Description and Application Research of Multiple Regression Model Optimization Algorithm Based on Data Set Denoising

Hao Kang* and Hailong Zhao

Army Academy of Armored Forces, 1 Huayuan Road, Economic and Technological Development Zone, ChangChun City, Jilin Province, China
Email: 29098545@qq.com

Abstract. Multiple regression model is based on a large number of data sample set in the prediction process, and the noise data in the data set will have a great impact on the results of the fitting equation, which makes the results unreliable and unreliable. This paper forwards fuzzy least square method based on fuzzy set theory, it can optimize the regression model algorithm and reduce the influence of noise data on the fitting equation. This algorithm is applied to the real estate price forecast, it would obtain the final fitting equation after repeating iterative calculation, which makes the price prediction and eliminates the influence of bad data on the fitting results and improves the reliability and availability of the forecast results.

1. Introduction
Regression analysis is a commonly used statistical analysis model, which can predict the value of a random variable through a variable or a group of variables, so as to help us make analysis and judgment on future events and obtain reasonable decision results. It is widely used in practical problems. However, in the actual application process, the collection and collection of data sample sets are very critical steps. If the sample points can reflect the real situation basically, then the establishment of the model is more ideal. On the contrary, there is a gap between the model and the real situation, so the reliability and authenticity of the prediction will be greatly reduced.

In this paper, fuzzy mathematics is used to optimize the typical multiple regression model, reduce the influence of noise data on the fitting equation, and improve the application efficiency of model prediction. The model is applied to the real estate price forecast for testing, through the correlation analysis of the factors affecting the housing price, the fuzzy regression model is established, and the reasonable forecast of the future housing price is made, and the good application effect is obtained.

2. Regression Analysis of Basic Processes and Optimization Strategies

2.1. The Basic Process
The application of regression analysis for prediction should follow a certain process [1-3]:

2.1.1. Select Independent Variables According to the Prediction Target. In general, a clear prediction of the specific goal, also determine the dependent variable. To screen independent variables, we should first analyze the correlation between each variable and the dependent variable, observe the manifestation and closeness of the correlation, and select the independent variables that are most closely related to the dependent variables.
2.1.2. Establish Reasonable Data Sample Set. The validity of the data sample set is very important and will have a direct impact on the predicted results. If there are noise data, ill-condition data and interference information in the sample set, the establishment of the regression model based on the predicted results will have a big deviation.

2.1.3. Determine the Regression Equation and Establish the Prediction Model. According to theoretical analysis and correlation analysis, if there are several important factors to predict phenomena have effect, at the same time and close relationship, with multiple regression equation can be determined, if it is a basic, difference, and other factors affect the effect is not very big or not close related, you can determine with regression equation with one unknown to forecast. If the distribution between independent variables and dependent variables shows a linear trend, the linear regression equation can be adopted; if it is the curve trend, the curve regression equation can be determined. The determination of multiple regression equations is the same, it refers to a number of independent variables and dependent variables have a correlation.

2.1.4. Use the Regression Model to Determine the Predicted Value and Estimate the Confidence Interval for the Predicted Value. If there is indeed a significant correlation between the predicted objects and the influencing factors, then the data laws of the past and present can continue into the future. The predicted value calculated by the regression equation is a specific number called a point estimate [4]. The point estimate is an average, the actual value may be higher or lower, and the range of its confidence interval must be guaranteed with a certain probability. The specific process is shown in figure 1 below.

![Figure 1. Regression analysis flow chart.](image)

2.2. Model Optimization Strategy
In the whole process of regression analysis, the selection of independent and dependent variables and the establishment of a reasonable sample set are very important, especially since the quality of the sample set directly determines the reliability of the prediction. When applying regression analysis to predict random variables, we are all based on the existing observed values of variables, and the data we get will be more or less affected by random factors. When all the data points can roughly reflect the real situation, the result we get is reasonable, which is a very ideal situation, which is rarely seen in practical application, or even not at all. Therefore, when a certain or some data points are greatly influenced by random factors, the results we get will be somewhat different from the real results, and the error of the decision results will be large.
Using fuzzy expectation and variance as the theoretical basis to use fuzzy least squares method to predict regression coefficient, set up a corresponding back to return model, namely fuzzy regression analysis, the model can make the pathological data and poor information interference to a minimum, the result of the decision to make decision accuracy and reliability of the corresponding increases, using the optimized model. This is shown in figure 2.

![Figure 2. Normal versus abnormal data points.](Image)

As shown in the figure above: for the simplest unary regression, if the data sample points are all ideal, then the fitting equation is also ideal. If the data sample point contains ill-conditioned data, then the fitting equation is not ideal, that is, it cannot truly reflect the relationship between variables. If there is a lot of ill data, then the establishment of the model has no real meaning. The fewer the data sample points, the more obvious this situation will be. So the best way to solve this kind of noise data, one is to increase the number of data points, another is to optimize the data model, reduce the impact of noise data on modeling, and the latter is a decisive method, is the end of our model optimization.

3. Establishment the Model

3.1. Theoretical Basis

The definition of fuzzy expectation is as follows [5]:

\[
FE(x) = \frac{\sum u_i x_i}{\sum u_i}
\]  

(1)

\(u_i\) is the membership function corresponding to the fuzzy set, corresponding to the value of the random variable. \(u_i\) is open domain in the fuzzy set, that is \(u_i \in (0,1)\).

The definition of fuzzy variance is as follows:

\[
FD(x) = \frac{\sum u_i (x_i - FE(x))^2}{\sum u_i}
\]  

(2)

3.2. The Establishment of Fuzzy Multiple Regression Model

3.2.1. Determine the Independent and Dependent Variables. The multiple regression model is a linear relationship between the random variable \(y\) and \(k\) independent variables \(x_1, x_2, \ldots, x_k\), namely:

\[
\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \epsilon
\]  

(3)

In the formula, \(\hat{y}\) represents the predicted value, \(y\) represents the observed value, \(x_1, x_2, \ldots, x_k\) is the regression coefficient, \(\epsilon\) is a random error, \(\epsilon \sim N(0, \sigma^2)\).
3.2.2. Set Up Samples. The values of the independent and dependent variables can be measured and observed here, or obtained from the actual data in the past, which is a group of known data. The more data points there are, the closer they are to the actual situation and the more realistic the results are. Therefore, a certain number of samples should be obtained, generally more than or equal to 100 samples.

3.2.3. Determine the Regression Coefficient. To determine the regression coefficient is to minimize the sum of squares between \( y \) and \( \hat{y} \), which is called the least square method, as shown in the following formula:

\[
\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min
\]

(4)

According to the typical regression analysis algorithm, the regression coefficient is obtained, that is, the value of \( x_1, x_2, \ldots, x_k \) is obtained, so that \( y \) and \( \hat{y} \) correspond to a certain set of observed values. Now define the fuzzy set, whose domain is all data points, and the fuzzy concept it describes is the degree of deviation between all data points and the fitting equation. The farther the data points are from the fitting equation, the smaller the membership degree will be, and vice versa [6]. The definition is as follows [3]:

\[
u(i) = \frac{1}{e^{(y_i - \hat{y}_i)^2}} \quad i = 1, 2, \ldots, n
\]

(5)

After adding the fuzzy domain, the least square method weakens the influence of ill-conditioned data and discrete data on the fitting equation. This algorithm is called fuzzy least squares and is defined as follows:

\[
\sum_{i=1}^{n} u_i (y_i - \hat{y}_i)^2 = \min
\]

(6)

According to the above equation, the solution equation is:

\[
s_{11} b_1 + s_{12} b_2 + \cdots + s_{1k} b_k = s_{1y} \\
s_{21} b_1 + s_{22} b_2 + \cdots + s_{2k} b_k = s_{2y} \\
\vdots \\
s_{1i} b_1 + s_{2i} b_2 + \cdots + s_{ki} b_k = s_{iy}
\]

(7)

The parameter meanings are as follows:

\[
s_i = \sum_a a x_i, x_j - \sum_a a x_i \sum a x_j \\
s_y = \sum_a a y - \sum_a a x_1 y - \sum_a a x_2 y - \cdots - \sum_a a x_k y
\]

(8)

The solution of \( b_1, b_2, \ldots, b_k \), that is, the solution of the regression coefficient, is obtained by using the above equations. Thus, the first fuzzy regression analysis is completed, that is, the first set of regression coefficient solutions is obtained. Then, the fuzzy set and membership function are redefined with the new coefficients, and the regression coefficient solution is determined by using the fuzzy least square method. After repeated calculation, the coefficient changes slightly, and the fitting equation is finally determined.

4. Application for Example

With the continuous development of domestic economy, is becoming more and more high, the price level in People’s Daily consumer goods, housing as a special kind of product is the stability of the ownership, that is to say, the basic all have personal housing resident, but with the improvement of people’s living standard, improve the housing of sex, sex, housing is becoming more and more be
favored by the people, commercial housing sales have been very popular, prices have become the most concern for people.

4.1. Correlation Analysis

According to the collection of samples of the housing sales in a certain area within a year, according to the factor analysis, the main factors affecting the housing price are area, number of bedrooms, location, orientation, daylighting, etc., the values of these indicators are collected. Here, the method of gray correlation analysis [7-8] is adopted to carry out correlation analysis on the indicators.

In the following table 1, $x_s$ represents the building area, $x_n$ represents the number of bedrooms, $x_a$ represents the location, and $x_d$ represents the orientation, $x_p$ represents the lighting. The ranking results of correlation degree of the sequence of constituent sample variables are shown in the following table 1:

| Subsequence | $x_s$ | $x_n$ | $x_a$ | $x_d$ | $x_p$ |
|-------------|-------|-------|-------|-------|-------|
| Correlation | 0.925 | 0.895 | 0.523 | 0.685 | 0.756 |
| Sort        | 1     | 2     | 5     | 4     | 3     |

The matrix table of variable correlation coefficient is shown in the following table 2:

|       | $y$       | $x_s$(Var-1) | $x_n$(Var-2) | $x_a$(Var-3) | $x_d$(Var-4) | $x_p$(Var-5) |
|-------|-----------|---------------|---------------|---------------|---------------|---------------|
| $y$   | 1.0000    | .631          | .448          | .585          | -.161         | -.555         |
| $x_s$ | .631      | 1.0000        | .996          | .239          | -.597         | -.251         |
| $x_n$ | .448      | .996          | 1.0000        | .209          | -.651         | -.208         |
| $x_a$ | .585      | .239          | -.209         | 1.0000        | -.311         | -.621         |
| $x_d$ | -.161     | -.597         | -.651         | -.311         | 1.0000        | .489          |
| $x_p$ | -.555     | -.251         | -.208         | -.621         | .489          | 1.0000        |

4.2. Establish Data Sample Set

According to the analysis, there is a certain correlation between the housing indicators, which have an impact on the housing price. Taking $x_s$, $x_n$, $x_a$, $x_d$, and $x_p$ as independent variables, $x_s$ variable represents the number of square meters of the house, $x_n$ variable represents the number of bedrooms, $x_a$ represents the location of the house, which conducts data processing on it, and uses the Numbers 5 to 1 to replace them respectively according to the order of advantages and disadvantages of CBD, sub-prosperous living area, good environment old city, downtown area, and urban and rural suburb. $x_d$ variable represents orientation, which is digitized according to the order of the orientation of due south, due north, due east, due west, southeast and southwest, and replaced by Numbers 6 to 1. $x_p$ represents daylighting, represented by the tangent of the daylighting rate [5]. Let’s take the house price as the dependent variable, and call it $y$. The sample data set is established as follows in table 3.

Since different independent variables have different influences on the results, the data should be weighted before use, that is, the proportion of influence should be set. In table 3, $w_s$, $w_n$, $w_a$, $w_d$ and $w_p$ are weighted values for area, number of bedrooms, location, orientation and daylighting. After data collection, data processing is carried out first, and then the data is weighted, so that the data can be used as a sample set in the algorithm. The weighted value comes from people’s attention to each factor. This is shown in the following figure 3.

Taking $x_s$ and $x_n$ as independent variables and dependent variables to draw a spatial scatter diagram, it can be seen that the sample points basically tend to be a plane and the spatial information is
clearly represented. If you have more than three variables, the model can’t be described in terms of spatial information, it’s a hyperplane, it could be a four-dimensional hyperplane, or a five-dimensional hyperplane. If there are noise data in the data sample points, sometimes also known as ill data, fitting according to the original data will have a negative impact on the fitting equation and affect the accuracy of the predicted results. The spatial fitting results are shown in the following figure 4, which only displays partial data. Figure 4 shows the diagram of the three-dimensional fitting equation.

Table 3. Data sample set.

| x_s*(w_s) | x_n*(w_n) | x_a*(w_a) | x_d*(w_d) | x_p*(w_p) | y    |
|----------|-----------|-----------|-----------|-----------|------|
| 60.5     | 1         | 2         | 6         | 7.42      | 3400.00 |
| 79.8     | 1         | 4         | 6         | 7.19      | 4200.00 |
| 90.2     | 2         | 4         | 4         | 5.08      | 5180.00 |
| 130.6    | 3         | 2         | 2         | 6.29      | 6080.00 |
| 122.5    | 3         | 1         | 6         | 8.80      | 4480.00 |
| 130.8    | 3         | 5         | 5         | 4.77      | 8600.00 |
| 92.7     | 2         | 3         | 3         | 5.27      | 7280.00 |
| 144.6    | 4         | 5         | 1         | 6.06      | 4020.00 |
| 88.7     | 2         | 3         | 5         | 8.91      | 5570.00 |

Other sample points are omitted.

Figure 3. Weighting of factor attention.

Figure 4. Double independent variable fitting equation diagram.

4.3. Fuzzy Regression Analysis
Before analysis, the data sample set should be standardized. Then, the first regression equation is obtained by using the fuzzy least square method as follows in table 4.

Table 4. First calculation result.

| Model | Unstandardized coefficient | Standardized coefficient |
|-------|---------------------------|--------------------------|
|       | B            | Standard error | Beta     |
| Constant | 6570.617            | 7503.927              | .341     |
| x_s    | 43.760            | 104.801              | -.457    |
| x_n    | -398.803           | 2184.387             | -.446    |
| x_a    | 155.028            | 442.743              | .314     |
| x_d    | 69.138             | 267.631              | -.114    |
| x_p    | -164.776           | 404.481              | .314     |
After the fourth and fifth iteration, the coefficient of the regression equation changed very little and was inclined to be stable. We think of this as the final regression equation, the final regression model, which can be used to predict can achieve the optimal effect, can eliminate the noise data to the maximum extent of the impact on the results. In general, if the ill-condition data is small, a satisfactory result can be obtained after about five iterations of calculation. The final calculation results are as follows in table 5:

| Model | Unstandardized coefficient | Standardized coefficient |
|-------|---------------------------|--------------------------|
|       | B    | Standard error | Beta |
| Constant | 6270.617 | 7403.927 | .289 |
| x5     | 33.760 | 99.801 | -.321 |
| x4     | -264.803 | 1998.387 | .146 |
| x3     | 125.028 | 302.743 | .214 |
| x2     | 77.138 | 229.631 | -.109 |

4.4. Forecast Application
After establishing a reasonable regression model, we can make a reasonable prediction according to the model, and predict the house price based on the factors such as the house area, the number of bedrooms, location, orientation and daylighting in the area. In fact, in the actual application process, validity verification and error interval estimation should be carried out to make the predicted results more reasonable, available and effective. These methods are not described here.

5. Conclusion
Regression analysis method is called factor analysis method in prediction. It is a process of finding out the quantitative relationship between a variable and some variables that can be regarded as the reason for change, establishing a mathematical model, and then predicting the results based on the existing factor data. In recent years, the fuzzy set theory has been applied to the data analysis algorithm, which has improved the scientificity and effectiveness of the algorithm, and has become a powerful mathematical tool to solve the corresponding problems.

References
[1] Xie Y 2018 Regression Analysis (Social Sciences Academic Press).
[2] He X Q and Liu W Q 2019 Applied Regression Analysis (Renmin University of China Press).
[3] Lin J Z 2018 Regression Analysis and Linear Statistical Models (Shanghai University of Technology Press).09.
[4] Wang L M 2019 Regression Analysis (Shanghai University of Finance and Economics Press).
[5] Kang H and Chen Y Y 2018 The establishment of decision model based on fuzzy regression and the application of chemical solid adsorption quantity International Conference on Physics, Computing and Mathematical Modeling pp 122-126.
[6] Smithson M and Verkuilen J 2016 Application of Fuzzy Set Theory in Social Science (Shanghai People’s Publishing Press).
[7] Chen H M 2017 Building Construction Engineering Evaluation Foundation (Capital University of Economics and Business Press).
[8] Li Z F and Jiang Y S 2015 An empirical study of housing price differences in China based on partial least squares regression analysis Journal of Systems Management 20 (1) 42-46.