

The Variability of Oxygen Saturation

Qing Li¹, Zekun Li² and Zhiqiang Yin³

¹Nerth China University of Science and Technology Mathematical Modeling Assciation,
Tangshan 063200, China

²North China University of Science and Technology, Tangshan 063200, China

³North China University of Science and Technology, Tangshan 063200, China

Abstract

Oxygen saturation is an important indicator to measure the oxygen-carrying capacity of human blood, and it is also an important physiological indicator for clinical diagnosis and family health care. The influence mechanism of oxygen saturation is a very necessary study and has far-reaching significance. First, in order to establish a model to describe the blood oxygen saturation model, perform data preprocessing, use Lagrangian interpolation to fill in the vacant blood oxygen saturation data, and measure gender, age, whether or not to smoke, and BMI index dimensional processing. Secondly, four parameters of gender, age, smoking and BMI index are selected to establish a multiple linear regression model. The R^2 of this equation is 0.258. In order to verify the accuracy of the division, a K-means clustering model was established, which was divided into five parameters based on gender, age, smoking, BMI index and blood oxygen saturation, and compared with the division results of the multiple linear regression model. The result is less error. Therefore, the clustering effect of the multiple linear regression model is very good. Third, in order to solve the problem of the low correlation coefficient of the multiple linear regression model, a multiple nonlinear regression model was established, and finally an equation of blood oxygen saturation with respect to gender, age, smoking and BMI index was obtained. The R^2 of this equation was 0.829, The correlation is high, and the regression effect has reached the ideal expectation. Finally, the pattern of blood oxygen saturation sequence is explored. Through Python data analysis and visualization, the relationship between blood oxygen saturation and age, smoking and BMI index is visually displayed in the form of bubble chart. Data analysis shows that blood oxygen saturation will show a downward trend within a certain range with age; smokers have lower blood oxygen saturation than non-smokers, and former smokers have lower blood oxygen saturation than people who have not smoked. It is concluded that smoking will reduce the blood oxygen saturation; the blood oxygen saturation of people with a normal BMI index is higher than that of people with a higher or lower BMI index. In summary, age, smoking, and BMI index will all have a dynamic impact on blood oxygen saturation.

Keywords

Blood oxygen saturation; Lagrangian interpolation; Multiple nonlinear regression; K-means clustering.

1. Introduction

In order to indicate the origin of problems, the following background is worth mentioning.

1.1. Background

Blood oxygen saturation is an important parameter to measure the ability of human blood to carry oxygen, and it is also an important physiological index in clinical diagnosis and family health care. In addition to heart rate, blood pressure, respiratory rate and temperature, pulse oximetry is considered to be the fifth most critical health indicator. Hemoglobin is an important part of blood cells. It is responsible for transporting oxygen from the lungs to other tissues in the body. The amount of oxygen contained in hemoglobin at any one time is called oxygen saturation. The oxygen saturation is expressed as a percentage, which is the ratio of the oxygen content of hemoglobin to the oxygen carrying capacity of hemoglobin. Blood oxygen saturation is an important physiological parameter that reflects whether the human body's respiratory function and oxygen content is normal. It is an important physiological parameter that shows whether our human tissues are healthy.

The main function of hemoglobin in human blood is to transport oxygen and carbon dioxide, and to balance the pH of the blood. There are four types of hemoglobin, namely oxyhemoglobin, reduced hemoglobin, carboxyhemoglobin and methemoglobin. During the contraction and relaxation of the heart, blood flows through the lungs in a pulsating state, causing a certain amount of reduced hemoglobin to combine with oxygen in the alveoli to produce oxygenated hemoglobin. These oxygen components pass through the arterial system all the way to the capillaries, and then are released, thereby maintaining the normal metabolism of human tissue cells.

In medical science, the physical quantity describing the size of oxygen in the blood is blood oxygen saturation, which refers to the percentage of the total oxygen capacity bound to reduced hemoglobin in the total blood volume, expressed by the mathematical formula as follows:

$$SaO_2 = \frac{HbO_2}{HbO_2 + Hb} \times 100\% \quad (1)$$

In clinical medicine, the monitoring of human blood oxygen saturation and peripheral blood circulation is a very important content. The peripheral blood circulation of the human body can be expressed by arterial pulse wave and venous return volume wave. As important information reflecting the state of the human body, its test results can be used to diagnose certain diseases, such as arteriosclerosis, hypertension, enteritis and peripheral circulatory disorders. The value of blood oxygen saturation is directly related to the cardiopulmonary function, breathing and circulatory system, and its monitoring plays a very important role in surgical anesthesia, medical monitoring, exercise and sleep process research, and oxygen therapy.

1.2. Restatement of Problems

Pulse oximeters are commonly used to monitor the blood oxygen saturation level of patients. In this study, a cross-sectional survey was used to obtain the oxygen saturation measured by 36 individuals over a period of time. At the same time, collect the physiological indicators of related personnel including age, BMI, gender, Smoking history and/or current smoking status, and any significant medical conditions that could affect reading. Now I want to find a model with physiological and medical significance. To represent the typical pattern of changes in oxygen saturation for each person. These characteristics should ideally be of biological or medical significance. Further analyze whether the pattern of oxygen saturation sequence is related to age, that is, which characteristics of the elderly have changed compared with young people. Find out the differences caused by related factors and study the influence of age on the observed variable, namely the blood oxygen level.

1.3. Our Work

First, analyze the data and find that the amount of measured data in 36 sets of data is not equal, so Lagrangian interpolation is used to fill in the data. Then the relevant indicators are processed to de-dimensionalize. Perform Min-max standardization processing on the impact index BMI and age, and obtain the processed data value. Quantify whether you smoke or not according to statistical methods, which is convenient for fitting operations in the model.

Based on known index data, select age, gender, smoking history, BMI index to establish a multiple linear regression model on blood oxygen saturation:

$$y = 0.35351x_1 - 0.5145x_2 + 0.0079x_3 - 0.0279x_4 + 98.6010 \quad (2)$$

The fitted R2 is 0.258 and the fitting effect is not good. Therefore, the model is improved in the following text and a multiple nonlinear regression model is established. The obtained R2=0.829, and the fitting accuracy after optimization has been greatly improved.

In this paper, we get the influence of changes in various indicators on blood oxygen saturation among people of different ages. Next, a preliminary statistical analysis of the age of one of the four indicators is carried out, in an attempt to explore the relationship between the difference in age and blood oxygen saturation. And use the K-means clustering algorithm to perform cluster analysis, trying to verify the results obtained from the statistical analysis. The verification shows that the results obtained by the statistical analysis are valid, namely:

The minimum and maximum blood oxygen saturation levels of young people, young people, and middle-aged and elderly people are successively decreasing, while the standard deviations of blood oxygen saturation of young people, young people, and middle-aged and old people are progressively increasing. With age, blood oxygen saturation will gradually decrease, and the fluctuation range of blood oxygen saturation will gradually increase, that is, the phenomenon of instability of blood oxygen saturation will appear.

2. Assumptions

All study subjects did not take vitamin E, vitamin C, angiotensin-converting enzyme inhibitors, calcium antagonists and other antioxidant drugs within 1 month before being tested;

It is assumed that there is no abnormality of the detection system, abnormal skin of the patient's detection site, poor detection environment, abnormality of the probe position of the detection instrument, hypoperfusion state of the detected site of the patient, and artifacts in the detection action throughout the test;

Assume that the 36 people tested are all measured under the same conditions;

Assuming that the 36 people tested are all health measurements, there will be no errors in blood oxygen saturation due to chronic obstructive pulmonary disease, etc.

3. Models

3.1. Data Pre-processing

Before analyzing each physiological factor, the known physical sign data can be de-dimensionalized, which can eliminate the influence of dimensions between features, and unify all the features into a roughly the same numerical range, so that different levels of The indicators can be further used for comparison and weighting in the model.

First, normalize the body mass index and age. For each variable value, set minA and maxA to be the minimum and maximum values of variable A respectively, and map the original value of A to y1 in the interval [0, 1] through Min-max standardization. The calculation formula is:

$$y_i = \frac{x_i - \min\{x_j\}}{\max\{x_j\} - \min\{x_j\}} \quad (1 \leq i, j \leq n) \quad (3)$$

Among them, $\min\{x\}$ is the minimum value of the sample data, and $\max\{x\}$ is the maximum value of the sample data.

Representative value analysis

Based on the field No. 1 and the coefficient of variation (CV) is 0.0e5, which is less than 0.15, there is a small probability of outliers in the current data. It is recommended to use the average value for descriptive analysis.

3.2. Quantitative Processing of Characteristic Indicators

This article uses Lagrangian interpolation to fill in the data.

For n known points on the plane, a polynomial of degree $n-1$ can be found.

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}. \quad (4)$$

Taking the coordinates of n points $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ polynomial function, we get:

$$y_1 = a_0 + a_1x_1 + a_2x_1^2 + \cdots + a_{n-1}x_1^{n-1}, y_2 = a_0 + a_1x_2 + a_2x_2^2 + \cdots + a_{n-1}x_2^{n-1} \quad (5)$$

.....

$$y_n = a_0 + a_1x_n + a_2x_n^2 + \cdots + a_{n-1}x_n^{n-1} \quad (6)$$

Solve the Lagrangian interpolation polynomial as:

$$\begin{aligned} L(x) = & y_1 \frac{(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)} \\ & + y_2 \frac{(x-x_1)(x-x_3)\cdots(x-x_n)}{(x_2-x_1)(x_2-x_3)\cdots(x_2-x_n)} \\ & + \cdots \cdots \end{aligned} \quad (7)$$

$$\begin{aligned} & + y_n \frac{(x-x_1)(x-x_2)\cdots(x-x_{n-1})}{(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})} \\ L(x) = & \sum_{i=0}^n y_i \prod_{j=0, j \neq i}^n \frac{x-x_j}{x_i-x_j} \end{aligned} \quad (8)$$

The point x corresponding to the missing function value is substituted into the interpolation polynomial to obtain the approximate value $L(x)$ of the missing value.

3.3. Multiple Linear Regression

The multiple linear regression analysis method refers to the method of establishing a predictive model for forecasting through the correlation analysis of two or more independent variables and a dependent variable. Since in the actual blood oxygen saturation test, based on the existing data, it is difficult for us to use an accurate function to link the variables together. In order to ensure that the subsequent data still meets such an accurate relationship, a regression model

is used to connect the two. At the same time, the specific methods for predicting the 36 sets of test data and establishing the model are as follows:

First establish the model equation:

$$y_i = \alpha_0 + \sum_{j=1}^n \alpha_j x_{ij} + u_i \quad i = 1, 2, \dots, n \quad (9)$$

Assuming that there are n sample observations, the multiple regression model can be expressed as a matrix:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad (10)$$

Applying the least squares method, that is, the estimation of the regression coefficient with the smallest residual sum of squares, the error u_i can be expressed as:

$$u_i = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n} \quad (11)$$

When the first-order partial derivative of $\hat{y}$ is zero, the optimal parameters are solved, and regression coefficients and regression equations are obtained.

Table 1. Table of target variable coefficients of blood oxygen saturation regression equation

	Mean blood oxygen saturation	
	Value	Standard error
intercept	98.6010	2.15043
Sex	0.35351	0.41910
SmokingStatus	-0.5145	0.31558
BMI	0.0079	0.08072
Age	-0.0279	0.01273

Get the linear regression equation of blood oxygen saturation:

$$y = 0.35351x_1 - 0.5145x_2 + 0.0079x_3 - 0.0279x_4 + 98.6010 \quad (12)$$

Among them: x_1 :Sex; x_2 :SmokingStatus; x_3 :BMI; x_4 :Age

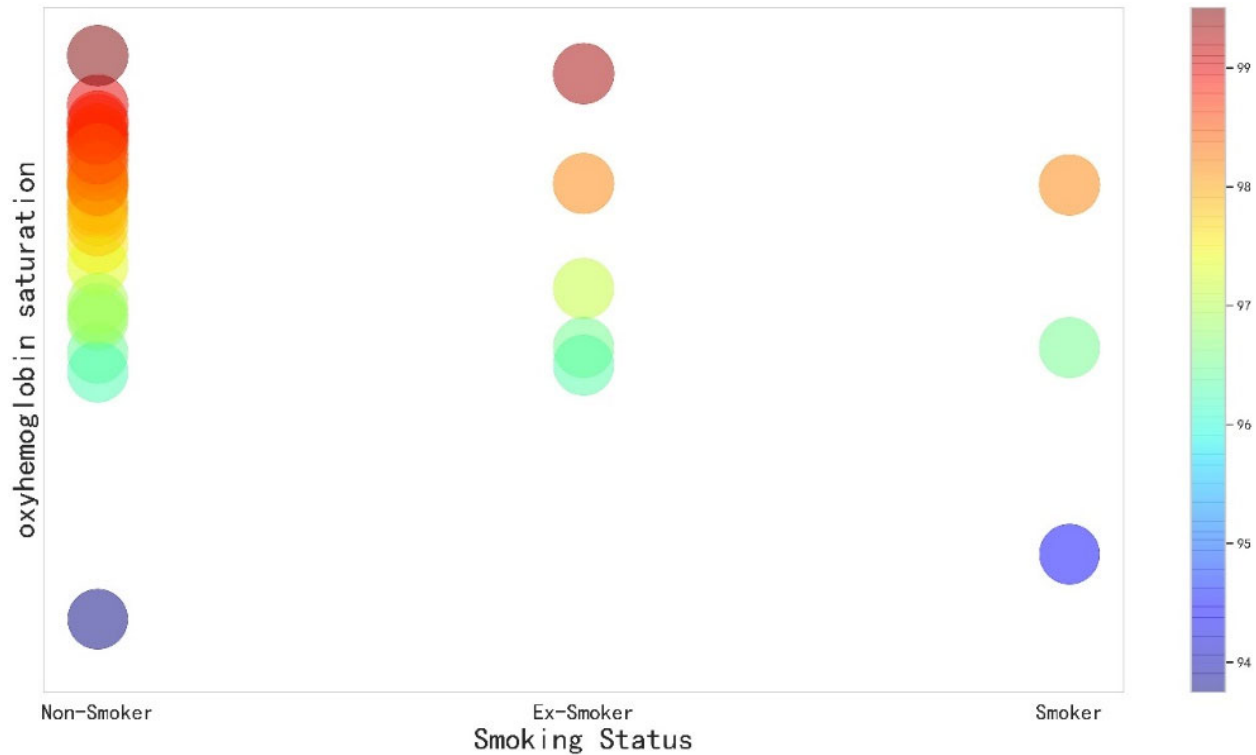


Figure 1. The connection between smoking and blood oxygen saturation

By observing the above figure, we can see that the blood oxygen saturation of non-smokers is significantly higher than that of smokers. The blood oxygen saturation of smokers is between smokers and non-smokers. The blood oxygen saturation of smokers is overall At a low level. get conclusion:

Smoking has a direct and greater impact on blood oxygen saturation.

Table 2. Statistics of blood oxygen saturation

Mean blood oxygen saturation	
Points	36
Degree of freedom	31
Residual sum of squares	40.38954
R square (COD)	0.25751
Adjusted R-squared	0.1617

In order to verify the accuracy of the linear regression equation of blood oxygen saturation, the blood oxygen saturation test data is substituted into the linear regression equation of blood oxygen saturation obtained by regression, and the obtained result is compared with the real value, through the following figure It can be seen that the fitting effect is better.

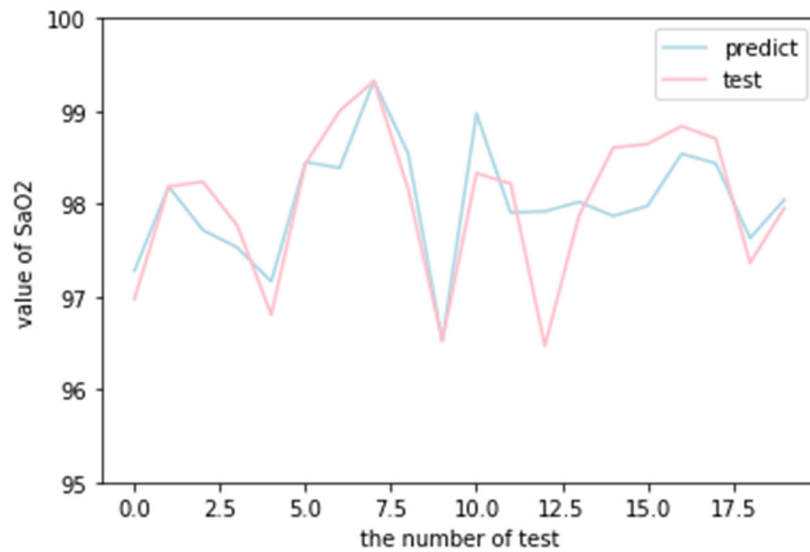
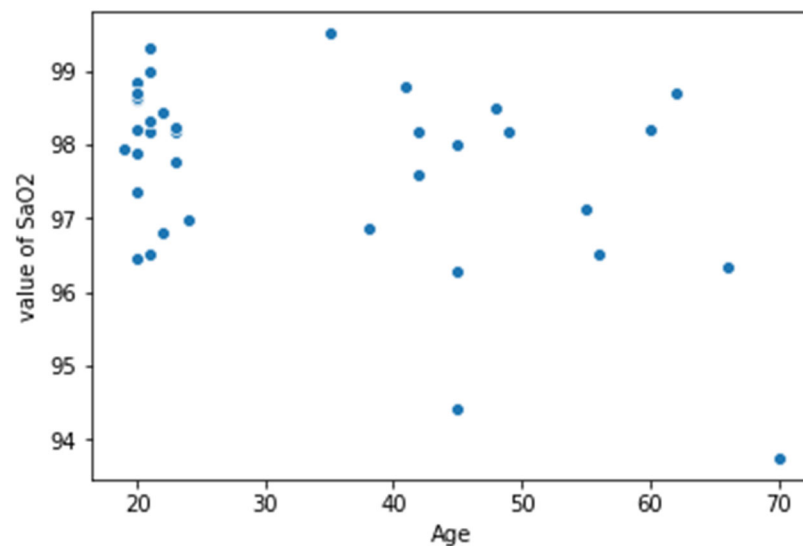


Figure 2. Comparison of the predicted value and the true value of the linear regression equation of blood oxygen saturation



Finger 3. The relationship between age and blood oxygen saturation

According to the above figure, we can observe that the value of blood oxygen saturation and its fluctuation range are related to age. Therefore, we divide the age group into three categories and perform statistical analysis to average the minimum, maximum, average, and standard deviation of blood oxygen saturation. The results are as follows:

Table 3. Blood oxygen saturation in people of different ages

age	Minimum blood oxygen saturation	Maximum blood oxygen saturation	Standard deviation of blood oxygen saturation
19-34	94.64845201	97.67036923	0.710492497
35-49	94.30784314	97.28011858	0.748785573
50-70	93.83333333	96.77490939	0.796214215

From the above statistical analysis, we can see that the minimum and maximum blood oxygen saturation levels of young people, young people, middle-aged and elderly people are successively decreasing, while the blood oxygen saturation standards of young people, young people, middle-aged and old people The difference is increasing sequentially. This shows that with age, the blood oxygen saturation will gradually decrease, and the fluctuation range of blood oxygen saturation will gradually increase, that is, the phenomenon of unstable blood oxygen saturation will occur.

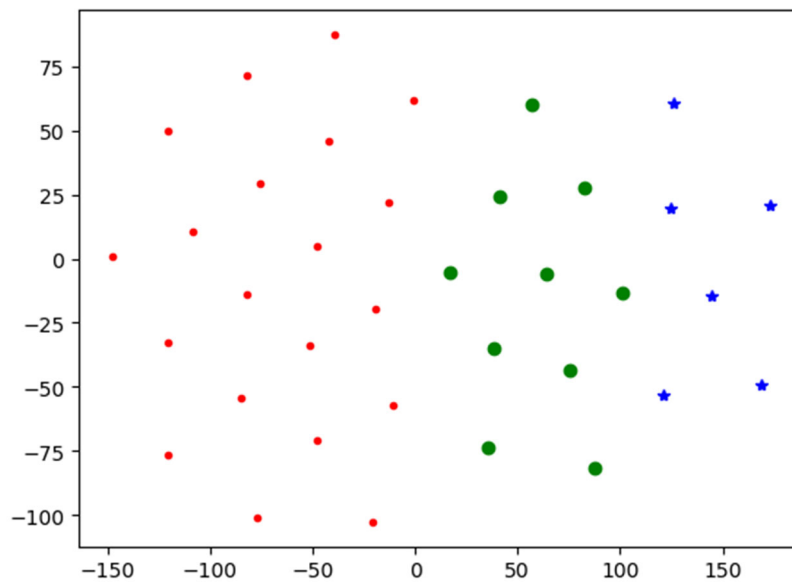


Figure 4. K-Means Cluster analysis

The analysis results show that when divided into three categories, the distribution is roughly the same as that shown in Figure 2. It further confirms the correctness of the above statistical analysis, and also shows that blood oxygen saturation does have an effect with age. The blood oxygen saturation of the elderly is lower than that of the young and fluctuates more.

3.4. Multiple Nonlinear Regression

It is known that the blood oxygen saturation rate is affected by many factors. Considering the complexity of the data relationship, the model is optimized into a multivariate binomial regression model. According to the characteristics of the data, there are two models:

Purequadratic:

$$y = \beta + \sum \beta_i x_i + \sum \beta_{jj} x_j^2$$

Interaction:

$$y = \beta_0 + \sum \beta_i x_i + \sum \beta_{jk} x_j x_k$$

Considering the calculation speed, a pure quadratic binomial regression model is selected, and the same main influencing factors are selected to obtain the prediction result of blood oxygen saturation.

$$\begin{aligned} y = & -1.455 \times 10^3 x_3 x_4 + 1.042 \times 10^3 x_4^2 + 2.026 \times 10^{-1} x_3^2 - 4.459 \times 10^{-2} x_1 x_3 x_4 \\ & - 2.592 \times 10^{-3} x_1 x_4^2 - 3.435 \times 10^{-3} x_3^3 + 4.688 \times 10^{-3} x_3^2 x_4 + 1.537 \times 10^{-3} x_3 x_4^2 \\ & + 5.195 \times 10^3 x_2 x_3 x_4 - 3.722 \times 10^3 x_2 x_4^2 \end{aligned}$$

x_1 :Sex; x_2 :Smoking Status; x_3 :BMI; x_4 :Age

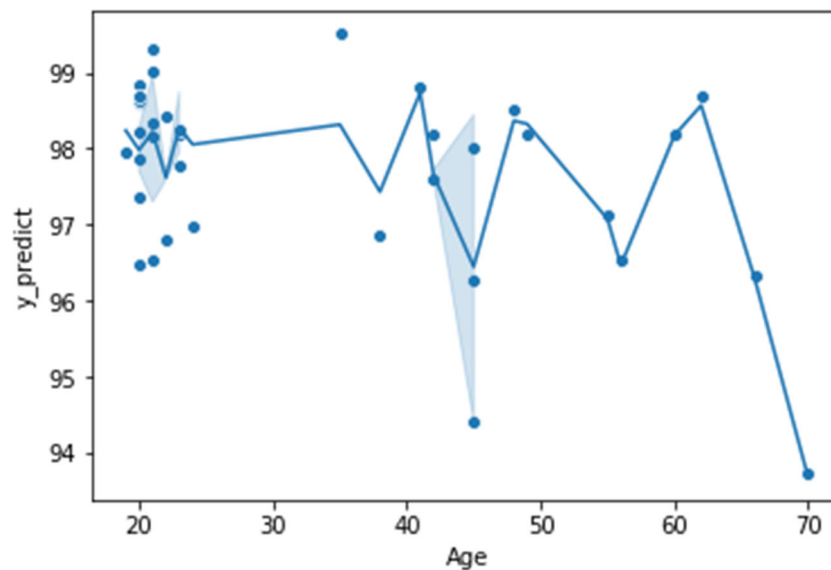


Figure 5. Multiple nonlinear regression

Goodness of Fit refers to how well the regression line fits the observations. The statistic that measures the goodness of fit is the coefficient of determination (also known as the coefficient of determination) R^2 . The maximum value of R^2 is 1. The closer the value of R^2 is to 1, the better the fit of the regression line to the observed value.

The R^2 of the multiple nonlinear regression equation fitted by the pure quadratic binomial regression model in this paper is 0.829, and the fitting effect is very good, in line with expectations. In addition, it can be seen from the figure above that the trend of the scattered points is the same as the trend of the broken line, which shows the accuracy of the multiple nonlinear regression model for fitting multiple nonlinear problems.

3.5. Explore the Pattern of Blood Oxygen Saturation Sequence

In order to explore the typical influencing factors of blood oxygen saturation and characterize a person with typical patterns derived from the exploration, the data visualization function of Python is used to visually display and analyze data. The main display form is bubble chart, and bubble chart. The size and shade of the color represent changes in the value.

In order to analyze whether the pattern of blood oxygen saturation is related to age, the relationship between blood oxygen saturation and age is visually displayed by analyzing the data and drawing a bubble chart.

As shown in the figure above, there is a strong relationship between age and blood oxygen saturation. It can be intuitively seen that the blood oxygen saturation of young people is generally higher than that of the elderly, and the blood oxygen saturation shows a random change within a certain range. With the trend of decreasing age.

Second, use Python to analyze the data and draw a visual bubble chart to explore the effect of smoking on blood oxygen saturation.

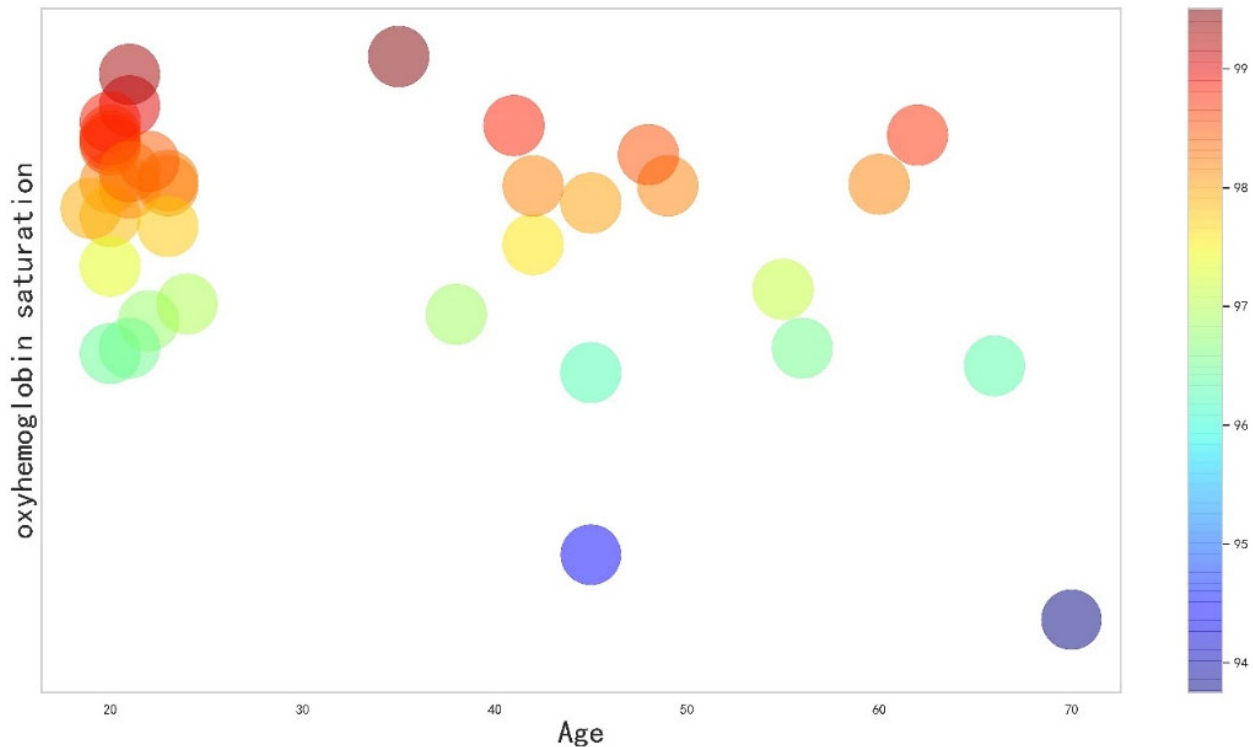


Figure 6. The relationship between age and blood oxygen saturation

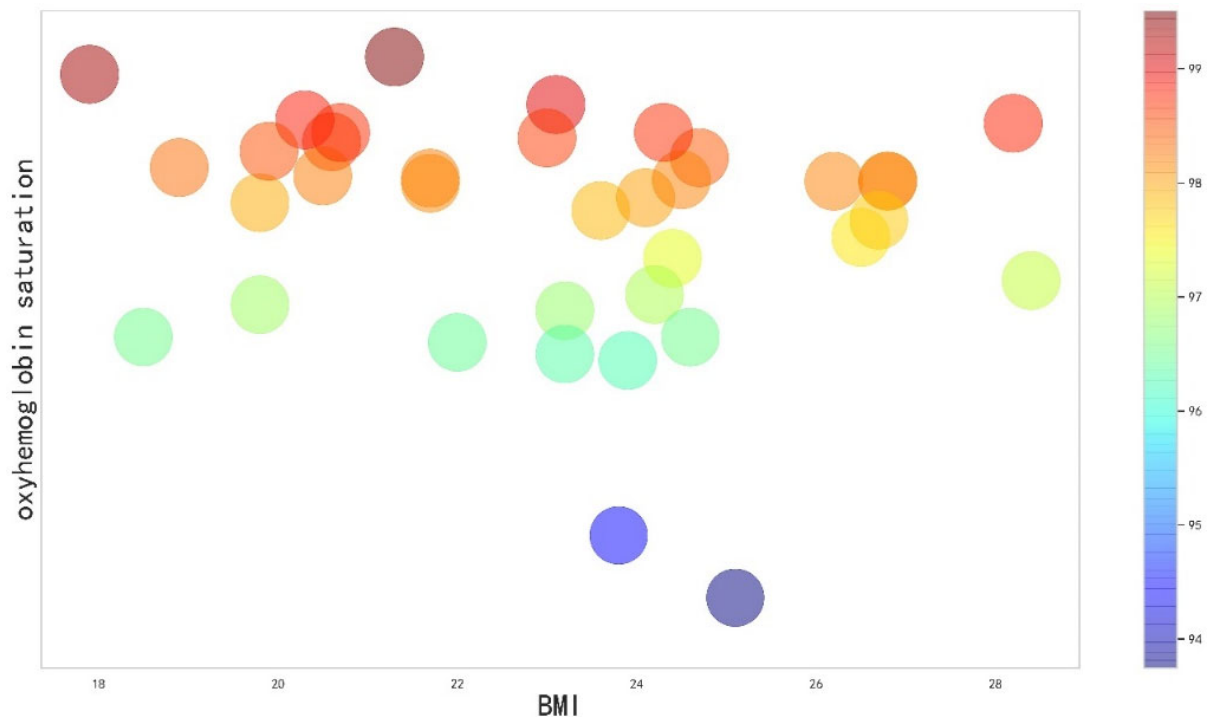


Figure 7. The relationship between BMI index and blood oxygen saturation

The picture above shows the relationship between the BMI index and blood oxygen saturation. We can directly see that when the human body's BMI index is at a normal level, the blood oxygen saturation reaches the optimal value; when the human body's BMI index is at a low or relatively high level At high abnormal levels, blood oxygen saturation will show a downward trend, and it does not reach the optimal value of human blood oxygen saturation.

4. Strength and Promotion

We have made perfect de-dimensionalization operations in data processing to make the analysis results more reliable.

Make the characteristics between different dimensions comparable, and eliminate the influence of the characteristic numerical magnitude caused by the dimension on the analysis results;

(2) After normalization, the numerical calculation problem is avoided without changing the change of the characteristic value.

Nonlinear regression prediction method/nonlinear regression analysis Nonlinear regression analysis is an extension of linear regression analysis, and it is also a structural model analysis of traditional econometrics. In social real economic life, the relationship between many phenomena is not linear. The analysis and prediction of this type of phenomenon generally requires nonlinear regression prediction. Through variable substitution, many nonlinear regressions can be converted into linear regressions. . Therefore, the linear regression method can be used to solve the problem of nonlinear regression prediction.

References

- [1] Information Technology - Data Processing; Findings in Data Processing Reported from Southern University of Science and Technology (SUSTech) (An adaptive QoS computation for medical data processing in intelligent healthcare applications)[J]. Information Technology Newsweekly,2020.
- [2] Patents; Researchers Submit Patent Application, "Systems and Methods for Medical Data Processing and Analysis", for Approval (USPTO 20150363560)[J]. Computers Networks & Communications,2016.
- [3] Information Technology - Information and Data Tabulation; Investigators at Ferdowsi University of Mashhad Report Findings in Information and Data Tabulation (An Efficient Storage-optimizing Tick Data Clustering Model)[J]. Computer Technology Journal,2020.
- [4] Information Technology - Data Caching; Studies from SNS College of Technology Reveal New Findings on Data Caching [Position-based Adaptive Clustering Model (Pacm) for Efficient Data Caching In Vehicular Named Data Networks (Vndn)][J]. Computers Networks & Communications,2020.
- [5] Comparing cotinine and NNAL verification of self-reported smoking status among Lung Cancer Screening eligible population from the 2007-2014 National Health and Nutrition Examination Survey (NHANES).[J]. Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals,2020.
- [6] Comparing cotinine and NNAL verification of self-reported smoking status among Lung Cancer Screening eligible population from the 2007-2014 National Health and Nutrition Examination Survey (NHANES).[J]. Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals,2020.