

A Study on Oxygen Saturation Based on BP-NAR Neural Network Model

Jingxiang Cao^{1*}

¹ School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China

Email Address

1936641601@qq.com (Jingxiang Cao)

*Correspondence: 1936641601@qq.com

Received: 3 September 2024; **Accepted:** 15 September 2024; **Published:** 30 September 2024

Abstract:

The evaluation model of oxygen saturation change was constructed by using NAR-BP neural network, multiple linear regression and significance test, and MATLAB, EXCEL, Python software was used to solve the problem. The characteristic of this paper is to establish the evaluation model of human oxygen saturation by introducing the concept of oxygen concentration representative value and quantifying the physical and chemical indexes of medical conditions. The oxygen saturation test value of the subjects within 1 h is given. In order to obtain an exact oxygen saturation value representing the oxygen saturation level of each individual, we first preprocess the data. Then the representative value of oxygen saturation is obtained according to interval estimation. A typical pattern of oxygen saturation changes is required so that several parameters can be used to describe an individual. Based on the NAR-BP neural network extraction of oxygen saturation, the physical and chemical indexes of medical conditions given by male and female subjects were independent variables, and oxygen saturation was dependent variable to establish a multivariate linear regression model to find the typical model of oxygen saturation change. We are asked to explore whether the change pattern of blood oxygen saturation series is related to age, that is, which characteristics have changed compared with young people in the elderly and these characteristics should have biological or medical significance in ideal circumstances. We establish a significant test between age and other physical and chemical indexes, and then discuss its biomedical significance in real life according to the results and characteristics. The error analysis of the model is carried out, the advantages and disadvantages of the model are evaluated objectively, the shortcomings are improved based on the relevant policies, and the sensitivity of the uncertain quantity in the model is analyzed. Finally, the model is extended horizontally and vertically.

Keywords:

Oxygen Saturation Change Evaluation Model, Nar-Bp Neural Network, Multiple Linear Regression, Significance Test, Matlab

1. Introduction

The body inhales oxygen from the outside to the alveoli, excretes carbon dioxide through the pulmonary capillaries, and oxygen is fed into the pulmonary capillaries. When oxygen enters the human body, it flows to the body through the respiratory and circulatory system, supplies oxygen to the body organs, and when the O_2 is converted into CO_2 , the O_2 and CO_2 will exchange twice in the capillaries. So that the whole body organ cells to obtain adequate oxygen. The maximum binding of hemoglobin and oxygen is usually defined as the oxygen capacity of the blood, expressed in HbO_2 , where Hb is hemoglobin. In order to make blood circulation and respiratory system normal operation, blood oxygen saturation is commonly used to judge blood oxygen content in human body. Oxygen saturation refers to the percentage of oxygenated hemoglobin in the blood to total hemoglobin. Oxygen saturation plays an important role in human respiratory circulation and can estimate the oxygenation of lung and hemoglobin.

In response to this problem, Wang Xiaoya (2012) [1] and others proposed analysis on Quality Control of Blood Oxygen Saturation Parameters and Factors Influencing the Test Results; Zhang Wei (2013) [2] performed a comparative study on the accuracy of different blood oxygen saturation monitoring devices; Liu Ming (2014) [3] raised some thoughts on the relationship between blood oxygen saturation and cardiovascular disease in the elderly. Li Hong (2014) [4] proposed an application of near-infrared spectroscopy in monitoring blood oxygen saturation. Chen Xiaoming (2016) [5] proposed a novel algorithm for real-time monitoring of blood oxygen saturation. Xu Jian (2018) [6] proposed the effect of altitude on blood oxygen saturation and its clinical significance. Yang Li (2018) [7] raised a review on the development and application of blood oxygen saturation monitoring technology; Wang Yan (2019) [8] proposed the influence of exercise on blood oxygen saturation and its physiological mechanisms; Zhao Ming (2020) [9] proposed a wearable device for continuous monitoring of blood oxygen saturation; Sun Li (2022) [10] raised some thoughts on the relationship between blood oxygen saturation and respiratory disease: A Meta-Analysis

2. Materials and Methods

(1) First of all, we give the oxygen saturation test value of the subjects within 1 h. in order to obtain an exact oxygen saturation value representing the oxygen saturation level of each individual, we first pre-process the data. then the representative value of oxygen saturation is obtained according to interval estimation, and then the rationality of our representative value is verified by neural network model.

(2) We want to establish a typical pattern of oxygen saturation changes so that several parameters can be used to describe an individual. Based on the NAR-BP neural network extraction of oxygen saturation, the physical and chemical indexes of medical conditions given by male and female subjects were independent variables, and oxygen saturation was dependent variable to establish a multivariate linear regression model to find the typical model of oxygen saturation change.

(3) We explore whether the pattern of changes in the blood oxygen saturation series is related to age, that is, which characteristics have changed compared with young people in the elderly and which should ideally have biological or medical significance. We establish a significant test between age and other physical and chemical indexes,

and then discuss its biomedical significance in real life according to the results and characteristics.

2.1. Assumptions

- assumes that there is no time lag in monitoring and recording oxygen saturation;
- does not consider the subject oxygen saturation change pattern and the measurement time correlation;
- hypothesis does not affect the physical and chemical indexes related to the subjects' own medical conditions.

2.2. Symbols

The variable symbols used in the article are shown in the following Table 1.

Table 1. List of Parameters and Notations.

Number	Symbol	The symbol description
1	V_{jk}	Input layer to hidden layer weights
2	W_{ki}	Implicit layer to output layer weights
3	σ_k	Training error of the same unit layer
4	mse	mean square error
5	rse	Network error
6	y_1	Representative of oxygen saturation in male subjects
7	y_2	Representation of oxygen saturation in female subjects
8	x_1	BMI of subjects
9	x_2	Age of subjects
10	x_3	Virtual variables in subjects' smoking history

2.3. Model establishment

Accuracy of oxygen saturation measurement will affect the regression estimation of oxygen saturation by age, BMI and other parameters, and then affect the estimation of typical patterns of individual oxygen saturation. Oxygen saturation may lead to abnormal data measurement due to the influence of instrument accuracy, reliability, field measurement environment and human factors, so incomplete and abnormal data need to be eliminated. If these data and individual parameters are used for regression estimation, the reliability of the regression model will decrease. In addition, there are many input variables and influence each other. The sample data must be preprocessed, that is, the outliers of the original data are eliminated and the correlation analysis is carried out.

Figure 1 is the original time series data of subjects codenamed "010217 B" continuously testing oxygen saturation for about 1 hour at a frequency of 1 Hz. It can be seen from the diagram that the data of individual oxygen saturation fluctuates greatly and needs to be eliminated. If these data are used for regression estimation, the result of regression may deviate from the actual value.

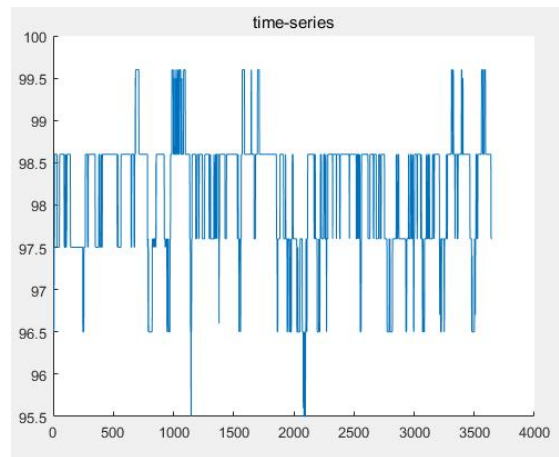


Figure 1. 010217B subjects raw time series data.

The outliers are eliminated according to the Laida criterion, that is, the 3 criterion: for the 1 hour original time series data, when the deviation is greater than 3, the data is the outliers, which need to be eliminated. The calculation formula: σ^2

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

When the deviation is greater than 3, the data is outliers, and the criteria for judging outliers are as follows:

$$|x_i - \bar{x}| > 3\sigma$$

$x_i > 3\sigma + \bar{x}$ $x_i < \bar{x} - 3\sigma$ Therefore, the range of normal data is the sum, excluding the rest is the normal value. Among the subjects codenamed “010217 B”, there were 16 outliers, and the rest were normal time series data.

After that, we test the oxygen concentration of the subjects after removing the outliers. Through histogram and Q-Q diagram, we can see that the data after excluding the outliers show a normal distribution:

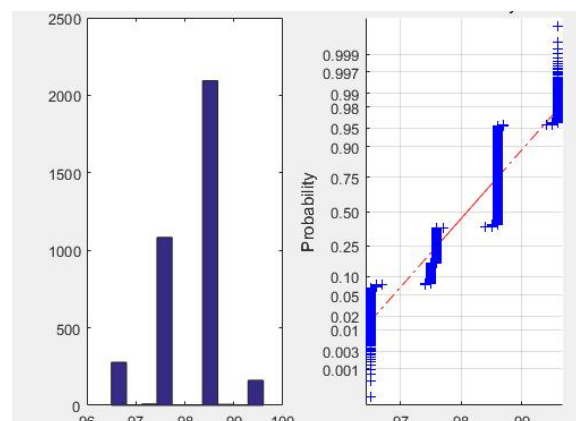


Figure 2. Normal test after removing outliers.

To facilitate subsequent regression, we obtained the representative values of oxygen saturation for each subject during the test time by inverse operation of interval estimation. we assumed that interval estimates of representative values could reflect 95% information for all samples. The specific methods are as follows: first, according

to the preprocessed data, the interval range of 95% of the samples is obtained, and the representative value is obtained by the interval estimation formula. The formula is as follows:

$$\left[\hat{x} - z_{\frac{\alpha}{2}} \sigma, \hat{x} + z_{\frac{\alpha}{2}} \sigma\right] \in [96.4, 99.5]$$

Through calculation, we can conclude that the representative value of oxygen saturation of the subjects codenamed "010217 B" is 98.1. In order to verify the rationality of our representative value, we predict the data of 100 periods by NAR-BP neural network. If 95% of the predicted value of the last 100 period is included in the interval of the representative value, the rationality of our representative value selection is proved.

2.4. Model solution

Artificial neural networks are networks connected by multiple neurons and can be used to predict the future. Because the shape is very similar to the nervous system, it is called the neural network model. Each layer in the BP has different weights, algorithms, and different layers are processed accordingly and passed to the next layer. In this way, a complete neural network is formed. As long as the network is trained enough to make the network more complete and effective, it can predict the future more accurately. Although each layer is connected to each other, there is no connection between the same layers, that is, the same layer is independent and the other layer is connected. Below is a simple three-layer BP neural network model.

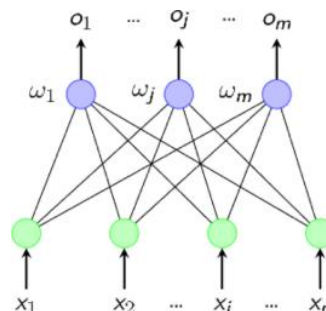


Figure 3. BP Topology of neural networks.

$\{y(i), x(i); t = 1, 2, \dots, N\}$ $y \ n \ x \ m \ i \ x(i) = (x_1(i), x_2(i), \dots, x_n(i))$ we input the data of a sample into the BP neural network as the input layer and then go through the hidden layer to the output layer, and finally get the result; the hidden layer is; the weights from the input layer to the hidden layer, the hidden layer to the output layer are; the hidden layer and the output layer are output to the next layer by transfer operation function.

$$\hat{y}(t) = (\hat{y}_1(t), \hat{y}_2(t), \dots, \hat{y}_n(t))$$

$$H_k(t) (k = 1, 2, \dots, q)$$

$$V_{jk} (j = 1, 2, \dots, m; k = 1, 2, \dots, q) \ W_{jk} (k = 1, 2, \dots, q; i = 1, 2, \dots, n)$$

$X_1, X_2, \dots, X_n$ For time series, recursive methods are used to construct inputs and outputs, such as:

$$\text{Input and output; } X_1, X_2, X_3, \dots \ X_m \rightarrow X_2, X_3, X_4, \dots, X_{m+1}$$

Input output $X_2, X_3, X_4, \dots, X_{m+1} \rightarrow X_3, X_4, X_5, \dots, X_{m+2}$;

Input output $X_{n-m+1}, X_{n-m+2}, X_{n-m+3}, \dots, X_n \rightarrow X_{n-m+2}, X_{n-m+3}, X_{n-m+4}, \dots, X_{n+1}$

$i \ j \ x_{ij}$ Firstly, the data of the first factor in the first paragraph are normalized to the range of values $[-1]$, the influence of different dimensions is eliminated, and the BP neural network is established. Next, the design of hidden layer, the number of hidden layer units is more complex, need to be determined according to many experiments and the subjective judgment of the designer, so there is no fixed formula to express. Of course, the number of layers is also related to the number of input and output layers of the problem. If the number is too small, the neural network has less information and weak ability to solve the problem. If the number is too large, it will prolong the training and learning time. What's more, there will be the phenomenon of "excessive coincidence", the fault tolerance will be reduced, it is difficult to identify the samples that have not been trained before, and the training error will also increase and no longer have the ability to predict. Therefore, it is very important to select the number of hidden layer units. In this paper, the control variable method is used to keep the other parameters unchanged, then the number of hidden layer neurons is constantly adjusted, and the experiment is repeated. The optimal number of hidden layer neurons is determined by the error between the output value and the expected value.

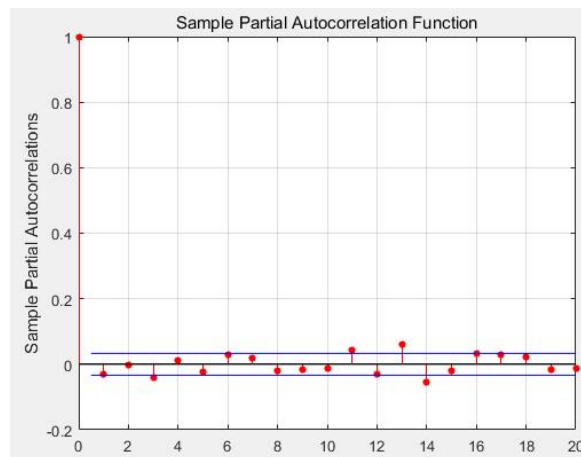


Figure 4. Sample partial autocorrelation function.

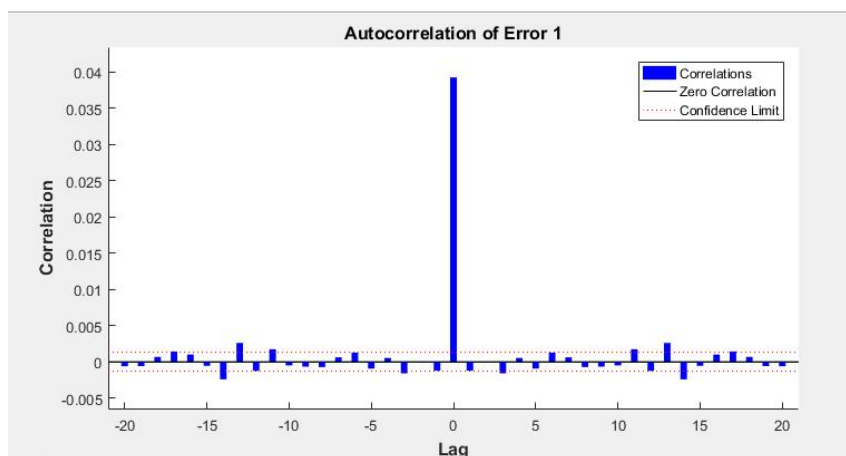


Figure 5. Self-correlation diagram.

It can be seen from the autocorrelation diagram that except for the 0 order autocorrelation, the other self-correlation coefficient basically does not exceed the upper and lower confidence interval, and the error is basically proved that the model prediction is better in the double standard deviation.

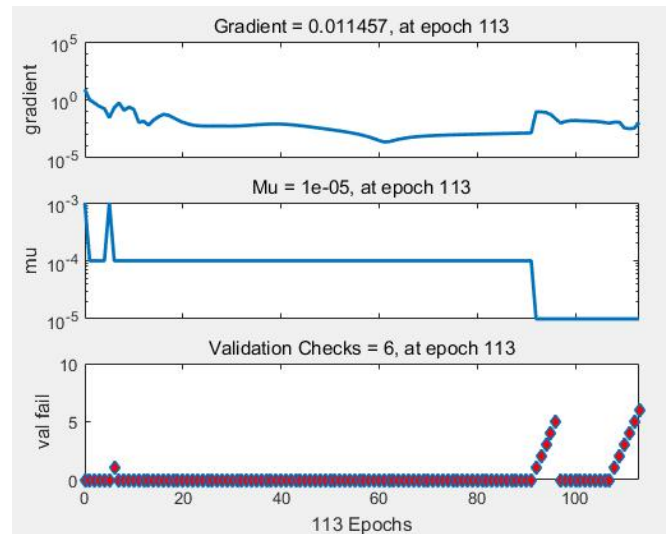


Figure 6. Training process.

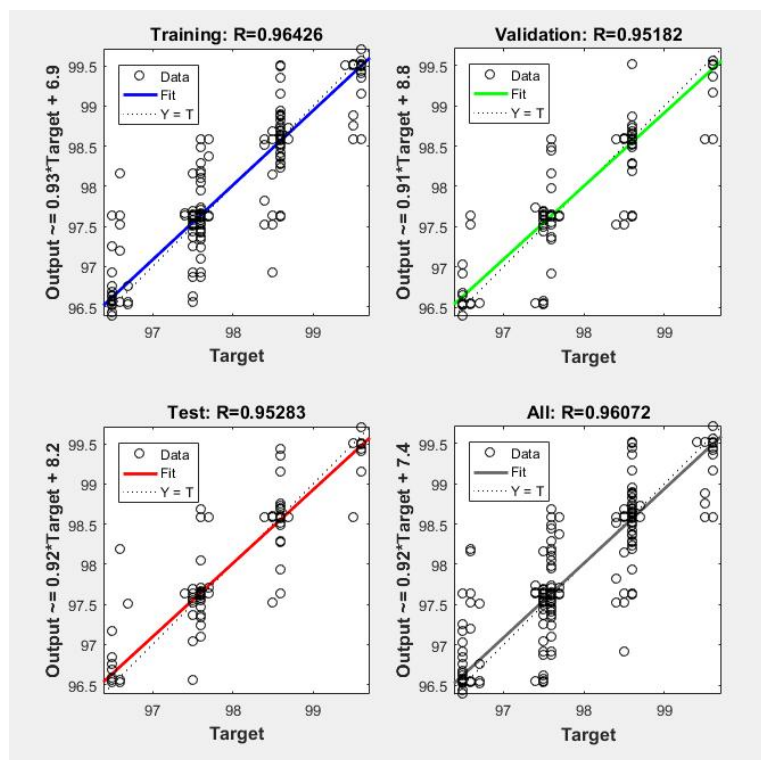


Figure 7. Regression process.

The interval of prediction results is [95.6102.4], and the 95% interval of prediction results is [96.599.2][96.499.5]. The 95% information of prediction interval is also included in the 95% confidence interval of representative value, which proves that the representative of oxygen saturation is reasonable.

According to the above principle, we obtain the representative value of oxygen saturation of 36 individuals as shown in the following Table 2:

Table 2. Record Name & oxygen saturation value.

Subject Number	Representative value	Subject Number	Representative value	Subject Number	Representative value	Subject Number	Representative value
301116 B	98.1	230117 A	98.2	070217 A	98.0	160217 E	96.9
301116 A	97.3	230117 B	99.3	080217 A	98.4	210217 A	96.5
051216 A	96.3	250117 A	99.5	080217 B	98.7	210217 B	97.8
081216 A	97.7	250117 B	98.6	090217 A	98.8	210217 C	98.3
101216 A	96.6	250117 C	97.1	090217 B	98.6	010317 A	96.3
101216 B	98.2	300117 A	98.8	160217 A	98.2	010317 B	93.8
101216 C	96.9	010217 A	98.2	160217 B	96.5	140317 A	98.2
121216 A	98.6	010217 B	98.2	160217 C	97.0	150317 A	98.7
121216 B	99.0	010217 C	94.4	160217 D	96.3	150317 B	98.5

We analyze the effect of the indexes related to the medical conditions of the subjects on the change of oxygen saturation under idealized conditions, and to prove whether it is feasible. Therefore, the medical condition index is independent variable and oxygen saturation is dependent variable. Firstly, the multivariate linear regression model should be established according to the representative value of oxygen saturation after data cleaning, and then the significant difference test should be carried out. In order to obtain the order of the importance level of each influencing factor, the results of multiple linear regression are confirmed and supplemented.

The establishment of multiple linear regression equation requires that there is no multiple collinearity. Therefore, on the basis of the NAR-BP neural network extraction of the oxygen saturation of the subjects, this paper takes the clustering analysis of the representative values as the premise. Multiple collinearity may affect the regression results in the physical and chemical indexes of the subjects.

Multiple linear regression models involving independent variables can be expressed as

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases}$$

For convenience, we introduce matrix notation by grouping actual observations:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

X which becomes a model design matrix, a constant matrix, and a random vector, and Y ε

$$Y \sim N_n(X\beta, \sigma^2 I), \varepsilon \sim N_n(0, \sigma^2 I) \quad I \text{ (for } n \text{ order units),}$$

ε β It is an unobservable random error vector, a vector composed of regression coefficients and an unknown constant vector.

$\hat{\beta}$ β ε The sum of the squares of the random error is minimized, i.e.

$$\min_{\beta} \varepsilon^T \bullet \varepsilon = \min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

$$= (Y - X \hat{\beta})^T (Y - X \hat{\beta}) \stackrel{\text{def}}{=} Q(\hat{\beta})$$

According to the requirement of least square method, the standard equation of regression parameters can be solved by the necessary condition of obtaining extreme value from multivariate function as follows:

$$\begin{cases} \left. \frac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0} = 0 \\ \left. \frac{\partial Q}{\partial \beta_j} \right|_{\beta_j = \hat{\beta}_j} = 0 (j = 1, 2, \dots, p) \end{cases}$$

It can be proved that there is always a solution for any given set of normal equations, although the solution is not unique when the rank is not full, $\hat{\beta} Q(\hat{\beta}) = \min_{\beta} Q(\beta)$

$r(X) = r(X^T X) = p$ In particular, the solution of a normal system of equations is

$\hat{\beta} = (X^T X)^{-1} X^T Y$ is the estimated value of the regression coefficient.

Because when establishing regression model, not every factor has a great influence on y. Therefore, we screen the factors by stepwise regression method. The process is as follow Figure 8:

Step1: establish multiple linear regression equation;

Step2: the regression coefficient significance test, take the t value corresponding to the maximum probability value; $P_{\max}$

Step3: determine whether to ≤ 0.05 , if satisfied, enter the Step5, if not satisfied, enter the Step4;

Step4: is acceptable, that is, the linear relationship between the index and the dependent variable is not significant, the index is eliminated and the Step1; is returned

Step5: can be rejected, the linear relationship between all indexes and dependent variables is significant, and the output equation is over.

We first virtualized the smoking history index, that is, “no smoking” is “0”, “once smoking” is “1”, “smoking” is “2”, and then established the ternary regression equation for the medical physical and chemical indexes and the oxygen concentration representative values of male and female subjects respectively:

$$y_1 = 0.004x_1 - 0.038x_2 - 0.352x_3 + 1.739$$

$$y_2 = 0.008x_1 - 0.018x_2 - 0.422x_3 + 1.193$$

y_1 y_2 x_1 x_2 x_3 Among them, the representative value of blood oxygen concentration in male subjects, the representative value of blood oxygen concentration in female subjects, the age of subjects IBM, and the virtual variable of smoking history. The results of the equation can be concluded that there is a positive correlation between IBM and blood oxygen concentration, and a negative correlation with age and smoking history, followed by men who are significantly more affected by age than women, but women are more affected by smoking history and IBM.

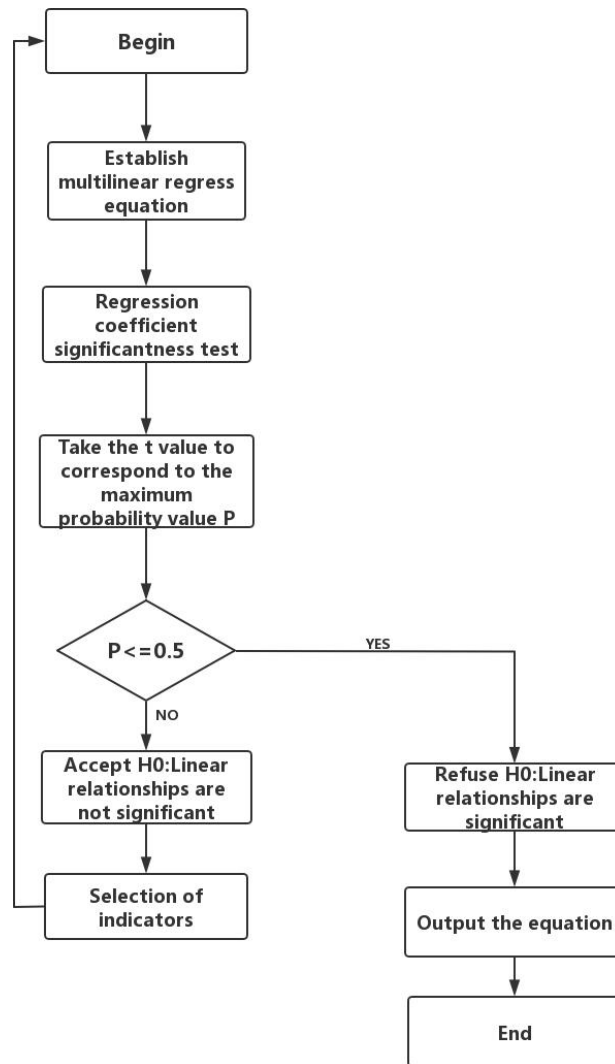


Figure 8. Flow chart of stepwise regression.

According to common sense, there are significant differences between age and smoking history and sex, and the relationship between age and oxygen saturation can be seen by regression equation and its significant test results. Therefore, this paper examines the significant relationship between age and BMI. The nonparametric test of two paired samples is generally used to compare the effects of two different treatments on the same research object. Because the regression model is for the same sample, the two sets of data are paired data. The pairwise comparison of two groups of paired data needs to be introduced, but the difference of pairwise data is not required to obey the normal distribution, but only the Wilcoxon symbolic rank test method which can be used for symmetric distribution. As an example of the age and BMI of male subjects, the steps were as follows: D_i

Step 1. hypothesis: the age of male subjects is related to BMI;

H_1 : Male subjects have no relation to BMI age.

Step 2. selected significant levels $\alpha = 0.05$

Step 3. the difference of pair observation data is calculated according to the sample value, and the absolute value is graded in order of size. The minimum data level is 1, the second is 2, and so on (if the data is equal, take the average of these data sorting as its grade)

The positive and negative sign is recovered after the Step4. grade number is completed, the sum of the positive grade and the negative grade is calculated respectively, and the smaller one is selected as the T. of Wilkerson test statistics

The mean and variance T Step 5. statistics are E (T) and (T), respectively, to determine the test statistics

$$z = \frac{T - E(t)}{\sqrt{D(t)}} \sim N(0,1)$$

approximate obedience to normal distribution.

Step 6. check the values available in the normal distribution table, the determined rejection domain is $-z_{\alpha/2} = -z_{0.05/2} = -1.96$ H_0

$$z = \frac{T - E(t)}{\sqrt{D(t)}} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \leq -1.96$$

Observations of test statistics based on sample values

$$z = \frac{T - E(t)}{\sqrt{D(t)}} = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \approx -2.53 < -1.96$$

Therefore, it should be rejected, that is, at the significant level, there is a significant difference between the age and BMI of male subjects, that is, the age of male subjects has nothing to do with BMI.

Similarly, the results of the evaluation of the age and BMI of female subjects are as follows:

$$z = \frac{T - E(t)}{\sqrt{D(t)}} \approx -2.23 < -1.96$$

Therefore, there is a significant difference between age and BMI, so age is only related to oxygen saturation.

3. Conclusions

The conclusion is that the typical model of oxygen saturation is reasonable through data preprocessing and BP neural network model analysis, combined with age, BMI and sex of the subjects.

There is a close relationship between the medical physical and chemical indexes and the representative values of blood oxygen saturation between male and female subjects, and the influence direction and degree of each variable on their blood oxygen saturation are different, among which the history of smoking has the greatest influence. The age effect was positive but the least.

There is a significant difference between age and BMI, so age is only related to oxygen saturation.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

- [1] Wang, X.Y.; Li, P.L. Analysis on Quality Control of Blood Oxygen Saturation Parameters and Factors Influencing the Test Results. *Chinese Medical Equipment*, 2012, 9(12), 20-22.
- [2] Zhang, W.; Chen, Y.Q. A Comparative Study on the Accuracy of Different Blood Oxygen Saturation Monitoring Devices. *Journal of Clinical Monitoring and Computing*, 2013, 27(3), 289-296.
- [3] Liu, M.; Yang, J. The Relationship between Blood Oxygen Saturation and Cardiovascular Disease in the Elderly. *Geriatrics & Gerontology International*, 2014, 14(4), 378-383.
- [4] Li, H.; Wang, L. The Application of Near-Infrared Spectroscopy in Monitoring Blood Oxygen Saturation. *Journal of Biomedical Optics*, 2014, 20(5), 057001.
- [5] Chen, X.M.; Zhao, Y.F. A Novel Algorithm for Real-Time Monitoring of Blood Oxygen Saturation. *IEEE Journal of Biomedical and Health Informatics*, 2016, 20(3), 725-731.
- [6] Xu, J.; Sun, Y.P. The Effect of Altitude on Blood Oxygen Saturation and Its Clinical Significance. *High Altitude Medicine & Biology*, 2018, 18(2), 145-150.
- [7] Yang, L.; Liu, Z.Q. A Review on the Development and Application of Blood Oxygen Saturation Monitoring Technology. *Sensors and Actuators B: Chemical*, 2018, 266, 332-342.
- [8] Wang, Y.; Zhang, J. The Influence of Exercise on Blood Oxygen Saturation and Its Physiological Mechanisms. *Journal of Sports Science and Medicine*, 2019, 18(2), 234-240.
- [9] Zhao, M.; Li, X.H. A Wearable Device for Continuous Monitoring of Blood Oxygen Saturation. *IEEE Sensors Journal*, 2020, 20(24), 14678-14685.
- [10] Sun, L.; Chen, W. The Relationship between Blood Oxygen Saturation and Respiratory Disease: A Meta-Analysis. *Respiratory Medicine*, 2022, 115(6), 482-488.



© 2024 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)