Logistic Regression Model based on heart disease and Its Potential Influencing Factors

Yu Qiu *

Department of Statistics, School of Mathematics and Statistics, Shandong University, Weihai, China

* Corresponding author: 202000810084@mail.sdu.edu.cn

Abstract. Heart disease is without doubt getting more and more popular in human society. According to the statistics of the World Heart Federation, one person dies of heart diseases for every 3 deaths in the world, and the number of deaths due to heart disease and stroke is as high as 17.5 million in the world every year. In this paper, 5 potential influencing factors and their data are selected to construct a logistic regression model to predict the possibility of one catching a heart disease so that early prevention may be achieved in time. During the construction of the model, some transformations are applied to the predictors to optimize the model. In the end, cross-validation method is used to test the final model, and the results show that the accuracy of the model is over 73%. In conclusion, this model can briefly predict the possibility of one catching a heart disease, and also reveal that the factors chosen do have some significant impacts on heart disease prediction.

Keywords: Heart disease, logistic regression, variables transformation, marginal model plot.

1. Introduction

Heart disease is a class of diseases that involve the heart or blood vessels and it is also the leading cause of death in all regions of the world except Africa. Heart disease is a clinical common and frequently-occurring disease with a high mortality rate and is known as the top one killer threatening human health. However, it is estimated that up to 90% of heart disease is preventable. Therefore, in this paper, a statistic model concerning some potential influencing factors will be built to help patients to get rapid and accurate diagnosis.

Babiker et.al conducted research related to risk factors of coronary heart disease. The research chooses factors such as age, gender, the measure of cholesterol, blood pressure, family history, smoking etc. and use some basic methods in logistic regression model to find out the relationship between the variables [1]. Thirugnanam et.al conducted research to find out the most influential factors as well as build an accurate prediction model using methods such as support vector machine, decision tree and logistic regression [2]. Khanna et.al conducted research intended to construct a model to predict the prevalence of heart diseases. It uses logistic regression and support vector machine as model and use F1 score and ROC curve as evaluative measures. A total of 14 variables including age, gender, chest pain type, blood pressure, cholesterol, maximum heart rate achieved and booelan measure are used [3].

Saw et.al conducted research in order to improve the accuracy of predicting the heart disease by using logistic regression model [4]. Khemphela and Boonjing conducted research in order to do heart disease prediction and compare several models. Artificial neural networks, logistic regression, decision trees are used as methods and a total 14 of variables including age, chest pain type, blood pressure, number of colored vessels and cholesterol are used. And through the comparison, it concluded that the artificial neural networks technique showed best goodness-of-fit to its dataset [5]. Desai et.al conducted research to access the accuracy of prediction model related to heart disease. It compares parametric model and nonparametric model such as back-propagation neural network and logistic regression and also use cross validation method to test these models [6]. Zhang, Diao and Ma conducted research to predict the presence of elder people catching heart disease using logistic regression model. It uses data mining to find out which are the most influential factors among 14

factors which are nearly the same as factors used in the research conducted by Khanna et.al. And construct the model using logistic regression and then compare the model to other algorithms [7].

Sharma and Sangal conducted research to build a hybrid model for heart disease prediction. It uses data mining to filter the dataset and get the most valuable and useful data to predict. After that, the random forest method and logistic regression model are introduced to build the model. The accuracy of the final model is estimated to be 95.08% [8]. Yaseliani and Khedmati conducted research to use techniques in machine learning to understand the symptoms of the heart disease and do some prediction. It uses a total of 22 predictor including age, gender, cholesterol, maximum heart rate achieved, resting blood pressure to fit the logistic regression model and access the impact of each predictor using Akaike information criterion and the p-value of the t-statistic. The final model is estimated to have a sensitivity and specificity of 84.21% and 90.38% respectively [9].

Rosalina et.al conducted research intended to build a prediction model based on completely natural data using multinomial logistic regression. Also, during the research, some other useful methods were applied to the same data set. Compare these results together, multinomial logistic regression shows the best accuracy against other computational techniques [10].

Based on the information obtained above, 5 most used predictor variables are considered to construct a new prediction model for the presence of heart disease using logistic regression in this paper. In this case, we can tell the probability of one suffering from the heart disease by his 5-dimension data including systolic blood pressure, family history, a measure of cholesterol, a measure of obesity and age. In this way, early prevention can be achieved, timely and effective treatment can be guaranteed, and human life can be extended.

2. Methods

2.1. Data Description

The data set chosen is downloaded from the data base of the book A Modern Approach to Regression in R. The data set contains 5 different variables concerning heart disease and 462 data points. Show the head of data set:

Table 1. The head of data set

x1	x2	x3	x4	x5	Y
160	5.73	1	25.3	52	1
144	4.41	0	28.87	63	1
118	3.48	1	29.14	46	0
170	6.41	1	31.99	58	1
134	3.5	1	25.99	49	1
132	6.47	1	30.77	45	0
142	3.38	0	20.81	38	0
114	4.59	1	23.11	58	1
114	3.83	1	24.86	29	0
132	5.8	1	30.11	53	1
206	2.95	0	26.81	60	1
134	4.44	1	23.09	40	1
118	1.88	0	21.57	17	0
132	1.87	0	23.63	15	0

From Table 1, x1 stands for systolic blood pressure, x2 for a measure of cholesterol, x3 for family history, x4 for a measure of obesity, x5 for age and Y for heart disease happens or not. As it can be seen, x3 and Y are both dummy variables whose values are either 0 or 1. (x3 = 1 indicates that patient has family history and Y = 1 indicates that heart disease happens).

2.2. Basic Numeric Characteristics

Table 2 shows the statistical characteristics of the variables used in this paper. For more details, seeing as in section 2.3.

Table 2. Numeric characteristics

Case	x1	x2	x3	x4	x5	Y
Mean	138.327	4.740	0.416	26.044	42.816	0.346
Var.	420.099	4.289	0.243	17.755	213.422	0.227
S.E.	20.474	2.069	0.493	4.209	14.593	0.476
Max.	218.000	15.330	1.000	46.580	64.000	1.000
Min.	101.000	0.980	0.000	14.700	15.000	0.000

2.3. Visualize the Data Set

Divide the data set into 2 parts according to Y equaling to 1 or 0 and draw the frequency histograms of these two parts about each variable within 1 graph.

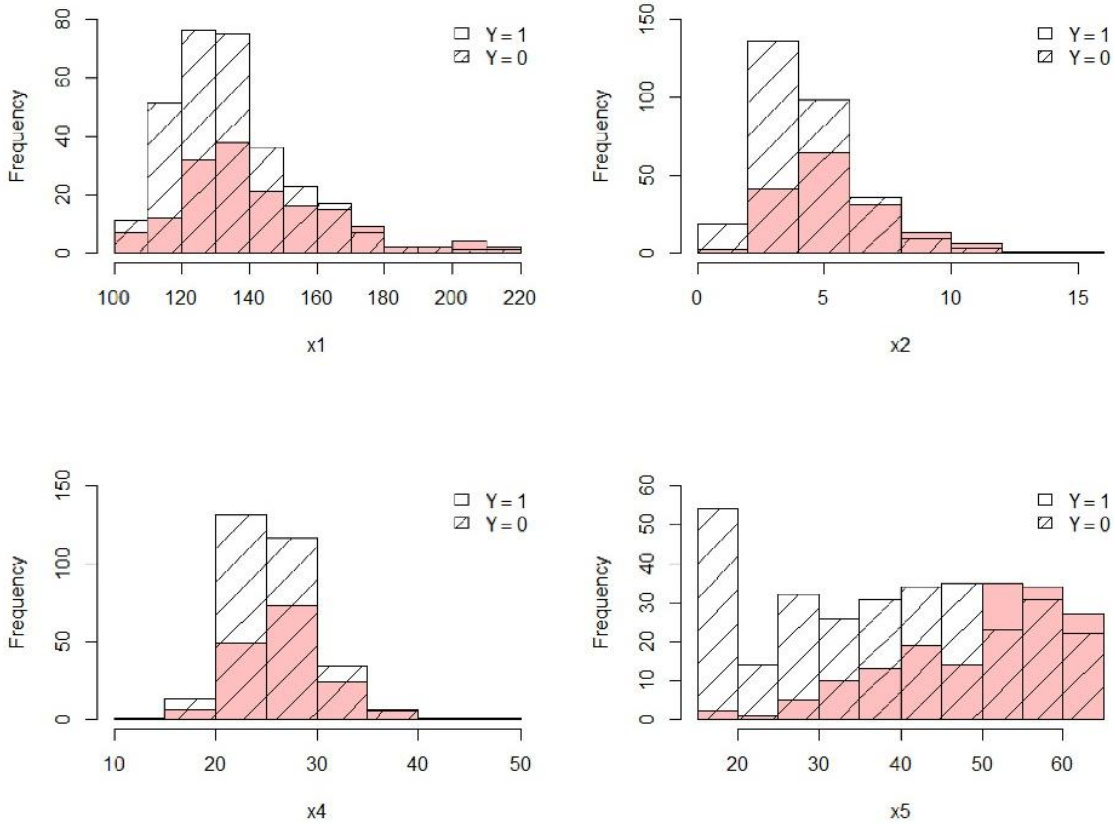


Figure 1. Frequency histogram

According to frequency histogram for x1 in Figure 1, within the interval from 110 to 140, when x1 takes the same value, and the proportion of Y=1 is smaller than that of Y=0. The farther away from this interval, the proportion of Y=1 tends to be larger, especially as x1 increases. However, within the interval where x1 less than 110, the proportion of Y=1 is relatively high, too. This may indicate that this model relates to some other forms of x1.

According to frequency histogram for x2 in figure 1, the proportion of Y=1 keeps increasing within whole range of x2 as x2 increases. This indicates that this variable may be suitable for a logistic regression. As for x4 in figure 1, same as figure 2, the proportion of Y equaling to 1 keeps increasing within the interval from 20 to 50 as x2 increases. However, within the interval where x4 less than 20, the proportion of Y equaling to 1 is relatively high, too. This may indicate that this model relates to some other forms of x4.

According to frequency histogram for x5 in figure 1, The proportion of Y equaling to 1 tends to increase roughly as x5 increases. This indicates that x5 may will be modelled correctly. However, the possibility of Y equaling to 1 is not so stable within the interval where x5 less than 50. This may indicate that this model relates to some other forms of x5. As for x3, because it is a dummy variable, it can be easily deduced from the row data that the possibility of Y equaling to 1 when x3 equals to 1 is much higher than that when x3 equals to 0. Therefore, there is strong evidence showing that this variable can be fit into the model.

2.4. Logistic Regression

Logistic regression is a generalized linear model, so its model forms can be expressed like $y = w^T x + b$, where w and b are the parameters to be obtained. Logistic regression uses function $\log\left(\frac{p}{1-p}\right)$ to be the dependent variable. Therefore, data of x can be used to estimate the of w^T and thus predict p. It is also fact that p is ranging from 0 to 1, so it is usually perceived as a kind of possibility.

2.5. Method of Calculating the Estimation of Parameters

First, write down the pmf of Y

$$P(y) = \left[\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdot\cdot+\beta_5 x_5)}}\right]^y \left[\frac{e^{-(\beta_0+\beta_1 x_1+\cdot\cdot+\beta_5 x_5)}}{1+e^{-(\beta_0+\beta_1 x_1+\cdot\cdot+\beta_5 x_5)}}\right]^{1-y} \quad (1)$$

Take the log of its likelihood function and then plug the data into this function and use Newton-Raphson method to calculate $\widehat{\beta_0}, \cdots, \widehat{\beta_5}$ to make this function as large as possible. In binary logistic regression, the deviance, which, in linear regression, corresponds to residual sum of squares, is defined like this:

$$G^2 = -2\sum_{i=1}^{n}[y_i log(\hat{y}_i) + (1-y_i) log(\hat{y}_i)] \quad (2)$$

Where $\hat{y}_i$ is the estimation value of $P(y_i = 1)$. Looking into another situation, let y_i directly be the estimation of $P(y_i = 1)$ and we get $logL_s = \sum_{i=1}^{n}[y_i log(y_i) + (1-y_i) log(1-y_i)]$; let $\hat{y}_i$ be the estimation of $P(y_i = 1)$ and we get $logL_m = \sum_{i=1}^{n}[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)]$. Also, we should assume that function $f(x) = xlogx$ is defined on $[0, +\infty)$ and has right continuity, so that we can say:

$$G^2 = 2[logL_s - logL_m] = -2logL_m \quad (3)$$

Noticing that the distribution of G^2 is no longer χ^2, but the subtraction of different deviances approximately follows χ^2 distribution. So, we can still use G^2 to do some model comparison and hypothesis test.

2.6. Cross Validation

Cross-Validation is a resampling procedure used to evaluate the accuracy of machine learning models. The goal of cross-validation is to split the available data into several parts or folds, such that each fold is used as a validation set for evaluating the performance of the model trained on the remaining data. The performance of the model is then averaged across all the folds to obtain a more robust estimate of the model's generalization performance. This process helps to reduce overfitting, as the model is trained and tested on different data, and it provides a more realistic evaluation of the model's performance on unobserved data.

3. Results and Discussion

3.1. Original Model

Plug the row data set into the R function, and get the first logistic regression model. The results are as follow:

Table 3. Original model (AIC:505.62)

Coef	Estimate	S.E.	z-value	Pr(>|z|)	
Intercept	-4.313	0.944	-4.570	0.000	***
x1	0.006	0.006	1.169	0.242	
x2	0.186	0.056	3.305	0.001	***
x3	0.904	0.221	4.090	0.000	***
x4	-0.036	0.029	-1.236	0.216	
x5	0.053	0.010	5.549	0.000	***

From table 3, there are some variables which are not obvious to the whole model. Therefore, some changes should be made.

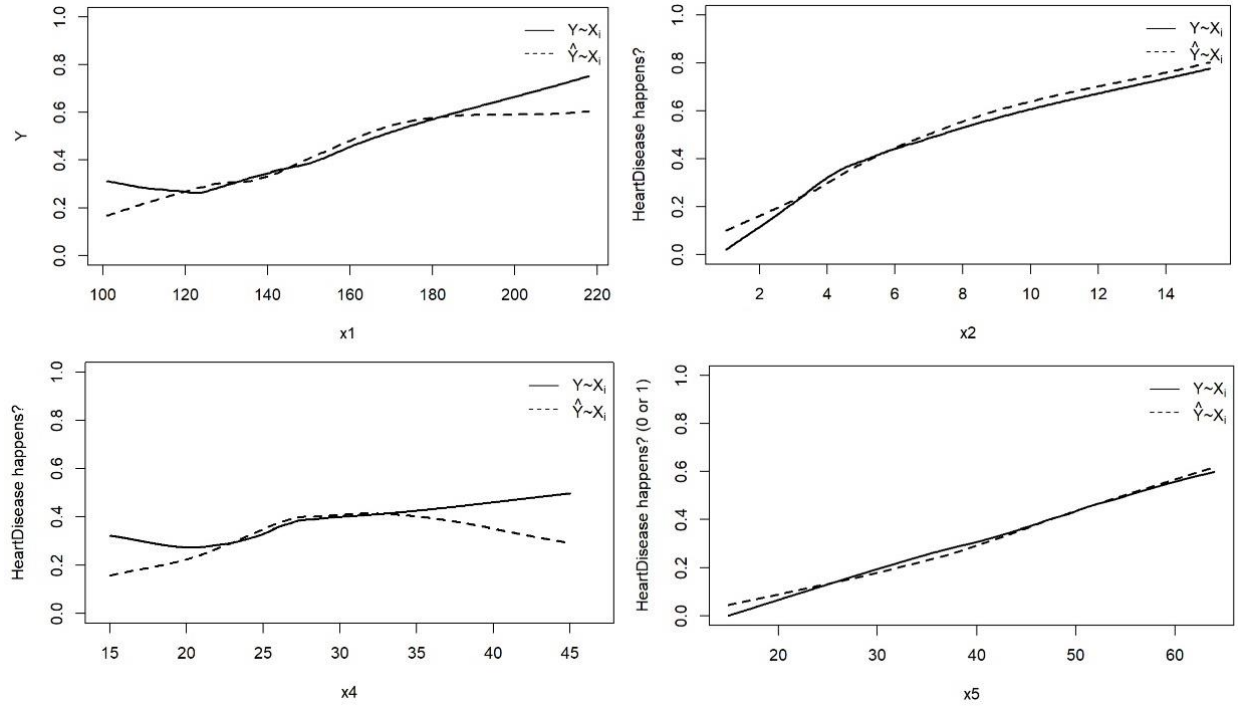


Figure 2. Marginal model plots for original model

Figure 2 shows the marginal model plots for original model, and a more detailed marginal model plots will be introduced in section 3.4.

3.2. Variables Transformation

Suppose that $\theta(x) = P(Y = 1|X)$. Through calculation, we can see that the log odds function can be transformed into these forms:

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + log\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)}\right) \quad (4)$$

Where X is a concrete variable and

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) = log\left(\frac{P(Y=1)}{P(Y=0)}\right) + log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right) \quad (5)$$

Where X is a continuous variable and $f(x|Y)$ is the density function of X given Y. Moreover, if we assume that the distribution of x given Y equals to j, $j = 0,1$ follow normal distribution $N(\mu_i, \sigma_j^2)$, the log odds function can be further transformed into: $log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$, where $\beta_1 = \frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}$, $\beta_2 = \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)$ and β_0 is also a constant.

Through further proof, it can be concluded that the log odds function is linear for predictor variable X, if X follows normal distribution with the same variance but different value of Y and if the covariance matrix of several specific predictor variables differs when Y equals to different values, then the log odds function will be the linear combination of $x_i, x_i^2 \; and \; x_i x_j$. And we give another conclusion without proving it: If $f(x|Y = j)$, $j = 0,1$ are pretty much skewed, both x and log(x) should be included in the long odds function.

3.3. Model Contains Log Form of Variables

Figure 3 can show a lot of things: x3 is a dummy variable, the log odds function is also a linear function of x3. Therefore, the boxplot for x3 is not displayed here.

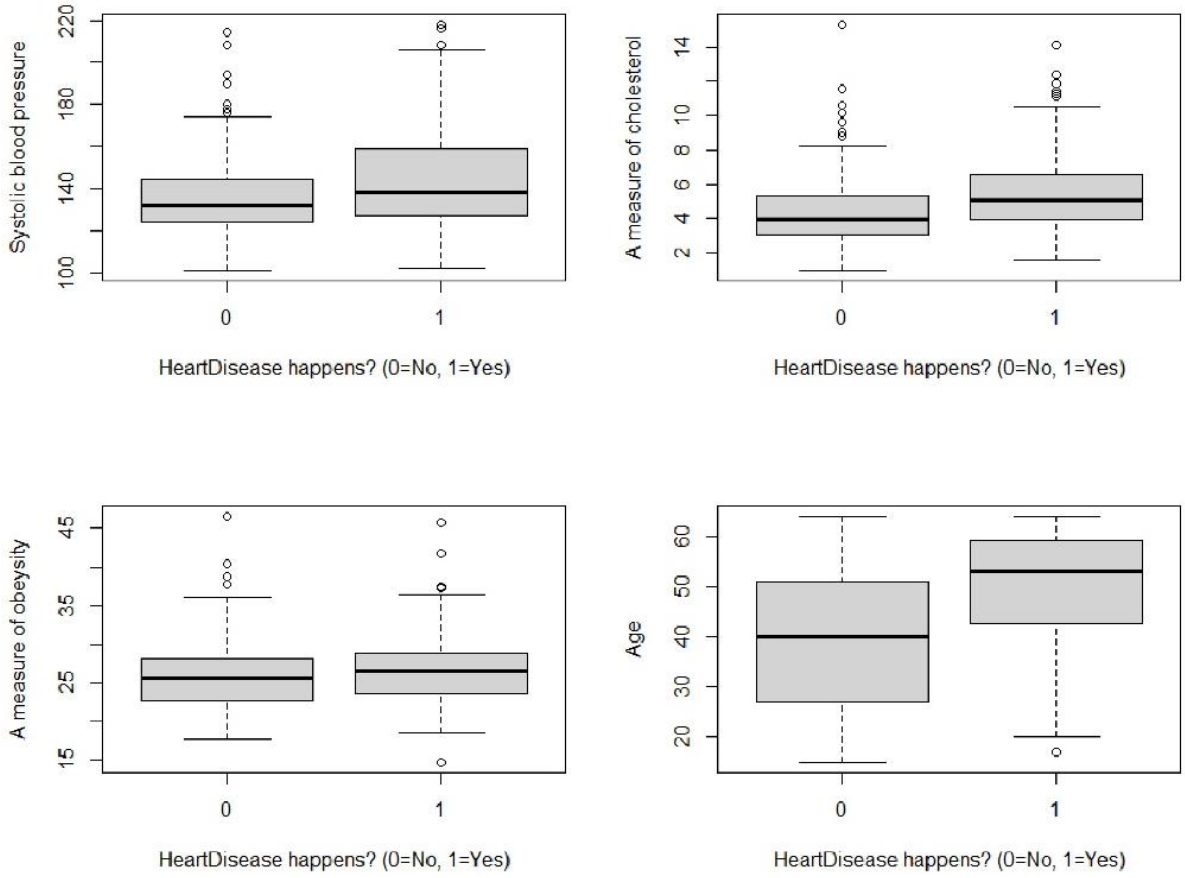


Figure 3. Boxplot of variables

For several other variables, we can see that, except for x5, the distributions of x1, x2 and x4 are skewed to some extent, so we add the log of them into the model. And the results are as follow:

According to table 4, some variables still do not do much attributes to the model. It indicates that maybe some new forms of variables should be introduced or some old variables should be filtered. However, first thing to do is to check if each of the variable is modelled correctly.

Table 4. Model contains log form of variables (AIC:503.64)

Coef	Estimate	S.E.	z-value	Pr(>|z|)	
Intercept	77.861	34.036	2.288	0.022	*
x1	0.101	0.053	1.912	0.056	.
log(x1)	-14.068	7.814	-1.800	0.072	.
x2	0.011	0.189	0.060	0.952	
log(x2)	1.055	1.018	1.036	0.300	
x3	0.929	0.225	4.136	0.000	***
x4	0.391	0.210	1.865	0.062	.
log(x4)	-11.711	5.755	-2.035	0.042	*
x5	0.055	0.010	5.671	0.000	***

3.4. Marginal Model Plot

There are a logistic regression model(M) and another nonparametric model(F), if one variable x_i is modelled correctly, then $\theta_M(x_i) = E_M(Y|x_i) \sim E_F(Y|x_i) = \theta_F(x_i)$. For $E_F(Y|x_i)$, it can be simply estimated by using a nonparametric fit to the plot of Y against x_i. For $E_M(Y|x_i)$, it can be estimated by an estimation of $E_M(\hat{Y}|x)$ and that is a nonparametric fit to the plot of $\hat{Y}$ against x_i.

Therefore, we draw the 'marginal model plot' where loess fitting curve for $\hat{Y}$ with respect to x_i and the loess fitting curve for Y with respect to x_i are both contained. Draw the Figure 4:

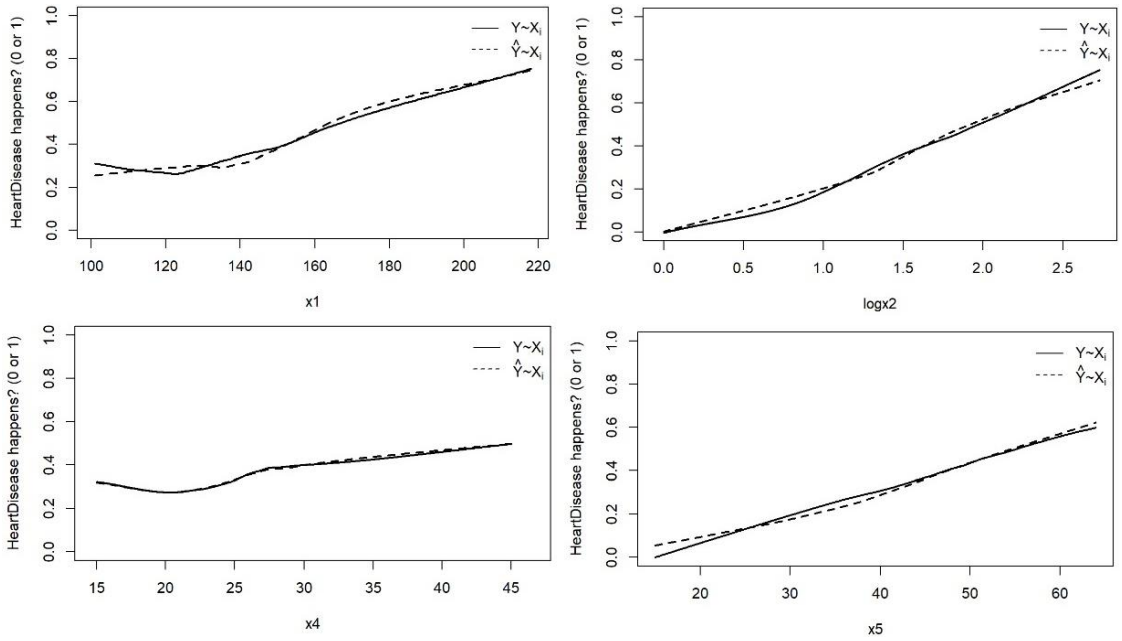


Figure 4. Marginal model plots

In Figure 4, X3 is a dummy variable, so two lines in its marginal model plot strictly agree. Thus, the marginal model plot for x3 is skipped. And the trend of marginal model plot for $logx$ and x are alike, so only one of them two is shown. From figure 3, it can be deduced that most of the variables fit better than the former model. However, there is still some slight unfitness in the graph of $x_1, \log x_2$ and x_5, so let's consider the interactive variable of them.

3.5. Interactive Variables and Quadratic Form of Variables

Recall the conclusion that we give formerly, the variable $x_i x_j$ should be contained in the model if the covariance matrix of x_i and x_j differs when Y equals to different values. And that means the linear regression of x_i on x_j has a different slope when Y equals to different values.

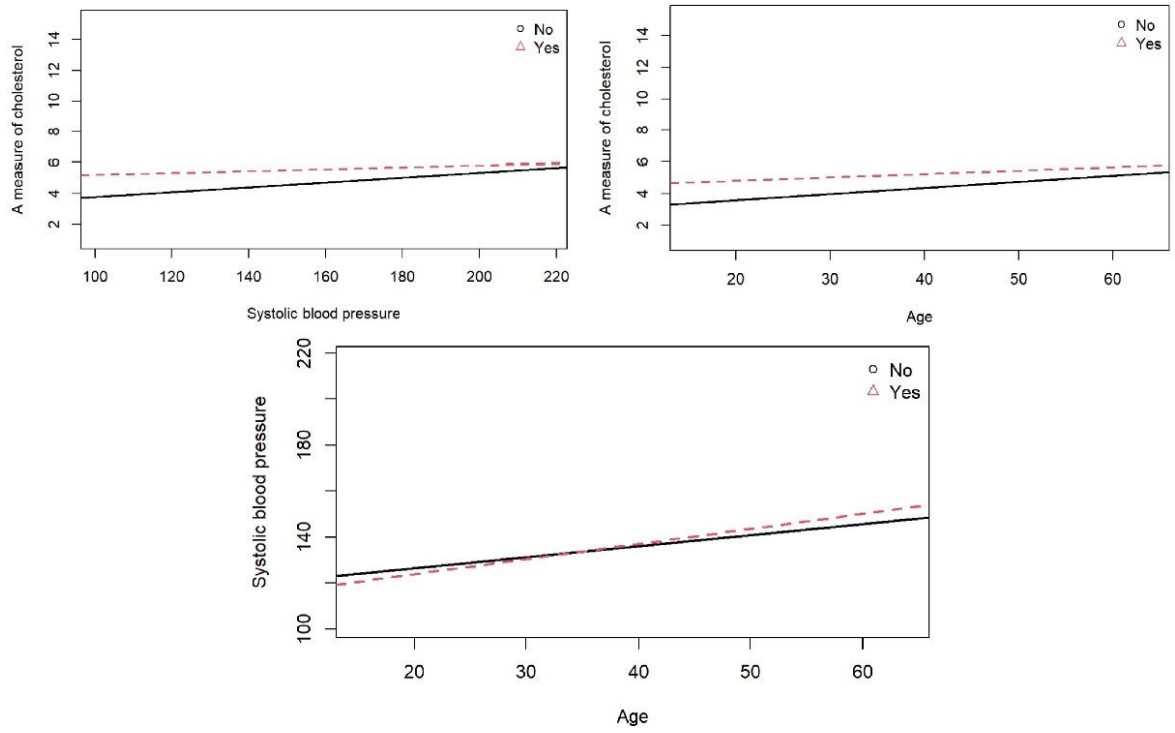


Figure 5. Slope comparison graph

From figure 5, we cannot perceive that the slopes in each plot different according to the x limit and y limit, so we did not add the interactive variable. (Actually these $x_i x_j$ have been added into the model for test, but the results are not good at all.)

At last, we consider to add the quadratic form of x_i. From all the boxplots drawn, it can be told the variance of x_1, x_2 *and* x_5 across the two values of Y is not likely the same, so we add them to the model and do a stepwise regression.

3.6. Final Model

Stepwise results are as follow. According to table 5 and figure 6 (same as figure 5, some plots are skipped), every coefficient is obvious at least in 90% confidence interval (x5.s refers to x5 squared), the AIC value is the least among all three models and every marginal model plots fit relatively well than before.

Table 5. Results of final model (AIC:500.71)

Coef	Estimate	S.E.	z-value	Pr(>|z|)	
Intercept	82.740	34.580	2.392	0.017	*
x1	0.107	0.053	2.006	0.045	*
logx1	-14.770	7.841	-1.884	0.060	.
x5.s	-0.001	0.001	-1.668	0.095	.
logx2	1.089	0.305	3.566	0.000	***
x3	0.904	0.225	4.025	0.000	***
x4	0.444	0.221	2.012	0.044	*
logx4	-13.430	6.084	-2.208	0.027	*
x5	0.160	0.064	2.487	0.013	*

The results of 10-fold cross-validation for original model and final model are 72% and 73% respectively, so the final model is relatively convincing.

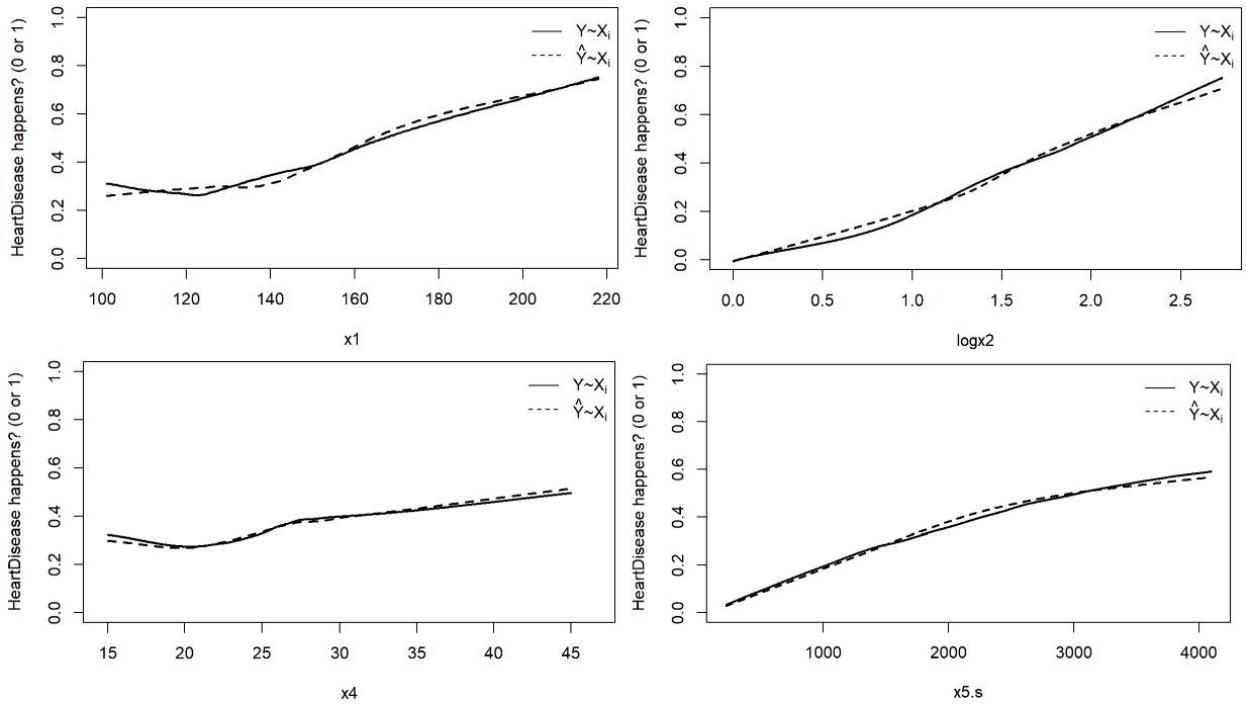


Figure 6. Marginal model plots for final model

4. Conclusion

In this study, 5 potential influencing variables are chosen to construct a logistic regression model to predict the possibility of a heart disease. During the research, some transformations have been done on some variables to optimize the model and cross-validation is used to test the accuracy of the model. Through the study, it has been found that, in logistic regression model, the possibility of Y equaling to 1 is related to x1, x3, x4, x5, logx1, logx2, logx4 and quadratic form of x5. Therefore, this model can be used to give a prediction only by the 5 factors used in the model, and the prediction results will have the accuracy of 73% in ideal situation. With the help of this model, early prevention may be achieved.

References

[1] Babiker S, Eltayeb Y, Sayed-Ahmed N, et al. Logit model in prospective coronary heart disease (CHD) risk factors prediction in Saudi population. Saudi Journal of Biological Sciences, 2021, 28 (12): 7027 - 7036.

[2] Mythili T, Mukherji D, Padalia N, et al. A heart disease prediction model using SVM-decision trees-logistic regression (SDL). International Journal of Computer Applications, 2013, 68 (16).

[3] Khanna D, Sahu R, Baths V, et al. Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease. International Journal of Machine Learning and Computing, 2015, 5 (5): 414.

[4] Saw M, Saxena T, Kaithwas S, et al. Estimation of prediction for getting heart disease using logistic regression model of machine learning. IEEE, 2020: 1 - 6.

[5] Khemphila A, Boonjing V. Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. 2010 international conference on computer information systems and industrial management applications (CISIM). IEEE, 2010: 193 - 198.

[6] Desai S D, Giraddi S, Narayankar P, et al. Back-propagation neural network versus logistic regression in heart disease classification. Advanced Computing and Communication Technologies: Proceedings of the 11th ICACCT 2018. Springer Singapore, 2019: 133 - 144.

[7] Zhang Y, Diao L, Ma L. Logistic Regression Models in Predicting Heart Disease. Journal of Physics: Conference Series. IOP Publishing, 2021, 1769 (1): 012024.

[8] Sharma H K, Sangal A L. Hybrid Model for Heart Disease Prediction Using Random Forest and Logistic Regression. Innovations in Computer Science and Engineering: Proceedings of 8th ICICSE. Springer Singapore, 2021: 657 - 662.

[9] Yaseliani M, Khedmati M. Prediction of Heart Diseases Using Logistic Regression and Likelihood Ratios. International Journal of Industrial Engineering & Production Research, 2023, 34 (1): 1 - 15.

[10] Ai M T, Sumiati S, Rosalina V. Predictive model for heart disease diagnosis based on multinomial logistic regression. Information Technology and Control, 2021, 50 (2): 308 - 318.