ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 4 No. 1 (2024) 205-216
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Analysis of Cardiovascular Disease Risk Factors in the United States using Logistic and Probit Regression

Teo Li Sha[1], Norziha Che Him[1]*, Mohd Saifullah Rusiman[1], Yusliandy Yusof[2]

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology*
UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA

*Corresponding Author: norziha@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2024.04.01.023

Article Info

Received: 27 December 2023
Accepted: 3 June 2024
Available online: 27 July 2024

Keywords

Cardiovascular Disease, Risk Factors, Binary Logistic Regression, Probit Regression, United States

Abstract

Cardiovascular Disease (CVD) is a non-communication disease that remains the most life-threatening disease worldwide, including the United States. Many research studies applied various statistical methods to identify the significant risk factors of CVD. This study aims to determine the significant risk factors of CVD among the residents in the United States in 2021 by constructing the Binary Logistic Regression model and Probit Regression model. Additionally, the performance of the logit model and probit model are compared by using Deviance, AIC and BIC. The dataset is collected from the Behavioural Risk Factor Surveillance System (BRFSS). The top five mortality rates in CVD of the U.S. states are chosen to represent the U.S. population, with a total sample size of 24932 respondents. The findings reveal that the logit model and probit model produced the same results, which the significant risk factors of CVD are male gender, age group from 6 to 13 (Age 45 and above), current smokers, non-heavy drinkers, underweight, high blood pressure, high cholesterol, and diabetes. These risk factors showed an increased probability of developing CVD among the residents in the U.S. Lastly, this study indicated that the probit model performed better than the logit model as it provides a slightly lower values of Deviance (11318.58), AIC (11350.58), and BIC (11475.47) compared to the logit model with Deviance (11330.19), AIC (11362.19), and BIC (11487.08).

1. Introduction

Cardiovascular disease (CVD) is a non-communicable disease that involves of heart and blood vessels. CVD encompasses coronary heart disease (CHD), cerebrovascular disease, rheumatic heart disease, myocardial infarction, stroke, and other heart disease [1]. The most life-threatening disease in the United States (U.S.) in 2021 was heart disease which recorded 695,547 deaths [2]. The top five CVD mortality rates of the states in the U.S. were Oklahoma (264.2), Mississippi (255.2), Alabama (247.5), Louisiana (235.5), and Arkansas (231) [3].

The growing epidemic of CVD is associated with various risk factors [1]. These risk factors can be classified into two groups which are modifiable factors and non-modifiable factors. Modifiable factors are the factors that can be controlled, while non-modifiable factors are the factors that cannot to be controlled [4][5]. In this study, the modifiable factors such as smoking status, blood pressure, cholesterol levels, alcohol are used, whereas non-modifiable factors such as gender and age are used for determining the CVD.

Binary logistic regression is widely used in the medical field for risk assessment of complex disease [6]. It is a method to determine the relationship between predictors and dichotomous response variable with two categories [7]. Based on the logit model, the changes in logit of the respondent variable is calculated [8]. In binary logistic regression output, the odd ratio for a risk factor contributing to a clinical result can be interpreted as indicating whether a person with a risk factor is more or less likely than a person without that risk factor to experience the event of interest [9]. Hence, by constructing binary logistic regression model and calculating the odd ratio, the risk factors of cardiovascular disease and its relationship can be determined in this study.

Probit regression is also used to identify the association between predictors and a binary outcome variable [10]. Probit model is more adequate to be used when sample size is large due to the Central Limit Theorem [11]. The coefficient of probit model is the effect on z of one-unit change in regressor when others are holding as constant [12]. In probit regression model, the marginal probability effect of a predictor is used to calculate the proportion of response variable either increase or deceases when predictor increases in value by one unit [11]. In this study, probit regression and marginal probability effect are applied. Other similar previous studies were done in using forecasting method with the statistics technique in worldwide according to various fields of studies [13, 14, 15, 16].

The purpose of the study is to determine the significant risk factors of cardiovascular disease among the residents in the United States in 2021. The first objective is to explore the relationship between risk factors and cardiovascular disease in the United States in 2021 by using descriptive analysis. Besides, the second objective is to construct binary logistic regression model and probit regression model in determining the significant risk factors of cardiovascular disease. After that, the performance of each method is compared to choose the best model by using Deviance, Akaike's information criteria (AIC), and Bayesian information criteria (BIC).

2. Materials and Methods

2.1 Data sources and data set

In this study, the data was obtained from Behavioural Risk Factor Surveillance System (BRFSS) official websites in 2021. Table 1 shows that the data description of this research.

Table 1 *Description of Variables*

Variable Category	Variable Name	Description	Data Type
Demographic	Gender	Sex of Respondent	Nominal
		1: Male	
		2 Female	
	Agecat_5years	13-level age category	Ordinal
		1: 18 ≤ Age ≤ 24	
		2: 25 ≤ Age ≤ 29	
		3: 30 ≤ Age ≤ 34	
		4: 35 ≤ Age ≤ 39	
		5: 40 ≤ Age ≤ 44	
		6: 45 ≤ Age ≤ 49	
		7: 50 ≤ Age ≤ 54	
		8: 55 ≤ Age ≤ 59	
		9: 60 ≤ Age ≤ 64	
		10: 65 ≤ Age ≤ 69	
		11: 70 ≤ Age ≤ 74	
		12: 75 ≤ Age ≤ 79	
		13: Age 80 or older	
Behaviours Factors	Smoking Status	Adults who are current smokers	Nominal
		0: No	
		1: Yes	
	Alcoholic Consumption	Heavy drinkers	Nominal

		0: No 1: Yes	
Medical Status	BMI	Body Mass Index	Continuous
	Blood Pressure	Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional 0: No 1: Yes	Nominal
	Cholesterol	Have you ever been told by a doctor, nurse, or other health professional that your cholesterol is high? 0: No 1: Yes	Nominal
Medical History	Diabetes	Ever told you have diabetes? 0: No 1: Yes	Nominal
Dependent Variable	Cardiovascular Disease (CVD)	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) 0: No 1: Yes	Binary

2.2 Dealing with Missing Data

In this study, the missing data is dealing by complete-case analysis, known as listwise deletion method will be used. It is a method of handling missing data by dropping the incomplete observations.

2.3 Descriptive Analysis

In descriptive analysis, all the variables are displayed in graphs. To have a clear image in analysis the relationship between BMI and CVD, the BMI value is categorized according the BMI values as shown as Table 2.

Table 2 *BMI classification*

BMI values	BMI categories
Less than 18.5 kg/m^2	Underweight
18.5 kg/m^2 to 24.9 kg/m^2	Normal weight
25.0 kg/m^2 to 29.9 kg/m^2	Overweight
More than 30.0 kg/m^2	Obese

Pie chart used for categorical data with less than five categories. Besides, bar chart is used for analysis with categorical data more than five level such as age variable [17].

2.4 Binary Logistic Regression

Binary logistic regression is a method to determine the relationship between independent variables and a dichotomous dependent variable with two categories. A logistic regression model uses the logit link (log odds) function to construct the relationship between the probability of an event of interest and a linear combination of independent variables. The general equation of binary logistic regression as shown in equation 1 [18].

$$\text{logit(p)} = \ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_i X_i \quad (1)$$

Once a logit model is formed, the odds ratio (OR) can be obtained by exponentiating model coefficients [19]. The odds ratios are used to measure the strength of the association between risk factors and outcomes, and it is related to the probability of a binary outcome [9]. The interpretation of OR was shown as Table 3 [20]. Hence, it is appropriate to predict whether CVD is present or absent based on the significant risk factors.

Table 3 *Description of Odds Ratio*

Odds Ratio	Description
OR < 1	Exposure associated with lower odds of outcome
OR = 1	Exposure does not affect odds of outcome
OR > 1	Exposure associated with higher odds of outcome

In logistic regression, Wald Test is used for testing the significance of each independent variable separately [21]. Moreover, likelihood ratio test is used to evaluate the significance of adding or dropping, a predictor or group of predictors from model [19]. For Pseudo R-squared including Cox & Snell R-Square and Nagelkerke R-squared were also used to calculate variation ratio for the dependent variable that can be explained by the model [7]. After constructing the models, the models of the logistic regression are evaluated using Hosmer and Lemeshow Test. The hypothesis of this test as shown as below [18]:

H_0 : There are no significant differences between observed and expected values.

H_1 : There are significant differences between observed and expected values.

2.5 Probit Regression

Probit regression is an alternative of logistic regression method used to determine the relationship between predictors and a binary dependent variable. The general probit model can be expressed as equation 2 [22].

$$\pi_i = \Phi\left(Z_i\right) = \Phi\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon_k\right) \tag{2}$$

Once a probit model is formed, the marginal probability effect of an independent variable is estimated as expressed as equation 3 [11].

$$\frac{dE\left[y \mid x\right]}{dx_i} = \frac{dp\left[y = 1 \mid x\right]}{dx_i} = \beta_i \Phi^{'}\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\right) \tag{3}$$

2.6 Model Selection

By chosen most adequate model, the researcher should conduct the goodness-of-fit test. The goodness of fit test describes how well the model fits the data. Deviance is also one of the goodness-of-fit tests for GLM, and the equation for deviance is as shown as equation 4 [23].

$$Deviance = 2(L_S - L_m) \tag{4}$$

In addition, Akaike's information criteria (AIC) and Bayesian information criterion (BIC) also are criteria of model selection. The formula for both were shown in equation 5 and equation 6, respectively [24].

$$AIC = -2\ln(L) + 2k \tag{5}$$

$$BIC = -2\ln(L) + k\ln(n) \tag{6}$$

3. Results and Discussion

3.1 Data Selection and Data Cleaning

The data was collected from BRFSS official website in 2021, which consist of 438693 observations. First, ten variables (State, Gender, Age, Smoke, Alcohol, BMI, Blood Pressure, Cholesterol, Diabetes, CVD) were chosen. Then, five U.S. states were selected to represent the U.S, which were the top five mortality rates due to CVD by states in the U.S. in 2021. The states were Oklahoma, Mississippi, Alabama, Louisiana, and Arkansas. After that,

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 1 (2024) p. 205-216

the data was dealing with the missing values, by using listwise deletion method. In R, `na.omit` function was used to remove missing values. Then, the observation becomes 18136.

3.2 Descriptive Analysis

In descriptive analysis, the analysis is undergoing into two sections.

3.2.1 Summary of Overall Data

The overall descriptive analysis has been investigated for fundamental understanding on the characteristics of the residents and the variables. The main focus on CVD is determined using pie chart.

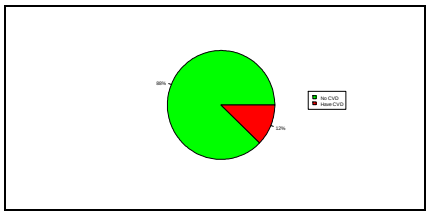

Fig. 1 *CVD variable*

Based on the pie chart in Fig. 1, the respondent that have not ever reported having coronary heart disease (CHD) and myocardial infraction (MI), known as CVD is recorded 15910 people. Only 12.27% of respondent indicated that they have ever reported having CVD.

3.2.2 Summary of Data with the Observations of Having CVD

For understanding on the relationship between risk factors and CVD, the residents that have ever reported having CHD or MI, namely CVD, and each of the risk factors have been determined.

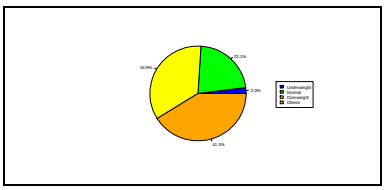

Fig. 2 *Body Mass Index (BMI) categories variable in CVD status*

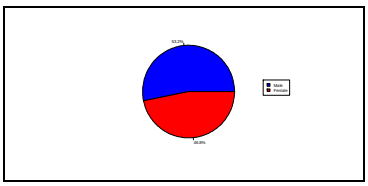

Fig. 3 Gender variable in CVD status

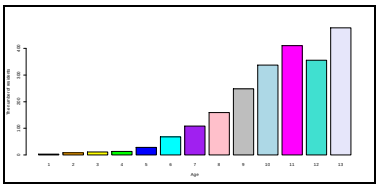

Fig. 4 *Age variable in CVD status*

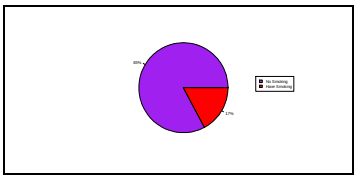

Fig. 5 *Smoking variable in CVD status*

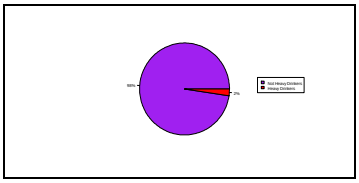

Fig. 6 *Alcohol variable in CVD status*

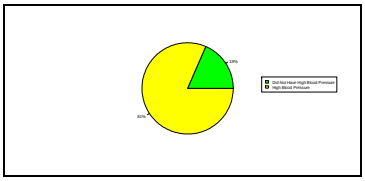

Fig. 7 *Blood Pressure variable in CVD status*

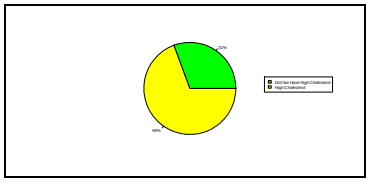

Fig. 8 *Cholesterol variable in CVD status*

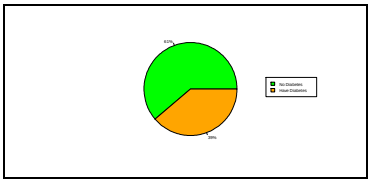

Fig. 9 *Diabetes variable in CVD status*

Based on Fig 2 to Fig 9, most residents with CVD are obese which recorded 41.11% of residents. Besides, more male has ever reported having CVD. Moreover, most of the respondents is from age group 13 (age 80 or older). Additionally, non-current smoking (83%) and not heavy drinkers (98%) recorded more percentage of residents with CVD. For the residents who have ever been told by doctor, nurse, or other health professional that they have high blood pressure (81%) and high cholesterol (69%) recorded more percentage of present CVD. However, more residents (61.23%) have not ever been told they have diabetes compare to having diabetes.

3.3 Binary Logistic Regression

Before constructing the binary logistic regression model, the dummy variables and references variables should be defined for the categorical variables. To build the logistic regression model, backward elimination method is used. The variable with the highest *p*-value which greater than 0.05 is removed from the model. This indicates that variables are not statistically significant to the model. The prosses will procced until all the variables are statistically significant to the model. Seven logit models are built and shown as Table 4.

Table 4 *Variables in Binary Logistic Regression Models*

Model	Variables in Model
1	female gender, age group (2-13), smoking, alcohol, BMI categories (underweight, overweight, and obese), blood pressure, cholesterol, and diabetes
2	female gender, age group (2-13), smoking, alcohol, BMI categories (underweight, and obese), blood pressure, cholesterol, and diabetes
3	female gender, age group (2, 3, 5-13), smoking, alcohol, BMI categories (underweight, and obese), blood pressure, cholesterol, and diabetes
4	female gender, age group (2, 3, 5-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes
5	female gender, age group (2, 5-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes
6	female gender, age group (5-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes
7	female gender, age group (6-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes

By using likelihood Test, the full model will be compared to reduced model is as shown as Table 5.

Table 5 *Likelihood Test*

Full Model	Reduced Model	*p*-value
1	2, 3, 4, 5, 6, and 7	> 0.1
2	3, 4, 5, 6, and 7	> 0.1
3	4, 5, 6, and 7	> 0.1
4	5, 6, and 7	> 0.1
5	6, and 7	> 0.1
6	7	> 0.05

Based on the Table 5, all the models exhibit the *p*-value greater than 0.05, that means all the reduced model is much better compared to full model. Hence, Model 7 is the best model since it showed all variables are significant and its likelihood test is greater than 0.05. Moreover, all the models are compared using Pseudo R-squared as shown as Table 6.

Table 6 *Pseudo R-squared*

Model	Cox & Snell R-squared	Nagelkerke R-squared
1	0.113345	0.215846
2	0.113345	0.215845
3	0.113322	0.215802

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 1 (2024) p. 205-216

4	0.113297	0.215754
5	0.113252	0.215669
6	0.113208	0.215584
7	0.113044	0.215273

Based on Table 6, all the models have the closely similar Pseudo R-squared values suggest that there is no significant difference in the proportion of variance explained by these models. To choose the best model for binary logistic regression model, four goodness of fit test are applied as shown in Table 7.

Table 7 *Goodness of Fit Test*

Model	Deviance	AIC	BIC	Hosmer-Lemeshow Test
1	**11324.04**	11368.04	11539.77	< 0.01
2	11324.05	11366.05	11529.97	< 0.01
3	11324.51	11364.51	11520.62	< 0.01
4	11325.02	11363.02	11511.33	< 0.05
5	11325.94	11361.94	11502.44	< 0.05
6	11326.85	**11360.85**	11493.54	< 0.05
7	11330.19	11362.19	**11487.08**	**> 0.05**

Based on Table 7, Model 1 has the lowest deviance value (11324.04), which means the model fit well. Besides, Model 6 have the lowest AIC value (11360.85), indicating the model captures the information in the data with the least loss. Moreover, Model 7 have the lowest BIC value (11487.08), meaning the model fit well. Additionally, the p-value of Hosmer-Lemeshow Test for Model 7 is 0.05442, which is greater than 0.05 indicate that there are no significant differences between observed and expected values. Based on Goodness of Fit Test, Model 7 perform well as it fulfilled two out of four criteria and the variables in Model 7 are statistically significant to the model. Therefore, Model 7 is selected as final model as shown as Table 8.

Table 8 *The Final Model for Binary Logistic Regression with Coefficient and Odds Ratio*

Variables	Coefficient Estimate	Standard Error	Wald	*p*-value	Odds Ratio
Intercept	-4.47588	0.13492		**< 0.001**	0.01138025
Gender (Female)	-0.62159	0.04885	161.8995	**< 0.001**	0.532708995
Age 6	0.97395	0.18135	28.8430	**< 0.001**	2.64837990
Age 7	1.20567	0.16525	53.2326	**< 0.001**	3.33898094
Age 8	1.25742	0.15553	65.3626	**< 0.001**	3.51632154
Age 9	1.53421	0.14782	107.7261	**< 0.001**	4.63767174
Age 10	1.75797	0.14414	148.7451	**< 0.001**	5.80062631
Age 11	1.97528	0.14266	191.7095	**< 0.001**	7.20865416
Age 12	2.26976	0.14522	244.3036	**< 0.001**	9.67704939
Age 13	2.53783	0.14243	317.5073	**< 0.001**	12.65222606
Smoking (1)	0.59251	0.06757	76.8885	**< 0.001**	1.80853009
Alcohol (1)	-0.53197	0.15139	12.3475	**< 0.001**	0.58744889
BMI (Underweight)	0.37156	0.18114	4.2074	**< 0.05**	1.44999966
Blood Pressure (1)	0.83631	0.06125	186.4061	**< 0.001**	2.30783024
Cholesterol (1)	0.66661	0.05240	161.8342	**< 0.001**	1.94762271
Diabetes (1)	0.63486	0.05230	141.3679	**< 0.001**	1.88676609

The final model for binary logistic regression is as shown as equation 7:

$$\ln\left[\frac{p}{1-p}\right] = -4.476 - 6.216\text{Gender\_female} + 0.974\text{age6} + 1.206\text{age7} + 1.257\text{age8} + 1.534\text{age9} + 1.758\text{age10} + 1.975\text{age11} + 2.270\text{age12} + 2.538\text{age13} + 0.593\text{Smoking\_1} - 0.532\text{Alcohol\_1} + 0.372\text{BMI\_Under} + 0.836\text{BloodPressure\_1} + 0.667\text{Cholesterol\_1} + 0.635\text{Diabetes\_1} \quad (7)$$

Based on Table 8, the estimated coefficient of female gender variable is -0.62159, meaning that the log odds of the CVD for female residents is lower compared to male gender. Besides, the odd ratio value for the female variables is 0.5327, meaning that the odd of having CVD for female residents are half the odds for male residents. For age categories, the reference category is age 1 (Age 18 to 24). Eight dummy age variables with positive coefficients, indicating that when the age group increase, the log-odds of CVD also increase. For the smoking variable with positive coefficient estimated value (0.59251), indicating the log-odds change in CVD for current smokers compared to non-current smokers. Besides, based on the odd ratio (1.8085), indicating current smoker have 80.85% larger odds of CVD present than non-current smokers. For alcohol, the coefficient estimated value is -0.53197, which indicates the log-odds change in CVD for heavy drinks compared to non-heavy drinkers.

Additionally, the BMI (underweight) variable with the coefficient (0.37156) illustrated that the log-odds change in CVD for underweight residents compared to those in normal weight. The odds ratio of this variable (1.4500) indicating the underweight residents have 45.0% greater odds to develop CVD than residents with normal weight. For the blood pressure (1) variable, the positive coefficient estimated value (0.83631) indicating there is an increase in the log-odds of CVD among the residents who have been told by a doctor, nurse, or other professional having high blood pressure compared to residents who have not been told by a doctor, nurse, or other professional having high blood pressure. For the cholesterol (1) variable, the odd ratio (1.9476), indicating residents who have been told by a doctor, nurse, or other professional having high cholesterol have 94.76% higher odds of having CVD compared to the residents who have not been told by a doctor, nurse, or other professional having high cholesterol. For disease such as diabetes (1) variable, the odds ratio value (1.8868) indicates residents who have been told they had diabetes have 88.68% higher odds of having CVD compared to the residents who have not been told they had diabetes.

Based on the binary logistic regression model, 13 variables with positive coefficient estimated values and with odds ratio larger than 1, indicate these variables have a positive association with developing CVD which means there is an increased probability of having CVD. These significant risk factors of CVD are residents from the age group (6-13), current smokers, underweight residents, residents with high blood pressure, residents with high cholesterol, and residents with diabetes. However, there are two variables (female, and heavy drinkers) with negative coefficient estimated values and with odds ratio less than 1, meaning these two variables show a negative association with developing CVD, indicating there is a decreased likelihood of CVD.

3.4 Probit Regression

Before building the probit regression model, the dummy variables and reference variables should be defined for the categorical variables. To build the probit regression model, backward elimination method is used. The variable with the highest *p*-value which greater than 0.05 is removed from the model. This indicates that variables are not statistically significant to the model. The prosses will procced until all the variables are statistically significant to the model. There are seven probit models are built and shown in Table 9.

Table 9 *Variables in Probit Regression Models*

Model	Variables in Model
1	female gender, age group (2-13), smoking, alcohol, BMI categories (underweight, overweight, and obese), blood pressure, cholesterol, and diabetes
2	female gender, age group (2-13), smoking, alcohol, BMI categories (underweight, and obese), blood pressure, cholesterol, and diabetes
3	female gender, age group (2, 3, 5-13), smoking, alcohol, BMI categories (underweight, and obese), blood pressure, cholesterol, and diabetes
4	female gender, age group (2, 3, 5-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes
5	female gender, age group (2, 5-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes

Enhanced Knowledge in Sciences and Technology Vol. 4 No. 1 (2024) p. 205-216

6	female gender, age group (5-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes
7	female gender, age group (6-13), smoking, alcohol, BMI categories (underweight), blood pressure, cholesterol, and diabetes

To select the best model for probit regression model, Deviance, AIC, and BIC are applied. The goodness-of-fit test for probit model is as shown in Table 10.

Table 10 *Goodness of Fit Test*

Model	Deviance	AIC	BIC
1	**11313.91**	11357.91	11529.64
2	**11313.91**	11355.91	11519.83
3	11314.06	11354.06	11510.17
4	11314.55	11352.55	11500.86
5	11315.67	11351.67	11492.17
6	11317.15	11351.15	11483.84
7	11318.58	**11350.58**	**11475.47**

Based on Table 10, Model 1 and Model 2 have the lowest deviance value (11313.91), which means the model fit well. Besides, Model 7 with the lowest AIC (11350.85) and BIC (11475.47) values, meaning the model fit well. Based on the Goodness of Fit Test, Model 7 perform well as it fulfilled two out of three requirements. Besides, the variables in Model 7 also show statistically significant to the model with *p*-value less than 0.05. Therefore, Model 7 is selected as the final model. The coefficient and marginal effect of final model as shown as Table 11.

Table 11 *The Final Model for Probit Regression with Coefficient and Marginal Effect*

Variables	Coefficient Estimate	Standard Error	*p*-value	Marginal Effect
Intercept	-2.34882	0.05740	**< 0.001**	
Gender (Female)	-0.33349	0.02668	**< 0.001**	-0.05721
Age 6	0.40106	0.08224	**< 0.001**	0.0688
Age 7	0.50373	0.07520	**< 0.001**	0.08641
Age 8	0.53381	0.07006	**< 0.001**	0.09157
Age 9	0.67667	0.06611	**< 0.001**	0.1161
Age 10	0.79672	0.06419	**< 0.001**	0.1367
Age 11	0.92465	0.06350	**< 0.001**	0.1586
Age 12	1.08910	0.06572	**< 0.001**	0.1868
Age 13	1.24614	0.06383	**< 0.001**	0.2138
Smoking (1)	0.33014	0.03692	**< 0.001**	0.05663
Alcohol (1)	-0.29265	0.07734	**< 0.001**	-0.0502
BMI (Underweight)	0.20050	0.10014	**< 0.05**	0.03439
Blood Pressure (1)	0.43916	0.03117	**< 0.001**	0.07533
Cholesterol (1)	0.36036	0.02791	**< 0.001**	0.06182
Diabetes (1)	0.35820	0.02955	**< 0.001**	0.06145

The final model for probit regression is as shown as equation 8:

$$\begin{aligned}\pi_1 = &-2.349 - 0.333\text{Gender\_female} + 0.401\text{age6} + 0.504\text{age7} + \\ &0.534\text{age8} + 0.677\text{age9} + 0.797\text{age10} + 0.925\text{age11} + 1.089\text{age12} + \\ &1.246\text{age13} + 0.330\text{Smoking\_1} - 0.293\text{Alcohol\_1} + \\ &0.201\text{BMI\_Under} + 0.439\text{BloodPressure\_1} + \\ &0.361\text{Cholesterol\_1} + 0.358\text{Diabetes\_1}\end{aligned} \quad (8)$$

Based on Table 11, the negative coefficient of female gender variable suggests that female residents have a lower risk to get CVD with a Z score of 0.33349 compared with male residents. For the marginal effect, female residents had a 5.72% lower chance of having CVD compared to male. For age categories, the reference category is age 1 (Age 18 to 24). All the coefficients of age variables are positive and with increasing values across age groups, indicating that when the age groups increase, the Z score of CVD also increase. For smoking variable with the positive coefficient estimated value (0.33014), indicating the residents who are current smoker have a higher chance to have CVD with a Z score 0.33014 compared to those non-current smokers. Based on the marginal effect value (0.05663), indicates that the residents who are current smokers has 5.66% higher chance of developing CVD compared to non-current smokers. For alcohol factors, the coefficient estimated value is -0.29265, which indicates the heavy drinkers versus non-heavy drinkers, decreases the Z score by 0.29265.

In additional, the BMI (underweight) variable with the coefficient (0.20050) illustrated that the residents who are underweight have a higher risk to have CVD with a Z score 0.20050 compared to those normal weight. Moreover, for blood pressure (1) variable with marginal effect value (0.07533), indicates the residents who have been told by a doctor, nurse, or other professional having high blood pressure have 7.53% higher risk of developing CVD compared to those who have not been told by a doctor, nurse, or other professional having high blood pressure. For the cholesterol (1) variable with positive coefficient estimated value (0.36036), indicates the residents who have been told by a doctor, nurse, or other professional having high cholesterol versus the residents who have not been told by a doctor, nurse, or other professional having high cholesterol, increases the Z score by 0.36036. Additionally, the disease such as diabetes (1) variable, the marginal effect value (0.06145) indicates residents who have been told they had diabetes have 6.15% higher chance of having CVD compared to the residents who have not been told they had diabetes.

Based on the probit model, 13 variables with positive coefficient estimated values and marginal effect values, indicate these variables have a positive association with developing CVD which means there is an increased probability of having CVD. These significant risk factors of CVD are residents from the age group (6 – 13), current smokers, underweight residents, and residents with high blood pressure, high cholesterol, and diabetes. Conversely, female and heavy drinkers with negative coefficient estimated values and negative marginal effect values, implies these two variables show a negative association with CVD.

3.5 Model Selection

After building the logit model and probit model, a best model between these two final models is selected using goodness of fit test. The goodness of fit test for these two models are as shown in Table 12.

Table 12 *Goodness of Fit Test*

Goodness of Fit Test	Binary Logistic Regression Model	Probit Regression Model
Deviance	11330.19	**11318.58**
AIC	11362.19	**11350.58**
BIC	11487.08	**11475.47**

Based on the Table 12, probit regression model has with a lower deviance (11318.58) and BIC (11475.47) compared to logistic regression model with deviance (11330.19) and BIC (11487.08), meaning the probit model slightly fit well than the logit model. Besides, probit model with a lower AIC (11350.58) compared to logit model (11362.19), indicating the probit model captures the information in the data with the least loss. Therefore, the probit regression model is selected as the best model in this study as it provides a slightly better fit compared to binary logistic regression model. The probit regression model equation is performed as equation 8.

4. Conclusion

The relationship between the risk factors and cardiovascular disease in the United States in 2021 was determined by using descriptive analysis. The findings showed the respondent that have ever reported having CVD is recorded 12.27%. Most residents with CVD are men, age group 13, obese residents, non-current smoking, not heavy drinkers, residents with high blood pressure, high cholesterol and no diabetes. Besides, the significant

risk factors of CVD were determined using binary logistic regression model and probit regression model. Both models produced the same results of the significant risk factors of CVD, which are male gender, age groups from 6 to 13 (Age 45 and above), current smokers, non-heavy drinkers, underweight, high blood pressure, high cholesterol, and diabetes. There is an increased probability of developing CVD among the residents in the U.S.. Lastly, probit model with lower Deviance (11318.58), AIC (11350.58), and BIC (11475.47) compared to logit model with Deviance (11330.19), AIC (11362.19), and BIC (11487.08). Thus, probit regression model is selected as the best model in this study as it provides a slightly better fit compared to binary logistic regression model.

Acknowledgement

The authors would thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Teo Li Sha, Norziha Che Him;* ***data collection:*** *Teo Li Sha;* ***analysis and interpretation of results:*** *Teo Li Sha, Norziha Che Him;* ***draft manuscript preparation:*** *Teo Li Sha, Norziha Che Him and Mohd Saifullah Rusisman. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] World Health Organization. (2021). *Cardiovascular diseases (CVDs)*. Retrieved May 10, 2023, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Centers for Disease Control and Prevention. (2023). *Leading Causes of Death*. Retrieved May 12, 2023, from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

[3] Centers for Disease Control and Prevention. (2022g). *Heart Disease Mortality by State*. Retrieved May 14, 2023, from https://www.cdc.gov/nchs/pressroom/sosmap/heart_disease_mortality/heart_disease.htm

[4] UCSF Health. (2022). *Understanding Your Risk for Heart Disease*. Retrieved May 10, 2023, from https://www.ucsfhealth.org/education/understanding-your-risk-for-heart-disease

[5] Indus Health Plus. (2022). *Heart Disease Causes & Risk Factors*. Retrieved May 16, 2023, from https://www.indushealthplus.com/heart-disease-causes-risk-factors.html

[6] Xu, W., Zhao, Y., Nian, S., Feng, L., Bai, X., Luo, X., & Luo, F. (2018). Differential analysis of disease risk assessment using binary logistic regression with different analysis strategies. *Journal of International Medical Research*, *46*(9), 3656–3664. https://doi.org/10.1177/0300060518777173

[7] Abdulqader, Q. A. (2017). Applying the Binary Logistic Regression Analysis on The Medical Data. *Science Journal of University of Zakho*, *5*(4), 330. https://doi.org/10.25271/2017.5.4.388

[8] Abd Elsalam, N. M. M. (2015). Binary Logistic Regression to Identify the Risk Factors of Eye Glaucoma. *International Journal of Sciences: Basic and Applied Research*, *23*(1), 366–376. https://doi.org/10.25271/2017.5.4.388

[9] Norton, E. C., Dowd, B. E., & Maciejewski, M. L. (2018). Odds Ratios—Current Best Practice and Use. *JAMA*, *320*(1), 84. https://doi.org/10.1001/jama.2018.6971

[10] Mauchant, D., Riley, M. A., Rice, K. D., Forster, A. L., Leber, D. D., & Samarov, D. V. (2011). Analysis of Three Different Regression Models to Estimate the Ballistic Performance of New and Environmentally Conditioned Body Armor. *NISTIR 7760*. https://doi.org/10.6028/nist.ir.7760

[11] Alhashmi, M. (2016). *Estimating the risk of non-communicable diseases based on behavioral risk factors and the social determinants of health: a case study of Canada* [Master of Science]. McGill University. https://escholarship.mcgill.ca/concern/theses/f4752k396

[12] Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2023). *Introduction to Econometrics with R*. https://www.econometrics-with-r.org/11.1-binary-dependent-variables-and-the-linear-probability-model.html

[13] Loo, K., & Asrah, N. M. (2022). Survey on Customer Satisfaction Towards Courier Services in Johor. *Enhanced Knowledge in Sciences and Technology*, *2*(2), 186-196.

[14] Zulkiflee, N. F., & Rusiman, M. S. (2021). Heart Disease Prediction Using Logistic Regression. *Enhanced Knowledge in Sciences and Technology*, *1*(2), 177-184.
[15] Lim, L. S. Y. S., & Ismail, T. D. I. B. (2022). A Study On Volatility and Tail Risk of Small-Cap Companies in Comparison to Big-Cap Companies. *Enhanced Knowledge in Sciences and Technology*, *2*(1), 241-249.
[16] Haron, N. A. A., & Kamardan, M. G. (2021). Queuing system of a busy restaurant using simulation software. *Enhanced Knowledge in Sciences and Technology*, *1*(2), 66-71.
[17] Faculty of Public Health. (2016). *Graphical methods in Statistics*. Retrieved May 25, 2023, from https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/graphical-methods-statistics
[18] Srimaneekarn, N., Hayter, A. J., Liu, W., & Tantipoj, C. (2022). Binary Response Analysis Using Logistic Regression in Dentistry. *International Journal of Dentistry*, *2022*, 1–7. https://doi.org/10.1155/2022/5358602
[19] Hilbe, J. M. (2016). Practical Guide to Logistic Regression. In *Chapman and Hall/CRC eBooks* (1st ed.). CRC Press Taylor & Francis Group. https://doi.org/10.1201/b18678
[20] Nishadi, A. S. T. (2019). Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab. *International Journal of Advanced Research and Publications*, *3*(8), 69–74. https://www.ijarp.org/published-research-papers/aug2019/Predicting-Heart-Diseases-In-Logistic-Regression-Of-Machine-Learning-Algorithms-By-Python-Jupyterlab.pdf
[21] Saif, M. A. M., & Raheem, S. H. (2020). Determine of The Most Important Factors That Affect The Incidence Of Heart Disease Using Logistic Regression Model. *ResearchGate*, 175–184. https://www.researchgate.net/publication/344324416
[22] Ling, N. T., Rusiman, M. S. B., S, S., Hamzah, F. M., & Ebas, N. A. (2022). Factors Affecting Customer Loyalty on Starbucks Malaysia using Binary Logistics and Probit Model. In *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. https://ieeexplore.ieee.org/document/9936751
[23] Agresti, A. (2018). *An Introduction to Categorical Data Analysis* (3rd ed.). John Wiley & Sons. https://www.wiley.com/en-us/Categorical+Data+Analysis,+3rd+Edition-p-9780470463635
[24] Sen, S., Maiti, S. S., & Chandra, N. (2016). The xgamma Distribution: Statistical Properties and Application. *Journal of Modern Applied Statistical Methods*, *15*(1), 774–788. https://doi.org/10.22237/jmasm/1462078200