Universe Scientific Publishing

Multiple Linear Regression Model of Blood Oxygen Saturation

Jinghao Yao, Yuhan Gao, Xueqing Dong

Shenyang Aerospace University, Shenyang 110136, China.

Abstract: We used standardized regression analysis to analyze the influence of age and recent smoking status on blood oxygen saturation. We found that blood oxygen saturation was negatively correlated with age and recent smoking status, but whether smoking had a greater influence on blood oxygen saturation than age. In order to further explore whether age and recent smoking status could jointly affect oxygen saturation of blood, we used k-means clustering algorithm and took age as the control variable to conduct clustering analysis. Is obtained by polynomial fitting, and then the blood oxygen saturation under all ages about age, smoking status of recent expression of partial derivatives of recent smoking status, get the following conclusion: smoking is oxygen desaturation, and compared with the young, the elderly smoking more effect on the blood oxygen saturation degree, are more likely to suffer from cardiovascular disease.

When the regression results were tested for significance, the significance of all variables was investigated, and it was found that only age and current smoking status were significant for oxygen saturation, while BMI and gender showed no difference from zero in T test, indicating that oxygen saturation was not affected by BMI and gender. Then, taking oxygen saturation as the dependent variable and age and current smoking status as independent variables, the spline interpolation method with the best fitting effect was found, and its expression was given in the text.

Keywords: Standardized Regression; Neural Network

1. Introduction

Pulse oximetry is often used to test the patient's blood oxygen saturation level, during the period of continuous testing, through the model description of oxygen saturation model, for data of 36 people, each subject 1 Hz frequency, about an hour of continuous measurement of oxygen saturation, recorded the information about the participants, including age, gender, smoking status at present.

Question: In order to understand whether oxygen saturation is related to age, that is, what characteristics have changed in older people compared with younger people, the characteristics should have biological or medical significance.

2. Problem analysis

In view of the problem, we can know from the first ask conclusion blood oxygen saturation mainly related to age and smoking status, age, and smoking then using standardized regression analysis for the influence of oxygen saturation degree, and found that smoking status and age of blood oxygen saturation were negatively correlated, but carries on the data analysis and found smoking more influence on the blood oxygen saturation, that is to say, the influence degree of the age less than smoking status, in order to further analyze the association between age and blood oxygen saturation, we classified by using the method of cluster analysis, ages, and then use matlab for polynomial fitting, In the same way, the data of three age groups were substituted, and then the partial derivative of smoking was obtained to analyze the influence of smoking at different age groups on blood oxygen saturation. To find out the corresponding biological or medical significance of smoking in the elderly in relation to cardiovascular disease.

3. The problem is to establish and solve the model

3.1 Identification of the degree of influence of oxygen saturation based on standardized regression analysis

Based on the multiple linear regression model established by the first question, we have learned that oxygen saturation is only associated with age and recent smoking status, but not with BMI and sex. Therefore, in order to further study the influence of Age and recent Smoking status on human blood oxygen saturation, we conducted standardized regression with blood oxygen saturation as the dependent variable and Age and recent Smoking status as the independent variable.

Standard regression refers to the regression analysis carried out after eliminating the influence of units taken by dependent variables and independent variables. The size of standardized regression coefficient reflects the influence degree of each independent variable on dependent variables. The comparison results of normalized regression coefficients are only applicable to a specific environment, and they may vary from time to time and place to place.

DE dimensionality treatment:

For the samplex_1,x_2, x_3……x_n, And then after dimension $x_{stdi} = \frac{x_{i-\mu}}{\sigma}$

(where μ is the sample mean and σ is the sample standard deviation)

After treatment with {Oximetry} _I, Age and Smoking by stata , normalization regression was made. The results are shown as follows

oximetry	Coef.	Std. Err.	t	P>|t|	Beta
smoking1	0	(omitted)			0
smoking2	-.3972196	.5860163	-0.68	0.503	-.1343429
smoking3	-1.965321	.8380077	-2.35	0.025	-.4418871
Age	-.0380873	.0127138	-3.00	0.005	-.4918497
_cons	99.46954	.8152098	122.02	0.000	.

Figure1 Standardized regression analysis table for blood oxygen saturation

As shown in figure, standardized regression coefficient absolute value reflect the Age, the effects of Smoking on the blood oxygen saturation degree, the size of the absolute value expressed the influence degree of the independent variable on the dependent variable, such as can be seen table Smoking3 standardized regression coefficients was significantly greater than the Age, Smoking2, Smoking1, suggesting that the recent Smoking status of blood oxygen saturation for Smoking is greatest, Age for small affect blood oxygen saturation, quit Smoking relative to Smoking almost to won't affect the dependent variable.

3.2 Age-based K-means clustering model

In order to further explore whether age combined with recent smoking status could affect blood oxygen saturation, we used k-means clustering algorithm and took age as the control variable to conduct clustering analysis.

The K-means clustering algorithm is a clustering analysis algorithm solved through iteration. Its steps are as follows: first, specify the number of clustering centers as K, then randomly select K objects as the initial clustering center, calculate the distance between each object and each sub-clustering center, and assign each object to the nearest clustering center. The cluster center and the objects assigned to it represent a cluster. For each sample assigned, the cluster center of the cluster reiterates the calculation based on the existing objects in the cluster. This process is repeated until the termination condition is met.

Its principle mind map is as follows:

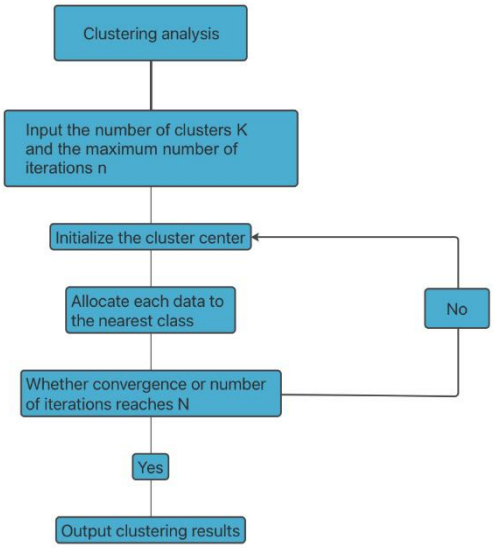

Figure2 Principle mind map

SPSS was used to conduct cluster analysis with age as classification variable and patient number as case labeling basis. The results are as follows:

Table1 Initial cluster center

Control variables	clustering		
	1	2	3
Age	19	49	70

Table2 Iteration history [a]

The iteration	Changes in clustering centers		
	1	2	3
1	2.050	3.917	5.500
2	.000	2.083	3.000
3	.000	.000	.000

The initial cluster centers were set as three, aged 19, 49, and 70 respectively. In each iteration, the samples were redistributed to the cluster centers according to the distance from each sample to each cluster center, and then the cluster center location was updated. After three iterations, the convergence is realized because there is no change or only slight change in the clustering center, and the minimum distance between the initial centers is 21.000.

3.3 Analysis model of the influence degree of blood oxygen saturation in different age groups based on polynomial fitting

Patients who are divided into three groups through k-means clustering algorithm adopt polynomial fitting method to fit the sample data of different age groups, and the equation obtained is shown in the following table:

Age	Fitting equation
[19,24]	$Oximetry_1(Age,Smoking) = 100.4 - 0.1034Age - 0.1257Smoking$
[35,49]	$Oximetry_2(Age,Smoking) = 101.7 - 0.0465Age - 1.738Smoking$
[55,70]	$Oximetry_3(Age,Smoking) = 112.6 - 0.1849Age - 2.426Smoking$

$Oximetry_i(Age,Smoking)$ Furthermore, to further discuss the effect of Oximetry on Smoking, the partial derivatives of Oximetry (Age, Smoking) are obtained from the three equations, and the partial derivatives of Oximetry (Age, Smoking) with respect to the variable Smoking are shown in the table below.

The age range	$\frac{\partial Oximetry}{\partial Smoking}$
[19,24]	-0.1257
[35,49]	-1.738
[55,70]	-2.426

As can be seen from the graph, the absolute value of Oximetry gradually increases with the increase of age, which indicates that the effect of Oximetry on blood oxygen saturation has increased with the increase of age. Therefore, the elderly should pay more attention to the effect of Smoking on cardiovascular disease.

Therefore, we note that although oxygen saturation decreases with age, increasing the risk of cardiovascular disease, smoking has a greater impact on oxygen saturation in older people, so older people should pay more attention to lifestyle to prevent cardiovascular disease.\

4. Model test

In question 1, we used a multiple linear regression model to analyze the independent variables that may influence oxygen saturation. The final conclusion was that oxygen saturation was only correlated with age and recent smoking status, but not with BMI and gender.

For problem 1, BP neural network is used to test the model.

BP neural network is chosen because it has the following advantages:

(1) Nonlinear mapping capability: BP neural network essentially realizes a mapping function from input to output. Mathematical theory proves that the three-layer neural network can approximate any nonlinear continuous function with arbitrary precision. This makes it especially suitable for solving complex internal mechanism problems, that is, BP neural network has strong nonlinear mapping capability.

(2) Self-learning and adaptive ability: DURING training, BP neural network can automatically extract the output and "legal rules" among output data through learning, and memorize the learned content in the network weight adaptably. It shows that BP neural network has highly self-learning and self-adapting ability.

(3) Fault-tolerant ability: BP neural network will not have a great impact on the global training results after local or partial neurons are damaged, that is to say, the system can still work normally even when local damage occurs. That is, BP neural network has a certain fault-tolerant ability.

We know from the analysis of problem number one. As one of the two important factors that mainly affect oxygen saturation, age and smoking status of 32 individuals (converted into dummy variables) were substituted into BP neural network, and 70% training volume and 15% validation15% Testing were set.

The neural network structure as shown in the figure below is used:

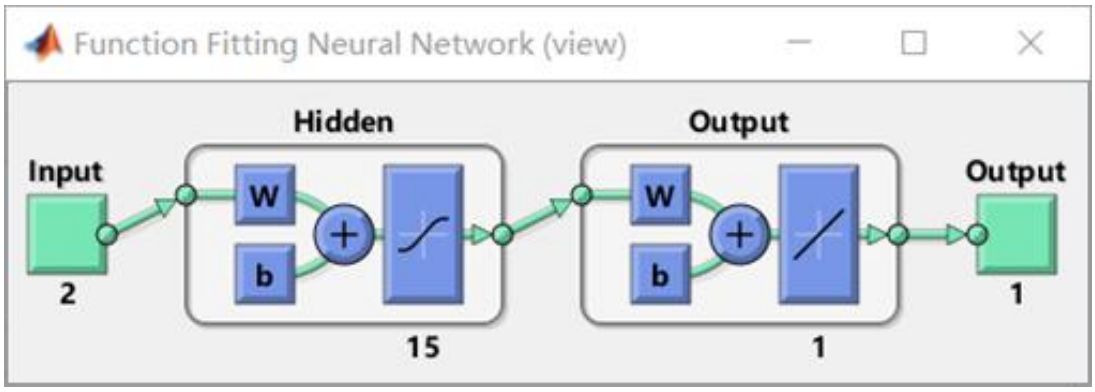

Figure3 Function Fitting Neural Network

After the Epoch after 1000 times, the Regression obtained is shown in the following figure

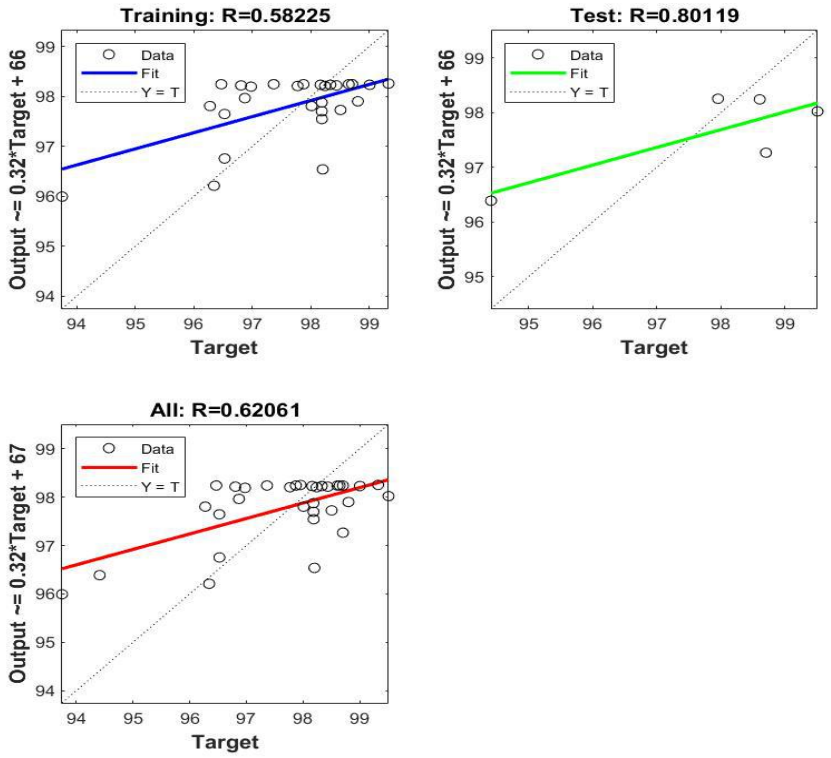

Figure4 The training data

We will use this neural network to predict the last four groups with SIM function, and obtain their relative errors as shown in the figure

It can be seen from the figure that the relative error is always below 2%, which indicates that the multiple regression model in question 1 is more reasonable. For this model test, we can see that under the same conditions, the difference between the results of multiple tests is small, which proves that the model establishment in question 1 in this paper is more in line with the actual situation.

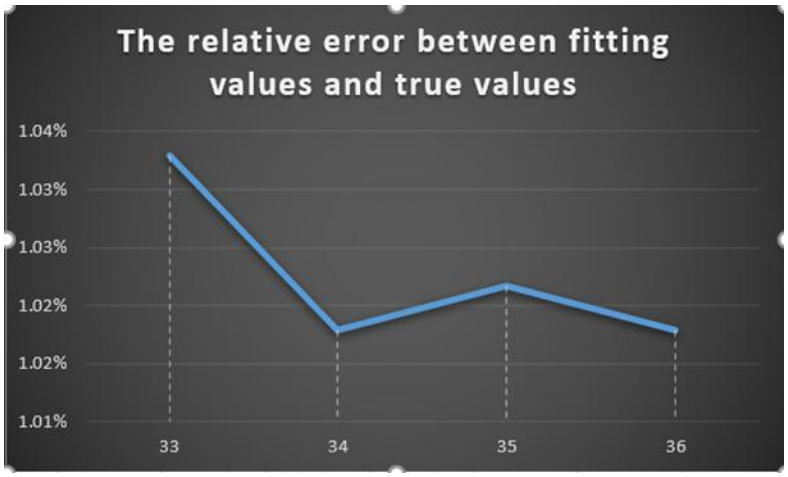

Figure5 The relative error between fitting value and true values

References

[1] Zhang Z, Pi Z, Liu B. TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise[J]. Biomedical Engineering, IEEE Transactions on, 2015, 62(2): 522-531.

[2] Sun X, Yang P, Li Y, et al. Robust heart beat detection from photo plethysmography interlaced with motion artifacts based on empirical mode decomposition [C] Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on. IEEE, 2012: 775-778.

[3] Wei, H.L., Billings, S.A., Liu, J.J., Time-varying parametric modelling and time-dependent spectral characterisation with applications to EEG signals using multiwavelets [J]. International Journal of Modelling, Identification and Control, 2010, 9(3): 215-224.

[4] Mallat, S.G., A theory for multiresolution signal decomposition: the wavelet representation[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1989, 11(7): 674-693.