Health Scope. 2013 Summer; 2(2): 67-72. DOI: 10.17795/jhealthscope-8622

Published Online 2013 July 06. **Research Article**

Comparison Between Applicability of the Logit and Probit Models to Diagnose Influencing Risk Factors of Cardiovascular Disease in Mashhad

Habibollah Esmaily [1], Mohammad Taghi Shakeri [2], Zeynab Avazzadeh [3,*], Hasan Doosti [4], Majid Ghayour Mobarhan [5]

[1] Health Sciences Research Center, Department of Biostatistics and Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, IR Iran
[2] Department of Community Medicine, Medical School, Mashhad University of Medical Sciences, Mashhad, IR Iran
[3] Department of Biostatistics and Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, IR Iran
[4] Department of Mathematics and Department of Mathematical Statistics, Kharazmi University of Tehran, Tehran, IR Iran
[5] Cardiovascular Research Center, Department of New Sciences, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, IR Iran

**Corresponding author*: Zeynab Avazzadeh, Department of Biostatistics and Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, IR Iran. Tel.: +985117640326, Fax: +98-5413512297, E-mail: zeynab.evazzadeh@gmail.com.

Received: October 23, 2012; **Revised:** December 10, 2012; **Accepted:** December 12, 2012

Background: Cardiovascular disease is considered as one of the most common diseases, causing many deaths worldwide annually.
Objectives: The current study was aimed at diagnosing influencing factors of cardiovascular disease by comparison between implicational Logit and Probit models.
Patients and Methods: This work is based on a study group approved by Mashhad University of Medical Sciences with the code of 85134 (Mashhad Study). Information was collected by the stratified-cluster sampling method and registering more than 7603 people randomly from Mashhad city. From them 682 were patients with cardiovascular disease. Data was analyzed using statistical Logit and Probit models.
Results: After analyzing linearity and interactional effects between variables, by considering the other variables unchanged, the Odd Ratio of sex, age, high density lipoprotein-cholesterol (hdl-c), systolic blood pressure (sys-bp), body mass index (BMI), smoking, total cholesterol, High-Sensitivity C-Reactive Protein (hs-CRP) and history on cardiovascular disease were 4.231, 2.70, 0.98, 3.8, 3.421, 2.014, 1.5, 1.7 and 7.215 respectively.
Conclusions: Independent variables such as sex, age, hdl-c, sys-bp, BMI, smoking, cholesterol and hs-CRP were considered as influencing variables, and education, marital status, disease history, height, waist circumference, diastolic blood pressure and low density lipoprotein were removed from the models.

Keywords: Logistic Models; Cardiovascular Diseases; Risk Factors

1. Background

Cardiovascular diseases are affected by several factors such as economic, social and cultural factors which can lead to morbidity and mortality; therefore, it is essential to recognize risk factors of these in each region (1). Cardiovascular Disease is a slow and complicated disease which starts from childhood and often with increasing age it progresses in people. This progress is faster for some people in third decade of their life. This disease has three important risk factors: increased cholesterol and triglycerides levels in blood, high blood pressure, smoking, and some others (2). Actually risk factors are defined as those factors that their presence increases the probability of starting disease in the future (3). Particularly, new biomarkers such as hs-CRP as the risk factors have been supposed in recent years and many researches have been focused on them. Some various ways have been investigated to determine the risk factors of the disease in some scientific literatures particularly Logit and Probit models. Which of the two models can better fit the data in a case study (3)? Logit and Probit models are the most widely used due to their nature. These models are very useful in medical studies. The purpose of this study was to compare Logit and Probit models by using a set of medical data. This work is based on a study group approved by Mashhad University of Medical Sciences with the code of 85134 (Mashhad Study). 7603 people were recruited randomly from Mashhad city, and information was collected by the stratified-cluster sampling method (4). Age, sex, occupation, smoking, history of disease, systolic blood pressure, diastolic blood pressure, BMI,

Implication for health policy/practice/research/medical education:
The current study was aimed at diagnosis of influencing factors of cardiovascular disease infection. Cardiovascular diseases are considered to be one of the most common diseases, causing death of many people worldwide annually. In order to reduce the prevalence of cardiovascular disease, more adherence to offers of this study are recommended. This project will help to Schedule policy regarding of influencing factors of cardiovascular disease of society.

hdl-c, low density lipoprotein-cholesterol (ldl-c), total cholesterol (TC), and hs-CRP are the risk factors of cardiovascular disease and also they are significant in both models. We could see that in Logit model estimated model deviance was equal to DL = 443.4289, and in Probit model estimated model deviance was equal to DP = 523.4293, and there is not a significant difference between them. So, none of them have advantage over each other, statistically (5).

2. Objectives

The purpose of this survey was to identify cardiovascular disease risk factors, estimating Logit and Probit coefficient and comparing these two models with each other.

3. Patients and Methods

Subjects: Mashhad city residents who are between 35 and 65years old. Information was collected by the stratified-cluster sampling method, and registering more than 7603 people randomly from Mashhad city. From them 682 were patients with cardiovascular disease.

3.1. Sampling Method

This study was based on a cohort survey approved by Mashhad University of Medical Science with the code of 85134 (Mashhad Study). This study was conducted on 7600 people who were selected randomly from Mashhad urban population and gathered with stratified-cluster sampling method. This information was gathered by going to the door of people whom were studied. These people were justified about the plan by the statistical agent of this study face to face and in the case of desiring to participate in this plan, a written consent was obtained. In this stage, the family list including residence address, cluster number, telephone number, age, and the number of people living in that house were recorded, and then a complete map of selected cluster was provided. Then, an invitation for entering the study was submitted to them and the time for visiting and filling questionnaire, and implementing required examination and experiments was specified.

3.2. Sample Size

Mashhad Study samples who were approximately 7600 people participated in this study. In first observations we saw that 682 people had cardiovascular disease which was enough for achieving research goals. Maximally two persons with opposed sex from each family whose age where more than 35 and lower than 65 were recruited from the study population. If the number of one gender was more than one, only one of them was assigned for the study randomly.

3.3. Plan Methodology

We identified important risk factors by reviewing the available literature. Then we studied Logit and Probit models and the estimating way of their parameters with different methods. We criticized and investigated these two methods theoretically, and we had in mind other methods that we could use for analyzing the data. Then we investigated Mashhad Study data and analyzed required variables with R software. Before estimation of the model, it was necessary to examine the linear independent variables. To check the linearity, we used two indices; tolerance is an indicator which shows how much the variability of the independent variable that we mean explained by the other independent variables in the model (4). In this study we investigated both quantitative and qualitative variables. Age, height, weight, BMI, Waist, systolic blood pressure, diastolic blood pressure, hdl-c, ldl-c, TG, total cholesterol (TC), hs-CRP were quantitative variables. Sex, occupation, occupation, Marital status, smoking, history disease were qualitative variables.

3.4. Sensitivity and Specificity

To evaluate and compare the variables that have been categorized in two states such as heart disease variable (being sick or not being sick), we used two indices, sensitivity and specificity. When we can divide data to positive and negative groups, accuracy of the results can be explained by using sensitivity and specificity indices. Sensitivity means the probability of a positive test among those who have the disease. Specificity means the probability of a negative test among those who do not have the disease.

True positive: the patient is correctly diagnosed.

False positive: the healthy person wrongly diagnosed as a patient.

True negative: the healthy person correctly diagnosed.

False negative: the patient wrongly diagnosed as healthy.

Sensitivity and specificity in a test depend on its nature and the test sample. However, the result of a test cannot be interpreted just with the sensitivity and specificity (5). For example, if the result of a blood test becomes positive and the test has 90% sensitivity and 96% specify, physician is not able to determine to what extend the patient is truly infected. For this purpose, we should use positive predictive value (or NPV if the test result is negative). Predictive value of a test depends on the prevalence of tested phenomena in the statistical population rather than the nature of the test and examples. Low values of the index (less than 0.1) for each variable represent multiple linearity with some other variables. Index variance inflation (VIF) is reverse to the tolerance (6). VIF values more than 10 indicate the linearity. The results of linearity are shown in Table 1. According to the Table 1, variables such as height, weight and body mass index had linearity problem so to solve this problem, it was necessary to remove one of these three variables. We examined the linearity among variables again with removing the weight variable.

Table 1. Logit Model Results

Coefficients	Odds Ratio	Confidence Interval of the Odds Ratio	
		Lower Bound	Upper Bound
Intercept	0.00		
Gender, male	4.23	3.99	4.52
Age	2.70	2.10	3.10
High-density lipoprotein	0.98	0.97	0.99
Systolic blood pressure	3.80	3.20	4.12
BMI	3.42	3.00	3.62
Smoking	2.01	1.83	2.23
Cholesterol	1.50	1.00	1.61
Reactive protein hs-CRP	1.70	1.10	2.10
Disease	7.21	7.00	7.60

3.5. Experimental Results

According to the Table 1, it can be seen that the linearity problem was solved by the removal of weight variable. Afterward sensitivity and specificity were %99 and %98 respectively. According to the Table 2 and the area under the ROC curve (0.745) in the Figure 1, provided classification scheme is accepted.

The Table 2 indicates that gender (sex), age, high-density lipoprotein (hdl-c), systolic blood pressure (sys-bp), body mass index (BMI), smoking, cholesterol (TC), hs-CRP have significant impacts on the risk of cardiovascular disease. Education, marital status, past medical history, height, waist circumference, diastolic blood pressure and lipoprotein with low density did not have any significant impact on cardiovascular disease. The hdl-c variable with negative coefficient indicated that increasing high density lipoprotein decreases the risk of cardiovascular disease. But BMI variable with positive coefficient showed that increasing BMI can increase the risk of cardiovascular disease.

Table 2. Results of Multicollinearity

Variables	Tolerance	VIF[a]
Gender	0.326	3.063
Age	0.665	1.504
Education	0.823	1.215
Job	0.739	1.354
Marital status	0.929	1.076
Stature	0.026	38.482
Weight	0.011	91.615
Waist	0.362	2.759
low-density lipoprotein	0.169	5.912
Systolic blood pressure	0.363	2.758
diastolic blood pressure	0.390	2.561
BMI	0.012	81.503
Disease	0.802	1.248
Smoking	0.691	1.447
High-density lipoprotein	0.762	1.312
Cholesterol	0.152	6.578
Reactive Protein	0.965	1.036

[a] Abbreviation: VIF, variance inflation

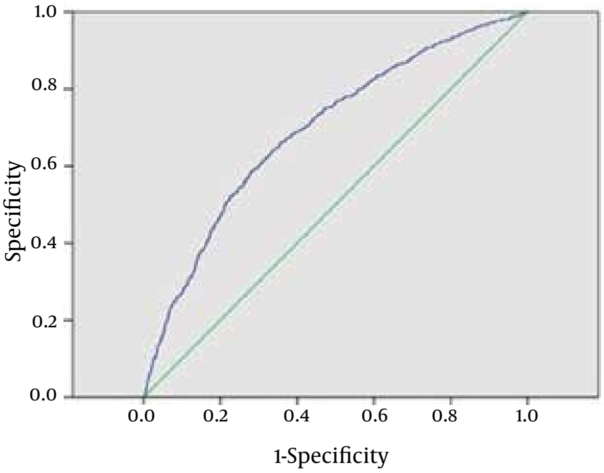

Figure 1. The ROC Curve

3.6. Estimating Logit Model to Identify Cardiovascular Disease Risk Factors in Mashhad City

Logit model was used to examine the impact of age, gender (sex), occupation, education, marital status, height, waist circumference, smoking, history of disease, systolic blood pressure, diastolic blood pressure, body mass index (BMI), hdl-c, ldl-c, and cholesterol (TC) and hs-CRP on cardiovascular disease. We should examine sensitivity and specificity indices and ROC curve to examine the model suitability (Figure 1, Table 2).

As mentioned earlier, interpretation of the chance ratio is as follows:

A. Values more than one, indicate a greater chance of success than failure.

B. Values less than one, indicate a lower chance of success than failure.

According to Logit model we can conclude that risk of heart disease is 2.7 times higher in older people compared to the younger. However, increasing high density lipoprotein reduces the risk of diabetes

3.7. Probit Model Estimation to Identify Factors for Cardiovascular Disease in Mashhad City

Multivariate Probit model was used to examine, the impact of age, sex, occupation, smoking, history of disease, systolic blood pressure, diastolic blood pressure, BMI, hdl-c, ldl-c, cholesterol (TC) and hs-CRP on developing cardiovascular disease. The results are presented in the Table 3.

Table 3. The Results of the Probit Model

Coefficients	Odds Ratio	Confidence Interval of the Odds Ratio	
		Lower Bound	Upper Bound
Intercep	3.083	3.02	3.12
Gender, male	4.217	4.002	4.307
Age	2.29	2.040	2.34
High-density lipoprotein	0.014	0.009	0.03
Systolic blood pressure	3.2	3.07	3.3
BMI	2.1	2.080	2.21
Smoking	2.015	1.8	2.03
Cholesterol	1.1	1.08	1.2
Reactive protein hs-CRP	1.4	1.08	1.6
Disease	4.034	3.28	5.81

Education, marital status, past medical history, height, waist circumference, diastolic blood pressure and lipoprotein with low density did not have any significant impact on cardiovascular disease. Considering the Table 3, if age variable increases one unit, the risk of heart disease would increase, too. So we can conclude that the risk of heart disease in people who are older compared to those who are younger, is 2.29 times higher.

3.8. Logit and Probit Comparing Model to Identify Risk Factors for Cardiovascular Disease in Mashhad City

In two previous sections, we examined cardiovascular disease risk factors with Logit and Probit models. Table 3 indicates the similarity of the both models results. Then, we examined deviance between the two models to compare them. We could see that in Logit model estimated model deviance was equal to D L = 443.4289 and in Probit model estimated model deviance was equal to D P = 523.4293, therefore there was not a significant difference between them. So, none of them have advantage over each other, statistically.

4. Results

In many studies, the response variable is measured by two states index, for example, death or life, being patient or healthy, success or failure of a treatment and some others. Therefore, we could not use the simple linear model (ordinary regression or analysis of variance) and we should find models that do not have requirement of normal distribution and consistency for being patient with cardiovascular disease. In this condition between dependent and independent variable (s), there is not a linear association and only through various transformations the relation could be linear so we could explain this through generalized linear models. Logit and Probit models are two important generalized linear models which have been compared in this study (6). In this study, we compared the application of Logit and Probit models in analysis of the cardiovascular risk factors of heart disease (7). Our review was made on 7603 individuals, in which approximately %42.5 had history of diseases, %25 were smoker, and %9 had history of cardiovascular disease. To achieve Logit and Probit models with the help of R software independent variables such as sex, age, hdl-c, sys-bp, BMI, smoking, cholesterol and hs-CRP were considered as influencing variable and education, marital status, disease history, height, waist circumference, diastolic blood pressure, low density lipoprotein were removed from the model. If other variables are considered to be unchanged, the effect of blood pressure on the risk of coronary heart disease is 0.68, that has more effects on the dependent variable (risk of cardiovascular disease) compared with other variables. The hdl-c variable has reverse effect on the risk of coronary heart disease, it means that if the hdl-c increases one unit, the risk of coronary heart disease decreases. We examined the deviance of the two models to compare them, deviation of Logit model was DL = 4289.443, and deviation of Probit model was DP = 4293.523. We see that the values do not differ very considerably. So, none of the models have advantages over each other, statistically (8-10). Therefore, the criterion for selecting one of the two models depends on the ease of applying computer programs and mathematical operation; however, Logit model is generally preferred. In a study performed in Jahrom, factors affecting low hearing of 152 patients were investigated. Using a probit model, association between the response variable (infection or disease) and the independent variables (the noise in environment, family history of diabetes, hypertension & hyperlipidemia, dizziness and tinnitus) were studied. They found that environment noise, family history, and tinnitus variables influence on low hearing. Also using logit model lead to the same result and there was not much differences between estimated values in both models. In Gorgan, a study was performed to determine the factors influencing work wearing and its influencing factors among nurses in Panjeazar health and education center. From 192 individuals who were observed, 47 persons (%24.5) had work wearing. They used logit and probit models to find association between work wearing and sex, age, marital status working hours per week. When they investigated work wearing with probit model they found marital status as an influencing factor. Work wearing for married nurses is higher than single ones and this result is compatible with logit model findings. We could conclude some points according to the result of this essay: cardiovascular disease risk is higher for men than women. Also, the risk of heart disease is 2.7 times higher in people who are older compared to those who are younger. Cardiovascular disease risk for obese people and also patients with high blood pressure is higher than the risk for other people. There is a direct link between high blood pressure and heart disease. People with blood pressure higher than 140 and 90 have hypertension and they are at the risk of heart disease (11-14). The risk of heart disease for obese people is 3.42 times more than skinny (15). People with high blood pressure have this disease 3.8 times more than normal population. Smoking is another risk factor for heart disease; smokers are more prone to develop heart disease and the risk for smokers is 2.014 times higher than the risk for nonsmokers. Increasing cholesterol level directly increases the risk of cardiovascular disease and heart disease. The risk of disease for people with high cholesterol is 1.1 times more than the risk for normal population.

5. Discussion

Keeping the blood pressure between 80 - 120 and even lower is of great importance. People who are overweight should reduce weight slowly by using medically approved methods. Doing exercises regularly, eating fiber foods and cereal rather than fats is recommended. Eating fiber foods such as vegetables and fruits could decrease blood cholesterol level. Increasing HDL and decreasing LDL to lower than 130 is of great importance. Changing the lifestyle like lowering salt intake, avoid stress and being upset could control the blood pressure. Quit smoking is never late, so smokers should quit smoking for preventing cardio-vascular disease, and even they should avoid polluted weather.

Acknowledgements

The authors would like to thank Mashhad Medicine University for granting permission to obtain relevant data for this study from their registered files.

Authors' Contribution

The authors would like to thank Mashhad Medicine University for granting permission to obtain relevant data for this study from their registered files.

Financial Disclosure

The overall implementation of this study including data extraction and analysis, report writing and manuscript preparation were the result of joint efforts by multiple individuals who are listed as coauthors of this paper.

Funding/Support

There is no financial disclosure.

References

1. Nadimi M. [Healthy Heart]. *Rahnama*. 2008.
2. Newman Dorland WA. Dorland's Illustrated Medical Dictionary. 32 ed.: Elsevier Saunders; 2011.
3. Mausner J, Kramer S. Epidemiology: An introductory text. 2 ed.: W.B. Saunders Co; 1985.
4. Eltzschig HK, Collard CD. Vascular ischaemia and reperfusion injury. *Br Med Bull*. 2004;**70**:71-86.
5. Setodji CM, Scheuner M, Pankow JS, Blumenthal RS, Chen H, Keeler E. A graphical method for assessing risk factor threshold values using the generalized additive model: the multi-ethnic study of atherosclerosis. *Health Serv Outcomes Res Methodol*. 2012;**12**(1):62-79.
6. Ayatollahi SM, Poorahmad S, Vakili M, Heydari T. Probit models & its application in medical data. *Andishe-ye Amari*. 2005;**10**:36-46.
7. Anastasopoulos PCh, Mannering FL. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. *Acc Anaysis Prevent*. 2011;**43**:1140-1147.
8. Litchfield J, Reilly B, Veneziani M. An analysis of life satisfaction in Albania: An heteroscedastic ordered probit model approach. *J Econ Behav Organ*. 2011;**81**(3).
9. Agresti A. An Introduction to Categorical Data Analysis. 2 ed.: John Wiley & Sons; 2007.
10. Slejko JF, Page RL, 2nd, Sullivan PW. Cost-effectiveness of statin therapy for vascular event prevention in adults with elevated C-reactive protein: implications of JUPITER. *Curr Med Res Opin*. 2010;**26**(10):2485-97.
11. Patil S, Geedipally SR, Lord D. Analysis of crash severities using nested logit model--accounting for the underreporting of crashes. *Accid Anal Prev*. 2012;**45**:646-53.
12. Sarmad Z, Bazargan A, Hejazi E. Research Methods in Behavioral Sciences. 13 ed. Tehran; 2007.
13. Kalantari M. Glossary of Statistics Research Methods and Psychometrics, Farhange Sabz, Tehran, . 2010.
14. Montgomery DC. Design and Analysis of Experiments. 7 ed.: Wiley; 2009.
15. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. John Wiley & Sons; 2012.