The Research of Influencing Factors in Diet that Related to Heart Disease

Jiasheng Zhou*

School of Public Health, Nantong University, Nantong, 226004, China

*Corresponding author: sammylee@alum.mit.edu

Abstract. Although previous studies have demonstrated that the incidence of heart disease is associated with obesity, sex, race, nationality, and age, there are still many unidentified factors that deserve to be studied. In this research, multi linear method is used to deal with the data from the Kaggle "COVID-19 Healthy Diet Dataset" which was published and updated in 2021 for 170 countries. It is concluded that although heart disease has no connection with Alcoholic Beverages, Animal Products, Aquatic Products Other, Fish Seafood, Fruits Excluding Wine, Miscellaneous, Milk Excluding Butter, Offals, Oil crops, Pulses, Spices, Starchy Roots, Stimulants, Sugar Crops, Sugar & Sweeteners, Tree nuts, Vegetal Products, Vegetable Oils, Vegetables, it has a relatively strong connection with Meat and a particularly strong relationship with Animal fats, Cereals Excluding Beer, Eggs. Many of the factors have a possibly strong connection with cancer have never appeared in previous studies. This paper provides some new perspectives to study the factors of heart disease and point to the way for further research afterward.

Keywords: Heart disease; obesity; unidentified factors; diet; multi linear method.

1. Introduction

Heart disease is the general name of all kinds of diseases about heart and is a kind of common circulatory diseases. Including rheumatic heart disease, congenital heart disease, hypertensive heart disease, coronary disease, carditis and so on. Research in China shows that the death rate of heart disease is increasing continuously. According to the statistics, the mortality of heart diseases in China fluctuated from 2004 to 2009. The rising rate is different in urban and rural areas: the rate average annually increased 2.70% in cities while 6.95 in rural areas. In urban areas, the mortality of senior citizens, especially the ones whom are over 85 years old, having the fastest rising speed. The mortality of young people between the ages of 10 and 29 increased the fastest in rural areas. And newborns' mortality is falling in both of the area [1]. So the age of people significantly affects the changes of mortality.

There have been many international studies on the influencing factors of heart disease (CHD) so far. CHD is a severe malformation which has a serious impact on newborns' survival and growth. Controlled experiment found that the influencing factors are taking folic acid, balanced diet, natural conception, elderly pregnancy, adverse pregnancy history, early infection, early mental stimulation, early harmful substances contact history, sickness and nutrition during pregnancy. Among them, elderly pregnancy, adverse pregnancy history, pregnancy infection, early infection, early mental stimulation and nutrition during pregnancy are independent risk factors for CHD, while taking folic acid, balanced diet, natural conception are protective factors [2, 3]. According to a 2019 study, 130 million people are born globally every year among whom 280,000 died in the neonatal period due to CHD [4, 5].

The morbidity of congenital heart disease is closely related to the environment, region and race [6]. For example, in Yunnan province, the epidemiological survey of congenital heart disease in 2013-2104 showed that the morbidity of congenital heart disease in Li Su nationality was higher than that of Han nationality and Tibetan nationality, and the morbidity of congenital heart disease in Yi nationality was higher than that of Tibetan nationality [7]. Epidemiological reports show the white race's CHD morbidity is lower than any other races [8]. Epidemiological reports in Korea also show

that the morbidity is about 9.21 per thousand for men and about 8.38 per thousand for women [9], indicating that the age of people also influences the morbidity [10].

When it comes to rheumatic heart disease, degenerative valve disorders are closely connected with age, and will be worth with population ages. Rheumatic heart disease is common around the world and more and more American catch this kind of disease. Due to research, the globalization of the society of USA is the key factor [11].

At present, there is a comprehensive study on the influencing factors of congenital heart disease at home and abroad, data collection and processing are also sufficient, and clear research results have been obtained. However, plenty of studies focus on CHD, sex or race, while there are few data studies and analysis of the influencing factors in our everyday life. However, the mortality rate of heart disease is increasing year by year, the situation is not optimistic. So reliable reports of the influencing factors should be created to guide people's lifestyle, so as to minimize the risk of heart disease. The focus is on finding whether there are key factors in diet that cause people without heart disease to develop heart disease.

It is expected to find statistics about heart disease in Kaggle, find whether there is a linear relationship between each variable and verify it by Rstudio, so as to obtain effective influencing factors.

2. Methods

Studies are found that obesity rate is closely connected to heart diseases such as heart failure or CHD. The team at Johns Hopkins Hospital found that, excluding the effects of hypertension, high cholesterol and blood sugar levels, morbid obese people were twice as likely to have heart failure as those with normal BMI [12]. Moreover, obesity has a huge impact on the morphology, structure and function of the heart [13], and can also increase the risk of coronary heart disease, hypertension and heart failure. For that obesity is closely connected with heart diseases, the factors(diet) have something to do with obesity can also influence the heart diseases.

2.1. Data Sources

The data is quoted from Kaggle “COVID-19 Healthy Diet Dataset”. This dataset includes. The last couple of columns also include undernourished, and COVID-19 cases as percentages of the total population for comparison purposes.

Data for different food group supply quantities, nutrition values, obesity, and undernourished percentages are obtained from Food and Agriculture Organization of the United Nations and the data for population count for each country comes from Population Reference Bureau.

2.2. Variable Selection

This paper quotes percentage of fat intake from different types of food in countries around the world and the counts of obesity. As table 1 shows. There are some missing values in the dataset which are difficult to trace, so this paper deletes the countries of the missing values. This paper finds the maximum and the minimum of each variable and calculates the mean value.

2.3. Research Protocol

The linear model is used to process the data and find the connections between the variables. The linear method is created to fit the models, which is helpful to find the reasonable factors of obesity. To look for the more accurate method, this paper also uses multi linear method. The tool Rstudio is used to complete the construction of these models and methods.

Table 1. Information for the 26 variables

Elements	Mean**	Maximum	Minimum
Country	-	-	-
Alcoholic Beverages	0.01	0.09	0.00
Animal Products	20.70	36.90	5.02
Animal fats	4.14	14.94	0.03
Aquatic Products Other	0.01	0.05	0.00
Cereals Excluding Beer	4.38	18.38	0.99
Eggs	0.95	3.28	0.06
Fish Seafood	0.85	8.40	0.02
Fruits Excluding Wine	0.54	9.67	0.04
Meat	9.49	26.42	0.90
Miscellaneous	0.06	0.46	0.00
Milk Excluding Butter	5.11	17.76	0.18
Offals	0.15	0.73	0.00
Oil crops	3.33	28.56	0.06
Pulses	0.26	2.69	0.00
Spices	0.28	2.69	0.00
Starchy Roots	0.23	2.18	0.01
Stimulants	0.64	3.38	0.00
Sugar Crops	0.01	0.19	0.00
Sugar & Sweeteners	0.01	0.08	0.00
Tree nuts	0.70	4.98	0.00
Vegetal Products	29.30	44.98	13.10
Vegetable Oils	18.57	36.42	4.95
Vegetables	0.31	1.15	0.03
Obesity*	18.71	2.10	45.60

*There are three missing values, this paper delete them.
**The results after removing the missing values.

3. Results and Discussion

There are some missing data in the dataset, for that the missing data are in some countries with a small population which have little effects and it is difficult to trace the data, so this paper decides to delete them. There are too much variables and their names are too long so we call the variables Alcoholic Beverages to Vegetables diet distribution. The first step is to find if there are close connections between some variables of diet distribution and obesity.

It can be seen from Figure 1 that the variables Animal fats, Eggs, Meat and Cereals Excluding Beer are effective variables for that each of them is in a linear relation with obesity.

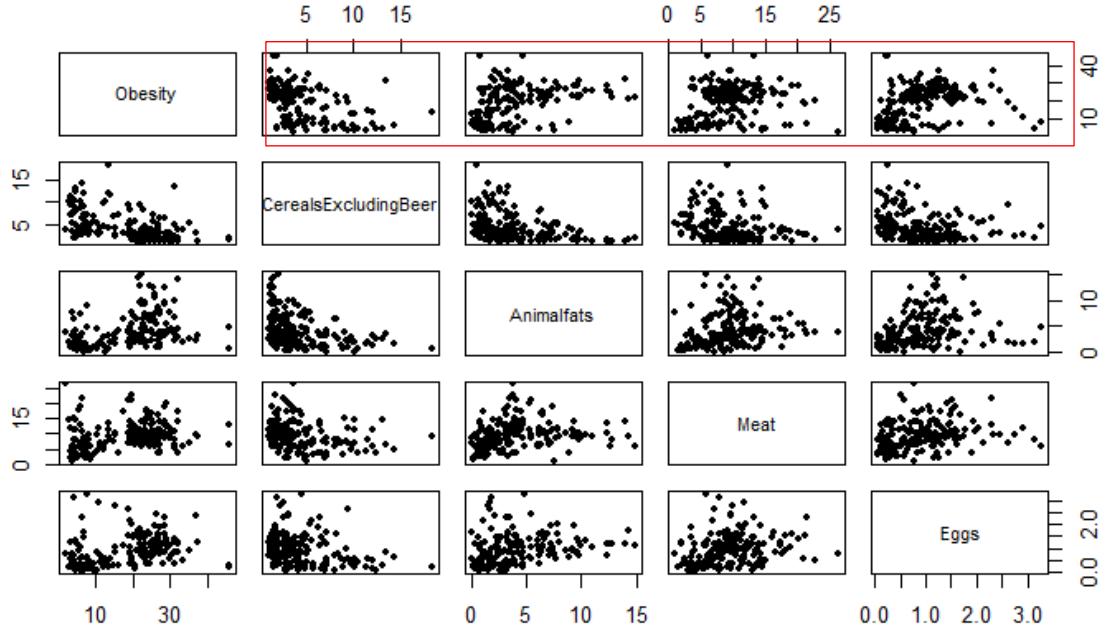

Fig. 1 Connections between variables and obesity

Then geom_point is used to view the distribution of points, identify the outliers and locate them in the dataset. The outliers are some meaningless extreme data and influence the linear method, this paper confirm that there is nothing wrong in processing data. So the outliers are finally deleted to ensure the accuracy. After removing the outliers, this paper creates a linear method and then simulate a linear relationship to see whether the method fit the distribution. The linear methods are summarized and the pictures are drown as shown in Figure 2.

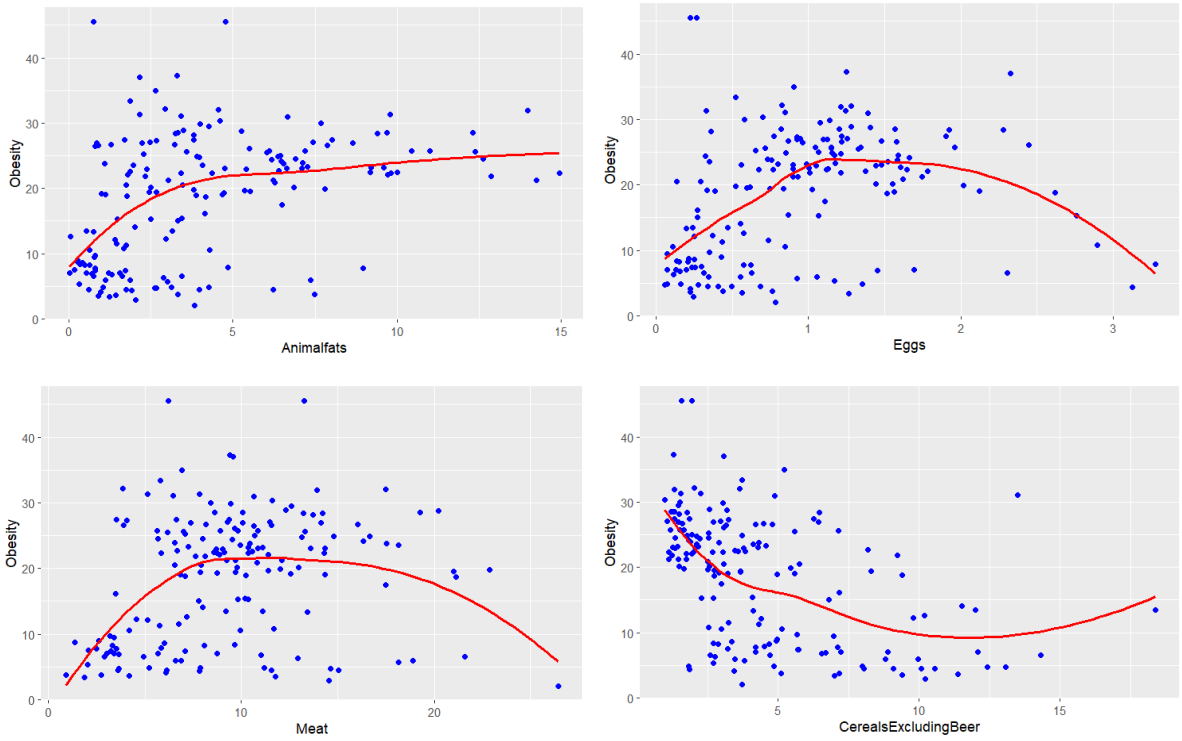

Fig. 2 Scatter gram and the fit of each method

Figure 2 shows some outliers, this paper locates the outliers and decides to delete the outliers. After deleting the outliers, this paper summarizes each method.

Table 2. Information for the linear methods

Variables	Estimate	Std. Error	t value	Pr(>|t|)	Relevance
Intercept(Eggs)	14.69	1.28	11.52	< 2e-16	***
Eggs	4.23	1.11	3.81	0.000196	***
Intercept(Animal fats)	13.95	1.10	12.65	< 2e-16	***
Animal fats	1.15	0.21	5.53	1.26e-07	***
Intercept(Meat)	14.44	1.64	8.78	1.96e-15	***
Meat	0.45	0.16	2.89	0.00429	**
Intercept	25.16	1.11	22.66	< 2e-16	***
Cereals Excluding Beer	-1.47	0.20	-7.18	2.21e-11	***

Relevance: Three types:“”, “**”, “***”.

Table 2 shows the summary of each method. The significance of linear relationship can be seen from the “*” behind the summary, so the Animal fats, Eggs and Cereals Excluding Beer are chosen to predict the obesity rate.

Then this paper predicts the method: delete the missing data and outliers, then 80% of the data are selected to be the training data and the rest to be the test data. Because each one was random selection, so the accuracy can be evaluated by repeating the above process many times. The predictions are observed and an analyzed by compared with the true value, while the predictions of all the three methods are significantly different from the true value. Finally multi linear method is chosen to predict the value.

Many permutations and combinations are tested such as Alcoholic Beverages and Animal Products; Animal fats, Aquatic Products Other and Cereals - Excluding Beer, or Eggs, Fish Seafood and Fruits - Excluding Wine. By comparing the accuracy of each combination, this paper find that the method reached by Animal fats, Eggs, Meat and Cereals Excluding Beer fit the model better than any other combinations (Table 3).

Table 3. Information for the multi linear method

Variables	Estimate	Std. Error	t value	Pr(>|t|)	Relevance
Intercept	18.29	2.34	7.83	6.13e-13	***
Meat	0.11	0.14	0.76	0.45	
Eggs	1.67	1.05	1.59	0.11	
Cereals Excluding Beer	-1.07	0.22	-4.77	4.00e-06	***
Animal fats	0.61	0.21	2.84	0.01	**

Then this paper repeats the process above, select 20% of the data to be the test data and use the multi linear method to predict the selected data. The result is that most of the predictions are accurate or close to the true value, but the predictions of some country are far from the true value.

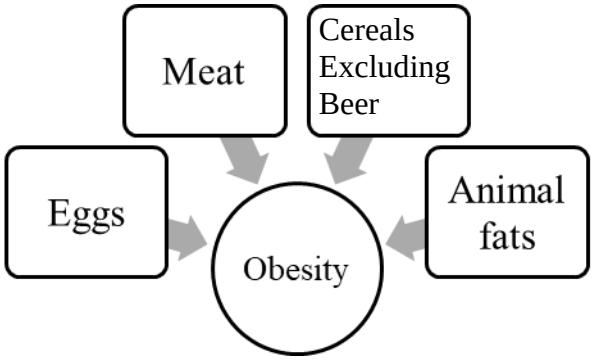

Fig. 3 Variable related obesity

As shown in Figure 3, the obesity rate is related to Meat, Eggs, Cereals Excluding Beer and Animal fats. Through calculations, this study yields the final linear regression equation:

$$\text{Obesity} = 1.108\text{Meat} + 1.671\text{Eggs} + 0.605\text{Animal fats} - 1.072\text{Cereals Excluding Beers} \quad (1)$$

After trying many times as well as verifying the accuracy of operation, this paper submits that the inaccurate predictions are influenced by some other factors. A gap of more than 15% of the true value is defined as inaccurate. To look for the other factors, this paper divide the inaccurate predictions into two groups, one is the countries whose predictions are higher than true value and the other is lower than true value. After repeating the predict process several times, this paper summarize and list the inaccurate country.

The result is that the countries whose predictions are higher: Cuba, Denmark, Indonesia, Madagascar, United Kingdom, New Zealand... And the countries whose predictions are lower: Cambodia, Fiji, Kuwait, Kiribati, Lesotho, Nicaragua, Yemen. It can be seen that the higher countries have archipelago or big islands. While some of the lower countries also have island, but they are small island states. The main characteristics of the higher countries are tropical and economically backward. This paper considers that the factors influence the predictions are as follows:

Different dietary structures: The small island like Fiji or Cambodia rely on fishing, so their diet is almost all based on seafood and fruit. This unbalanced diet may affect the results for that eating a moderate amount of seafood contributes to good health while eating too much seafood can lead to obesity.

Climate: Some big island such as Cuba, UK or Denmark, they have abundant food sources like other accurate countries. So this paper considers that the possible factor is climate. Most of them are temperate marine climate or tropical rainforest climate. So they have rich animal husbandry or fruit and vegetable industry, which help their citizens reduce the risk of obesity.

Economic situation: Countries like Cambodia, Fiji, Kuwait, Kiribati, Lesotho, Nicaragua, Yemen whose economy is relatively backward. This paper considers that to eat as much boring staples as possible, the people there use large amount of oil and seasoning, which are not concluded in the dataset but are important factors of obesity.

4. Conclusion

The current study selects diverse data and focuses on influencing factors that may be associated with developing heart disease. From the above study, it can be concluded that Animal fats, Eggs, Meat and Cereals Excluding Beer have direct linear relationship with obesity, most of which have not been paid much attention to before. But in some countries, the relationship is not obvious, which may due to the dietary structure, climate and economic situation.

It can not be denied that due to the limited types of diet, this method may have errors in addition to the factors, and the data which was used to prove that obesity is related to heart disease and the data about diet and obesity are not from the same group, causing possible differences, which may also affect the persuasiveness of the results. However, this method also has some advantages. First of all, instead of signal linear method, this paper uses the multi linear method, which not only is more effective but also avoid special values, thus ensuring the accuracy. What's more, it is helpful to the prevention and treatment of heart disease. Though some previous studies have proven that several factors are associated to heart disease such as sex, race, age or obesity, these factors are difficult to contribute to the guideline of people' daily life. While this paper concretizes the factors into daily diet, which enables people to prevent heart disease to a certain extent by reasonable diet. This paper also remains a problem about the factors that influence the accuracy of method in certain countries, which requires further investigation, pointing the way to further relevant research in the future. Once a new causative factor other than the one already identified is discovered, the risk of heart disease could be reduced and the quality of people's live would be improved.

References

[1] Wei Qin, Weiwei Shi, Jinchai Gao, Xiaoyan Ma, Yajing Cao, Jixin Sun, Ze Cui. A Joinpoint regression analysis of heart disease death trends in 2004-2019 in urban and rural China. The Chinese Cardiovascular Journal, 2022, 27(04): 371-376.

[2] Ye Pan, et al. Analysis of the characteristics and influencing factors of congenital heart disease. Journal of Xinxiang Medical University, 2019, 36(3): 281-284.

[3] Yali Zhang, Zhuochao Lyu, Juhong Lan, et al. Analysis on prevalence and influencing factors of congenital heart diseases in neonates born from 2018 to 2020. Maternal and Child Health Care of China, 2023, 38(01): 137-141.

[4] Jeena P M, et.al. Clinical profile and predictors of severe illness in young South African infants. South African medical journal, 2008, 98(11): 883-888.

[5] Lawn J E, Cousens S, Zupan J. 4 Million Neonatal Deaths: When?Where? Why? The Lancet, 2005, 365(9462): 891-900.

[6] Bernier P L, Stefanescu A, Samoukovic G, et al. The Challenge of Congenital Heart Disease Worldwide:Epidemiologic and Demographic Facts. Seminars in Thoracic & Cardiovascular Surgery Pediatric Cardiac Surgery Annual, 2010, 13(1): 26-34.

[7] Cilin Han, Lihong Jiang, Zongliu Hou, et.al. Epidemiological characteristics of congenital heart disease between 3 to 18 years in different regions of Yunnan Province. The Journal of Kunming Medical University, 2017, 38(08): 30-34.

[8] Knowles R L, Ridout D, Crowe S, et al. Ethnic and socioeconomic variation in incidence of congenital heart defects. Archives of Disease in Childhood, 2016, 102(6): 496-502.

[9] Jang S Y, Seo S R, Moon J R, et al. Prevalence and mortality of congenital heart disease in Korean adults. Medicine, 2018, 97(27): 11348.

[10] Kunlin Li, Lihong Jiang. Report on the incidence of congenital heart disease at home and abroad. Yunnan Medicine, 2020, 41(03): 286-289.

[11] Kameswari M M D, et al. Valvular Heart Disease: Diagnosis and Management. Mayo Clinic Proceedings, 2010, 85(05): 483-500.

[12] Jian Zhang, Xiaohong Huang. Advance of the effects of Obesity on the Cardiovascular System. The Chinese Journal of Molecular Cardiology, 2020, 20(03): 3417-3420.

[13] Qingxia Wang, Guangming Zhang. The Journal of Modern Integrated Traditional Chinese and Western Medicine. Progress in obesity and cardiovascular disease, 2020, 29(20): 2277-2280.