Automatic report

Escola Nacional de Saúde Pública

A. Peralta-Santos

16 de setembro de 2023

Índice

# Introduction

This is a smart report generated with R on the date 16 September 2023. We used a dataset in the MLDataR called Heart Disease. In this report we will combine what we have learnt so far in the R course.

# Methods

## Study Design and Data Source

This study employs a retrospective analysis design using the Heart Failure Dataset from MLDataR. This dataset is an open-source, publicly available resource that compiles data from heart failure patients, offering comprehensive insights into clinical, demographic, and laboratory features that may be associated with heart failure outcomes.

## Data Collection

The Heart Failure Dataset from MLDataR was downloaded and parsed into a format amenable to our analysis pipeline.

We created a logistic regression model, with the following equation

# Results

## Descriptive Statistics

Our data set as 918 patients with a mean age of 53.5

1. \*\* Patient Characteristics \*\*

|  | | | **Heart Disease** | |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Overall**, N = 9181 | **No**, N = 4101 | **Yes**, N = 5081 | **p-value**2 |
| **sex** | 918 |  |  |  | <0.001 |
| F |  | 193 (21%) | 143 (35%) | 50 (9.8%) |  |
| M |  | 725 (79%) | 267 (65%) | 458 (90%) |  |
| **age** | 918 | 54 (47, 60) | 51 (43, 57) | 57 (51, 62) | <0.001 |
| **resting\_ecg** | 918 |  |  |  | 0.004 |
| LVH |  | 188 (20%) | 82 (20%) | 106 (21%) |  |
| Normal |  | 552 (60%) | 267 (65%) | 285 (56%) |  |
| ST |  | 178 (19%) | 61 (15%) | 117 (23%) |  |
| **resting\_bp** | 918 | 130 (120, 140) | 130 (120, 140) | 132 (120, 145) | <0.001 |
| **cholesterol** | 918 | 223 (173, 267) | 227 (197, 267) | 217 (0, 267) | <0.001 |
| 1n (%); Median (IQR) | | | | | |
| 2Pearson's Chi-squared test; Wilcoxon rank sum test | | | | | |

## Model

We fitted a logistic model (estimated using ML) to predict heart\_disease with age, sex, resting\_ecg, resting\_bp and cholesterol (formula: heart\_disease ~ age + as.factor(sex) + relevel(factor(resting\_ecg), ref = “Normal”) + resting\_bp + cholesterol). The model’s explanatory power is moderate (Tjur’s R2 = 0.20). The model’s intercept, corresponding to age = 0, sex = F, resting\_ecg = LVH, resting\_bp = 0 and cholesterol = 0, is at -4.46 (95% CI [-5.80, -3.15], p < .001). Within this model:

* The effect of age is statistically significant and positive (beta = 0.06, 95% CI [0.04, 0.07], p < .001; Std. beta = 0.55, 95% CI [0.39, 0.71])
* The effect of sex [M] is statistically significant and positive (beta = 1.47, 95% CI [1.10, 1.86], p < .001; Std. beta = 1.47, 95% CI [1.10, 1.86])
* The effect of relevel(resting ecg, ref = “Normal”)LVH is statistically non-significant and positive (beta = 0.15, 95% CI [-0.23, 0.53], p = 0.451; Std. beta = 0.15, 95% CI [-0.23, 0.53])
* The effect of relevel(resting ecg, ref = “Normal”)ST is statistically non-significant and positive (beta = 0.21, 95% CI [-0.19, 0.61], p = 0.306; Std. beta = 0.21, 95% CI [-0.19, 0.61])
* The effect of resting bp is statistically significant and positive (beta = 8.48e-03, 95% CI [2.64e-04, 0.02], p = 0.044; Std. beta = 0.16, 95% CI [4.88e-03, 0.31])
* The effect of cholesterol is statistically significant and negative (beta = -3.76e-03, 95% CI [-5.27e-03, -2.30e-03], p < .001; Std. beta = -0.41, 95% CI [-0.58, -0.25])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.

The plot bellow is the relationship between age and resting blood pressure by heart disease

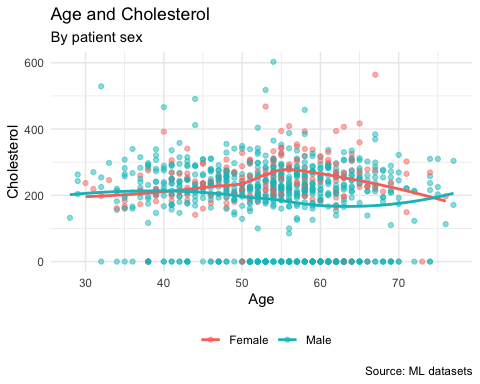


Table 1: **Odds Ratio for Heart Failure**

*Based on a binomial GLM adjusted for confounding*

| Risk factor | Odds ratio | p value | Low 95%CI | High 95%CI |
| --- | --- | --- | --- | --- |
| Intercept | 0.012 | 0.000 | 0.003 | 0.043 |
| Age | 1.060 | 0.000 | 1.042 | 1.078 |
| Male | 4.355 | 0.000 | 3.001 | 6.408 |
| LVH ECG, ref. Normal*1* | 1.157 | 0.451 | 0.792 | 1.692 |
| ST ECG, ref. Normal*1* | 1.232 | 0.306 | 0.828 | 1.844 |
| Blood preassure | 1.009 | 0.044 | 1.000 | 1.017 |
| Cholesterol | 0.996 | 0.000 | 0.995 | 0.998 |
| *1*ECG was performed with a resting patients | | | | |
| Source: MLDataR dataset | | | | |

# Discussion

This study utilizes a logistic model to predict heart disease using multiple factors such as age, sex, resting electrocardiogram (ECG) results, resting blood pressure (BP), and cholesterol levels. The logistic model is a form of regression analysis used when the dependent variable is a binary outcome - in this case, the presence or absence of heart disease.

The logistic model has a moderate explanatory power with a Tjur’s R^2 of 0.20. This means that approximately 20% of the variability in the presence of heart disease can be explained by the variables included in this model.

The model’s intercept is at -4.46 with a 95% Confidence Interval (CI) ranging from -5.80 to -3.15. This indicates that at the baseline level (i.e., when age, sex, resting ECG, resting BP, and cholesterol are zero), the log-odds of having heart disease is -4.46. This value is statistically significant, with a p-value less than 0.001.

Among the predictors, age and sex are statistically significant and positively associated with heart disease. The positive beta values of 0.06 and 1.47 respectively suggest that with each additional year of age, and being male, the log-odds of heart disease increase (1).

Resting ECG, when relevelled to “Normal”, shows two categories LVH (Left Ventricular Hypertrophy) and ST (ST wave abnormality) are not statistically significant in predicting heart disease. This means the odds of heart disease do not significantly change with these specific ECG results as compared to a normal ECG result (2).

Resting BP is statistically significant and positively associated with heart disease. The positive beta value of 0.00848 suggests that for each unit increase in resting BP, the log-odds of having heart disease slightly increase.

It’s important to note that these results should be interpreted in the context of the data used, and any clinical implications should be considered cautiously. Other factors not included in the model may also be significant predictors of heart disease, and the relationship between cholesterol and heart disease may need further investigation to understand why it is negatively associated in this model. Additionally, the logistic model, like any other model, is a simplification of reality and should be used as one piece in a larger puzzle when making predictions about heart disease.

# References

1. Cowin P, Rowlands TM, Hatsell SJ. Cadherins and catenins in breast cancer. Current opinion in cell biology. 2005;17(5):499–508.

2. Litaker D, Koroukian SM, Love TE. Context and healthcare access: looking beyond the individual. Medical care. 2005;531–40.