STA303 FINAL REPORT

Li Quan Soh

Student Number: 1003357565

April 12, 2024

Introduction

Heart disease is recognized as one of the most prevalent health issues in the world, taking an estimated 17.9 million lives each year ("Cardiovascular diseases", n.d.). Studies have been done to determine biomarkers that can predict individuals at high risk of developing heart diseases. Currently, a number of health markers has been linked to increased risk of heart diseases including cholesterol level, triglyceride level, blood sugar level, blood pressure, and several protein biomarkers. Despite extensive research on this subject, many studies have primarily examined them individually. Furthermore, these studies have been conducted using small and selected samples. The key health markers examined in this study will be cholesterol level, blood sugar level and blood pressure. The purpose of this study is to examine the associations between the risk of developing heart disease and the key metabolic health markers in the context of potential confounders: age and sex. Additionally, this study is conducted using a sample size of 918 observations, and thus has bigger scale than many existing studies.

Methods

Exploratory Data Analysis

Data from a total of 918 observations were taken from the UCI Machine Learning Repository. This dataset has 12 variables: age, sex, chest pain type, resting blood pressure, fasting blood sugar level, resting ECG state, maximum heart rate, having exercise angina, oldpeak, slope of the peak exercise ST segment, and having heart disease. Literature reviews were conducted to select appropriate predictor of interests and confounders. Missing values and duplicate rows were removed from the data set. Exploratory data analysis was conducted on the predictors and response variables to ensure there are no weird patterns, such as summary statistics, histograms, boxplots and 2x2 contingency tables. Invalid data points were removed, and the categorical variables were factorized in preparation for modelling. The three assumptions of GLM were verified. Observations are independent as the likelihood of having heart disease is independent of another individual. Logistic model using logit link was chosen as the research question pertains to the odds of having heart disease given a set of health markers. Furthermore, the response variable is a binary nominal variable.

Modelling

Logistic regression model was fitted with all variables as the initial model. Multicollinearity was then assessed and visualized via *ggcorrplot*. Predictors with high correlations are noted for downstream analysis. 4 model selection techniques that were used to select significant variables included stepwise AIC, stepwise BIC, LASSO and Elastic Net. Mixing parameter $\alpha = 1$ and $\alpha = 0.5$ were used for LASSO and Elastic Net, respectively. Then, a logistic regression model was fitted for each model selection technique.

Diagnostics

The following diagnostics are performed for each of the 4 models. Multicollinearity in the regression models were assessed using Variance Inflation Factor (VIF). Variables with high VIF (> 5) may affect their significance and considered redundant, and thus will be removed. Influential points were assessed using DFFits and Cook's Distance. These influential points were noted but not removed. DFBetas was considered but ultimately not included in our diagnostics as many of our variables are categorical and thus the plots would be difficult to interpret.

Model Validation

10-fold cross-validation was used to assess the risk of overfitting and to obtain the Mean Squared Error (MSE). ROC Curve is used to illustrate the predictive power of the model by comparing the True Positive Rate (TPR) and False Positive Rate (FPR). The AUC score was also noted.

Final Model Selection

The final model was selected based on their variable selections, number of influential points, multicollinearity, MSE, and AUC score. Predictor of interests not selected in the chosen model are included in the final model as they are crucial to the study's objective. Diagnostics and validations are then performed on this final model to ensure it does not perform egregiously.

Results

172 observations were removed as they had invalid measurements in cholesterol level and blood pressure. Table 1.1 illustrates the summary statistics of the variables in the final model.

Table 1.1 Summary statistics of variables

Variables	Types of variables	Range/ Number of data in each level	Mean (if applicable)
Have Heart Disease (1: Yes, 0: No)	Binary nominal	0: 390 1: 356	
Resting Blood Pressure (mmHg)	Continuous	92 – 200	133
Cholesterol Level (mm/dL)	Continuous	85 – 603	237.0
Fasting Blood Sugar Group (1: ≥120mg/dL, 0: otherwise)	Binary Nominal	0: 621 1: 125	
Age (years)	Continuous	28 – 77	52.88
Sex (1: Male, 0: Female)	Binary Nominal	0: 182 1: 564	
Chest Pain Type (0: ASY, 1: NAP, 2: ATA, 3: TA)[1]	Polytomous Nominal	0: 370, 1: 169, 2: 166, 3: 41	
Have Exercise Angina (1: Yes, 0: No)	Binary Nominal	0: 459 1: 287	
ST Segment Slope: (1: Abnormal, 0: Normal)[2]	Binary Nominal	0: 349 1: 397	

[1] ASY: asymptomatic. NAP: Non-Anginal Pain. ATA: Atypical Angina. TA: Typical Angina. [2] Slope of the peak exercise ST segment [Abnormal: Flat or Downsloping, Normal: Upsloping]

Model selection process

Model with variables selected by stepwise-AIC (AIC-model) has 8 variables. Model with variables selected by stepwise-BIC (BIC-model) has 5 variables. Model with variables selected by LASSO (LASSO-model) and Elastic Net (Elastic-Net-model) both has the same 3 variables. Both AIC-model and BIC-model has similar number of influential points according to DFFITS and Cook's Distance plot. They also have similar MSE value (approx. $2e^{-4}$) and AUC score (0.93). Both LASSO-model and Elastic-Net-model shared similar number of influential points and AUC score (0.83). LASSO-model has slightly lower MSE value than Elastic-Net-model

(3.3e-4 and 5.2e-4 respectively). The variables in all of these models have acceptable VIF value of 1 – 1.3. VIF value below 4 suggests that the predictors are not really correlated with each other.

Final model

The final model is fitted with variables selected by stepwise BIC. Out of the 4 models, it is closest to a parsimonious model. It has a high AUC score, low MSE value, acceptable number of influential points, while selecting 5 variables. Since none of the 4 models selected the predictors of interest, they will be added to the final model as they are crucial in answering the research question. The final model consists of 8 predictors [Table 1.2]. None of the 3 metabolic health markers are significant in predicting whether someone has heart disease or not. However, the confounders (age and sex) are significant. It has 68 influential points according to the Cook's Distance plot (9% of sample size) [Figure 1.1]. It has MSE value of $2.4e^{-4}$, which means the model is a good fit for the data. An AUC score of 0.93 suggests that the model's performance in predicting individuals with or without heart disease is very good [Figure 1.2]. All the predictors have VIF value of 1 – 1.2, which suggests there is no concern for multicollinearity [Table 1.3]. Likelihood Ratio Test (LRT) is used to compare the final model to the full-variable model. The p-value of the test is 0.1 > α=0.05. This means that we cannot reject the null hypothesis and thus the full-variable model does not provide a significantly better fit to the data than the final model.

Table 1.2 Regression output

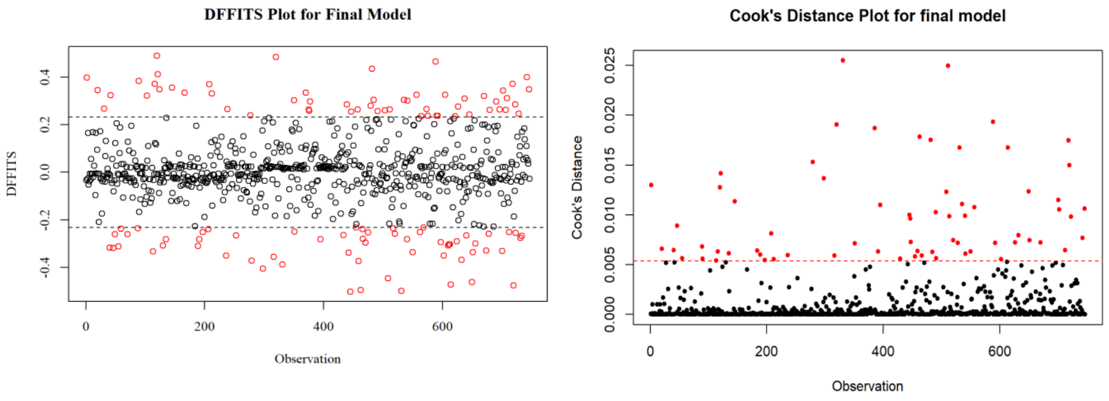
Variable	Coefficient Estimate	Std. Error	z value	Pr(> |z|)	Significance Code
Intercept	-7.06829	1.253264	-5.623	1.88e-08	***
Age	0.037042	0.013226	2.801	0.005100	**
Sex: Male	1.821966	0.305763	5.959	2.54e-09	***
CPT: NAP	-1.581053	0.293000	-5.396	6.81e-08	***
CPT: ATA	-1.751893	0.342856	-5.110	3.23e-07	***
CPT: TA	-1.568068	0.468562	-3.347	0.000818	***
RBP	0.013171	0.007123	1.849	0.064462	.
Cholesterol	0.002651	0.001969	1.346	0.178264	
FBS Level	0.195186	0.322459	0.605	0.544977	
EA: Yes	1.039939	0.253606	4.101	4.12e-05	***
ST Slope: Abnormal	2.686071	0.253289	10.605	< 2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. CPT: Chest Pain Type. NAP: Non-Anginal Pain. ATA: Atypical Angina. TA: Typical Angina. RBP: Resting Blood Pressure. FBS: Fasting Blood Sugar Level. EA: Have Exercise Angina. ST Slope: slope of the peak exercise ST segment.

Table 1.3 Multicollinearity table with VIF score

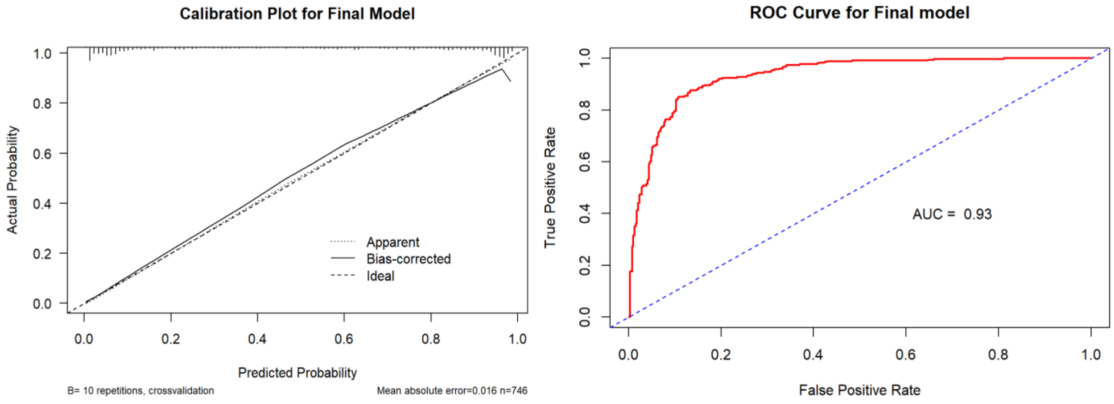
Variables (and factor levels)	Variance Inflation Factor
Age	1.090395
Sex	1.153915
Chest Pain Type 1:	1.188702
Chest Pain Type 2:	1.145991
Chest Pain Type 3:	1.161011
Resting Blood Pressure	1.096983
Cholesterol Level	1.054622
Fasting Blood Sugar Level	1.073707
Have Exercise Angina	1.126015
Slope of ST Segment	1.160116

Figure 1.1 Influential points in the final model



Left: DFFITS plot for final model (number of influential points = 113). Right: Cook's Distance plot for final model (number of influential points = 72).

Figure 1.2 Calibration plot and ROC Curve of the final model



Left: Calibration plot illustrating goodness of fit of the final model. MSE=0.016. Right: ROC Curve illustrating the predictive power of the final model (TPR vs FPR). AUC=0.93.

Discussion

The odd of developing heart diseases for individuals with abnormal fasting blood sugar level (>120mg/dL) compared to individuals with normal blood sugar level (≤120mg/dL) increases by 1.2 times. However, changes in cholesterol level and blood pressure neither increase or decrease the odd of developing heart diseases for an individual.

Results from this study indicated that none of the 3 metabolic health markers were significant (α = 0.05) in determining whether an individual has heart disease, contrary to existing research. Consequently, although the model has good predictive power and is a good fit for the data, the lack of significance in the predictors of interest may limit our understanding of the relationship between the key metabolic health markers and the probability of developing heart diseases. Additionally, the limitation also raises questions regarding the generalizability of this model to other populations or medical contexts. Literature review suggests that this limitation may be caused by small sample size or multicollinearity, but this is not applicable to our study since we do not have those issues.

Citation

fedesoriano. (2021). Heart Failure Prediction Dataset [Data set].Kaggle. https://www.kaggle.com/fedesoriano/heart-failure-prediction

Roeters van Lennep, J., Westerveld, H., Erkelens, D., & E van der Wall, E. (2002). Risk factors for coronary heart disease: Implications of gender. Cardiovascular Research, 53(3), 538–549. https://doi.org/10.1016/s0008-6363(01)00388-1

Shin, J., Ham, D., Shin, S., Choi, S. K., Paik, H.-Y., & Joung, H. (2019). Effects of lifestyle-related factors on ischemic heart disease according to body mass index and fasting blood glucose levels in Korean adults. PLOS ONE, 14(5). https://doi.org/10.1371/journal.pone.0216534

World Health Organization. (n.d.). Cardiovascular diseases. World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases

Tyroler, H. A. (1971). Blood pressure and cholesterol as coronary heart disease risk factors. Archives of Internal Medicine, 128(6), 907–914. https://doi.org/10.1001/archinte.1971.00310240061007