Assignment 2 - Data Analysis 2 and Coding with R

David Utassy

01/01/2021

Introduction

In this project I address the question, how to predict from some measurable variables whether someone has heart disease. This is an important question as according to the World Health Organization, cardiovascular diseases are the leading cause of death globally. Of course, this topic requires broad knowledge on the topic. To get familiar with the problem I recommend this article, which describes the theory in an understandable way.

Data

This dataset is hosted on Kaggle (Heart Disease Dataset)[https://www.kaggle.com/johnsmith88/heart-disease-dataset], it is from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long each. The original data source is the UCI Machine Learning Repository, which contains 76 attributes, but all published experiments refer to using a subset of 14 of them, so as mine.

In the dataset, every row is a person with the given parameters in each column. The number of the observations is 1025 originally, which will decrease to 1000 after omitting some with missing values. This dataset was pretty clean already, only a few modifications were needed to make the analysis more convenient.

The most challenging part of this project is to understand somehow the variables I have. In order to get the reader somewhat familiar with them, I introduce them in a few sentences in an easy to understand form. For further details, I recommend the previously mentioned online sources.

1. Target: (binary) whether the patient has heart disease or not (originally 0: yes, 1: no) but I made a modification, therefore 0: no, 1: yes
2. Age: (quantitative) the age of the patient
3. Gender: (binary) The gender of the patient (0:female, 1: male)
4. Chest pain type (qualitative)
    - Value 0: asymptomatic (no symptoms)
    - Value 1: atypical angina
    - Value 2: pain without relation to angina
    - Value 3: typical angina
5. Resting blood pressure in mm Hg (qualitative)
6. Serum cholesterol in mg/dl: The level of cholesterol in the blood
7. Fasting blood sugar > 120 mg/dl: (binary) Whether the level of blood sugar is under or above the threshold (0: under, 1:above)
8. Resting electrocardiographic results (qualitative)
    - Value 0: probable left ventricular hypertrophy

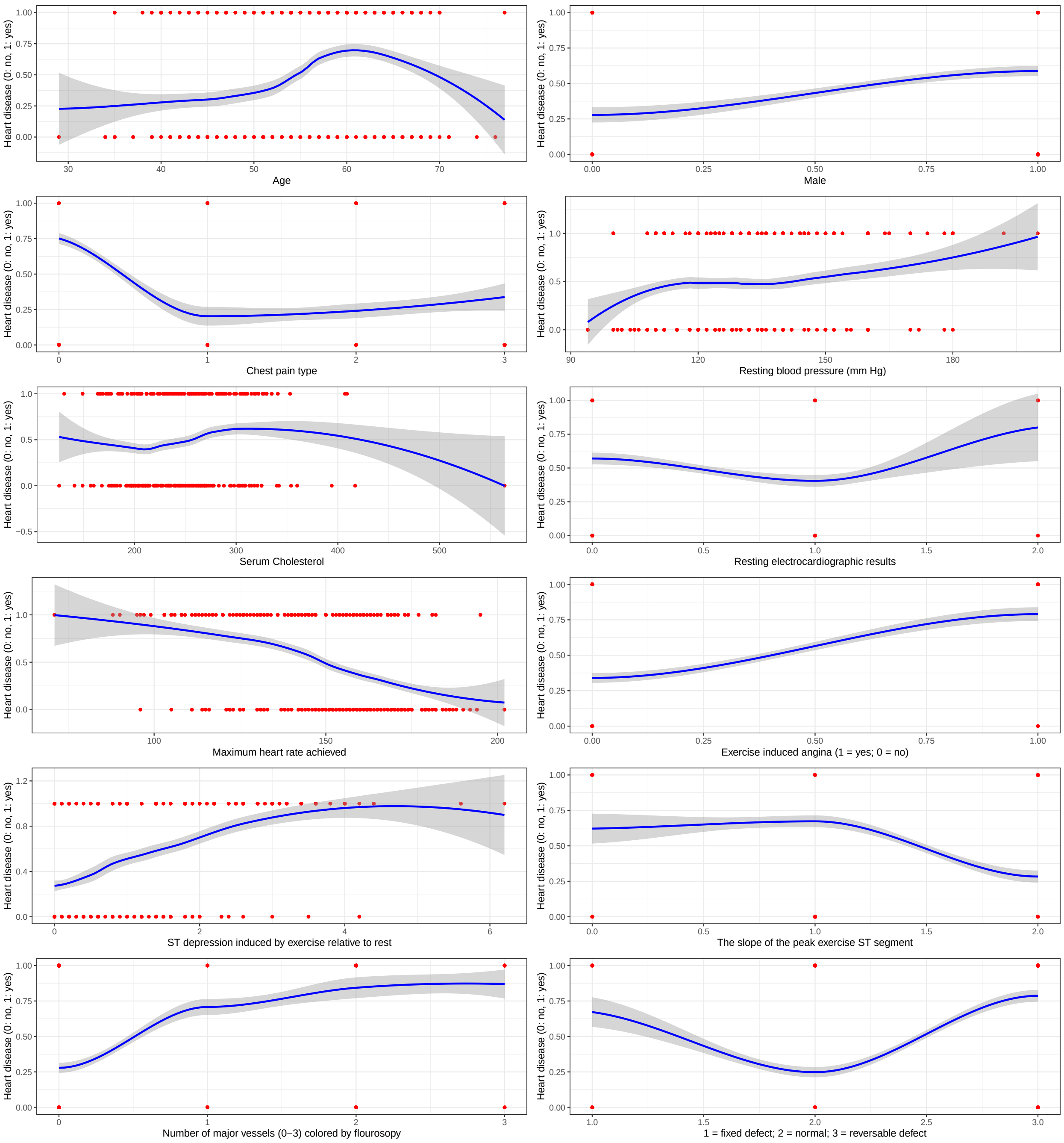
    - Value 1: normal
    - Value 2: abnormalities in the T wave or ST segment

9. Maximum heart rate achieved (qualitative)
10. Exercise-induced angina: (binary) whether pain occurred during exercise (0: no, 1: yes)
11. Oldpeak: (quantitative) Decrease of the ST segment during exercise according to the same one on rest. (0 is optimal)
12. The slope of the peak exercise ST segment: (qualitative) the slope of a specified line segment on the ECG
    - Value 0: descending
    - Value 1: flat
    - Value 2: ascending
13. Number of major vessels (0-3) colored by fluoroscopy: (qualitative, ordered) Number of narrow vessels, ideally should be 0.
14. Results of the blood flow observed via the radioactive dye
    - Value 1: fixed defect (no blood flow in some part of the heart)
    - Value 2: normal blood flow
    - Value 3: reversible defect (a blood flow is observed but it is not normal)

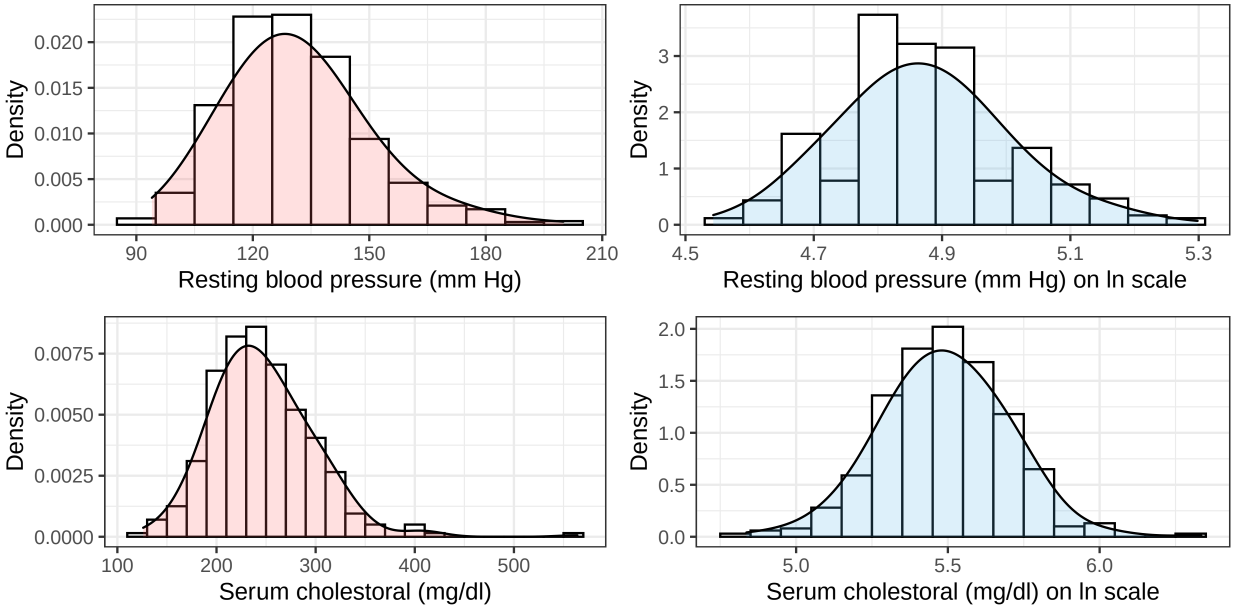
variable	mean	median	std	min	max	skew
Heart disease (0: no, 1: yes)	0.49	0.0	0.50	0	1.0	0.03
Age	54.61	56.0	9.04	29	77.0	-0.26
Gender (0:female, 1:male)	0.69	1.0	0.46	0	1.0	-0.83
Resting blood pressure (mm Hg)	131.59	130.0	17.71	94	200.0	0.74
Cholesterol level (mg/dl)	247.00	240.5	51.70	126	564.0	1.06
Fasting blood sugar > 120 mg/dl (0: under, 1:above)	0.15	0.0	0.35	0	1.0	2.01
Maxium heart rate during the stress test	148.98	152.0	23.08	71	202.0	-0.50
Exercise-induced angina (0: no, 1: yes)	0.34	0.0	0.47	0	1.0	0.68
Oldpeek	1.09	0.8	1.18	0	6.2	1.18
Number of main blood vessels	0.70	0.0	0.94	0	3.0	1.12

In order to get to know my variables, I have made some basics histograms. From those, it was visible that some variables might need a log transformation. In the appendix, we can see the result of some possible transformations, but in the end, I decided to stay with the original scales as the modifications probably would not pay off.

The table above contains the summary statistics of the variables I used in the analysis. To highlight some patterns in our data, the mean of the presence of heart disease is almost 50%, which could be good for prediction, however, 70% of the observations are male, which is not representing the population well. The age variable also shows that the sample is not really representing a whole population, but probably a subset of people who has to do something with cardiovascular problems. The other variables are more like medical terms, therefore I would not go into detailed preconceptions.


Heart disease (0: no, 1: yes)
1.00
0.75
0.50
0.25
0.00
30
40
50
60
70
Age
Heart disease (0: no, 1: yes)
0.00
0.25
0.50
0.75
1.00
Male
Heart disease (0: no, 1: yes)
0
1
2
3
Chest pain type
Heart disease (0: no, 1: yes)
1.0
0.5
0.0
90
120
150
180
Resting blood pressure (mm Hg)
Heart disease (0: no, 1: yes)
−0.5
200
300
400
500
Serum Cholesterol
Heart disease (0: no, 1: yes)
0.0
0.5
1.0
1.5
2.0
Resting electrocardiographic results
Heart disease (0: no, 1: yes)
100
150
200
Maximum heart rate achieved
Heart disease (0: no, 1: yes)
Exercise induced angina (1 = yes; 0 = no)
Heart disease (0: no, 1: yes)
1.2
0.8
0.4
0.0
0
2
4
6
ST depression induced by exercise relative to rest
Heart disease (0: no, 1: yes)
The slope of the peak exercise ST segment
Heart disease (0: no, 1: yes)
Number of major vessels (0−3) colored by flourosopy
Heart disease (0: no, 1: yes)
1.0
1.5
2.0
2.5
3.0
1 = fixed defect; 2 = normal; 3 = reversable defect

Appendix



Possible variable transformations