A Pretty Resonance: Prediction for Heart Disease Diagnosis

Shan Jiang (sj2921)

Department of biostatistics, Columbia University

ABSTRACT

Based on a pooled analysis of 920 subjects from America and E.U, this study harmonizes empirical evidence to build a multinomial model and logistic regression model for prediction of the accuracy of diagnosis of heart disease. Adopting data from UCI machine learning repository, results show that among three proposed models, the nominal logit model serves the best fit for prediction of heart disease. Among all risk factors, chest pain type, serum cholesterol level, stress level, physically related warning signs significantly affect the heart disease status of subjects in all models, together with the impact of age and sex.

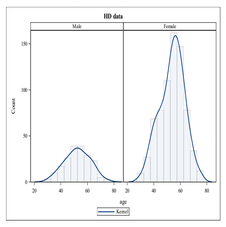
OBJECTIVE

Complicated elements are involved in the heart disease diagnostics, which need to be further studied. In order to integrate data from various places and provide a supportive application for doctors and general practitioners to make more precise decisions, and establishing specific links for different risk factors in cardiovascular disease, this paper explores two types of model for fitting categorical response to estimate the effect of some well-known risk factors of heart disease.

METHOD

For gauging the accuracy of diagnosis of heart disease across EU and U.S, UCI data originated from 4 sources including Hungary, Switzerland, Cleveland and Long beach in U.S are requested. This international dataset dates from 1988 and contains 920 subjects and 14 variables. With no missing values for organize datasets, univariate distribution are firstly examined in this study, subjects in this study retains an average age of 53.5, ranges between 28 to 77, there is no significant difference in female vs. male in age distribution.

The primary outcome were assessed with the presence of narrow vessels, for constructing of model, we adjusted for age, sex, stress, heart rate, in total all 13 covariates are then prespecified to potentially affect the response variable.



To investigate possible relations between independent variables and outcome, Pearson chi-square independence test and paneled bar plots are conducted to test and visualize whether two categorical variables are independent.

Then follows the stepwise procedure, the binomial logistic model are built by transforming response variable into binary, and multinomial logistic model are subcategorized as two specific models targeting at different outcome variable Types—ordered and unordered, through mapping on generalized logit link and logit link, and adding interaction terms for cholesterol and Thallium stress test result (which are categorized into four groups). Finally, model diagnosis and a comprehensive interpretation and detailed practical meaning are conducted for comparison and picking up final model.

RESULT

After a general overview of correlation between covariates and 2 by 2 table visualization comparison of diseased subjects vs non-disease presence individuals , it is clear that there are no multicollinearity exist which is indicated by VIF > 10, while a majority of variables have association with the outcome. Specifically, formally test results suggest, for all 6 categorical variables indicated, the chi-square test of independence implies none is independent of the heart disease status, so we cannot rule out any variables indeed by hypothesis testing.

Multinomial models are initiated for modelling the severity of heart disease. To achieve this goal, two types of links are built: one is ordered model using logit link, while the other treats the heart disease variable as the nominal variable, and fit a generalized logit link. By using the Hosmer-Lemeshow goodness-of-fit of 71.6499 with d.f of 35, multinomial ordered model is proved to be a not adequate fit (p = < 0.05). The alternative unordered generalized logit link model attains 38.02 in terms of Hosmer-Lemeshow statistic(p = 0.2141 > 0.05), implying a good fit.

Then by recoding the diagnosis variable into the binary outcome, we fit a binomial logistic regression model again, which also acquires a p-value of 0.1829 concerning goodness of fit. Again, we use common model comparison criteria for finalizing our model selection, here is the tabular form comparison of AIC, BIC and log-likelihood.

Model Fit Statistics (Intercept and Covariates)			
Criterion	logistic model	Multinomial ordered model	Multinomial unordered model
AIC	636.759	1900.689	1928.885
SC	704.300	2209.449	2020.548
-2 Log L	608.759	1772.689	1890.885

Model fitting statistics results from two categories of model analysis for intercept and covariates are presented above, the first binomial logistic model is picked for final model in this study since this model has the lowest AIC, BIC and log-likelihood value.

When it comes to the final binominal logistic model, the logit link was pre-specified for transformation, the construction of binomial logistic model using stepwise iteration finally incorporated 9 variables:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot Sex + \beta_2 \cdot Age + \beta_3 \cdot chest\ pain + \beta_4 \cdot cholest + \beta_5 \cdot thal + \beta_6 \cdot exang + \beta_7 \cdot oldpeak + \beta_8 \cdot slope + \beta_9 \cdot ca$$

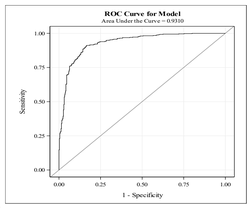
In the final model, 9 out of 14 covariates are finally chosen, including sex and age. The coefficients show: compared with male, female individuals is 1.6 times more likely to have the heart disease when adjusted for all other factors. For one year increase in age, the odds of having the heart disease will increase by 2.7%. Besides, cholesterol level also serves a crucial risk factor, with one unit increase in the serum cholesterol, the odds of having heart disease will increase by 1.6 times when adjusted for all other covariates.

It is also found that in this study ST depression induced by exercise relative to rest shows a positive correlation with the heart disease presence. Opposite to the positive correlation with the heart disease, this model also informs us some physical symptoms and stress which intervenes the heart disease status significantly. Compared with typical angina subjects, people who suffered from atypical chest pain have 55% decrease in their odds of getting heart disease, for these Non-anginal pain patients, they have 27.1% decrease in odds of getting heart disease, while for asymptomatic chest pain individuals, they bear a high risk of getting the heart disease since they are 2.03 times more likely to be diagnosed as heart diseased ones compared with the typical angina patients.

Similarly, thallium stress test result shows a negative impact for heart disease. Contrasted with the reversible defect ones, people who behave normal or fixed defect have their odds lowered 45.3% and 81.3% respectively.

CONCLUSION

Our study has confirmed some well-known demographical risk factors for heart disease, such as gender and age, corresponds to the existing studies that women are high-risk bearers for heart disease and aging enjoys a positive correlation with the disease prevalence, which is reasonable as the aging process is one of the most important determinant (Brian & David, 2012). Meanwhile, it is already been confirmed in the previous study that the ST-segment depression is a strong indicator for increasing the risk of getting heart disease (Laukkanen et al, 2012). Besides, the maximum heart rate and chest pain are gradually widely-known for its warning role of heart disease for general public, which also proved in this study.



Given the consistence of our results with the existing studies, this binomial logistic model is justifiable for its Predictability(as shown in its ROC curve).

This study contributes to the epidemiological studies for a more accurate prediction of the heart disease, as indicated by a universal warning signs such as maximum heart rate, typical chest pain and thallium stress test results.

However, some limitations are also unavoidably exist in this study. First, the valid information for distinguishing severity of heart disease are omitted in this binomial logistic model as there are only dichotomous outcomes. Second, the observations in this dataset are subject to the selection bias as the original dataset had lots of missing values and these variables only contains sex and age for demographic profile, which is not sufficient for mapping the whole picture of the race, ethnicity and other important determinants of health. Third, the binomial logistic model per se relied on some assumptions such as large sample size and the error terms to be independent are not realized in this study.

BIBLIOGRAPHY

[1]Shephard, R.J. and Balady, G.J., 1999. Exercise as cardiovascular therapy. *Circulation*, *99*(7), pp.963-972.

[2]Babič, F., Olejár, J., Vantová, Z. and Paralič, J., 2017, September. Predictive and descriptive analysis for heart disease diagnosis. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 155-163). IEEE.

[3]Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S. and Froelicher, V., 1989. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, *64*(5), pp.304-310.

[4]North, B.J. and Sinclair, D.A., 2012. The intersection between aging and cardiovascular disease. *Circulation research*, *110*(8), pp.1097-1108.

[5] Siu, S.C., Sermer, M., Colman, J.M., Alvarez, A.N., Mercier, L.A., Morton, B.C., Kells, C.M., Bergin, M.L., Kiess, M.C., Marcotte, F. and Taylor, D.A., 2001. Prospective multicenter study of pregnancy outcomes in women with heart disease. *Circulation*, *104*(5), pp.515-521.

[6] Castelli, W.P., Garrison, R.J., Wilson, P.W., Abbott, R.D., Kalousdian, S. and Kannel, W.B., 1986. Incidence of coronary heart disease and lipoprotein cholesterol levels: the Framingham Study. *Jama*, *256*(20), pp.2835-2838.

[7] Laukkanen, J.A., Mäkikallio, T.H., Rauramaa, R. and Kurl, S., 2009. Asymptomatic ST-segment depression during exercise testing and the risk of sudden cardiac death in middle-aged men: a population-based follow-up study. *European heart journal*, *30*(5), pp.558-565.