MEDS Clinical Medicine (2022)
Clausius Scientific Press, Canada
DOI: 10.23977/medsc.2022.030315
ISSN 2616-1907 Vol. 3 Num. 3

Statistical-based approach to the analysis of blood oxygen saturation

Lixin Liu

Intelligent Equipment Academy, Shandong University of Science and Technology, Tai'an, Shandong 271001, China

Keywords: Multiple regression linear analysis, Digital quantification, Statistical analysis, MATLAB, SPSS

Abstract: Blood oxygen saturation refers to the percentage of the volume of oxygenated hemoglobin bound by oxygen to the total volume of hemoglobin that can be bound in the blood. Blood oxygen saturation can reflect the oxygen carrying capacity of human blood. It is clinically believed that blood oxygen saturation is closely related to the patient's blood oxygen content. The physiological signal of blood oxygen saturation occupies a very important position in measuring our physical health indicators. Monitoring blood oxygen saturation can not only reduce the incidence of cardiovascular and cerebrovascular diseases, but also keep abreast of the current physical condition. we used quantification methods such as average calculation and digital quantification to quantify the four types of information of participants. First, by integrating the data given by the topic, we obtained relevant data about the participants, including the age, BMI, gender, smoking history (current smoking status), blood oxygen saturation, average blood oxygen saturation, and standard deviation of blood oxygen saturation. We use MATLAB to calculate the average blood oxygen saturation and the standard deviation of blood oxygen saturation by using the blood oxygen saturation per hertz given to each person in one hour. We will take the average oxygen saturation of each person as the dependent variable, age, BMI, gender, and smoking history as the independent variables, bring the quantified data into the model, and use MATLAB for multiple linear regression calculations to obtain the corresponding regression coefficients, and the correlation coefficient, P value, error, etc., to analyze the regression results. Thus, the regression equation is obtained, and the relevant parameters are obtained to characterize a person.

1. Introduction

We need to study whether blood oxygen saturation is related to age. The title prompts us to study which characteristics have changed compared with the elderly and young people. First, we look up the information to determine the definition of young people and old people. We need to discuss the elderly and young people separately according to age, and look at the changes in blood oxygen saturation concentration under the age factor. At the same time, we have to analyze the characteristics of different factors. We searched for information and found more samples for analysis. We analyzed gender, smoking history, etc., and used SPSS and other software for statistical analysis to obtain physical characteristics. What changes have taken place, and on this basis, speculate what medical significance these features should have under ideal circumstances. By analyzing the above two issues, we have analyzed the important role of blood oxygen saturation in the human body. At the same time, we must always pay attention to our health, and always pay attention to whether the various data indicators of the body meet the standards, so that we can live a healthy life. On the basis of analyzing the problem, we explained the advantages and disadvantages of the multiple linear regression model we established, and on this basis, we extended the multiple linear regression model.

1.1 Background

In recent years, with the continuous development of the society and economy, people's living standard has been significantly improved. With the increasing convenience of food, clothing, housing and transportation, people's life has become increasingly busy. The incidence of "modern civilization diseases" such as hypertension and diabetes is increasing, and people's health problems are becoming more and more serious. Oxygen saturation plays an important role in the measurement of our health. Low blood oxygen levels in arrhythmias often lead to dizziness, lack of concentration and anxiety. In severe cases, it can be fatal to the heart and can even lead to heart failure, especially in vulnerable people such as the elderly. Therefore, monitoring pulse oxygen saturation not only plays an important role in our health, but also can greatly reduce the incidence of cardiovascular and cerebrovascular diseases.

(1) Using the information given to participants, including age, BMI, sex, smoking history, and other data, typical patterns of changes in blood oxygen saturation were found, and several parameters were used to characterize a person.

(2) Look for age-related patterns of oxygen saturation. Compare what changes occur in the physical characteristics of the young and the old. According to these changes, what medical significance should they have.

1.2 Problem analysis

1.2.1 Generality

This is a multiple regression analysis problem using data from 36 individuals. Before the establishment of the model, we should first analyze the four indicators given, then give corresponding quantitative methods according to the characteristics of each indicator, and finally get the quantitative results and conduct analysis. Then, the model is built, and the quantitative results are substituted into the multiple linear regression model built by us, and then the problem is analyzed according to the results obtained.

1.2.2 Faced problem

We quantified the given data. To find an appropriate quantification method, we should first analyze the characteristics of each indicator. MATLAB and other software were used to run the code given by the title, calculate the mean value and standard deviation of each person's blood oxygen saturation concentration, and then numerically quantify the gender and age. Establish a multiple linear regression model, put the quantized data into the model, and calculate the corresponding regression coefficient, correlation coefficient, P value, error, etc. The regression results were analyzed. So we get the regression equation, and we use the regression equation to characterize a person.

Second, we first need to look for the definitions of the elderly and the young. We need to separate the elderly and the young according to age to see the changes of blood oxygen saturation concentration under the age factor. At the same time, we also analyze the characteristics of different factors, we find the materials, to find more samples for analysis, we from the perspectives of sex, smoking history, carries on the analysis, using SPSS statistical analysis software, such as to what changes have taken place in physical characteristics, and based on speculation ideally these characteristics what medicine should have meaning.

2. Model assumptions

1) It is assumed that in the process of data calculation, if the error is within a reasonable range, the data with great influence on the results can be ignored and not discussed;

2) When the statistical results are much different from the actual results, the influence of this factor can be ignored;

3) It is assumed that the test process frequency always remains at 1HZ without any change;

4) Suppose that the various instincts truly and objectively reflect each person's physical condition and oxygen saturation;

5) It is assumed that the quantitative results are relatively objective and can reflect the oxygen saturation of blood;

6) It is assumed that the sum of squares error of the model after training reaches the minimum.

3. Implement

3.1 Quantitative treatment of indicators

As shown in Table 1, quantitative treatment table of indicators:

Table 1 Quantitative treatment table of indicators

The serial number	Name of index	Quantitative treatment	Calculated on the basis
1	Gender	Digital quantification	0 for Female; 1 for Male
2	Smoking history	Digital quantification	a for Non-Smoker; b for Smoker; c for Ex-Smoker
3	Mean oxygen saturation	take the average	Using MATLAB to take the average value of oxygen saturation per Hertz of each person in one hour
4	Standard deviation of oxygen saturation	take the average	Use MATLAB to take the standard deviation of oxygen saturation per Hertz of each person in one hour

3.2 Data Analysis

Through quantitative methods such as digital quantization and average calculation, we quantified four kinds of information of participants.

First, by integrating the data given in the title, we obtained relevant data for the participants, including age, BMI, sex, smoking history (current smoking status), oxygen saturation, mean oxygen saturation, and standard deviation of oxygen saturation. We calculated the mean o2 saturation and the standard deviation of O2 saturation by MATLAB, using the oxygen saturation per Hertz given to each person in an hour. We will take the average oxygen saturation of each person as the dependent variable and age, BMI, gender and smoking history as independent variables, and use MATLAB for multiple linear regression calculation to obtain the results, so as to obtain relevant parameters to characterize a person, as shown in Table 2.

Table 2 Relevant data Table

Record Name	Gender	Smoking Status	BMI	Age	Mean oxygen saturation	The standard deviation
070217A	F	Non-Smoker	19.8	19	97.9552	0.5413
301116B	M	Non-Smoker	21.7	20	98.2181	0.5001
051216A	M	Non-Smoker	22	20	96.4688	0.8642
081216A	M	Non-Smoker	23.6	20	97.8733	0.8281
121216A	M	Non-Smoker	20.6	20	98.6062	0.5272
250117B	M	Non-Smoker	23	20	98.6455	0.5353
300117A	F	Non-Smoker	20.3	20	98.8383	0.4699
080217B	F	Non-Smoker	20.7	20	98.7039	0.3727
090217B	F	Non-Smoker	24.4	20	97.3643	1.0892
210217A	M	Smoker	24.6	21	96.5237	0.7423
121216B	F	Non-Smoker	23.1	21	99.0019	0.5084
010217B	F	Non-Smoker	21.7	21	98.1634	0.734
210217C	F	Non-Smoker	18.9	21	98.3307	0.6312
230117B	F	Ex-Smoker	17.9	21	99.3245	0.6314
101216C	M	Non-Smoker	23.2	22	96.8059	1.2923
080217A	F	Non-Smoker	24.7	22	98.4351	0.5565
230117A	M	Smoker	26.8	23	98.1868	0.7253
210217B	M	Non-Smoker	26.7	23	97.7705	0.5404
010217A	F	Non-Smoker	20.5	23	98.2395	0.7262
160217C	F	non Smoker	24.2	24	96.9763	0.7019
250117A	F	Non-Smoker	21.3	35	99.5081	0.3422
160217E	F	Non-Smoker	19.8	38	96.8704	0.7335
090217A	F	Non-Smoker	28.2	41	98.802	0.4395
301116A	M	Non-Smoker	26.5	42	97.5857	0.9263
140317A	M	Non-Smoker	26.2	42	98.1847	0.5431
160217D	M	Non-Smoker	23.9	45	96.2751	1.0183
010217C	F	Smoker	23.8	45	94.4123	1.2859
101216A	F	Non-Smoker	24.1	45	98.0098	0.669
150317B	M	Non-Smoker	19.9	48	98.5044	0.4052
101216B	M	Non-Smoker	26.8	49	98.185	0.7928
250117C	M	Ex-Smoker	28.4	55	97.1268	0.7311
160217B	F	Ex-Smoker	18.5	56	96.5249	0.6731
160217A	M	Ex-Smoker	24.5	60	98.2009	0.7689
150317A	F	Non-Smoker	24.3	62	98.7045	0.4357
010317A	F	Ex-Smoker	23.2	66	96.3442	1.2223
010317B	M	Non-Smoker	25.1	70	93.7483	0.9462

3.3 Multiple linear regression

The multiple linear regression analysis model is as follows:

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \quad (1)$$

Type of β_0 , β_1 ,⋯, β_m, σ^2 are x_1, x_2,⋯, x_m ,independent of the unknown parameters, including β_0 , β_1 ,⋯, β_m called regression coefficient.

Now n independent observation data are obtained $(y_i, x_{i1}, \cdots, x_{im})$, $i = 1, \cdots, n, n > m$, According to Equation:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n \end{cases} \quad (2)$$

Denote as:

$$X = \begin{bmatrix} 1 & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nm} \end{bmatrix}, \qquad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad (3)$$

$$\varepsilon = [\varepsilon_1 \ \cdots \ \varepsilon_n]^T, \ \beta = [\beta_0 \ \beta_1 \ \cdots \ \beta_m]^T$$

Equation is expressed as:

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 E_n) \end{cases} \quad (4)$$

Where E_n is an n-order matrix.

The parameters in the beta type 5-1 β_0 , β_1,⋯, β_m still use the least squares estimation, namely should select estimate $\widehat{\beta_J}$, make when $\beta_J = \widehat{\beta_J}, j = 0,1,2, \cdots, m$, the error sum of squares:

$$Q = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2 \quad (5)$$

To the minimum. To this end, let:

$$\frac{\partial Q}{\partial \beta_j} = 0, \ \ j = 0,1,2, \cdots, n$$

Get:

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im}) = 0 \\ \frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im}) = 0 \ \ j = 1,2, \cdots, m \end{cases} \quad (6)$$

This system is reduced to the following normal equations:

$$\begin{cases} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i2} + \cdots + \beta_m \sum_{i=1}^{n} x_{im} = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} {x_{i1}}^2 + \beta_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \beta_m \sum_{i=1}^{n} x_{i1} x_{im} = \sum_{i=1}^{n} x_{i1} y_i \\ \beta_0 \sum_{i=1}^{n} x_{im} + \beta_1 \sum_{i=1}^{n} x_{im} x_{i1} + \beta_2 \sum_{i=1}^{n} x_{im} x_{i2} + \cdots + \beta_m \sum_{i=1}^{n} {x_{im}}^2 = \sum_{i=1}^{n} x_{im} y_i \end{cases} \quad (7)$$

The matrix form of the normal system is:

$$X^T X \beta = X^T Y \quad (8)$$

When the matrix X has full rank, $X^T X$ is an invertible square matrix, and the solution of equation is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (9)$$

Substitute $\hat{\beta}$ back into the original model to get an estimate of y:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_m} x_m \quad (10)$$

And the fitting value of this set of data is$\hat{Y} = X\hat{\beta}$,error of fitting$e = Y - \hat{Y}$ Referred to as the residual error, and can be used for random error ε estimates, and Q can be represented as:

$$Q = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \quad (11)$$

Is the residual sum of squares (or the residual sum of squares),namely$Q(\hat{\beta})$.

The following analysis results are given without proof:

1) $\hat{\beta}$ is the linear unbiased minimum variance estimate of β. Refers to $\hat{\beta}$ is a linear function of Y;The expectation of $\hat{\beta}$ is equal to β;In the linear unbiased estimation of β, the variance of $\hat{\beta}$ is the smallest.

(2) $\hat{\beta}$ obeys normal distribution

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}) \quad (12)$$

Referred to as$(X^T X)^{-1} = (c_{ij})_{n \times n}$.

(3) Sum of squared residuals$Q$,$EQ = (n - m - 1)\sigma^2$,And:

$$\frac{Q}{\sigma^2} \sim X^2(n - m - 1) \quad (13)$$

This gives an unbiased estimate of σ^2:

$$s^2 = \frac{Q}{n-m-1} = \hat{\sigma}^2 \quad (14)$$

s^2 is the residual variance (the variance of the residual), s is called the residual standard deviation.

(4) Decompose the total square sum $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$, there are:

$$SST = Q + U,\ U = \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2 \quad (15)$$

Where, Q is the residual sum of squares defined by Formula 5-5, reflecting the impact of random error on Y, and U is called the sum of regressive squares, reflecting the impact of independent variable on Y. The above decomposition uses a system of normal equations.

Hypothesis testing of regression models:

Whether there is a linear relationship between the dependent variable y and the independent variable $x_1, \cdots, x_m$, as shown in Equation, needs to be tested. Obviously, if all $|\hat{\beta_J}|$ $(j = 1, \cdots, m)$ are small, the linear relationship between y and $x_1, \cdots, x_m$ is not obvious. Therefore, the null hypothesis can be set as:

$$H_0: \beta_J = 0 (j = 1, \cdots, m)$$

When H_0 is true, U and Q defined by the decomposition formula (34) satisfy:

$$F = \frac{U/m}{Q/(n-m-1)} \sim F(m, n - m - 1) \quad (16)$$

Under at significance level α. The α quantile$F_\alpha(m, n - m - 1)$, if $F < F_\alpha(m, n - m - 1)$, accept H_0; Otherwise, refuse.

Note that the acceptance of H_0 only indicates that the linear relationship between y and $x_1, \cdots, x_m$is not obvious, and there may be a nonlinear relationship, such as the square relationship.

There are also some indicators to measure the degree of correlation between y and $x_1, \cdots, x_m$, such as defining the complex determination coefficient by the ratio of the regressive sum of squares in the total sum of squares:

$$R^2 = \frac{U}{S} \quad (17)$$

$R = \sqrt{R^2}$ is called the complex correlation coefficient, The larger R is, the closer the correlation between y and $x_1,\cdots,x_m$is. Generally, R greater than 0.8(or 0.9) is considered to be valid.

Hypothesis testing and interval estimation of regression coefficients

When the above H_0 is rejected, β_i is not all zero, but some of them are equal to zero. Therefore, the following m tests should be further performed$(j = 0,1,\cdots,m)$:

$$H_0{}^{(j)}: \beta_j = 0$$

By 5-12, $\widehat{\beta_j} \sim N(\beta_j, \sigma^2 c_{jj})$, $c_{jj}$is$(X^T X)^{-1}$ the first elemen(j,j), with s^2 instead of σ^2, by type 5-12, 5-13 and 5-14, when $H_0{}^{(j)}$ was founded:

$$t_j = \frac{\widehat{\beta_j}/\sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}} \sim t(n-m-1) \tag{18}$$

For a given α, if$|t_j| < t_{\frac{\alpha}{2}}(n-m-1)$,accept $H_0{}^{(j)}$;Otherwise, refuse.

Type 5-18 also can be used to$\beta_j$for interval estimation$(j = 1,\cdots,m)$, under the confidence level $1-\alpha$, β_j confidence interval is:

$$\left[\widehat{\beta_j} - t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}},\ \ \widehat{\beta_j} + t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}}\right] \tag{19}$$

Among $s = \sqrt{\frac{Q}{n-m-1}}$.

The regression model is used for prediction

When the regression model and coefficient pass the test, may be given by$x_0 = (x_{01},\cdots,x_{0m})$ to predict y_0, y_0 is random. Phase predicted value (point estimation) is:

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1}x_{01} + \cdots + \widehat{\beta_m}x_{0m} \tag{20}$$

Given α, the prediction interval fory_0 (interval estimation) can be calculated. The results are complex, but when n is large and x_{0i} is close to the average $\overline{x_i}$ the prediction interval for $y_0$can be simplified as follows:

$$\left[\widehat{y_0} - z_{\frac{\alpha}{2}}s, \widehat{y_0} + z_{\frac{\alpha}{2}}s\right] \tag{21}$$

Where $z_{\frac{\alpha}{2}}$ is $\frac{\alpha}{2}$ quantile of the standard normal distribution.

For y_0 interval estimation method can be used to give the known data of residual $e_i = y_i - \hat{y}_i$ $(i = 1,\cdots,n)$ of the confidence interval, e_i obey the normal distribution with zero mean, so if a e_i confidence interval does not include the zero point, argues that this data is abnormal, can eliminate.

3.4 Linear regression results

We incorporated four data -- gender, smoking history, BMI and age -- into the regression model.Firstly, the outliers are analyzed to obtain outliers.By analyzing the outliers, we can assume that these data are correct.Then we conducted multiple regression analysis, and we obtained the regression coefficient β, the interval estimation of the regression coefficient, the residual and the confidence interval values, which were listed in the Table 3, Table 4 and Table 5.

Table 3 Values obtained by regression model

Record Name	Outlier analysis	Residual	Confidence interval	
070217A	2.2694	-0.4042	-2.7545	1.9462
301116B	2.7399	0.2652	-2.0676	2.5981
051216A	4.2491	-1.4828	-3.7587	0.7931
081216A	2.0122	-0.0715	-2.4313	2.2884
121216A	3.9976	0.6486	-1.6368	2.9341
250117B	2.4724	0.6982	-1.6446	3.0410
300117A	2.0225	0.5107	-1.8487	2.8702
080217B	1.7779	0.378	-1.9904	2.7465
090217B	3.6841	-0.9458	-3.2432	1.3516
210217A	5.5218	-1.322	-3.5487	0.9047
121216B	2.2534	0.7159	-1.6350	3.0668
010217B	1.4655	-0.1286	-2.5084	2.2513
210217C	2.7601	0.0268	-2.3053	2.3589
230117B	11.5082	1.1466	-0.8326	3.1257
101216C	2.7915	-1.0813	-3.4122	1.2497
080217A	3.0382	0.1856	-2.1361	2.5073
230117A	5.7261	0.4098	-1.8090	2.6285
210217B	3.1916	-0.0721	-2.3880	2.2439
010217A	1.5818	0.0017	-2.3739	2.3773
160217C	3.4825	-1.2161	-3.5211	1.0889
250117A	3.6161	1.6295	-0.6705	3.9294
160217E	3.5242	-0.9257	-3.2291	1.3777
090217A	8.0986	1.1307	-0.9930	3.2544
301116A	2.3193	0.3055	-2.0430	2.6540
140317A	2.7402	0.9032	-1.4296	3.2361
160217D	3.0192	-0.9272	-3.2496	1.3952
010217C	9.8215	-3.0941	-5.1460	-1.0421
101216A	2.6607	0.4396	-1.8962	2.7754
150317B	9.0333	1.3739	-0.7111	3.4590
101216B	4.0562	1.1136	-1.1696	3.3969
250117C	9.0969	0.3704	-1.7120	2.4528
160217B	9.5066	-0.6128	-2.6781	1.4524
160217A	8.7353	1.5761	-0.5214	3.6735
150317A	8.274	1.6391	-0.4774	3.7556
010317A	7.7668	-0.477	-2.6143	1.6602
010317B	14.1852	-2.7078	-4.5656	-0.8500

Table 4 Obtained the four values of RANT

Decision coefficient r^2	F statistic	P values	Variance error
0.1950	1.8776	0.1394	1.4125

Table 5 Regression coefficient values

symbol	The numerical	Interval estimation of regression coefficients
ε	98.3347	93.6818 102.9877
$\widehat{\beta_0}$	0.3687	-0.5211 1.2585
$\widehat{\beta_1}$	-0.0651	-0.6915 0.5612
$\widehat{\beta_2}$	-0.0043	-0.1770 0.1685
$\widehat{\beta_3}$	-0.0296	-0.0592 -0.0001

Through the regression model, we also obtained the residual leverage diagram, as shown in Figure 1:

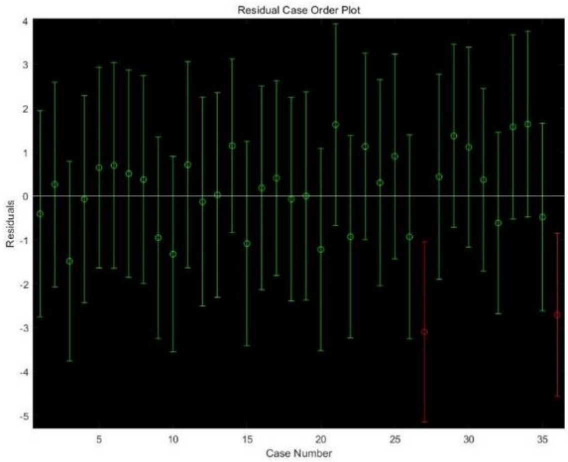

Figure 1 Residual lever diagram

Thus, we get the regression equation:

$$Y = 98.3347 + 0.3687x_1 - 0.0651x_2 - 0.0043x_3 - 0.0296x_4$$

In addition, the residual lever diagram shows that the residual is uniformly distributed near the 0 point line. Among the four values returned by STAT, the decision coefficient r^2= 0.1950, indicating that the model is well fitted.F statistic = 1.8776>0.000, which meets the requirements.However, the P value associated with the significance probability was 0.1394>0.05, indicating that some variables in the regression equation could be eliminated.

According to the regression model, the results can be obtained according to the regression equation we obtained:

$$Y = 98.3347 + 0.3687x_1 - 0.0651x_2 - 0.0043x_3 - 0.0296x_4$$

Represents a typical pattern of changes in oxygen saturation, thereby characterizing a person.

4 Question two

We need to study whether oxygen saturation is related to age. The question suggests that we should study what characteristics have changed in the elderly compared to the young.First, we consulted materials to determine the definitions of the young and the old. According to the National Bureau of Statistics of China and The Law on the Protection of the Rights and Interests of the Elderly of China, the age standard for the young is 15-34 years old, and the starting age standard for the old is 60 years old.Therefore, we will study the relationship between the oxygen saturation of people of different sex, smoking history and BMI in these two age groups, so as to compare which physical characteristics have changed in older people compared with younger people.

4.1 Comparison analysis diagram

We will smoking history and gender as the independent variable, the average blood oxygen saturation as the dependent variable, separate F - Non - Smoker, F - Smoker, F - the EX - Smoker, M - Non - Smoker, M - Smoker, M - the EX - Smoker, calculate the average oxygen saturation in the six categories, MATLAB drawing histogram, as shown in figure 2:

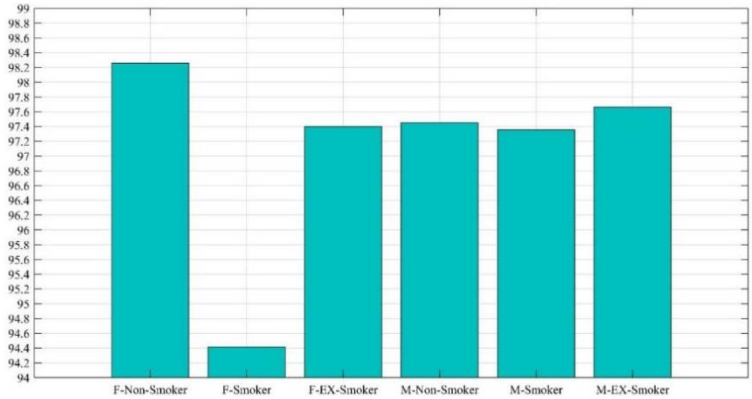

Figure 2 Comparison of smoking and gender

According to Figure 2, we found that female smokers' oxygen saturation was lower than that of the other five groups, while female non-smokers' oxygen saturation was higher than that of the other five groups. For men, smoking, non-smoking or previous smoking did not have a significant effect on oxygen saturation.This is not the final word, and we will continue to analyze it.

We also took age as an independent variable and mean oxygen saturation as a dependent variable to find out whether different ages would have an impact on oxygen saturation, as shown in Figure 3.

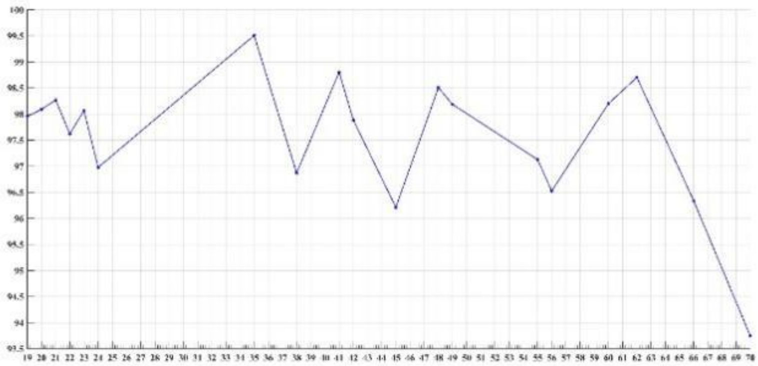

Figure 3 A broken line diagram of age and oxygen saturation of 36 individuals

According to Figure 3, we found that for people of different ages, blood oxygen saturation fluctuated, and only participants aged 70 years had much lower blood oxygen saturation than participants of other ages.

Us through the above two figure shows linear relationship, we found the sample data analysis for 36 individuals too one-sided, the result of the sample, we need a more data we through access to information, find a 1200 people accept blood oxygen saturation test data, including male 520, female 680 people, aged between 17-89, in line with the requirements of ontology.We used SPSS software for statistical analysis of blood oxygen saturation and other data. χ^2 test was used for count data, and P<0.05 indicated that the difference was statistically significant.

First, we studied the influence of age on blood oxygen saturation, as shown in Table 6.

Table 6 Influence of age on blood oxygen saturation

Age	Number of people	Oxygen saturation<90%	Oxygen saturation>90%
Age < 60 years	1000	96	904
Age ≥ 60 years	200	44	156
Total	1200	140	1060

Through the above data and SPSS statistical analysis, we know, the age < 60 years old in patients with low blood oxygen saturation 90% accounted for 9.6% (96/1000), oxygen saturation in patients aged 60 or higher accounted for 22.0% (44/200) was less than 90%, the age of 60 or more patients significantly lower blood oxygen saturation of < 60 patients, difference was statistically significant (χ^2= 24.86, P = 0.00).

We found that the oxygen saturation of older people was lower than that of younger people after taking age into account only, but not age alone. We needed to further analyze gender and smoking history.

We studied the effect of sex on oxygen saturation, as shown in Table 7

Table 7 Influence of blood oxygen saturation

Gender	Number of people	Oxygen saturation<90%	Oxygen saturation>90%
Male	520	28	492
Female	680	75	605
Total	1200	103	1097

Through statistical analysis of the above data and SPSS, we know that the blood oxygen saturation of 11.03% (75/680) female patients is lower than 90%, and 5.38% (28/520) male patients is lower than 90%, the difference is statistically significant (χ^2=11.97, P=0.00).

We found that women had higher oxygen saturation below 90 percent than men.

We also need to analyze the history of smoking. We refer to the "2015 Xiamen Adult Tobacco Survey Bulletin" commissioned by Xiamen Center for Disease Control and Prevention, and we get the bar chart of the proportion of smokers in different age groups in Xiamen, as shown in Figure 4:

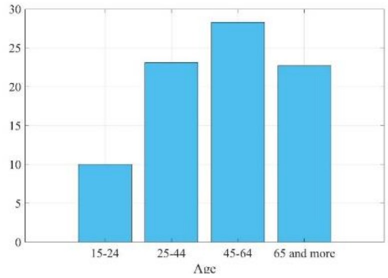

Figure 4 Bar Chart of the proportion of smokers in different age groups in Xiamen

From this figure, we find that the proportion of middle-aged and elderly smokers is larger than that of young people.

4.2 Results

Through the above analysis, we found that the blood oxygen saturation of the elderly is lower than that of the young if only taking age into account. Meanwhile, we also found that the blood oxygen concentration of the female is lower than that of the male, and that of the smokers is lower than that of the non-smokers.

Blood oxygen saturation is one of the important indicators of body health. The blood oxygen saturation of normal healthy people should be kept between 95% and 100%. If it is lower than 90%, it has entered the range of hypoxia, and if it is lower than 80%, it will be severe hypoxia, which will cause great damage to the body and endanger life. If the body's blood oxygen is too low, the body experiences the following characteristic changes:

(1) Shortness of breath: One of the common symptoms of low oxygen saturation is shortness of breath. As soon as shortness of breath occurs, it means the patient must be treated immediately.

(2) Lack of energy: The body needs a lot of energy for survival activities. Oxygen is one of the sources of energy for tissues and cells, so low oxygen saturation will affect the energy source of cells. Lack of energy can lead to fatigue, slower thinking, reduced concentration, reduced resistance and other symptoms.

(3) Brain changes: The brain is the most sensitive organ to hypoxia, once brain cells lack oxygen, it will seriously affect health, and even cause irreversible brain damage endangering life. Persistent low oxygen saturation can lead to memory loss, personality changes and other mental problems.

(4) High altitude pulmonary edema: Low oxygen saturation caused by environmental changes. At high altitudes, the air is thin and oxygen content is much lower than in plain areas. Headaches, swelling and shortness of breath can occur when the body is deprived of oxygen during exercise.

4.3 Medical significance of blood oxygen saturation

For human body, if blood oxygen saturation is too low, it will cause harm. Therefore, the following medical significance can be obtained from our research data:

(1) Improve measurement methods

Many clinical diseases will cause the lack of oxygen supply, which will directly affect the normal metabolism of cells, and even threaten people's lives. Therefore, real-time monitoring of arterial oxygen concentration is very important in clinical rescue.

The traditional measurement method of blood oxygen saturation is to collect human blood first, and then conduct electrochemical analysis with blood gas analyzer to measure blood oxygen partial pressure PO2 and calculate blood oxygen saturation. This method is cumbersome and does not allow continuous monitoring.

Refers to the set of photoelectric sensor, measurement, simply by placing sensors on one finger, using fingers as hemoglobin transparent containers, use red light wavelength of 660 nm and 940 nm of near-infrared light as light source, determine the intensity of light transmission through tissue bed, to calculate the hemoglobin concentration and oxygen saturation, the instrument can display the human body blood oxygen saturation, for clinical provides a continuous non-invasive blood oxygen measuring instrument.

(2) Disseminate reference values to the public

Normal human blood oxygen saturation should not be lower than 94%, below 94% is insufficient oxygen supply. Some scholars defined oxygen saturation <90% as the standard for hypoxemia, and believed that when oxygen saturation was higher than 70%, the accuracy could reach ±2%, while when oxygen saturation was lower than 70%, there could be errors. Clinically, we have compared spo2 values with SPO2 values in a number of patients and concluded that SPO2 readings can reflect the patient's respiratory function and reflect spo2 changes to a certain extent. In addition to blood gas analysis for patients with clinical symptoms and numerical values inconsistent in individual cases after thoracic surgery, routine application of pulse blood oxygen saturation monitoring can provide meaningful indicators for clinical observation of changes in the condition, avoid repeated blood collection of patients, and reduce the workload of nurses. Therefore, it is worthy of promotion. Clinical generally more than 90% is ok, of course, to different departments.

(3) Promote the dangers of hypoxia

Hypoxia is the imbalance between oxygen supply and oxygen consumption, that is, the tissue cell metabolism is in hypoxia state. Whether the body is hypoxic or not depends on whether the oxygen transport volume and oxygen reserve accepted by each tissue can meet the requirements of aerobic metabolism. The harm of anoxia is related to the degree, speed and duration of anoxia. Severe hypoxemia is a common cause of death under anesthesia, accounting for about one-third to two-thirds of deaths due to cardiac arrest or severe brain cell damage. Clinically, general PaO_2<80mmHg is hypoxia, which is basically equivalent to severe hypoxemia.

Hypoxia has a huge impact on the body. For example, the effect on CNS, liver and kidney function. The first symptoms of hypoxia are compensatory heart rate acceleration, increased cardiac beat and cardiac output, and the circulatory system compensates for oxygen deficiency in a highly dynamic state. Redistribution of blood flow and selective dilation of brain and coronary vessels are also produced to ensure adequate blood supply. However, in severe hypoxic conditions, ATP synthesis is reduced due to subendocardial lactic acid accumulation, resulting in myocardial inhibition, leading to bradycardia, pre-phase contraction, decreased blood pressure and cardiac output, as well as ventricular fibrillation and other arrhythmias and even cardiac arrest. In addition, hypoxia and the patient's own disease may have important effects on homeostasis.

5. Evaluation and promotion

5.1 Model Evaluation

5.1.1 Advantages of the model

(1)Multiple linear regression model can solve any complex linear relationship well;

(2)In regression analysis, if there are two or more independent variables, it is called multiple regression. In fact, a phenomenon is often associated with multiple factors, and it is more effective and practical to predict or estimate dependent variables jointly by the optimal combination of multiple independent variables than to predict or estimate only one independent variable. Therefore, multiple linear regression has more practical significance than unitary linear regression.

(3)The multivariate linear regression analysis is the most basic and simple one in the multivariate regression analysis.

(4)Using regression models, a unique result can be calculated using standard statistical methods as long as the model and data are the same.

5.1.2 Shortcomings of the model

(1) Sometimes, in the regression analysis, it is only a conjecture which factor is selected and which expression is adopted for this factor, which affects the diversity of electrical factors and the unpredictability of some factors, making the regression analysis limited in some cases.

(2) The basic principle and basic calculation process of multiple linear regression are the same as that of unary linear regression. However, due to the large number of independent variables, the calculation is quite troublesome. Generally, statistical software is needed in practical application. Only the basic problems of multiple linear regression are introduced here.

5.2 Model Generalization

Through the interpretation of this title, we can easily find that this is a multiple regression analysis problem. We have established a multiple linear regression model. Through careful observation of the model established by us, it is not difficult to find that this model is not only applicable to the eigenvalue problem of blood oxygen saturation concentration, but also can play a guiding role in solving analytical problems.

This model is built to solve the problem of comprehensive evaluation under a certain index. Our model covers a wide range of aspects, such as actual economic problems, stock market analysis, household consumption expenditure and so on... These kinds of problems can be well solved. Multiple linear regression model can be widely used in education, economy, administration and other fields.

References

[1] Li JC, Sun YUE, Nyima QUTSO, Sun ZONGKE, Zhang W, Wang ZE, Yao XIAOYUAN, Zhang B, Xu N. Study on the relationship between blood oxygen saturation and heart rate of emergency personnel and the occurrence of plateau reaction symptoms[J]. Modern preventive medicine, 2020,47(21):3871-3875.

[2] Zhang M, Shen G.L., Qian Feiyang. Analysis of risk factors and prevention strategies for decreased oxygen saturation during in-hospital transfer of critically ill children[J/OL]. Chinese general medicine, 2020(11):1864-1868 [2020-12-05].

[3] Mao JX, Xiao D, Luo Y, Zhang Y, Wang XUL. Exploration of near-infrared spectroscopy in determining cerebral oxygen saturation and determining the degree of intestinal necrosis in neonatal necrotizing small intestinal colitis[J]. Lingnan Modern Clinical Surgery, 2020, 20(05):562-566+572.

[4] Huang C, Wang X. Effect of endoscopic pneumatic mask oxygenation on oxygen saturation in patients undergoing painless gastroscopy[J]. Chinese clinical research, 2020, 33(10):1442-1445.

[5] Wei ZQ, Pang XN. Research on the calculation of emergency oxygen supplementation in passenger cabin based on blood oxygen saturation[J]. Aerospace Medicine and Medical Engineering, 2020, 33(05):422-426.

[6] Zeng Wen. Effect of central negative pressure suction closed suction on oxygen saturation and psychological state of patients with emergency mechanical ventilation[J]. Medical Equipment, 2020, 33(19):155-156.