Study Applicable for Multi-Linear Regression Analysis and Logistic Regression Analysis

Ju Wu*

College of Mathematics and Information Science, Neijiang Normal University, Neijiang 641112, China

Abstract: Current study focus on using method of multi-linear regression analysis and logistic regression analysis, and discuss about the condition and scope of multi-linear regression analysis and logistic regression analysis. A modeling method has been introduced keeping in the basic principles of multi-linear regression analysis and logistic regression analysis. The modeling method and two forms of analytic methods have been analyzed, based on two clinic test data of diabetes and Model-2 diabetes as objects of study in combination with the analytic methods of multi-linear regression and logistic regression. Analysis result indicate that glycosylated hemoglobin, glycerin trilaurate, total cholesterol of serum and blood sugar concentration present obvious positive relation ($P < 0.05$), whereas insulin and blood sugar present negative relation($P < 0.05$); body mass index (BMI) and relative factors are dangerous; physical excise and relative factors are protective. In conclusion, multi-linear regression analysis and logistic regression analysis respectively have their own emphasis; for example, multi-linear regression analysis emphasizes on analyzing linear dependent relation with an dependent variable and multiple independent variables, whereas logistic regression analysis emphasizes on analyzing the relation between probability of occurring an incident and independent variables.

Keywords: Body mass index (BMI), diabetes, glycosylated hemoglobin, logistic regression analysis, model-2 diabetes, modeling method, multi-linear regression analysis.

1. INTRODUCTION

There is always studies on relation between two variables while handling measuring data. Such a relation generally exists in two types: one is complete determination relation, namely functional relation; and the other is correlativity, namely close relation between two variables, but in this case one cannot calculate outvalue of the other variable from one or more variables [1, 2]. In studies done in the fields of medicine, economy and engineering, usually the applicable method of studying relation between one variable and two or more influential factors are multiple linear regression and logistic regression analysis [3, 4].

Diabetes is one of metabolic diseases characterized by high blood sugar. High blood sugar results either from defective insulin secretion or its damaged biological action or both. High blood sugar permanently exists in diabetes resulting in damaging each tissue, especially chronic damage or functional disorder of eyes, kidney, liver, blood vessel and nerve. In clinical diagnosis, it usually makes definite diagnosis of patient with diabetes by inspecting patient's blood sugar, urine sugar, urine acetone bodies, glycosylated hemoglobin (HbA1c), glycosylated serum protein, serum insulin and C peptide level, blood fat, immunity index and urinary albumin excretion and so forth.

Model-2 diabetes is a kind of typical diabetes with complex pathogenesis, accounting for more than 90% among total diabetic patents. With the advancement of people's living standard and aging of population, diabetes has become one of the most serious public health issues in the world. It therefore, needs to be discussed the factors affecting diabetes especially with reference to Model-2 diabetes based on SIFT and examination, so as to provide with scientific principle while formulating preventive measures and community actions [5, 6]. This article mainly discusses pathogenic factors and risk factors affecting diabetes after respectively applying multi-linear regression and logistic regression analyses according to two aforesaid aspects. So, study of pathogenic factors and their pathogenicity, as well as social dangerous factors influencing diabetes, conducted with respect to comparison of results of two analytic methods, will help find full scope of these two analytic methods applicable for the industry.

2. MULTI-LINEAR REGRESSION ANALYSIS

Multi-linear regression analysis is one of analytic methods that quantitatively describe linear co-existence relation between one dependent variable and multiple independent variables by a regression equation. It is also called multi-linear regression. Under such a qualitative method we will deeply cognize analysis conclusion, and understand quantitative dependent relation between each two factors, which further discloses inherent rules between two factors [7]. Generally speaking, multi-linear regression procession shall synchronously provide with several backup function relations, as well as with understanding of the capacity of test data for each relation, which the researcher shall make selection according to self-expectation theory. In general, relation between relating variables shall be either linear or non-linear;

*Address correspondence to this author at College of mathematics and Information Science, Neijiang Normal University, Neijiang 641112, China; Tel: +8613825421189; E-mail: wuju@126.com

this article only elaborates study on the multi-linear regression analysis. A general multi-linear regression mathematic model is given by Formula 1:

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_3 \end{bmatrix}=BX=\begin{bmatrix} b_{01} & b_{11} & \cdots & b_{p1} \\ b_{02} & b_{12} & \cdots & b_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ b_{01} & b_{1n} & \cdots & b_{pn} \end{bmatrix}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_1 & \cdots & x_1 \\ \cdots & \cdots & \cdots & \cdots \\ x_p & x_p & \cdots & x_p \end{bmatrix} \tag{1}$$

Formula 1 shows a multi-linear regression process and synchronously provides with several optional function relations, where; yi stands for dependent variable, used to represent change of identical matter, xi stands for independent variable factor, acts as a group of factors influencing change of y, and B stands for parameter to be estimated. A general regression model selected by the researcher suitable for this study is as follows:

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_px_p \tag{2}$$

Where, b_0 stands for constant, representing total mean estimation for all dependent variables without their value being influenced by the independent variable, such as the estimation of total mean value for the dependent variable while the independent variable remains zero; b_1, b_2, …, b_p are called as partial regression coefficient, used to represent rate of change occurring along with the change of independent variable while the dependent variable is only influenced by the corresponding independent variable, namely rate of change of dependent variable while independent variable adds up a unit. There is a positive correlation between dependent variable and independent one when bi is positive; whereas there is a represents negative correlation between dependent variable and independent one when bi is negative.

B stands for parameter to be estimated in formulas 1 & 2. To calculate parameters b_0, b_1, b_2, …, bp from the function attributed to its mathematic model, this generally applies to minimum two squares, namely to find one approximate function out of functions attributed to its mathematic model so that such approximate function possibly gets access to real function based on well-known corresponding data.

In addition, preferred multiple non-linear regression analysis include: polynomial model, index model, power function model, growth curve model and so forth. However, in practice, multiple non-linear regression model is usually converted into multi-linear regression model through relating methods. As for conversion method, refer to literature [8], as no detail of analysis is given here.

3. LOGISTIC REGRESSION ANALYSIS

Linear regression model and generalized linear regression model require that the dependent variable is a continuous and normal distributive variable, and that the independent variable and the dependent one are linear in relationship. Assumption or condition for linear regression model is violated while independent variable and dependent factor have a non-linear relation. In the study on epidemiology, this condition usually meets when the dependent variable is a discrete variable. For example, medical treatment as none-effective, obvious medical effect, recovery; survival or death of little white rat under different toxic agents; and paroxysm or none paroxysm under some exposure. The most sought condition is that dependent variable is dichotomous. Herein, logistic regression model is the best, which has no requirement for distribution of dependent variable. Viewing from studies in mathematics, logistic regression model shall skilfully avoid distribution of classified variables, and make up and remove some defects from linear regression model and generalized linear regression. Viewing from studies in medical, logistic regression model can handle a number of practical issues and plays a crucial role in the development of medical science.

In the daily practice, we usually come across the dependent variable that is a classified one. To study the relation between a classified variable and a set of independent variables, manly occurring in epidemiology, we usually probe into risk factors of some diseases, as well forecast probability of disease occurrence according to the risk factors. For example, we selected two groups of people to discuss the risk factors for stomach cancer: one group is positive for stomach cancer, and the other is negative for stomach cancer, and subjects from both the groups definitely have different physical signs and life style and so forth. Herein, dependent variable is just "if there is stomach cancer", namely Yes or No, acting as two classified variables. Dependent variable may include a number of variables, such as age, sex, habit of diet, *Helicobacter pylori* infection and so forth. Dependent variable is thus consecutive and classified.

Logistic regression model and multi-linear regression model are actually mostly identical, the biggest difference lies on different dependent variables, and the rest is basically the same. Logistic regression analysis is one probability-mode regression analysis method, which may be used to analyze the relation between the probability of the occurrence of some dependent and independent variable, applicable to data when the independent variable is classified, especially applicable to such condition when the independent variable has been assigned a two-item classification. Independent variable in the model should either have a qualitative discrete value or a computation observation value.

For logistic regression model, refer to formula 3.

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\beta_1x_1+\cdots+\beta_nx_n \tag{3}$$

Hereby it concludes;

$$p=\frac{\exp\left(\beta_0+\beta_1x_1+\beta_1x_1+\cdots+\beta_nx_n\right)}{1+\exp\left(\beta_0+\beta_1x_1+\beta_1x_1+\cdots+\beta_nx_n\right)} \tag{4}$$

In practical application, logistic regression analysis is usually used to identify risk factors, seek forecast and judgment and so forth.

Table 1. Estimation and test result of multi-linear regression coefficient.

Variable	Partial Regression Coefficient	Standard Error	T Value	P Value
Constant item (b0)	6.3214	3.012	2	0.048
Glycosylated hemoglobin	0.7012	0.254	2.678	0.015
Glycerin trilaurate	0.3751	0.214	1.7	0.1
Total cholesterol of serum	0.1524	0.355	0.39	0.701
Insulin	-0.2861	0.101	-2.301	0.015

Note: R2=0.6072, F=8.315, P=0.00029.

4. APPLICATION OF MULTI-LINEAR REGRESSION ANALYSIS METHOD

To study the pathogenesis of diabetes and the correlation between two clinical pathogenetic test indexes, along with the discussion on the correct use of multi-linear regression analysis, we selected 54 subjects having definite clinical diagnosis, as well as corresponding to its clinical diagnosis case, we selected a few representative influence factors such as: total cholesterol of serum, glycerin trilaurate, insulin and glycosylated hemoglobin content. Then, an analysis has been carried out for each index under multi-linear regression model, so as to study the relation between each index and blood sugar level of diabetes patient. Multi-linear regression analysis results are referred in Table **1**.

From Table **1**, we can obtain a multi-linear regression equation as follows:

$$y = 6.3214 + 0.7012 * x_1 + 0.3751 * x_2$$
$$+0.1524 * x_3 - 0.2861 * x_4$$

Where, y stands for blood sugar content of diabetes patient, blood sugar contents of diabetes patient without any other influential factors are all 6.3214, x1, x2, x3, x4 respectively stand for glycosylated hemoglobin, glycerin trilaurate and total cholesterol of serum and insulin concentration. And if x1, x2, x3, x4 per unit concentration, patient blood sugar concentration increases to 0.7012, 0.3751, 0.1524 and -0.2861 respectively. If partial regression coefficient b4 is negative, that is concentration of blood sugar and insulin in a patient have a negative relation, and that insulin enables to some extent, the reduction in the blood sugar concentration, which is also consistent with the insulin function. Viewing from partial regression coefficient, glycosylated hemoglobin, glycerin trilaurate and total cholesterol of serum all relatively improve blood sugar concentration, including glycosylated hemoglobin that has maximum effect on blood sugar concentration, and total cholesterol of serum that has minimum effect on blood sugar concentration. In addition, upon achieving a decisive coefficient R2, we may consider that the change in blood sugar concentration level can be mostly (by the rate of 60.72%) explained by the change in the concentration levels of glycosylated hemoglobin, glycerin trilaurate and total cholesterol of serum and insulin. P=0.00029 indicates the statistical meaning of preset multi-linear regression model, thus it is known that such a model has obvious statistical meaning.

5. APPLICATION OF LOGISTIC REGRESSION ANALYSIS METHOD

In this article, we analyzed the risk factors of Model-2 diabetes, proposed preventive measures related to community and highlighted the usage scope of logistic regression analysis method. For the purpose, we selected 327 cases of positive patients with Model-2 diabetes in three hospitals during the period from Feb 2014 to Sept 2014, and 327 negative patients after undergoing two continuous tests of urine sugar and blood sugar, as the study object. Then, the two groups have been divided into classified case group and contrast case group according to the new diagnosis standard for diabetes formulated by WHO in 1999: wherein, case group age (49.7±5.1 years old), sex ratio: male: female (172:155); contrast group age (50.5±6.1 years old), sex ratio: male: female (175:152), no obvious difference exists between two groups for study object, or statistical meaning exists. Logistic regression analysis for the study object should be done under logistic regression analysis method.

5.1. Result of One-Factor Analysis

This study selected 18 study objects, applied to non-conditional logistic regression analysis method to make quantization for each factor endowment; endowment method is shown in literature [9, 10], and analyzed for paroxysm relation between each factor and Model-2 diabetes; analysis result is shown in Table **2**.

The table above shows that 14 factors (such as educational degree and income status) have obvious relation with Model-2 diabetes (P<0.05) among the 18 study factors selected. Body mass index (BMI) has the most close relation with Model-2 diabetes. Analysis result is consistent with the clinical data.

5.2. Multi-Factor Analysis

After identifying the 18 factors possibly influencing Model-2 diabetes, apply for SIFT variable and analyze logistic gradual regression. Variable is sifted when α=0.05 is set as standard; analysis results are shown in Table **3**.

Table 2. Model-2 diabetes one-factor non-conditional logistic regression analysis result.

Study factor	B	OR	Study Factor	B	OR
Educational degree	-0.5214*	0.593	History of diabetes family	0.6325*	1.882
Income status	-0.3458*	0.704	Favor of dessert	0.8621**	2.368
History of smoking	0.2142*	1.230	Favor of meat	-0.0021	0.997
History of drinking	-0.0921	0.912	Favor of vegetable	-0.8914*	0.410
Body Mass Index（BMI）	1.5724**	4.818	Salt content of food	0.0251	1.025
Labor intensity	-0.0991	0.905	High glycosylated hemoglobin	1.0214*	2.770
Physical	-0.7124*	0.490	High glycerin trilaurate	0.8662**	2.377
History of high blood pressure	0.7012*	2.016	High total cholesterol of serum	0.2145*	1.239
History of coronary heart disease	0.5614*	1.753	Low insulin	-0.8612*	0.422

Note: "*" stands for P<0.05, "**"stands for P<0.01.

Table 3. Analysis result of Model-2 diabetes logistic multi-factor regression.

Variable	White tower	OR	95%CI
（BMI）Body mass index	1.5126	4.538	2.75-7.58
History of diabetes family	1.2721	3.568	1.42-6.65
Favor of dessert	0.9641	2.622	1.02-3.01
Physical exercise	-0.5142	0.597	0.49-0.89
History of high blood pressure	0.9614	2.615	1.01-5.03
Glycosylated hemoglobin	1.2131	3.363	0.56-1.79
High glycerin trilaurate	0.6218	1.862	1.71-3.52
Favor of vegetable	-0.1562	0.855388	1.84-3.26

As viewed from the aforesaid table, and after applying SIFT, gradual regression for 18 potential influential factors was calculated. Variables accessing to regression model include: BMI including BMI-c, family history of diabetes, favor of dessert, history of high blood pressure, high glycosylated hemoglobin and high glycerin trilaurate are high risk factors (OR>1); physical exercise and favor of vegetable are protective factors (OR<1).

CONCLUSION

In studies conducted in the field of modern science, there are a number of methods handling the relation between one dependent variable and multiple independent variables. This article analyzed and studied multi-linear regression analysis method and multi-linear logistics analysis method, and also analyzed their respective emphasis as well as difference between them. Principle and basic calculation process of multi-linear regression is the same as mono-linear one. However, relative complex calculation is required because of numerous of independent variables, and also because the calculation complexity for estimation rapidly rises along with adding up dependent variables. Initial studies established that multi-linear regression analysis manages to find out mathematical expression which almost always represents their relations even if there is no strict and definite functional relation existing between independent and dependent variables, by which this should make relatively accurate qualitative and quantitative analysis. Accordingly, multi-linear regression analysis is highly applicable in some fields of science such as medical science. Multi-linear regression model and generalized linear regression model require continuous positive distributive variables, and that independent and dependent variables have a linear relation. If the dependent variable is classified, independent and dependent variables are non-linear, the assumption of linear regression model is violated. At this point, logistic regression model is the best among all regression models, which does not require dependent variable distribution. From the mathematics point of view, logistic regression model should skillfully avoid classified variables distribution, and make up and correct some defects existing in the linear and generalized linear regression models. For medical study, logistic regression model can handle various practical issues and plays crucial roles in the development of medical sciences. For example, in an study of epidemiology, this condition usually meets when the dependent variable is discrete, such as, medical treatment as none-effective, obvious medical effect, recovery; survival or death of little white rat under different toxic agents; and paroxysm or none paroxysm under some exposure. Such issues should be handled with logistic regression analysis.

This article uses diabetes as the study object, and discusses with respect to the selection and correct use of methods and scope of multi-linear regression and logistic analyses. In fact, multi-linear regression analysis and logistic regression analysis are still applicable to clinic, industrial technology, aerospace and chemical engineering, and relative industries. Diabetes data study and analysis are one of the fields applicable to multi-linear regression and logistic analyses. Therefore, theoretical study of multi-linear regression and logistic analyses warrants relative scholars conducting further studies.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge the support from National Natural Science Foundation of PR China (51208056), Foundation of Loess Region Key Laboratory of Transportation Ministry (KLTLR-Y12-7) and Shanxi Transportation Science and Technology Foundation (11-2-24).

REFERENCES

[1] Z. Q. Sun, *Medical statistics*. Beijing: People's Medical Publishing House, 2006, pp. 313-333.

[2] Q. Q. Liu, *Study for judgment analysis and application of logistic regression model during distinguishing type of crude oil and fuel one*. Qingdao: Chinese Marine University, 2013.

[3] Z. Y. Sun, *Study for application of multi-linear regression analysis and Logistic one*. Nanjing: Nanjing University of Information Science and Technology, 2008.

[4] W. Wang, *Complete logistic regression analysis and evaluation by Excel.* Shenzhen: Jinan University, 2008.

[5] W. H. Xu and J. S.Fan, "Dangerous factor of Model-2 diabetes and Logistic regression analysis", *Reasonable pharmacy as clinic,* vol. 2, no. 21, pp. 8-9, 2009.

[6] T. Li and M. Li, "Correct use of multi-linear regression and logistic one", *Clinic collections*, vol. 24, no. 15, pp. 4-5, 2009.

[7] Y. Chen, L. Zhou, and Y. C. Xu, "Study for correlation with Model-2 diabetes and inheritance and environmental factor", *Chinese Journal of Preventive Medicine*, vol. 36, no. 3, pp. 191-194, 2002.

[8] C. L. Zhang and B. C. Wei, "Partial influence analysis and its application for purely distributive non-linear model", *Journal of applicable mathematics in colleges and universities edition A*, vol. 21, no. 2, pp. 148-156, 2006.

[9] R. A Johnson, D. W. Wiehem, and X. Y. Lu, *Practical multiply statistic analysis*. Beijing: Tsinghua University Press, 2001.

[10] H. X. Gao, *Applicable multiply statistic analysis*. Beijing: Beijing University press, 2005.

Received: August 25, 2014 Revised: November 29, 2014 Accepted: December 11, 2014