on the quasi linear function (QRM), and discussed the parameter estimation of QRM strategies, and this paper gives the parameters estimation method based on the genetic algorithm and the least squares estimation method, and the error test method based on the residual.

From the above reviews, we know that the current regression analysis theory has tended to be mature. Its research focuses on the selection of regression function, but it is worth noting that these results hold under the precondition that the sample data are completely reliable. In real life, completely reliable data is difficult to obtain, which makes the sample data cannot satisfy the assumption of classical regression. Therefore, the mechanism of the classical regression model has many defects. In addition, the innovation of the existing prediction methods mostly concentrates on improving the prediction accuracy, and gives less consideration on data itself; while the basis of reliable prediction is the reliability of the data. Therefore, it is necessary to make up for the shortcomings of the classical regression method. In this paper, for the shortcomings of the current regression, we mainly do the work as follows. In Section 2, we analyze three defects of the classical regression model through describing the characteristic. In Section 3, for the first shortcoming, we add the reliability dimension to the general description of the samples; for the second shortcoming, we propose the concept of basic effect function; then we establish a regression model based on the sample effect. In Section 4, we analyze the characteristics of this model. In Section 5, using a concrete case, we verified the two methods, and make a difference between them. Finally, conclusions are derived in Section 6.

2. **Characteristic Analysis of Regression Problems.** Regression analysis is a kind of observation data-based statistical method, which is used to find the relationship between explanatory variable and explained variable on the basis of a certain hypothesis, and the basic form is:

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

Here, x denotes explanatory variable, y denotes explained variable, $\mu(x)$ (called ***regression function***, and it denotes the mathematical expectation of $y(x)$ intuitively) is the deterministic relationship of x, and $\varepsilon(x)$ is the error term.

Current regression analysis theories are mostly established on the basis of $\varepsilon(x)$ obeying normal distribution $N(0, \sigma^2)$, and the basic process is as follows.

Step 1 Determine the fundamental regression function $\mu(x)$ according to the scatter diagram distribution characteristics of sample $Z = \{(x_i, y_i)|\ i = 1, 2, \cdots, n\}$ (for example, linear function, quadratic function).

Step 2 Determine the parameters value of regression function $\mu(x)$ combining with the given sample points and least square method (that is, $\min \sum_{i=1}^{n} [y_i - \mu(x_i)]^2$), and then get the estimates $\hat{\mu}(x) \triangleq \hat{y}(x, Z)$.

Step 3 Test the rationality of regression function $\hat{y}(x, Z) = \hat{\mu}(x)$ under a certain reliability.

It is easy to see that the above processes are established based on the assumption that the sample data are all reliable (that is, sample inconsistencies are caused by randomness). However, collecting samples in the actual problem is often influenced by many subjective and objective factors, and different sample's reliability is often different. Due to the fact that the reliability of regression equation depends on the reliability of the samples (the higher (lower) the reliability of the sample data is, the higher (lower) the reliability of corresponding regression results is), the current methods often cannot be directly used for practical problems. Its shortcomings can be summarized as the following three aspects.

1) **The description way of the sample data is not perfect.** The current data description only contains the observed values of explanatory variables x_i and explained

variables y_i, and lacks the reliability description of (x_i, y_i). However, the reliability of the sample is closely related to the reliability of the regression equation. Thus, if r_i denotes the reliability of the observed value (x_i, y_i), we can reasonably describe the sample as (x_i, y_i, r_i).

2) **The mechanism of sample effect is not perfect.** Different samples with different reliability have a different effect (or utility) in the regression process, therefore, the established regression equation should be a synthesis of a series of regression equations under different reliability sample data sets.

3) **The description system of regression results is not perfect.** The current description way of the regression results is hypothesis-test based on the random theory. It does not consider the reliability description of regression results caused by the sample reliability.

The above analyses show that the existing regression methods need be improved further. In the following, we will focus on the shortcomings of current regression methods and discuss the synthesizing effect-based regression model.

3. **Synthesizing Effect-Based Regression Model.** In this section, surrounding the effect characteristics of the sample reliability, we will discuss the construction strategy of the regression model based on the sample effect. For convenience, 1) $\Omega = \{(x_i, y_i, r_i)|i = 1, 2, \cdots, n\}$ denotes the sample data set (here, (x_i, y_i) denotes the observed values of explanatory variables and explained variables and r_i denotes the reliability of (x_i, y_i)); 2) $\Omega_\lambda = \{(x_i, y_i)|(x_i, y_i, r_i) \in \Omega \text{ and } r_i = \lambda\}$ denotes the sample data set whose reliability is λ; 3) $\hat{y}(x, \Omega)$ and $\hat{y}(x, \Omega_\lambda)$ are corresponding shortcomings for the regression equation based on the data set Ω and Ω_λ.

It is easy to see, $\hat{y}(x, \Omega_\lambda)$ (the regression equation of the classical sense) has the exact meaning. However, $\hat{y}(x, \Omega)$ varies with the reliability of the sample. For the sample data set $\Omega = \{(x_i, y_i, r_i)|\ i = 1, 2, \cdots, n\}$, if all the subsets with the various reliability: $\Omega_{\lambda_1}, \Omega_{\lambda_2}, \cdots, \Omega_{\lambda_m}$ contain enough elements (here, $\lambda_1, \lambda_2, \cdots, \lambda_m$ denote all of the different values of $r_1, r_2, \cdots, r_n$ and meet $\lambda_1 < \lambda_2 < \cdots < \lambda_m$), we can understand $\Omega_{\lambda_1}, \Omega_{\lambda_2}, \cdots, \Omega_{\lambda_m}$ as a kind of decomposition of Ω and $\{\hat{y}(x, \Omega_{\lambda_k})|k = 1, 2, \cdots, m\}$ as the basic factors reflecting the local feature of $\hat{y}(x, \Omega)$. So, regression problem $\hat{y}(x, \Omega)$ can be interpreted as a synthesis problem of $\{\hat{y}(x, \Omega_{\lambda_k})|\ k = 1, 2, \cdots, m\}$.

In the prediction problem, the higher (lower) the reliability of the sample is, the larger (smaller) the credibility of the corresponding regression equation is. So in the process of synthesizing of $\{\hat{y}(x, \Omega_{\lambda_k})|k = 1, 2, \cdots, m\}$, the effect of $\hat{y}(x, \Omega_{\lambda_k})$ increases with λ_k. If $W(\lambda)$ denotes the effect of $\hat{y}(x, \Omega_{\lambda_k})$, we can understand $W(\lambda)$ as a mapping (***called basic effect function***) from $[0, 1]$ to $[0, +\infty)$ and $W(\lambda)$ should satisfy the following principles.

Principle 1: The effect of λ is monotonic, that is, $W(\lambda)$ is monotonic non-decreasing on $[0, 1]$.

Principle 2: The effect of λ is continuous, that is, $W(\lambda)$ is continuous on $[0, 1]$.

Principle 3: The effect of λ is existence, that is, $W(\lambda) > 0$ always holds for any $\lambda \in (0, 1]$.

Principle 4: The effect of λ is normalization, that is, $W(0) = 0$, $W(1) = 1$ always hold.

Here, Principle $1 \sim 3$ must be satisfied, and they respectively correspond to the following facts in real problems: 1) the higher the sample reliability is, the larger the credibility of the corresponding regression equation is; 2) when the sample data reliability changed little, so does the credibility of its corresponding predicted results; 3) when sample data has certain reliability, its corresponding predicted results have a certain credibility. Principle 4 is set to maintain a consistence with conventional processing models. $W(0) = 0$ can be better reflecting the intuitive fact that when sample data is completely unreliable, its corresponding predicted results cannot be trusted. It is easy to verify:

$$W_1(\lambda) = \lambda^\alpha, \ 0 < \alpha < \infty, \tag{2}$$

$$W_2(\lambda) = Q_L((a_0, c_0), (a_1, c_1), \cdots, (a_n, c_n))(\lambda) \tag{3}$$

are basic effect functions. Here, $0 = a_0 < a_1 < \cdots < a_n = 1$, $0 = c_0 < c_1 \leq c_2 \leq \cdots \leq c_n = 1$, $Q_L((a_0, c_0), (a_1, c_1), \cdots, (a_n, c_n))$ denotes the following rules (***called quasi-linear corresponding rules*** based on $(a_0, c_0), (a_1, c_1), \cdots, (a_n, c_n)$): when $a_{k-1} \leq x \leq a_k$, $Q_L((a_0, c_0), (a_1, c_1), \cdots, (a_n, c_n))(x) = c_{k-1} + (c_k - c_{k-1})(x - a_{k-1})/(a_k - a_{k-1})$, $k = 1, 2, \cdots, n$.

For a given basic effect function $W(\lambda)$, if we regard $W(\lambda_1), W(\lambda_2), \cdots, W(\lambda_m)$ as the basic factor of the effectiveness value of $\hat{y}(x, \Omega_{\lambda_1})$, $\hat{y}(x, \Omega_{\lambda_2}), \cdots, \hat{y}(x, \Omega_{\lambda_m})$, then

$$\sum_{k=1}^{m} w_k \cdot \hat{y}(x, \Omega_{\lambda_k}) \triangleq \hat{y}(x, \Omega \oplus W(\lambda)) \tag{4}$$

is a systematic comprehensive model of $\{\hat{y}(x, \Omega_{\lambda_k})|\ k = 1, 2, \cdots, m\}$ by considering the effect of sample (called the ***regression model based on synthesizing effect***, shorted for **BSE-RM**). Here, $w_k = W(\lambda_k)/\sum_{i=1}^{m} W(\lambda_i)$, $k = 1, 2, \cdots, m$.

Obviously, 1) $W(\lambda)$ is a kind of parameter reflecting decision-making notion. In fact, it is a kind of processing mechanism on uncertain information. Its purpose is to quantify the importance of the reliability. 2) $\hat{y}(x, \Omega \oplus W(\lambda))$ varies with $W(\lambda)$. The inconsistency is caused by the fact that there do not exist generally accepted uncertainty processing methods. 3) The selection of $W(\lambda)$ is the core problem of (4). In practice, the concrete form of the $W(\lambda)$ should be determined by the characteristics of the prediction problem, the use of the regression results and so on. 4) Different $W(\lambda)$ reflects different decision-making consciousness, and even the difference is remarkable. For $W(\lambda) = \lambda^\alpha$, α is the parameter which concentratively describes processing consciousness on reliability. Its role characteristics are stated as follows: a) when $\alpha = 1$, the effect of the reliability increases linearly along with the increase of reliability; b) when $\alpha \neq 1$, although the effect of the reliability still increases with the increase of reliability, the change way will occur fundamentally, and the smaller α is, the smaller the difference effect of the reliability is (especially, when $\alpha \to 0$, $\lambda \in (0, 1]$, $\lambda^\alpha \longrightarrow 1$ will always hold, and this implies that the effect of various reliability will tend to be 1). The larger α is, the more remarkable the core status of the effect reliability will be.

4. The Characteristic Analysis of BSE-RM.

In Section 3, we analyzed the characteristics of reliability and proposed the regression model of BSE-RM based on samples effect. In this section we will further discuss the value rule of $\hat{y}(x, \Omega \oplus W(\lambda))$ from the angle of quantification.

Theorem 4.1. *Let $\Omega = \{(x_i, y_i, r_i)|i = 1, 2, \cdots, n\}$ denote the sample data set, and $\Omega_{\lambda_1}, \Omega_{\lambda_2}, \cdots, \Omega_{\lambda_m}$ denote subsets with all the various reliability in Ω. Then $\hat{y}(x, \Omega \oplus W_1(\lambda)) = \hat{y}(x, \Omega \oplus W_2(\lambda))$ always holds if and only if $W_1(\lambda)$ and $W_2(\lambda)$ are $r_1 = r_2 = \cdots = r_n = r$ or $\hat{y}(x, \Omega_{\lambda_1}) = \hat{y}(x, \Omega_{\lambda_2}) = \cdots = \hat{y}(x, \Omega_{\lambda_m})$.*

Proof: Sufficiency. By (4) and the definition of the basic effect function, we can know: 1) When $r_1 = r_2 = \cdots = r_n = r$, $\hat{y}(x, \Omega_{\lambda_1}) = \hat{y}(x, \Omega_{\lambda_2}) = \cdots = \hat{y}(x, \Omega_{\lambda_m})$, always holds for any $W(\lambda)$; 2) When $\hat{y}(x, \Omega_{\lambda_1}) = \hat{y}(x, \Omega_{\lambda_2}) = \cdots = \hat{y}(x, \Omega_{\lambda_m})$, $\hat{y}(x, \Omega \oplus W(\lambda)) = \sum_{k=1}^{m} w_k \hat{y}(x, \Omega_{\lambda_k}) = \hat{y}(x, \Omega_{\lambda_1}) \sum_{k=1}^{m} w_k = \hat{y}(x, \Omega_{\lambda_1})$ always holds for any $W(\lambda)$. That is to say, when $r_1 = r_2 = \cdots = r_n = r$ or $\hat{y}(x, \Omega_{\lambda_1}) = \hat{y}(x, \Omega_{\lambda_2}) = \cdots = \hat{y}(x, \Omega_{\lambda_m})$, $\hat{y}(x, \Omega \oplus W_1(\lambda)) = \hat{y}(x, \Omega \oplus W_2(\lambda))$ always holds for any $W_1(\lambda)$ and $W_2(\lambda)$.

Necessity. In the following, we prove the necessity. Suppose $r_1 = r_2 = \cdots = r_n = r$ and $\hat{y}(x, \Omega_{\lambda_1}) = \hat{y}(x, \Omega_{\lambda_2}) = \cdots = \hat{y}(x, \Omega_{\lambda_m})$ do not hold, then there will be $t_1, t_2 \in \{\lambda_1, \lambda_2, \cdots, \lambda_m\}$ making $t_1 \neq t_2$ and $\hat{y}(x, \Omega_{t_1}) \neq \hat{y}(x, \Omega_{t_2})$ hold at the same time, and we can assume $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m$ and $\hat{y}(x, \Omega_{\lambda_1}) \neq \hat{y}(x, \Omega_{\lambda_2})$. In the following, we verify that $\hat{y}(x, \Omega \oplus W_1(\lambda)) = \hat{y}(x, \Omega \oplus W_2(\lambda))$ does not always hold for any $W_1(\lambda)$ and $W_2(\lambda)$ in two cases.

Case 1. If $m = 2$, then for $W_1(\lambda) = \lambda$ and $W_2(\lambda) = Q_L((0,0),(\lambda_1,1),(1,1))(X)$, we have $W_1(\lambda_1) = \lambda_1$, $W_1(\lambda_2) = \lambda_2$, $W_2(\lambda_1) = W_2(\lambda_2) = 1$, $\hat{y}(x, \Omega \oplus W_1(\lambda)) = \lambda_1 \hat{y}(x, \Omega_{\lambda_1})/(\lambda_1 + \lambda_2) + \lambda_2 \hat{y}(x, \Omega_{\lambda_2})/(\lambda_1 + \lambda_2)$, $\hat{y}(x, \Omega \oplus W_2(\lambda)) = (\hat{y}(x, \Omega_{\lambda_1}) + \hat{y}(x, \Omega_{\lambda_2}))/2$, $\hat{y}(x, \Omega \oplus W_1(\lambda)) - \hat{y}(x, \Omega \oplus W_2(\lambda)) = (\lambda_1 - \lambda_2)[\hat{y}(x, \Omega_{\lambda_1}) - \hat{y}(x, \Omega_{\lambda_2})]/[2(\lambda_1 + \lambda_2)] \neq 0$.

Case 2. If $m > 2$, we can assume $m = 4$ (the rest should be considered in the same way), and then for $W_1(\lambda) = Q_L((0,0),(\lambda_1,0.2),(\lambda_2,0.8),(\lambda_3,1),(1,1))(\lambda)$ and $W_2(\lambda) = Q_L((0,0),(\lambda_1,0.4),(\lambda_2,0.6),(\lambda_3,1),(1,1))(\lambda)$, we have $W_1(\lambda_1) = 0.2$, $W_1(\lambda_2) = 0.8$, $W_1(\lambda_3) = W_1(\lambda_4) = 1$, $W_2(\lambda_1) = 0.4$, $W_2(\lambda_2) = 0.6$, $W_2(\lambda_3) = W_2(\lambda_4) = 1$, $\hat{y}(x, \Omega \oplus W_1(\lambda)) = 0.2\hat{y}(x, \Omega_{\lambda_1})/3 + 0.8\hat{y}(x, \Omega_{\lambda_2})/3 + \hat{y}(x, \Omega_{\lambda_3})/3 + \hat{y}(x, \Omega_{\lambda_4})/3$, $\hat{y}(x, \Omega \oplus W_2(\lambda)) = 0.4\hat{y}(x, \Omega_{\lambda_1})/3 + 0.6\hat{y}(x, \Omega_{\lambda_2})/3 + \hat{y}(x, \Omega_{\lambda_3})/3 + \hat{y}(x, \Omega_{\lambda_4})/3$, $\hat{y}(x, \Omega \oplus W_1(\lambda)) - \hat{y}(x, \Omega \oplus W_2(\lambda)) = -0.2(\hat{y}(x, \Omega_{\lambda_1}) - \hat{y}(x, \Omega_{\lambda_2}))/3 \neq 0$.

According to Theorem 4.1, we know that: 1) if we set the reliability of the samples to be 1 in BSE-RM, (4) will be the regression function of current methods, and this implies that (4) is a generalization of the existing regression method; 2) only when the reliability of the sample is not completely the same, can the effect of the samples be reflected; 3) the different reliability of sample data sets can get the same regression function.

In regression problems, the number of samples should be enough. Therefore, the above discussion is only a thought method, and it cannot be simply used in practice. We can construct BSE-RM by the following steps (here, $\Omega_{\lambda_1}, \Omega_{\lambda_2}, \cdots, \Omega_{\lambda_m}$ denotes the various reliability of the data set and $\lambda_1 < \lambda_2 < \cdots < \lambda_m$; $|\Omega_{\lambda_k}|$ denotes the number of samples in Ω_{λ_k}):

Step 1 Based on the requirement on the least sample size, we should integrate $\Omega_{\lambda_1}, \Omega_{\lambda_2}, \cdots, \Omega_{\lambda_m}$ into $\cup_{k=1}^{m_1}\Omega_{\lambda_k}, \cup_{k=m_1+1}^{m_2}\Omega_{\lambda_k}, \cdots, \cup_{k=m_{k-1}+1}^{m}\Omega_{\lambda_k}$.

Step 2 Based on the sample size, determine the comprehensive reliability $\alpha_1, \alpha_2, \cdots, \alpha_s$ of the $\cup_{k=1}^{m_1}\Omega_{\lambda_k}, \cup_{k=m_1+1}^{m_2}\Omega_{\lambda_k}, \cdots, \cup_{k=m_{k-1}+1}^{m}\Omega_{\lambda_k}$, that is

$$\alpha_i = \sum_{k=m_i-1}^{m_i} \frac{|\Omega_{\lambda_k}|}{|\cup_{k=m_{i-1}+1}^{m_i} \Omega_{\lambda_k}|} \cdot \lambda_k, \quad (\text{here}, i = 1, 2, \cdots, s;\ m_0 = 1,\ m_s = m). \tag{5}$$

Step 3 Based on $\Omega_{\alpha_1} = \cup_1^{m_1}\Omega_{\lambda_k}, \Omega_{\alpha_2} = \cup_{k=m_1+1}^{m_2}\Omega_{\lambda_k}, \cdots, \Omega_{\alpha_s} = \cup_{m_{s-1}+1}^{m}\Omega_{\lambda_k}$, determine $\hat{y}(x, \Omega \oplus W(\lambda))$ by using (4) and (5).

5. **Application Example.** In this section, we will combine a concrete case to further illustrate the effectiveness and the specific implementation process of BSE-RM.

Case description: In investment market, the benefits and risks coexist, and market environment directly affects the correlation relationship between them. How to develop an investment scheme is important in academics and applications. In order to develop a better investment plan, a financial investment company decided to summarize the correlation between investment rate and investment risk based on the previous investment records. The data of 192 investment results are shown in Table 1. Here, x_i denotes the risk value of one investment; y_i denotes real income rate of one investment; z_i denotes expected income of one investment; $r_i = |y_i - z_i|$ denotes the deviation of real income and expected income; $r_i^* = 1 - r_i/\max\{r_1, r_2, \cdots, r_n\}$ denotes the satisfaction degree.

Since the deviation of real income and expected income is caused by investment risk, $r_i = |y_i - z_i|$ can be regarded as a quantitative index describing the investment risk, which reflects the satisfaction degree on the investment plan. This shows that the correlation of investment rate and investment risk can be summarized as a regression problem based on data sample $\{(x_i, y_i, r_i^*)|i = 1, 2, \cdots, n\}$. Combining with several different satisfaction processing methods, we will use BSE-RM to determine the regression function of investment rate on investment risk. The specific process is stated as follows.

Step 1 Take the first 180 data in Table 1 as the regression sample set Ω, and the last 12 as the test sample data.

Step 2 Divide Ω into several sub-sample data set $\{\Omega_0, \Omega_{0.1}, \Omega_{0.2}, \Omega_{0.3}, \Omega_{0.4}, \Omega_{0.5}, \Omega_{0.6}, \Omega_{0.7}, \Omega_{0.9}, \Omega_1\}$, according to the sample satisfaction.

Step 3 We integrate $\{\Omega_0, \Omega_{0.1}, \Omega_{0.2}, \Omega_{0.3}, \Omega_{0.4}, \Omega_{0.5}, \Omega_{0.6}, \Omega_{0.7}, \Omega_{0.9}, \Omega_1\}$ into $\{\Omega_0 \cup \Omega_{0.1} \cup \Omega_{0.2}, \Omega_{0.3} \cup \Omega_{0.4} \cup \Omega_{0.5}, \Omega_{0.6} \cup \Omega_{0.7}, \Omega_{0.9} \cup \Omega_1\}$, according to the regression sample not less than 30.

Step 4 Compute the comprehensive satisfaction of the integrated four sample data sets: $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} \triangleq \{0.160, 0.411, 0.676, 0.951\}$.

Step 5 Determine the form of the regression function of each group by combining with scatter plots of $\Omega_0 \cup \Omega_{0.1} \cup \Omega_{0.2}, \Omega_{0.3} \cup \Omega_{0.4} \cup \Omega_{0.5}, \Omega_{0.6} \cup \Omega_{0.7}, \Omega_{0.9} \cup \Omega_1$; then use the least square method to obtain the regression function of each group: $\hat{y}(x, \Omega_{\lambda_k})$, $k = 1, 2, 3, 4$. The specific regression functions and their test values are shown in Table 2 (here, we assume that the total residual sum of squares is $S_T = \sum(y_i - \overline{y})^2$, the residual sum of squares is $S_R(\hat{\mu}(x)) = \sum(y_i - \hat{\mu}(x_i))^2$ and the regression sum of squares is $S_e(\hat{\mu}(x)) = \sum(\hat{\mu}(x_i) - \overline{y})^2$; $R^2(\hat{\mu}(x)) = 1 - S_R(\hat{\mu}(x))/S_T$ is goodness of fit; $F(\hat{\mu}(x)) = S_e(\hat{\mu}(x))/[S_R(\hat{\mu}(x))/(n-2)]$ is F test value).

Step 6 Determine the regression model based on synthesizing effect: $\hat{y}(x, \Omega \oplus W(\lambda))$, combined with the given $W(\lambda)$ and $\hat{y}(x, \Omega_{\lambda_k}), k = 1, 2, 3, 4$ and Formula (4). In Table 3, we give the regression functions under several concrete $W(\lambda)$.

From Table 2, we can know that the R^2 and the F of $\hat{y}(x, \Omega_{\lambda_k})$, $k = 1, 2, 3, 4$ are better than that of classic regression function $\hat{y}(x, \Omega_{\lambda_\infty})$. This shows that reasonable layer can improve the quality of regression function and eliminate the shortage that regression model is difficult to determine in the classical regression methods in a sense (for example, in this case, although all samples scatter plot is roughly as parabola, there is part of the sample clearly presenting linearity. So it is not reasonable to fit it only with parabolic or linear).

In order to demonstrate performance of BSE-RM, we will analyze the regression functions in Table 3 with the selected 12 samples and the test indexes: prediction precision and predicted residual sum of squares. The selected sample values, the predicted values of the functions and the precision and the predicted residual sum of square of the regression functions are shown in Table 4 (here, $Q(\hat{\mu}(x)) = 1 - \sum_{i=1}^{12}(y_i - \hat{\mu}(x_i))^2 / \sum_{i=1}^{n} y_i^2$ denotes the precision of $\hat{\mu}(x)$ and $S_R(\hat{\mu}(x)) = \sum_{i=1}^{12}(y_i - \hat{\mu}(x_i))^2$ denotes the predicted residual sum of square of $\hat{\mu}(x)$).

From Table 4 we can know: 1) The prediction precision values of $\hat{\mu}_i$ $(i = 0, 1, 2, 3, 4)$ are all greater than the universal threshold 0.85. This shows that $\hat{\mu}_i$ $(i = 0, 1, 2, 3, 4)$ are all feasible as the basis of prediction; 2) The S_R of $\hat{\mu}_i$ $(i = 1, 2, 3, 4)$ is much better than that of the $\hat{\mu}_0$. Namely, the information of the BSE-RM is more comprehensive and closer to the real value.

All of above analysis and discussion show that: 1) On the basis of sample reliability research, prediction method based on the data has extensive practical background; 2) BSE-RM has good structure characteristics and interpretability. The basic idea of BSE-RM has reference significance to many complex decision problems (namely, we can group the sample data according to certain strategy, which can reduce the computational complexity and avoid the problem that regression model is difficult to determine to a certain extent); 3) Regression function is different under different decision-making consciousness (that is the basic effect function), which dovetails beautifully with the realistic decision problem.

Table 1. The sample data information: $x_i(\%)$, $y_i(\%)$, $z_i(\%)$, r_i, and r_i^*

x_i	11.1	11.3	12.5	12.7	13.4	13.6	15.6	17.1	18.4	19.9	15.3	15.5	16.1	16.3	16.7	29.5
y_i	16.3	16.1	16.7	17.1	17.7	18.1	18.3	19.5	20.3	22.1	18.8	19.2	20.1	20.3	20.4	23.6
z_i	17.2	17.1	17.6	18	18.5	18.9	19.2	20.3	21.2	22.8	19.4	19.7	20.9	20.9	21.2	16.7
r_i	0.9	1	0.9	0.9	0.8	0.8	0.9	0.8	0.9	0.7	0.6	0.5	0.8	0.6	0.8	6.9
r_i^*	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	1	0.9	0.9	0.9	0.3
x_i	16.8	17.2	17.5	17.7	18.4	18.6	19.5	19.6	21.3	19.9	20.6	21	21.4	21.6	21.5	29.3
y_i	20.1	20.3	22.1	22.2	20.5	20.6	21.2	21.1	21.4	21.1	21.7	21.4	21.6	21.7	21.8	23.8
z_i	20.8	20.6	23.2	23.6	21.8	21.7	22.5	20.8	20.8	20.5	20.6	20.5	20.3	20.3	20.5	17
r_i	0.7	0.3	1.1	1.4	1.3	1.1	1.3	0.3	0.6	0.6	1.1	0.9	1.3	1.4	1.3	6.8
r_i^*	0.9	0.9	1	0.9	0.9	0.9	0.9	0.9	1	0.9	0.9	1	0.9	0.9	0.9	0.3
x_i	21.7	21.9	21.9	22.4	22.7	23.2	23.6	24	24.6	25.7	26.3	26.7	27.2	27.8	28.6	28.9
y_i	22	21.9	21.9	22.2	22.5	22.7	22.8	22.6	22.8	23.1	22.7	22.8	22.5	22.2	22	24.3
z_i	20.8	20.5	21.2	21.3	21.1	21.4	21.6	21.5	22.2	22.7	22.4	21.9	22	21.8	21.3	17.6
r_i	1.2	1.4	0.7	0.9	1.4	1.3	1.2	1.1	0.6	0.4	0.3	0.9	0.5	0.4	0.7	6.7
r_i^*	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	1	1	0.9	1	1	0.3
x_i	22.5	23.5	23.8	24.2	35.9	35.1	35.3	35.6	35.7	35.8	36.5	36.6	37.5	37.6	37.7	33.7
y_i	24.4	25	25.1	25.3	29.1	13.1	12.6	13.3	13.2	13	12.2	12.1	11	10.9	10.8	17.2
z_i	21.6	27.9	28	28.4	18	20	19.4	20	20.7	20.3	19.6	18.3	17.1	18.1	17.9	8.8
r_i	2.8	2.9	2.9	3.1	2.9	6.9	6.8	6.7	7.5	7.3	7.4	6.2	6.1	7.2	7.1	8.4
r_i^*	0.7	0.7	0.7	0.7	0.7	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.3	0.3	0.2
x_i	37.9	38.1	38.3	38.7	39.2	38.1	38.3	38.4	38.6	38.7	38.9	40	40.2	40.3	40.7	33.5
y_i	10.5	10.3	10	9.6	9	13.4	15.3	14.2	14.5	14.5	13.9	12.1	12.9	12.9	12.5	17.3
z_i	17.4	17.1	16.3	16.2	15.2	20.1	21.7	20.8	19.6	20.5	19	17.9	18.2	18.4	18	8.1
r_i	6.9	6.8	6.3	6.6	6.2	6.7	6.4	6.6	5.1	6	5.1	5.8	5.3	5.5	5.5	9.2
r_i^*	0.3	0.3	0.4	0.3	0.4	0.3	0.4	0.3	0.5	0.4	0.5	0.4	0.5	0.5	0.5	0.1
x_i	25.1	25.2	25.4	25.7	26.3	26.4	26.6	26.9	27.4	27.5	27.7	28.2	28.3	28.4	28.6	33.3
y_i	28.9	28.8	28.5	28.2	27.4	27.3	27.1	26.7	26.1	26	25.8	25.2	25	24.9	24.7	17.6
z_i	34.2	34.1	34	33.6	32.8	32.5	32.3	31.8	21	21	20.9	20.4	19.9	19.7	19.8	9.7
r_i	5.3	5.3	5.5	5.4	5.4	5.2	5.2	5.1	5.1	5	4.9	4.8	5.1	5.2	4.9	7.9
r_i^*	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.2
x_i	31	31.2	31.7	31.8	32.6	32.9	33.4	38.8	39.2	39.5	39.8	40.1	40.6	40.9	11.5	32.8
y_i	21.5	21	20.8	20.5	20.2	19.8	19.3	13	12.2	12.9	11.8	11.5	11.1	10.9	18.1	17.8
z_i	20.8	20.4	20.1	19.4	18.9	18.4	17.9	12.5	11.5	11.6	10.7	12.6	12.2	12.1	21.7	8.7
r_i	0.7	0.6	0.7	1.1	1.3	1.4	1.4	0.5	0.7	1.3	1.1	1.1	1.1	1.2	3.6	9.1
r_i^*	0.9	0.9	0.9	0.9	0.9	0.9	0.9	1	0.9	0.9	0.9	0.9	0.9	0.9	0.6	0.1
x_i	11.8	12	11.6	11.7	12.1	12.3	13.4	14.5	15.3	16.9	17.4	18.3	19.1	19.6	20	32.5
y_i	18.5	19	18.4	18.5	18.7	18.8	19.5	20.1	20.6	21.5	21.8	22.4	22.9	23.1	23.9	17.6
z_i	22.8	23.3	21.3	21.7	22.2	22.1	22.6	22.9	23.5	24.6	25.3	25.3	26.5	26.6	27.3	9.2
r_i	4.3	4.3	2.9	3.2	3.5	3.3	3.1	2.8	2.9	3.1	3.5	2.9	3.6	3.5	3.4	8.4
r_i^*	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	1	0.7	0.7	0.7	0.2
x_i	21.7	22.3	22.7	23.1	20.1	20.3	20.5	20.7	21.4	21.5	21.6	22.3	29.3	29.9	30.5	31.7
y_i	23.6	23.3	23.5	25	23.2	23.3	23.4	23.5	23.8	23.9	23.9	24.3	21.9	21.6	21.8	16.8
z_i	26.9	26.4	26.4	28.8	26.9	26.4	26.6	26.8	27.3	27.3	17.5	28	22.8	22	22.7	7.9
r_i	3.3	3.1	2.9	3.8	3.7	3.1	3.2	3.3	3.5	3.4	3.6	3.7	0.9	0.4	0.9	8.9
r_i^*	0.7	0.7	0.7	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.9	1	0.9	0.1
x_i	29.6	29.9	30.2	30.3	31.4	31.8	32.7	32.9	33.6	33.9	34.1	34.7	21.3	21.5	21.4	31.6
y_i	23.5	23.1	25.1	24.5	24.1	23.9	22.1	20.9	20.1	21.2	21.1	20.2	20.9	20.9	20.9	17.8
z_i	29.9	29.6	32.1	31.4	29.5	29.4	27.2	26.2	25	26.7	27.8	27.3	30.9	30.8	29.7	9.0
r_i	6.4	6.5	7	6.9	5.4	5.5	5.1	5.3	4.9	5.5	6.7	7.1	10	9.9	8.8	8.8
r_i^*	0.4	0.4	0.3	0.3	0.5	0.5	0.5	0.5	0.5	0.5	0.3	0.3	0	0	0.1	0.1
x_i	21.6	21.8	21.8	22.3	22.6	23.1	23.5	23.9	24.5	25.6	26.2	26.6	27.1	27.7	28.5	31.1
y_i	20.8	20.8	20.7	20.6	21.5	19.7	20.2	20.1	20	19.6	19.4	19.3	19.9	20	18.9	18.0
z_i	29.5	29.4	28.8	28.9	29.7	28.2	28.1	28	27.7	27.4	11.3	11.2	11.7	11.7	10.7	9.4
r_i	8.7	8.6	8.1	8.3	8.2	8.5	7.9	7.9	7.7	7.8	8.1	8.1	8.2	8.3	8.2	8.6
r_i^*	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1
x_i	29.2	29.8	30.4	30.9	11.3	12.7	13.1	17.8	18.6	20.1	24.3	25.4	26.8	29.2	32.5	37.2
y_i	18.5	18.4	18.2	18	21.7	22.9	24.1	24.5	24.1	23.9	24.5	24.8	23.5	23.8	21.8	19.0
z_i	10.4	10.3	10	9.5	20.3	21.4	25.7	25.1	26.2	24.3	29.6	28.1	22.2	21.3	19.8	17.5
r_i	8.1	8.1	8.2	8.5	1.4	1.5	1.6	0.6	2.1	0.4	5.1	3.3	1.3	2.5	2.0	1.5
r_i^*	0.2	0.2	0.2	0.2	0.9	0.9	0.8	0.9	0.8	1.0	0.5	0.7	0.9	0.8	0.8	0.9

Table 2. $\hat{y}(x, \Omega_{\lambda_k})$ and their test numbers

$\hat{y}(x, \Omega_{\lambda_k})$	$R^2(\hat{y}(x, \Omega_{\lambda_k}))$	$F(\hat{y}(x, \Omega_{\lambda_k}))$
$\hat{y}(x, \Omega_{\lambda_1}) = -0.301x + 27.402$	0.926	348.164
$\hat{y}(x, \Omega_{\lambda_2}) = -1.264x + 61.028$	0.901	480.250
$\hat{y}(x, \Omega_{\lambda_3}) = -0.008x^2 + 0.809x + 10.086$	0.986	1075.848
$\hat{y}(x, \Omega_{\lambda_4}) = -0.043x^2 + 2.073x - 2.563$	0.971	975.985
$\hat{y}(x, \Omega_{\lambda_\infty}) = -0.046x^2 + 2.173x - 2.447$	0.701	207.624
Here, $\hat{y}(x, \Omega_{\lambda_\infty})$ denotes regression function based on the 180 samples without considering the reliability.		

Table 3. Regression functions under some different decision-making consciousness

Regression model	Regression function	
The classical model	$\hat{\mu}_0(x) = -0.046x^2 + 2.173x - 2.447$	
BSE-RM	$W_1(\lambda) = \lambda$	$\hat{\mu}_1 \triangleq \hat{y}(x, \Omega \oplus W_1(\lambda)) = -0.021x^2 + 0.868x + 15.698$
	$W_2(\lambda) = \lambda^2$	$\hat{\mu}_2 \triangleq \hat{y}(x, \Omega \oplus W_2(\lambda)) = -0.027x^2 + 1.267x + 9.054$
	$W_3(\lambda) = \sin(0.5\pi\lambda)$	$\hat{\mu}_3 \triangleq \hat{y}(x, \Omega \oplus W_3(\lambda)) = -0.021x^2 + 0.883x + 15.398$
	$W_4(\lambda) = Q_L((0,0),(0.5,0.4),(1,1))(\lambda)$	$\hat{\mu}_4 \triangleq \hat{y}(x, \Omega \oplus W_4(\lambda)) = -0.022x^2 + 0.967x + 13.955$

Table 4. The predictive value of the corresponding functions and the test number of the functions

x_i	11.3	12.7	13.1	17.8	18.6	20.1	24.3	25.4	26.8	29.2	32.5	37.2	$Q(\hat{\mu}_i(x))$	$PRSS(\hat{\mu}_i(x))$
y_i	21.7	22.9	24.1	24.5	24.1	23.9	24.5	24.8	23.5	23.8	21.8	19.0		
$\hat{\mu}_0(x_i)$	16.2	17.7	18.1	21.7	22.1	22.6	23.2	23.1	22.8	21.8	19.6	14.7	0.979	138.558
$\hat{\mu}_1(x_i)$	22.8	23.3	23.5	24.5	24.6	24.7	24.4	24.2	23.9	23.1	21.7	18.9	0.999	3.631
$\hat{\mu}_2(x_i)$	19.9	20.8	21.0	23.1	23.3	23.6	23.9	23.8	23.6	23.0	21.7	18.8	0.996	21.934
$\hat{\mu}_3(x_i)$	22.7	23.2	23.4	24.5	24.6	24.7	24.5	24.3	24.0	23.3	21.9	19.2	0.999	3.257
$\hat{\mu}_4(x_i)$	18.0	18.5	18.6	19.4	19.4	19.4	18.6	18.3	17.8	16.7	14.7	10.9	0.999	3.032

6. Conclusions.

Classical regression model is a statistical tool. It is not only widely applied in many fields, but also is the basis of the theory and method of econometrics. However, classical regression model does not consider data reliability, which restricts its extensive application greatly. With reliability of the sample as carrier, BSE-RM makes up for its shortcomings to a large extent. Theoretical analysis and example calculation show that BSE-RM has not only good structure characteristics and interpretability, but also good regression effect. Therefore, the paper enriches the existing theory of regression analysis, provides a basic method to make forecast and decision with different characteristics of the sample data, and has broad application prospects in many fields.

In Section 2, we put forward three shortcomings for the classical regression model, and BSE-RM solved two of them only. The third one (the description system of regression results is not perfect) still exists. The sample set with different reliability maybe derive the same regression function (for example, when $r_i \equiv r > 0$ holds in $\Omega = \{(x_i, y_i, r_i)|i = 1, 2, \cdots, n\}$, the form of $\hat{y}(x, \Omega \oplus W(\lambda))$ has nothing with $W(\lambda)$ and r). This shows that just relying on $\hat{y}(x, \Omega \oplus W(\lambda))$ cannot fully describe the characteristics of the regression results, which lacks the quantitative index about reliability of the regression results. So we will discuss the descriptive system of the regression results combining with $\hat{y}(x, \Omega \oplus W(\lambda))$ in the future work.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (71371064, 71540001) and the Natural Science Foundation of Hebei Province (F2015208099, F2015208100).

REFERENCES

[1] B. J. Huang and J. S. Lin, The prediction for civil airport passenger throughput based on multiple linear regression analysis, *Mathematics in Practice and Theory*, vol.43, no.4, pp.172-178, 2013.

[2] R. K. Jain, K. M. Smith et al., Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, *Applied Energy*, vol.123, pp.168-178, 2014.

[3] W. C. Hong, Chaotic particle swarm optimization algorithm in a support vector regression electric load forecasting model, *Energy Conversion and Management*, vol.50, no.1, pp.105-117, 2009.

[4] Y. Sai, F. T. Zhang and T. Zhang, Research of Chinese stock index futures regression prediction based on support vector machines, *Chinese Journal of Management Science*, vol.21, no.3, pp.35-39, 2013.

[5] Y. Guo, E. Nazarian et al., Hourly cooling load forecasting using time-indexed ARX models with two-stage weighted least squares regression, *Energy Conversion and Management*, vol.80, pp.46-53, 2014.

[6] A. Y. Sun, D. B. Wang and X. L. Xu, Monthly streamflow forecasting using gaussian process regression, *Journal of Hydrology*, vol.511, pp.72-81, 2014.

[7] D. Li and X. F. Shi, Estimates of pedestrian crossing delay based on multiple linear regression and application, *Procedia – Social and Behavioral Sciences*, vol.196, no.611, pp.1997-2003, 2013.

[8] D. G. Kleinbaum and L. L. Kupper, *Applied Regression Analysis and Other Multivariate Methods*, Duxery Press, Boston, Massachussetts, 1978.

[9] K. Hariharan, R. V. Prakash and M. S. Prasad, Weighted error criterion to evaluate strain-fatigue life prediction methods, *International Journal of Fatigue*, vol.33, pp.727-734, 2011.

[10] M. Dobber, R. van der Mei and G. Koolea, A prediction method for job runtimes on shared processors: Survey, statistical analysis and new avenues, *Performance Evaluation*, vol.64, pp.755-781, 2007.

[11] F. Li, C. Jin, Y. Shi and K. Yang, Study on quasi-linear regression methods, *International Journal of Innovative Computing, Information and Control*, vol.8, no.9, pp.6259-6270, 2012.