Application of Generalized Linear Model (GLM) in Heart Disease Diagnosis

Kai Zhang (kaz4006)

Department of Health Policy and Research

ABSTRACT

Cardiovascular disease are the number 1 cause of death in adults in the United States. In the present, multiple tests are available for potential cardiovascular disease in diagnosis processing. Doctor are with the help of these tests, and make final diagnosis based on their experience and knowledge. However, individual ability is limited and the diagnosis may not accurate. In order to improve the diagnosis accuracy and efficiency, multiple statistical techniques are used to assist doctors and physicians. In this poster, three statistical models are applied on a patients data, so as to exploring the applicability of generalized linear model in cardiovascular disease diagnosis. Specifically, logistics regression model, proportional odds (PO) model, and partial proportional odds model are employed on potential cardiovascular patients data collected from two the US hospitals and two European hospitals. Based on the multiple clinical tests, the generalized linear models are supposed to predict the presence as well as the severity of heart disease.

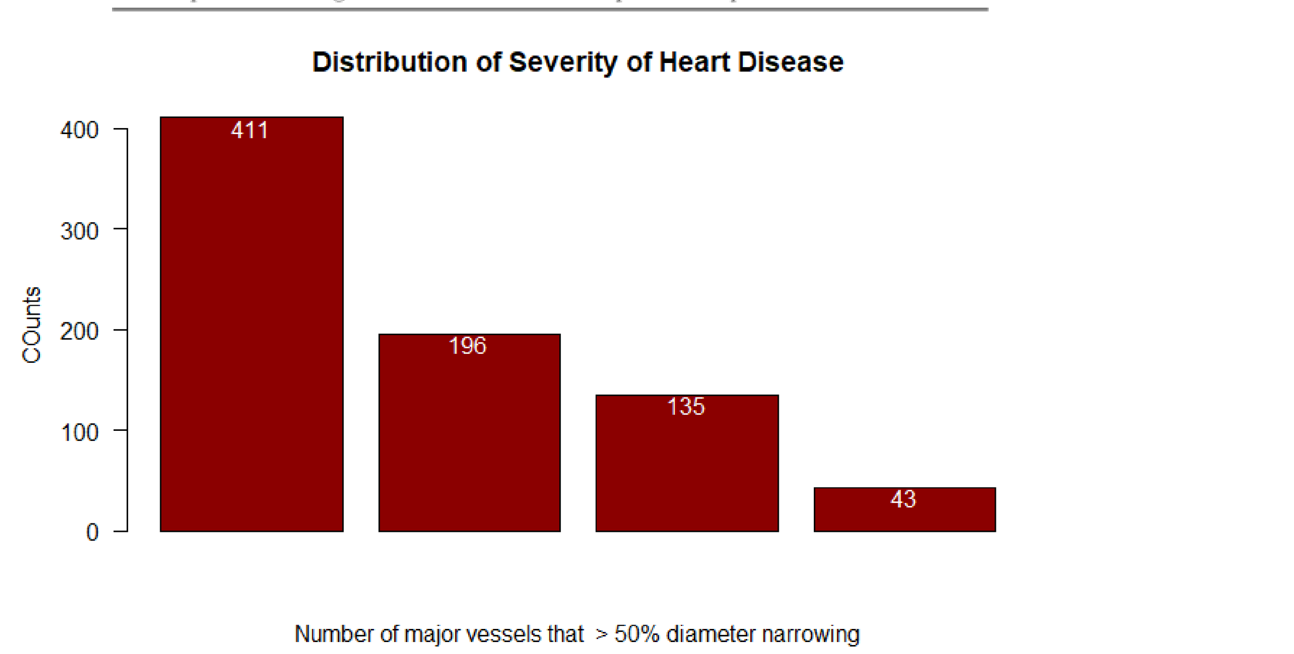
OBJECTIVE

Base on the test data in Table 1, there are two main objectives in this project:

1. To predict the presence of narrowing vessel.
2. To predict the severity of heart disease.

Table 1: Data Dictionary

Variable Name	Type	Variable Description
age	numeric	age
sex	categorical	sex
cp	categorical	Chest pain type
trestbps	numeric	Resting blood pressure
fbs	categorical	Fasting blood pressure > 120 mg/dl
restecg	categorical	Resting electrocardiographic results
thalach	numeric	Maximum heart rate achieved
exang	categorical	Exercise induced angina
oldpeak	numeric	ST depression induced by exercise relative to rest
slope	categorical	Slope of the peak exercise ST segment
ca	categorical	Number of major vessels
thal	categorical	Thalassemia
daig	categorical	Number of major vessels that > 50% diameter narrowing
hospital	categorical	hopital of the patients



METHOD

- Analysis of Binary Response

To predict on the presence of narrowing vessel, logistic regression with backwards selection is used. The coefficients are estimated by maximum likelihood estimation. The performance of the fitted model is evaluate by Receiver operating characteristic (ROC) and the area under the ROC (AUC). To rank the importance of features in linear model, the absolute value of the t-statistic for each model parameter is used. Higher absolute value indicates higher ranks of importance of the corresponding feature in the model.

- Analysis of Ordinal Response

To predict on the severity of the heart disease, proportional odds (PO) model is employed. The key assumption of PO model is that the effects of the covariate are the same for all categories on the log odds scale. The nominal test is used to test the PO assumption.

Partial proportional odds model relaxes the PO assumption, it allow the some covariates to have different effects for different categories on log odds scale. Validated set is created to evaluate the model performance.

RESULT

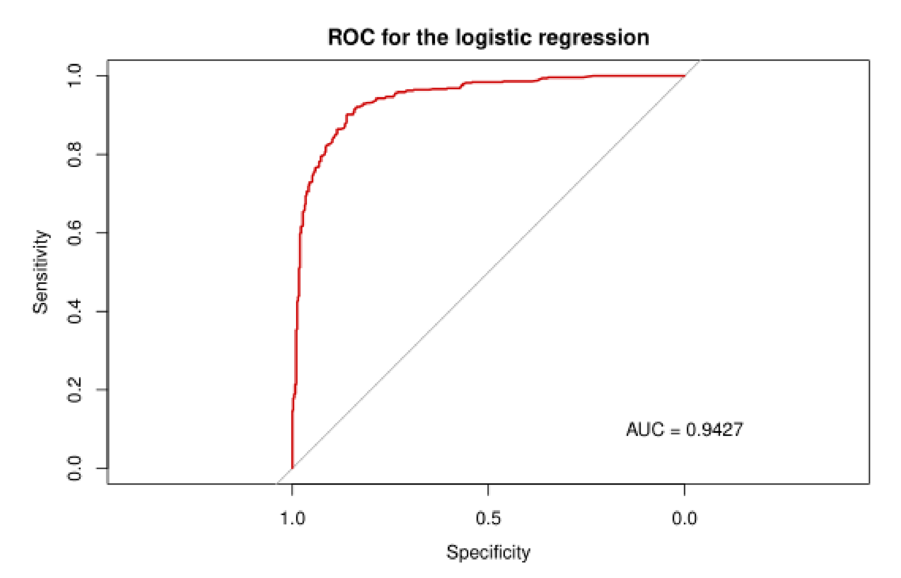
- Analysis of Binary Response

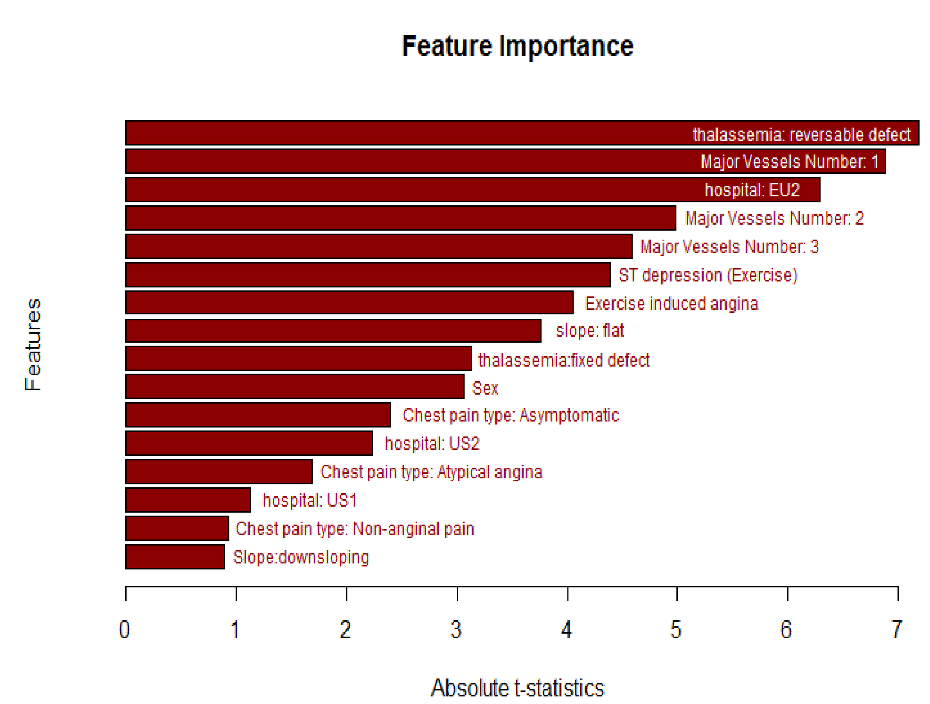
- Model Evaluation

Table 2: Summary Table of Analysis of Binary Response

Category	Beta(SE)	OR(95% CI)	P-value
Intercept			
-	-4.22 (0.56)	-	<0.001
Oldpeak			
-	0.54 (0.12)	1.71 (1.35, 2.19)	<0.001
Sex			
Female	-	-	-
Male	0.88 (0.29)	2.40 (1.38, 4.24)	0.002
cp			
Typical angina	-	-	-
Atypical angina	-0.81 (0.48)	0.44 (0.17, 1.14)	0.09
Non-anginal pain	-0.41 (0.44)	0.67 (0.28, 1.58)	0.35
Asymptonmatic	1.00 (0.42)	2.71 (1.20, 6.20)	0.02
exang			
no	-	-	-
yes	1.00 (0.25)	2.73 (1.68, 4.44)	<0.001
slope			
upsloping	-	-	-
flat	0.93 (0.25)	2.55 (1.57, 4.16)	<0.001
downsloping	0.37 (0.42)	1.45 (0.64, 3.28)	0.37
ca			
0	-	-	-
1	1.95 (0.28)	7.02 (4.08, 12.39)	<0.001
2	1.83 (0.37)	6.26 (3.11, 13.22)	<0.001
3	2.11 (0.46)	8.29 (3.51, 21.69)	<0.001
thal			
normal	-	-	-
fixed defect	1.19 (0.38)	3.29 (1.58, 7.04)	0.002
reversable defect	1.68 (0.23)	5.39 (3.42, 8.59)	<0.001
hospital			
EU1	-	-	-
EU2	3.02 (0.48)	20.40 (8.35, 55.45)	<0.001
US1	0.31 (0.27)	1.36 (0.80, 2.34)	0.26
US2	0.70 (0.32)	2.02 (1.09, 3.77)	0.03

- Model Evaluation



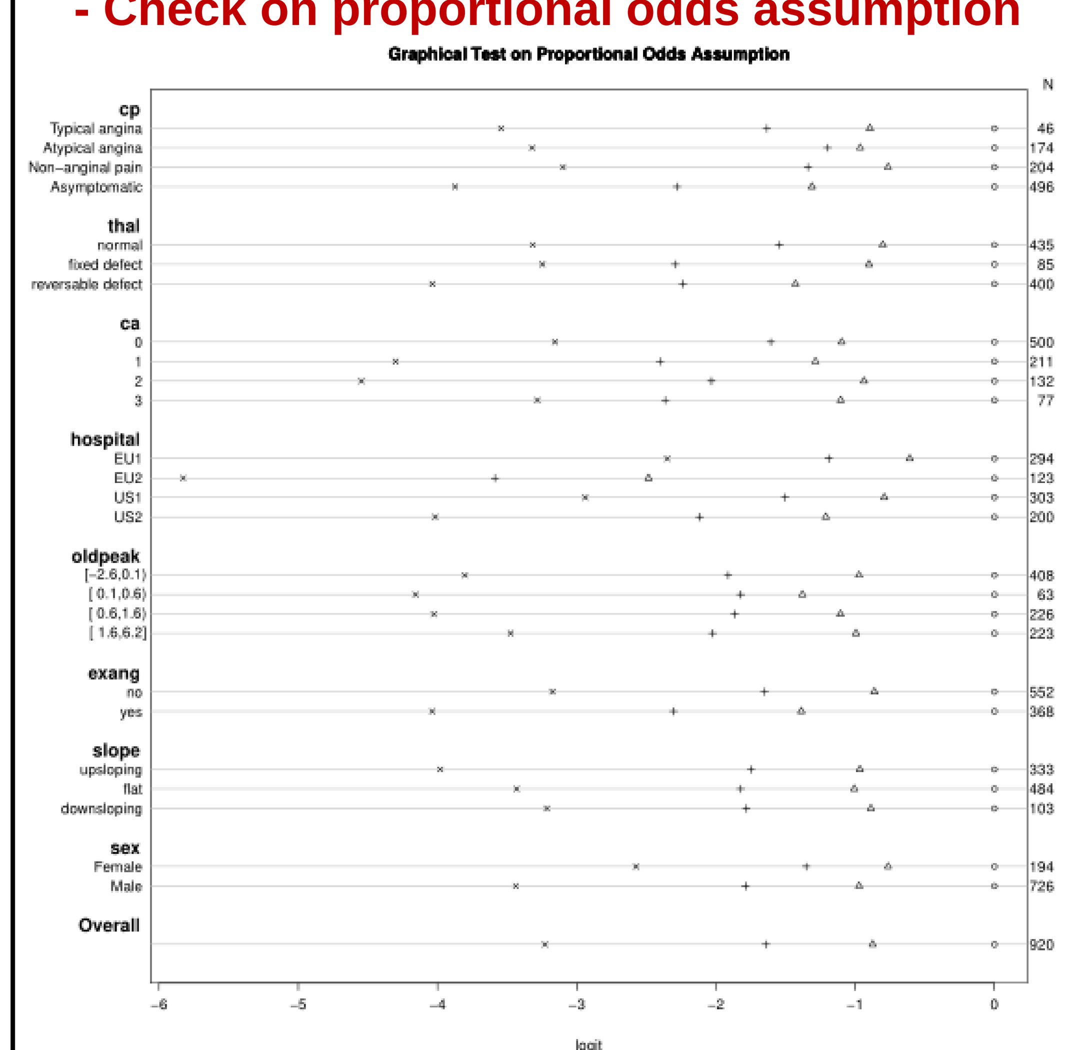


- Interpretation

According to Table 2, clinical features (chest pain type, thalassemia, number of major vessels), exercise related features (exercise induced angina, slope of the peak ST segment, ST depression induced by exercise), demographical characteristics (sex, hospital) are potential features useful in diagnosing presence of heart disease. The AUC of the fitted model is 0.943, indicating good accuracy.

- Analysis of Ordinal Response

- Check on proportional odds assumption



Assumption of proportional odds (PO) is both graphically and statistically. According to the figure in the left, distance between the symbols are more greatly different across levels of chest pain type (cp), thalassemia (thal), number of major vessels colored by fluoroscopy (ca), and hospital. While in other four dependent variables, alignments of the symbols are similar across levels. Also by nominal test, PO is violated on predictor cp, thal, ca and hospital (see Table 3).

Table 3: Summary Table of Nominal Test

	Df	logLik	AIC	LRT	Pr(>Chi)
	NA	-925.351	1890.701	NA	
cp	9	-909.921	1877.842	30.859	0.0003
thal	6	-910.518	1873.035	29.666	<0.0001
ca	9	-904.835	1867.671	41.031	<0.0001
hospital	9	-905.221	1868.442	40.259	<0.0001
oldpeak	3	-923.689	1893.379	3.323	0.345
exang	3	-922.287	1890.575	6.127	0.106
slope	6	-924.194	1900.388	2.314	0.889
sex	3	-923.747	1893.495	3.207	0.361

- Partial Proportional Odds Model

The following model relaxes proportional assumption and allows the effects of some covariates (cp, thal, ca and hospital, in this case) different across categories of severity of heart disease on the log odds scale, while the remain covariates keep assumption of PO.

- Fitted Model

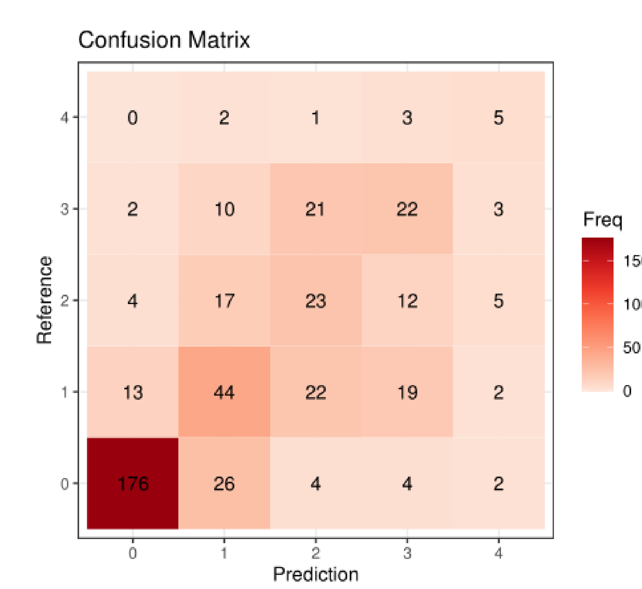
Table 4: Summary Table of Partial Proportional Model

Coefficients	OR (95%CI)	P-value
Interpret		
0|1.(Intercept)	39.47 (15.56-100.12)	<0.001
1|2.(Intercept)	66.27 (25.25-173.96)	<0.001
2|3.(Intercept)	83.33 (27.94-248.46)	<0.001
3|4.(Intercept)	386.23 (40.52-3681.17)	<0.001
Chest Pain Type(ref: typical angina)		
0|1.cpAtypical angina	2.48 (1.03-5.97)	0.042
1|2.cpAtypical angina	2.80 (0.97-8.04)	0.0559
2|3.cpAtypical angina	0.73 (0.21-2.54)	0.624
3|4.cpAtypical angina	1.26 (0.07-22.81)	0.873
0|1.cpNon-anginal pain	1.40 (0.64-3.08)	0.4
1|2.cpNon-anginal pain	1.29 (0.55-3.05)	0.561
2|3.cpNon-anginal pain	0.99 (0.37-2.68)	0.987
3|4.cpNon-anginal pain	0.97 (0.11-8.86)	0.978
0|1.cpAsymptomatic	0.39 (0.18-0.83)	0.015
1|2.cpAsymptomatic	0.76 (0.34-1.71)	0.509
2|3.cpAsymptomatic	0.89 (0.34-2.28)	0.801
3|4.cpAsymptomatic	1.07 (0.13-8.74)	0.946
Thalassemia (ref: normal)		
0|1.thalfixed defect	0.29 (0.14-0.58)	<0.001
1|2.thalfixed defect	0.36 (0.2-0.67)	0.001
2|3.thalfixed defect	0.92 (0.46-1.83)	0.802
3|4.thalfixed defect	0.22 (0.07-0.73)	0.0127
0|1.thalreversable defect	0.23 (0.15-0.35)	<0.001
1|2.thalreversable defect	0.47 (0.31-0.7)	<0.001
2|3.thalreversable defect	0.58 (0.36-0.92)	0.0207
3|4.thalreversable defect	0.87 (0.33-2.28)	0.772
Slope (ref: upsloping)		
slopeflat	2.01 (1.42-2.85)	<0.001
slopedownsloping	1.89 (1.13-3.19)	0.016
Sex (ref: Male)		
sexMale	2.24 (1.45-3.48)	<0.001

Coefficients	OR (95%CI)	P-value
Number of major vessels colored (ref: 0)		
0|1.ca1	0.18 (0.11-0.3)	<0.001
1|2.ca1	0.28 (0.18-0.43)	<0.001
2|3.ca1	0.66 (0.4-1.08)	0.101
3|4.ca1	0.62 (0.23-1.66)	0.339
0|1.ca2	0.17 (0.09-0.33)	<0.001
1|2.ca2	0.13 (0.08-0.22)	<0.001
2|3.ca2	0.25 (0.15-0.42)	<0.001
3|4.ca2	0.74 (0.27-2.06)	0.567
0|1.ca3	0.14 (0.06-0.31)	<0.001
1|2.ca3	0.12 (0.06-0.22)	<0.001
2|3.ca3	0.30 (0.16-0.55)	<0.001
3|4.ca3	0.10 (0.04-0.27)	<0.001
Hospital e(ref: EU1)		
0|1.hospitalEU2	0.07 (0.03-0.17)	<0.001
1|2.hospitalEU2	0.46 (0.27-0.79)	0.005
2|3.hospitalEU2	0.64 (0.36-1.15)	0.138
3|4.hospitalEU2	1.62 (0.56-4.72)	0.373
0|1.hospitalUS1	0.74 (0.46-1.2)	0.221
1|2.hospitalUS1	1.12 (0.69-1.82)	0.651
2|3.hospitalUS1	1.26 (0.73-2.17)	0.402
3|4.hospitalUS1	1.16 (0.49-2.75)	0.737
0|1.hospitalUS2	0.56 (0.32-0.96)	0.0354
1|2.hospitalUS2	0.96 (0.58-1.58)	0.859
2|3.hospitalUS2	1.25 (0.73-2.16)	0.418
3|4.hospitalUS2	1.53 (0.6-3.92)	0.371
ST depression induced by exercise		
oldpeak	1.75 (1.52-2.02)	<0.001
Exercise induced angina (ref: NO)		
exangyes	1.86 (1.34-2.58)	<0.001

- Model Evaluation

By dividing training set and validation set, the model accuracy is 61.09% (prediction result see confusion matrix below).



- Interpretation

Chest pain type and hospital are useful in predicting presence of heart disease, but have small power in predicting the severity. Similarly, feature thalassemia is also more useful in predicting relative mild heart disease. In comparison, features slope, sex, ST depression, number of major vessels and exercise induced angina have significant effect both on presence and severity of heart disease.

DISCUSSION

Using logistic regression model, prediction on presence of heart disease can attain very high accuracy (87.8%). If response of multiple classes is of interest, partial proportional model is used because PO assumption is violated. However, partial PO model make it difficult to interpret the coefficients, especially for our case. We may consider keep PO model with high predictivity even the PO assumption violated.

In heart disease diagnosis, if patient want to test the presence of heart disease, tests on number of major vessels (0-3) colored by fluoroscopy and also thalassemia are recommended (especially ca). Though their cost is relatively high, but they will provide highest power in predictivity. In addition, test on chest pain type is encouraged to performed, because is free of cost but with information. Plus, test on fasting blood sugar and resting electrocardiographic are not necessary. When severity of heart disease is of interest, tests on number of major vessels and thalassemia is still recommended.

CONCLUSION

Heart disease can be well-predicted by logistic regression and partial PO model.