

Analysis of influencing factors of urban residents' physical health based on logistic regression method

Chengxian Zhou*, Zhiqi Zhang, Kaiwen Xian

Department of International Education, Changchun University Of Technology, Changchun, China,
130012

* Corresponding Author Email: nataliagaveskaya@gmail.com

Abstract. This article aims to conduct an in-depth study of residents' eating habits, lifestyles and their close connection with physical health to provide a strong scientific basis to deal with the increasing problem of chronic diseases in society. In order to solve this problem, this article uses a variety of sophisticated data analysis methods and models to explore the following core issues. Firstly, the credibility and validity of the research data were ensured by conducting reliability and validity tests on the questionnaire data; secondly, in the data processing, manual judgment and statistical principle judgment were used to identify and process outliers, thereby The quality of the data was improved; then, the logistic regression model was used to analyze the relationship and degree of correlation between the hypertension prevalence data and various indicators, and to explore in depth whether these factors were significantly related to the occurrence of chronic diseases. Finally, by comparing the research results, this article summarizes the connection between diet structure, lifestyle and other related factors and common chronic diseases, providing strong support for improving the health status of residents, which is an important contribution of this study.

Keywords: Reliability and validity test, 3σ principle to eliminate outliers, Logistic regression.

1. Introduction

Chronic non-communicable diseases (hereinafter referred to as chronic diseases) represented by cardiovascular and cerebrovascular diseases, hypertension, malignant tumors and chronic obstructive pulmonary disease have become important problems affecting the health of Chinese residents. The prevalence of chronic diseases continues to rise as people change their lifestyles. As is well known, health status is closely related to age, eating habits, physical activity, occupation, etc^[1]. How to achieve the purpose of promoting physical health by reasonable arrangement of diet, moderate physical exercise and practicing a healthy lifestyle is a common concern of the whole society.

On the basis of the collected questionnaire data, the reliability and validity tests of the questionnaire should first be performed. Through this critical step, the credibility and validity of the data can be ensured, thus providing a solid and reliable basis for subsequent analysis. This initial processing phase not only helps to eliminate potential measurement bias and data error, but also lays a solid analytical foundation for the present study.

This paper first preprocesses the data to ensure the quality and reliability of the data. For index data following normal distribution, the method based on 3σ principle is used to identify and eliminate outliers. Subsequently, the standards of various dietary indicators were properly quantified. By comparing these criteria with the data, the compliance and rationality of the data can be determined. Finally, this paper builds a logistic regression model for the disease data of hypertension, aiming to deeply explore the relationship and correlation degree between chronic diseases and various indicators. Thus, it is possible to quantitatively analyze the effects of different factors on the occurrence of hypertension and further understand the potential mechanisms of influence^[2].

2. Model establishment and solution

2.1. Validity and reliability test of the questionnaire

Test of the validity of the questionnaire^[3]Used to ensure that the questionnaire measurement tool can accurately measure the concepts or characteristics to be measured. The reliability test is used to assess the stability and consistency of the questionnaire measurement tool, to ensure that similar results are obtained in different situations.

The KMO test (Kaiser-Meyer-Olkin) and the Bartlett spherical test were used to verify sufficient correlation between the variables.

$$KMO = \frac{\sum \sum_{i \neq j} r_{ij}}{\sum \sum_{i \neq j} r_{ij}^2 + \sum \sum_{i \neq j} a_{ij}^2} \quad (1)$$

Where r_{ij} represents the simple correlation coefficient and $\alpha_{ij}. 1, 2, 3, \dots, k$ represents the partial correlation coefficient.

KMO, test and Bartlett sphere test of the questionnaire using SPSS, and the results are shown in Table 1:

Table 1. KMO and Bartlett spheroid tests

Number of KMO sampling suitability quantities		0.663
Bartlett Spherical test	Approximate chi square	28494.733
	free degree	703
	conspicuousness	0

According to the results in Table 1, this paper shows that the number of KMO sampling suitability = 0.663 > 0.6, and the Bartlett sphere test is <0.05, indicating that the validity of the questionnaire is good.

The clone Bach coefficient method (Cronbach's alpha) was used to measure the correlation between the questionnaire items to determine the reliability of the questionnaire.

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum S_i^2}{S_x^2} \right) \quad (2)$$

Where K is the number of questionnaire questions, S_i^2 represents the score variation of all subjects on question i, and S_x^2 is the variance of the total score of all subjects.

SPSS tested the data in the questionnaire. The results are shown in Table 2:

Table 2. Reliability statistics

Clone of Bach, Alpha	Clone Bach Alpha based on the normalization term
0.741	0.741

According to the results in Table 2, we can see that the clone Bach Alpha = 0.741, between 0.7-0.8, which indicates that the questionnaire has good internal consistency and acceptable reliability. Considering the results of the above two methods, this paper believes that the data of this questionnaire are valid.

2.2. Data preprocessing

There are many missing values in the questionnaire, and not processing of data containing missing values may bias the later analysis. Select the appropriate missing value processing method^[4]Usually consider the research question, the nature of the data and the analytical purpose.

For completely random missing data, such as height, weight, etc., the occurrence of missing values was not associated with any other variable, including the missing value itself. In this case, the distribution of missing values is consistent with the non-missing part of the data, which is filled by means or means. For some data that fit to a normal distribution, some edge values can be removed using the 3a principle. Normal distribution is a very common continuous probability distribution. If

the random number follows a normal distribution with a mean of μ and a standard deviation of σ , it is written down as

$$X \sim N(\mu, \sigma^2) \quad (3)$$

Then its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (4)$$

The probability of the value appearing in the $(\mu - \sigma, \mu + \sigma)$ range is 68.268949%

The probability of the value appearing within $(\mu - 2\sigma, \mu + 2\sigma)$ is 95.449974%

The probability of the value appearing in the range $(\mu - 3\sigma, \mu + 3\sigma)$ is 99.730020%

Under normal distribution, the vast majority of data points will be concentrated near the mean, while relatively few data points far from the mean, and the possibility of the value appearing outside the range $(\mu - 3\sigma, \mu + 3\sigma)$ is less than 0.3%. Therefore, the value falling outside the interval $(\mu - 3\sigma, \mu + 3\sigma)$ can be regarded as outliers and removed^[5].

An important premise for applying the 3σ principle is that the data of the variables should conform to the normal distribution, so this paper first uses SPSS to draw P-P maps and Q-Q maps for the 110 variables that may conform to the normal distribution. Because there are too many variables, height is taken as an example to detail the selection of outliers in height.

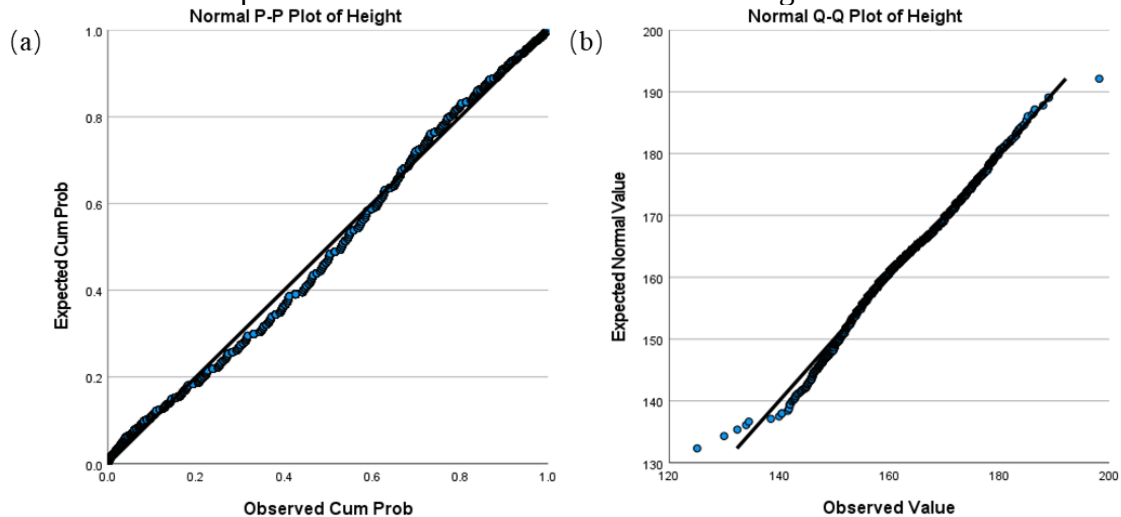


Fig. 1 (a)Normal P-P plot of height(b)Normal Q-Q plots for height

In the P-P and Q-Q plots, the diagonal represents the case where the actual data agrees perfectly with the theoretical distribution. Observe whether the data points are distributed along the diagonal line. If the data points are roughly along the diagonal, it indicates that the cumulative distribution of the data is similar to the theoretical distribution. As can be seen from Fig 1, the P-P map of height is highly fitted, and it can be considered that the data of height variables basically fit the normal distribution. In order to make the data processing more rigorous, this paper uses SPSS and one-sample Kolmogorov-Smirnov test to further judge the 11 data sets with good fit. The results are shown in Table 3:

Table 3. Describes the results

variable name	average value	standard deviation	skewness	kurtosis	K-S checkout
stature	162.264	8.017	0.189	-0.222	4.8975E-42
weight	60.767	11.895	1.246	5.132	2.862E-89
waistline	79.534	10.331	0.247	0.5	6.4442E-20
hip	93.816	7.798	-0.211	6.833	3.5311E-40
systolic pressure	112.906	15.39	0.981	2.066	4.6309E-207
diastolic pressure	74.123	10.197	0.62	1.001	7.543E-209
pulse	73.363	7.929	0.531	0.947	9.579E-157
cholesterol	4.806	0.975	0.673	1.513	1.5465E-28

As can be seen from the results of Table 3, the P-value of KS significance of 11 variables is < 0.05 . According to the data, it is difficult to have a strictly normal distribution when the data amount is more than 5000, but when the absolute value of the kurtosis is less than 10 and the absolute value of the skewness is less than 3, this paper can be considered that this variable is basically consistent with the normal distribution.

For some data that do not conform to the normal distribution, the data can be analyzed visually^[6]. The distribution of the observed data, the location of the outliers. If some data points are indeed considered outliers, and eliminating them does not affect the analysis results, these data points can be removed from the data set.

2.3. Rational analysis

2.3.1 Variety of foods, reasonable collocation

According to the eight Dietary Guidelines for Chinese Residents (2022), a cereal-based balanced diet should be maintained, and the daily diet should include cereals, potato, vegetables and fruits, livestock, poultry, fish, eggs, milk, and beans. On average of more than 12 foods per day and more than 25 per week. Do a good job of meat and vegetable collocation, the main and auxiliary collocation. According to the questionnaire data after dimension reduction, the bar chart can be obtained as shown in Fig 2:

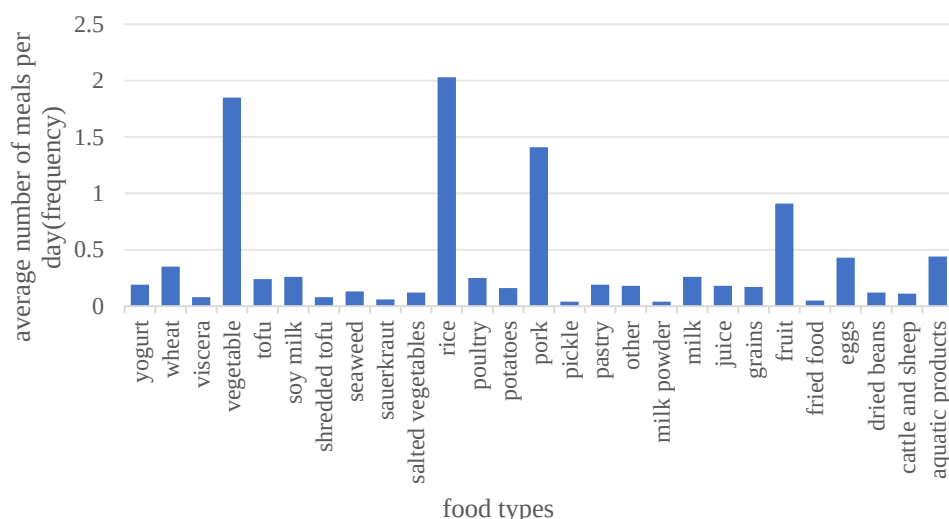


Fig. 2 Average daily use of each food item

In Fig 2, on average, residents eat only rice, vegetables, and meat that are guaranteed to be eaten at least once a day, while other categories of food are consumed less frequently. In grain and potato, only rice is eaten more per day, while wheat, grains and potatoes are eaten less, indicating that residents' intake of grain and potato food is not diversified enough and too single. Vegetables and fruits can basically be eaten at least once a day. Among livestock, poultry and fish, pork is eaten the most often.

Combined with the consumption frequency of various foods, most residents did not achieve the diversity of food intake, Each category of food may only often eat a small part, although the meat and vegetable collocation, but should also increase the number of consumption of a variety of different foods.

2.3.2 Diverse diet

In the Chinese residents dietary guidelines (2022), the third of the eight balanced dietary guidelines mentioned vegetables and fruits, whole grains and dairy products is an important part of a balanced diet, to ensure the daily intake of not less than 300g of fresh vegetables, but also to ensure that the daily intake of 200~350g of fresh fruit, the average daily intake of food data visualization as follows.

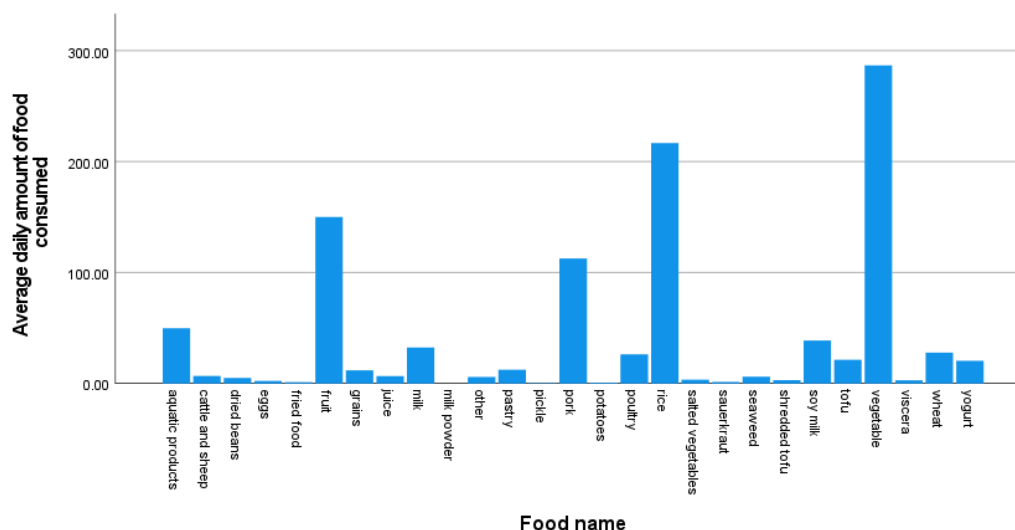


Fig. 3 Meverage daily consumption of each food

In **Fig 3**, the average daily intake of cereal is about 250g, which basically meets a person's daily requirements for cereal food. The average daily intake of fresh vegetables is about 290g, which is close to the balanced diet guidelines. The requirement is no less than 300g per day, indicating that residents basically eat enough vegetables, but it should also be improved appropriately enough to meet the requirements in the guidelines. The average daily intake of fresh fruits is about 150g, which is lower than the 200 to 350 g required in the balanced dietary guidelines.

Considering the above analysis of food types, the intake of most foods is lower than the balanced dietary guidelines, especially the intake of fruit is far below the normal standard.

2.3.3 Reduce salt and oil, and control sugar and wine

By analyzing the relevant data in the questionnaire, the results obtained are shown in Table 4:

Table 4. Statistical analysis of the total daily consumption (g) of food such as oil and salt

	plant oil	animal oil	salt	soy sauce	vinegar	Sauce	aginomoto	drink
average value	121.53	6.18	15.24	19.75	6.74	2.08	1.56	47.42
median	116.67	0	13.33	16.67	5	0	0	0
mode	166.67	0	16.67	16.67	0	0	0	0
crest value	1333.33	666.67	1166.67	558.33	166.67	83.33	53.33	3228.57
25 centile	83.33	0	8.33	8.33	0	0	0	0
50 centile	116.67	0	13.33	16.67	5	0	0	0
75 centile	166.67	0	16.67	25	8.33	3.33	1.67	0

Through the indicators of eating at home on the questionnaire, the results are shown in Table 5:

Table 5. Statistical analysis of the number of people eating at home on weekends

	Number of people eating breakfast at home on weekends	Number of people eating lunch at home on weekends	Number of people eating dinner at home on weekends
average value	3.2	3.56	3.65
median	3	3	3
mode	3	3	3

In Table 5, the average number, number and median of people eating at home on weekends are almost 3. Therefore, in the following analysis of food addition data, the default number of each family is 3.

In Fig 3, the average daily consumption of high-salt foods, such as pickles, is about 4g. According to Table 4, the average daily intake of 15g of salt per person and the average daily intake of 5g of salt

per person, which meets the standard of no more than 5g of salt per person in the balanced dietary guidelines per day.

In Fig 3, the average daily consumption of fried pasta is about 1g, and the intake is small, which meets the normal demand index. As can be seen from Table 4, the average household intake of 121g of vegetable oil per day, while the median, number and quartile of animal oil are 0g. It is known that most residents almost do not eat animal oil, so the average total daily oil intake is about 121g, and the average per person is about 40.3g. Above the requirement of 25 to 30 g in the balanced dietary guidelines.

In Table 4, the average daily total consumption of alcohol is about 47g, but the median, number and quartiles of alcohol statistics are 0, indicating that all the residents did not drink alcohol and only a small number of residents consumed alcohol. But the average daily intake of residents exceeded the guideline standard of 15g.

2.4. Establish a logistic regression model and solve the hypertension data

Dichotomous logistic regression is a logistic regression^[7]An application to solve the dichotomization problem, dividing samples into two categories (e. g. 0 and 1). Its goal is to predict the probability that a sample belongs to a positive category based on the input features. The hypertension index is a binomial categorical variable, which has only two values, 1 for hypertension and 0 for no hypertension, meeting the use conditions of dichotomy logistic regression.

2.4.1 Construct a logistic regression model

Bitaxonomic logistic regression model The dependent variable Y is whether hypertension was diagnosed, and its value is coded as 0 and 1.

$$Y = \begin{cases} 1(\text{ Sick }) \\ 0(\text{ Not sick }) \end{cases} \quad (5)$$

There are 95 indicators that affect the value of Y (including the basic information, living habits and eating habits in the table, etc.), namely X_1 、 X_2 、 X_3 、...、 X_{95} , The probability of $Y = 1$ is written as p, and the probability of $Y = 0$ is 1-p. The following equation can be established:

$$\logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_m X_m \quad (6)$$

M is the number of influencing factors, namely 95, β_0 Is the constant term, and the β_1 、 β_2 、 β_3 、...、 β_m . It is the regression coefficient corresponding to the influencing factors (independent variables). After a simple transformation, the logistic regression prediction model can be obtained as follows:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_m X_m}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_m X_m}} \quad (7)$$

When analysis with real data, maximum likelihood estimation (MLE) is usually used for the regression coefficient β_i To estimate. The corresponding sample joint probability, that is, the likelihood function is:

$$L = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{1-Y_i} \quad (8)$$

p_i Indicates the probability of a positive result in the corresponding case. If what actually appears is a positive result, $Y_i = 1$, otherwise, $Y_i = 0$.

2.4.2 Model evaluation

According to the above formula, when the relevant evaluation data is calculated and the ROC curve is drawn to evaluate the established model^[8]. The results are shown in Table 5:

Table 6. The model evaluation data

P	precision	recall	Accuracy rate	F1	AUC
0	0.95	0.95	0.934	0.935	0.879

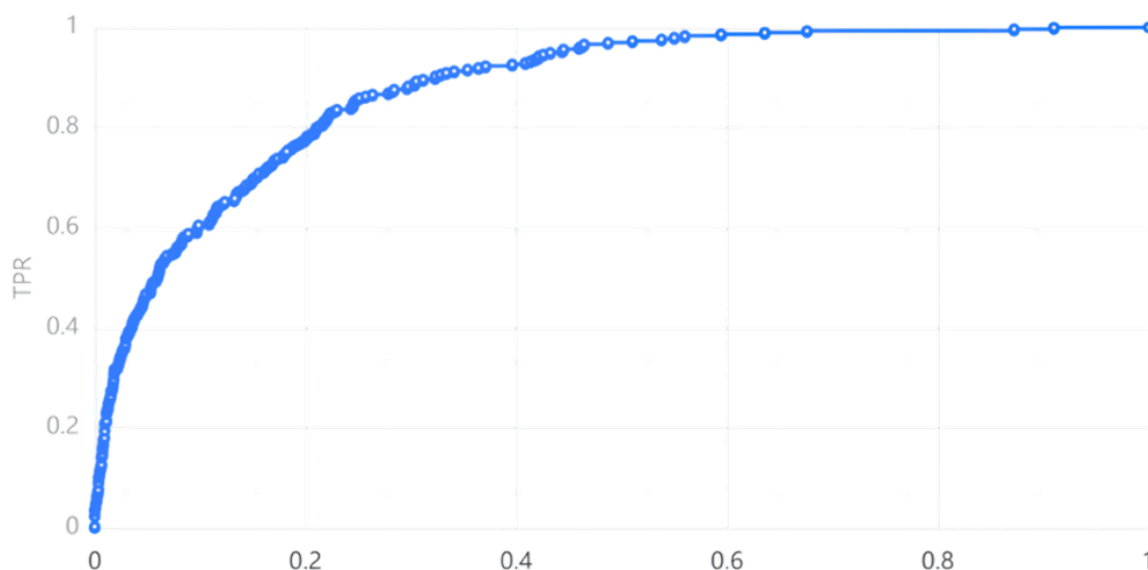


Fig. 4 The ROC curve

In Table 6, the P-value is 0, which is much less than 0, indicating that the model is valid. At the same time, the AUC is 0, but the closer the AUC is to 1, the better the effect of the model. In Fig 4, the ROC curve^[9] has the area enclosed by the positive direction along the x axis, indicating that the model has good performance at various thresholds.

2.4.3 Interpretation model

SPSS for all indicators, and the results are shown in Table 7:

Table 7. Regression coefficient of the basic information class index

Item	regression coefficient
year of birth	0.0034
sex	0.4805
degree of education	0.2156
marital status	-0.1855
occupation	-0.1224

Table 8. Some regression coefficients of lifestyle habits category indicators

Item	regression coefficient
Whether to smoke	-0.1013
Days of passive smoking	0.1188
Total number of smoking	-0.0081
Days to eat breakfast	-0.0765
Days to eat Chinese food	0.2086
Whether to take part in physical exercise	-0.0962
...	...

According to the analysis in Table 7, among the basic information indicators, except birth year, other indicators have certain correlation with the prevalence of hypertension, among which gender was the most correlation, and marital status, educational level and occupational nature also had weak correlation. According to Table 8 and the analyzed data, proper physical exercise and reasonable dietary arrangement can reduce the prevalence of hypertension to some extent^[10].

3. Conclusions

This study clarified the important relationship between residents' dietary patterns and the impact of multiple factors on chronic diseases. The findings of this article indicate that although many urban residents prefer rice, vegetables, and meat, there is a lack of fruits and dairy products, which may lead to deficiencies in certain nutrients. In addition, this article also found that some residents drink alcohol in excess of health recommendations, posing potential risks to long-term health.

Hypertension is an important chronic disease that this article focuses on. The logistic regression model in this article clarifies the relationship between multiple factors and hypertension. Factors such as gender, marital status, and cultural background are all associated with hypertension. In addition, the study in this article highlights that the risk of hypertension can be effectively reduced by increasing physical activity and adjusting dietary structure, especially increasing the intake of vegetables and fruits.

In short, through clear data analysis, this study highlights the impact of different factors on chronic diseases and reveals the close connection between these factors and the dietary structure of residents. These findings provide practical guidance and recommendations for improving residents' health, emphasizing the importance of dietary patterns to reduce the risk of chronic diseases.

References

- [1] Liu Beibei, Tian Qingfeng, Guo Jinling. Analysis of chronic diseases status and comorbidity patterns in middle-aged and elderly people in China [J]. *Medicine and Society*, 2022,35 (08): 58-61 + 66..
- [2] Deng Qijuan, Liang Jingdi, Xu Deli. Analysis of lifestyle and dietary habits among people at high risk of hypertension [J]. *Grassroots Medical Forum*, 2021,25 (20): 2930-2931.
- [3] Duan Yuanyuan. Basic principles of educational questionnaire design and reliability and validity test methods [J]. *Modernization of Education*, 2019,6 (28): 101-103.
- [4] Kong Qin, Ye Changqing, Sun Yun. Research on data preprocessing methods under big data [J]. *Computer Technology and Development*, 2018,28 (05): 1-4.
- [5] Zhuo Lin, Zhao Houyu, Zhan Siyan. Review of anomaly detection methods and their applications [J]. *Computer Research*, 2020,37 (S1): 9-15.
- [6] Wang Xiaokai, Zhu Xiaowen. Three statistical methods for identifying and eliminating outliers in metrological verification [J]. *China Test*, 2018,44 (S1): 41-44.
- [7] Fei Yunli. A computer logistic regression analysis [J]. *Hunan Industrial Vocational and Technical College*, 2020,20 (01): 14-17.
- [8] Wu Saier, Chen Jian, ZHOU Wendy, et al. Evaluation and test of debris flow susceptibility based on logistic regression model: Take Benzi bar- -Changbo river section in the upper reaches of Jinsha River as an example [J]. *Modern Geology*, 2018,32 (03): 611-622.
- [9] Wang Yanguang, Zhu Hongbin, Xu Weichao. Overview of the ROC curves and their analytical methods [J]. *Journal of Guangdong University of Technology*, 2021,38 (01): 46-53.
- [10] Yuan Jiao, Wu Qingsong, Lei Shu, etc. Study on the prevalence status and influencing factors of hypertension in middle-aged and elderly people in China [J]. *General Practice in China*, 2020,23 (34): 4337-4341.