

Study on the correlation between residents' habits and chronic diseases based on multiple regression model

YiFan Fu^{1*}, XiWen Bao²

¹ Department of Economics and Finance, Sanya University, Sanya, China, 572022

² School of Saxo Fintech Business, Sanya University, Sanya, China, 572022

* Corresponding Author Email: 15501938775@163.com

Abstract. With the change of people's lifestyle, the prevalence of chronic diseases continues to rise, and the health status is closely related to age, diet, physical activity, occupation and so on. In order to further analyze the relationship between common chronic diseases (such as hypertension and diabetes) and smoking, drinking, eating habits, living habits, work nature, exercise and other factors, as well as the degree of correlation, this paper conducted dimensionality reduction on two variables of living habits and eating habits, and entered the data after dimensionality reduction into pearson correlation coefficient test to analyze the correlation. The second dimension reduction is carried out again, and fewer indexes are obtained. According to the index after the second dimension reduction, the multiple regression analysis model is established to analyze the correlation. This study has reference value for residents to choose healthy daily life style.

Keywords: Data Processing, Multiple Regression, Principal Component analysis.

1. Introduction

With the change of people's lifestyle, the prevalence of chronic diseases continues to rise. As we all know, health is closely related to age, eating habits, physical activity, occupation and so on. How to achieve the purpose of promoting physical health through reasonable arrangement of diet, appropriate physical exercise and practice of healthy lifestyle is a common concern of the whole society.

In order to deeply analyze the relationship between common chronic diseases (such as hypertension, diabetes, etc.) and factors such as smoking, drinking, eating habits, living habits, nature of work, exercise, etc., and the degree of correlation, this paper demonstrates based on survey data. Li Ting used Chi-square test, rank sum test and binary logistic method to analyze the influencing factors in the analysis of chronic disease status of rural residents in minority areas, and concluded that education is the main influencing factor of residents' health [1]. Both Chen Xin and Gong Chuanhui used binary logistic model for regression analysis in their research. Chen Xin conducted basic data analysis in the study on influencing factors of physical activity of pregnant women based on health belief model. On the premise that the data conform to normal distribution, independent T test and binary Logistic model are adopted to analyze the data, and it is concluded that motion is the most important influencing factor [2]. With maturity scale and dimension as data support, Gong Chuanhui conducted an empirical analysis on the influencing factors of physical health, mental health and comprehensive health of the elderly population in China by using the binary Logit model and the generalized ordered Logit model, and concluded that education, exercise and security system are the most important influencing factor [3]. Wang Yudan used Chi-square test and Fisher exact probability method to study the factors affecting the physical health of students in Changchun city. Statistically significant variables in the univariate analysis were incorporated into the multivariate logistic analysis, and the OR value was used for screening [4]. According to Yao Chenziah's research on the social and economic factors affecting residents' health, based on the mature data scale as the research data, ordered Logistic regression was adopted for regression analysis. Furthermore, grouping regression was used to compare and analyze the existing differences, and it was concluded that education level, social and economic status and income were the most important influencing factors [5]. Ning Xiaochun used binary logistic regression to analyze the physical health status and influencing factors

of the elderly, and concluded that physical exercise was the most important influencing factor.[6]Wu Haiyun also used this method to analyze the factors of chronic diseases in Guangzhou. It is concluded that the smoking habits of middle-aged and elderly residents are related to gender, drinking habits, passive smoking time and educational level [7]. Ding Ning adopted stratified random sampling method combined with single factor unconditional logistic regression analysis and concluded that old age, overweight and obesity, and smoking are important influencing factors of chronic diseases [8]. Han Jinlin also used multi-factor unconditional Logistic to establish regression analysis and conduct factor analysis [9]. Han Peng used stratified cluster random sampling combined with CHAID decision tree method for analysis and concluded that it was an important influencing factor [10].

In order to seek the relationship between chronic diseases and factors such as smoking, drinking, eating habits, living habits, work nature and exercise, as well as the degree of correlation, this paper establishes the exact function expression, and then sets up independent variables to construct multiple linear regression. The dimensionality of 29 indicators of eating habits was reduced for the first time and then for the second time to meet the requirements of linear regression. A multiple regression analysis model was established according to the indicators after the second dimensionality reduction and correlation analysis was carried out.

2. Data dimension reduction

In order to obtain the correlation between residents' living habits and chronic diseases, this paper needs to construct multiple linear regression for analysis. However, there are many indicators in the original data, so direct analysis shows that the indicator data is too excessive, too miscellaneous and does not meet the requirements of linear regression. Therefore, dimensionality reduction of indicators is required.

As for lifestyle habits, we believe that smoking, drinking and physical activity are included. Considering the workload, correlation analysis cannot be conducted directly, and direct correlation analysis will lead to more overlap among indicators, so it is necessary to reduce the dimension of indicators. The data in this paper are from: www.mcm.edu.cn. It is authoritative and accurate to select the questionnaire data of residents' life and health.

For smoking status, only the number of cigarettes smoked in a week and the number of passive smoking days were counted to indicate the smoking status of the population. For alcohol consumption, the number of drinks consumed per week was measured. For physical activity, the average time spent per day by residents was measured. The original indicators are represented as these four indicators to replace people's living habits. Some results are shown in Table 1:

Table 1. Partial results of lifestyle dimension reduction

ID	Total smoking	Days of passive smoking	Capacity for alcohol	Average time spent exercising each day
10001	0	0	0	0
10002	0	0	0	20
10003	140	5	0	0
10004	0	2	0	0
10005	70	0	0	0
10006	0	7	0	20
10007	0	7	0	60
10008	0	2	0	0
10010	0	1	0	0

For eating habits, the 109 indicators given are too large and need to be reduced in dimension. A brief dimension reduction is done here, and the overlapping indicators are replaced by dimension reduction.

In order to avoid overlap among indicators, the dimensions of overlapping indicators were reduced here, and 109 diet-related indicators were reduced to 39 indicators. For example, we believe that there is a certain overlap in the selection of indicators for dining places, and indicators such as the number of days you don't eat breakfast and eating breakfast at home can be included. The Original questionnaire data are shown in Table 2:

Table 2. Original questionnaire data

	Meal number	without eating	Eat at home	Bring meal	Office canteen	restaurant	Number of people eating at home at weekdays	Number of people who eat at home on weekdays
D1	breakfast							
D2	lunch							
D3	dinner							

Therefore, we selected six indicators to reduce the dimension of the dining place. The following six indicators were selected to describe the dining place (Number of people who eat breakfast at home on weekdays, Number of people who eat breakfast at home on weekends, Number of people who eat Chinese food at home on weekdays, Number of people who eat Chinese food at home on weekends, Number of people eating dinner at home on weekdays, Number of people who eat dinner at home on weekends).

The same processing was performed on other dietary data, and the final processing results were shown in Table 3:

Table 3. Partial results of dimensionality reduction

ID	Rice consumption	Amount of wheat flour consumed	The amount of edible grains	Amount of edible tubers	Amount of fried noodles eaten	Pork consumption
10001	1	4	0	1	0	1
10002	3	3	2	2	0	1
10003	4	0	0	0	0	4
10004	4.5	8	2	1	0	1
10005	1.5	4	0	0	0	5
10006	4	6	0	4	0	2
10007	2	2	2	2	0	0.5
10008	4	0	1.5	1	4	1
10010	1.5	2	3	1	0	1
10011	2	2	0	0	4	2

For the 29 indicators of eating habits after initial dimensionality reduction, linear regression still could not be performed, so secondary data dimensionality reduction was carried out. KMO test and Barritt sphericity test were used to determine the overlap between indicators, and principal component dimension reduction was carried out through the test, and t-sne was used to reduce dimensionality for those that passed.

Firstly, KMO test and sphericity test are carried out to judge the correlation of these secondary indexes under the primary indexes, and then the dimensionality reduction method is selected. In this paper, the principal component analysis method is used to reduce the dimensionality of multidimensional indicators. Generally speaking, the greater the correlation between indicators, the better the dimensionality reduction effect. The test results are show in Table 4:

Table 4. Results of KMO test and Bartlett spherical test

Index name		diet
Bartlett Sphericity test	KMO value	0.839
	Approximate chi-square	57700.152
	df	741
	P	0.000**

(Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively)

The P-value is 0.000($P \leq 0.00$ **) for this test result, we used the principal component analysis to reduce the dimension of the level index of diet.

Principal component dimension reduction requires the following steps: 1.Standardization processing, and then eliminate the difference between the dimensions and the value of each index; 2.Construct and calculate the correlation coefficients collected among various indicators, and construct the correlation coefficient matrix to facilitate subsequent calculation; 3.calculate the eigenvalue eigenvector; 4.select the principal component; 5.Calculate the overall score.

The specific results of calculating the eigenvalues according to the third step are show in Table 5:

Table 5. Eigenvalue contribution rate table

Total variance interpretation				Total variance interpretation			
Characteristic root				Characteristic root			
ingredie nt	Characteris tic root	Variance interpretati on rate (%)	Cumulati ve variance explanati on rate (%)	ingredie nt	Characteris tic root	Variance interpretati on rate (%)	Cumulati ve variance explanati on rate (%)
1	5.25	13.461	13.461	21	0.774	1.984	74.049
2	3.948	10.122	23.583	22	0.761	1.951	76.001
3	1.484	3.806	27.389	23	0.746	1.913	77.913
4	1.456	3.734	31.123	24	0.732	1.877	79.791
5	1.398	3.586	34.709	25	0.725	1.859	81.649
6	1.336	3.425	38.134	26	0.704	1.805	83.454
7	1.193	3.058	41.192	27	0.684	1.754	85.208
8	1.095	2.807	43.999	28	0.68	1.742	86.95
9	1.062	2.723	46.722	29	0.655	1.68	88.63
10	1.024	2.627	49.349	30	0.644	1.652	90.282
12	0.978	2.507	54.427	32	0.612	1.57	93.461
13	0.939	2.407	56.834	33	0.569	1.459	94.92
14	0.926	2.375	59.209	34	0.549	1.407	96.327
15	0.905	2.321	61.531	35	0.519	1.33	97.657
16	0.857	2.197	63.728	36	0.485	1.243	98.9
17	0.847	2.173	65.901	37	0.188	0.481	99.38
18	0.808	2.072	67.972	38	0.147	0.377	99.757
19	0.808	2.071	70.043	39	0.095	0.243	100
20	0.789	2.022	72.065				

Based on the consideration of contribution rate, the first 11 variables are selected, and the dimensionality reduction of principal components can obtain a better interpretation of the original variables. Therefore, the first 11 principal components were selected for the subsequent dimensionality reduction processing. The results are shown in Table 6:

Table 6. Principal component selection results and weights

name	Variance interpretation rate (%)	Cumulative variance explanation rate (%)	Weight (%)
Principal component1	0.135	13.461	25.99
Principal component2	0.101	23.583	19.544
Principal component3	0.038	27.387	7.345
Principal component4	0.037	31.119	7.206
Principal component5	0.036	34.702	6.919
Principal component6	0.034	38.127	6.612
Principal component7	0.03	41.174	5.884
Principal component8	0.028	43.966	5.39
Principal component9	0.027	46.674	5.23
Principal component10	0.026	49.277	5.024
Principal component11	0.025	51.791	4.855

Finally, the data of smoking, drinking, eating habits, living habits, work nature, exercise and other factors are summarized are show in Table7:

Table 7. Summary of final dimensionality reduction data

Total smoking	Days of passive smoking	Capacity for alcohol	Average time of physical activity per day	occupation	Diet score
0	0	0	0	9	-0.422
0	0	0	20	13	0.195
140	5	0	0	4	-0.201
0	2	0	0	12	-0.029
70	0	0	0	9	-0.598
0	7	0	20	9	0.039
0	7	0	60	10	-0.253
0	2	0	0	10	-0.342
0	1	0	0	10	-0.340
0	7	0	60	1	-0.352
0	0	0	30	10	-0.393

3. Establish multiple linear regression model

For the analysis of common chronic diseases (such as hypertension, diabetes, etc.), only one indicator is selected to represent the relevant chronic diseases, such as y1 for hypertension and y2 for diabetes. Multiple linear regression equations were established between y1 and smoking, drinking, eating habits, living habits, work nature, exercise and other factors (x1,x2,x3,x4,x5,x6). The same goes for Y2. Here, take hypertension as an example, for the relevant indicators of hypertension, too

miscellaneous, choose to reduce dimension treatment. KMO and Bartlett tests were carried out, and the results were as follows.

Table 8. KMO and Bartlett test results

Sample the Kaiser-Meyer-Olkin measure of adequacy.		0.827
Bartlett Sphericity test	Approximate chi-square	18686.587
	df	15
	Sig.	0.000

It is found that the results are good and can be directly used for dimensionality reduction by principal component analysis. Therefore, the hypertension dataset was reduced to one data, are show in Table9.

Table 9. Characteristic values of hypertension

Total variance interpretation			
Characteristic root			
ingredient	Characteristic root	Variance interpretation rate (%)	Cumulative variance explanation rate (%)
1	3.486	58.103	58.103
2	1.01	16.829	74.932
3	0.604	10.072	85.004
4	0.429	7.15	92.154
5	0.277	4.617	96.771
6	0.194	3.229	100

Finally, a list of indicators used to describe the data related to hypertension was obtained. With this indicator as y, a multiple linear regression model was established. Through the dual analysis of the results of image and Person test in the process of data processing, we can see the relationship between these variables. Therefore, we build a multiple linear regression model as follows:

$$\begin{cases} y_{\text{hypertension}} = \alpha_0 + \alpha_1 x_{\text{smoking}} + \alpha_2 x_{\text{Passivesmoking}} + \alpha_3 x_{\text{xtipple}} + \alpha_4 x_{\text{Exerciseduration}} + \alpha_5 x_{\text{diet}} \\ y_{\text{diabetes}} = \beta_0 + \beta_1 x_{\text{medical}} + \beta_2 x_{\text{aid}} + \beta_3 x_{\text{often}} + \beta_4 x_{\text{Exerciseduration}} + \beta_5 x_{\text{diet}} \\ \xi \sim N(0, \sigma^2) \end{cases} \quad (1)$$

Where, ξ following normal distribution, $\beta_1, \beta_2, \dots, \beta_4$ are the respective variable coefficients

By substituting the data into MATLAB and using the regress function for calculation, the results are show in Table10:

Table 10. Results of hypertension regression analysis

	Nonnormalized coefficient		Standardization coefficient	t	P	VIF	R ²	F
	B	Standard error	Beta					
constant	-0.169	0.027	-	6.303	0.000**	-		
Total smoking	0.001	0	0.036	2.59	0.010**	1.214		
Days of passive smoking	-0.01	0.004	-0.032	2.301	0.021**	1.159		
Capacity for alcohol	0	0	-0.011	0.815	0.415	1.098	0.015	F=15.425 P≤0.000***
Average daily physical activity	0.002	0	0.076	5.917	0.000**	1.013		
occupation	0.017	0.003	0.079	6.056	0.000**	1.037		
Diet score	0.035	0.03	0.015	1.168	0.243	1.006		

The significance P value is 0.000, rejecting the null hypothesis that the regression coefficient is 0, so the model basically meets the requirements. For the collinearity of variables, VIF is all less than 10, so the model has no multicollinearity problem, and the model is well constructed.

The same treatment can be done for diabetes, and the results are show in Table11:

Table 11. Results of regression analysis of diabetes mellitus

	Nonnormalized coefficient		Standardization coefficient	t	P	VIF	R ²	F
	B	Standard error	Beta					
Capacity for alcohol	-0.157	0.033	-	4.694	0.000**	-		
Total smoking	0.001	0	0.024	1.695	0.090*	1.218		
Days of passive smoking	-0.007	0.005	-0.02	1.432	0.152	1.16		
Capacity for alcohol	0	0.001	-0.011	0.822	0.411	1.1	0.008	F=7.973 P≤0.000**
Average daily physical activity	0.002	0	0.055	4.236	0.000**	1.013		
Diet score	0.016	0.004	0.059	4.486	0.000**	1.037		
Diet score	0.019	0.038	0.006	0.501	0.616	1.006		

Multivariate linear regression model was established to analyze the results. The results showed that hypertension was significantly related to the total number of smoking, the number of passive

smoking days, the average time of physical exercise per day and occupation, and had little correlation with other indicators. There was a significant relationship between diabetes and the total number of cigarettes smoked, the average time of physical exercise per day, and occupation, and there was little correlation with other indicators.

4. Conclusion

Chronic non-communicable diseases (hereinafter referred to as chronic diseases), represented by cardiovascular and cerebrovascular diseases, diabetes, malignant tumors and chronic obstructive pulmonary disease, have become an important problem affecting the health of Chinese residents.

On the basis of the stipulated Dietary Guidelines for Chinese Residents (2022), this paper processed the data information of the questionnaire and analyzed the correlation between the two variables of living habits and eating habits and other data indicators given. Firstly, the two variables of living habits and eating habits were treated with dimensionality reduction, and the data after dimensionality reduction was entered into pearson correlation coefficient to analyze the correlation.

Chronic diseases have become an important health problem of Chinese residents. With the change of people's lifestyle, the prevalence rate of chronic diseases continues to rise. Based on the multiple regression model, this paper studies the relationship between residents' living habits and chronic diseases, and analyzes that smoking, physical exercise and occupational factors lead to chronic diseases and physical health problems of residents. In daily life, residents should maintain good living habits, reduce the number of smoking and passive smoking, and strengthen physical exercise to ensure good health. Through the analysis of the relationship between chronic diseases and daily living habits, this paper can make some guiding suggestions for residents to prevent chronic diseases.

References

- [1] Li Ting, Hou Weiwei, Kang Hua et al. Minority areas rural residents chronic illness condition analysis [J]. Journal of health education and health promotion, 2021 (03) : 241-244.
- [2] Chen X. Study on influencing factors of physical activity in pregnant women based on health belief model [D]. Jilin university, 2020. DOI: 10.27162 /, dc nki. Gjlin. 2020.005222
- [3] GONG C H. Study on health status of elderly population in China and its influencing factors [D]. Northwest normal university, 2021. DOI: 10.27410 /, dc nki. Gxbfu. 2021.002247
- [4] Wang Yudan. Status quo and influencing factors of common physical health problems of middle school students in Changchun [D]. Jilin university, 2021.
- [5] Yao C. Social and Economic factors affecting residents' health [D]. Zhongnan University of Economics and Law, 2019. (in Chinese)
- [6] Ning Xiaochun; Lai Yinjuan; Wu Changtao; Huang Zhibi; Mo Haijuan; Li Cuixia; Sun Qiufen. Physical health status and influencing factors of 60 ~ 69 years old elderly in Guangxi [J]. Modern Preventive Medicine, 2020, 47 (11): 2030-2034.
- [7] Cui Mingming, Nie Changhong. Based on the index evaluation system of our country food security evolution research [J]. Proceedings of the Chinese Academy of Sciences, 2019 (8) : 910-919.
- [8] Ding Ning, Li Yuxia. Wuwei in elderly residents status and influencing factors of chronic illnesses [J]. China sanitation engineering, 2023, 22 (5) : 628-629 + 632.
- [9] HAN Xilin, Wang Rui, Li Yujiao et al. Study on self-reported chronic disease prevalence and influencing factors of physicians in Yunnan Province [J]. Soft Health Sciences, 2019,37(10):80-84+88.
- [10] Han Peng, Gao Yu, Sun Xiaochen, et al. Some residents in shandong hypertension status and influence factors analysis [J]. Journal of preventive medicine BBS, 2023 (8) : 567-573.