Project 1 - Heart Disease data

Andreas Artemiou

1 Dataset

The dataset consists of 4 files. Each file contains a different set of observations collected at 4 different locations on a heart disease study. There are 75 variables measured for each patient which are the following:

- id: patient identification number
- ccf: social security number (I replaced this with a dummy value of 0)
- age: age in years
- sex: sex (1 = male; 0 = female)
- painloc: chest pain location (1 = substernal; 0 = otherwise)
- painexer (1 = provoked by exertion; 0 = otherwise)
- relrest (1 = relieved after rest; 0 = otherwise)
- pncaden (sum of 5, 6, and 7)
- cp: chest pain type - Value 1: typical angina - Value 2: atypical angina - Value 3: non-anginal pain - Value 4: asymptomatic
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- htn
- chol: serum cholestoral in mg/dl
- smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
- cigs (cigarettes per day)
- years (number of years as a smoker)

- fbs: (fasting blood sugar ¿ 120 mg/dl) (1 = true; 0 = false)
- dm (1 = history of diabetes; 0 = no such history)
- famhist: family history of coronary artery disease (1 = yes; 0 = no)
- restecg: resting electrocardiographic results - Value 0: normal - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of ¿ 0.05 mV) - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- ekgmo (month of exercise ECG reading)
- ekgday(day of exercise ECG reading)
- ekgyr (year of exercise ECG reading)
- dig (digitalis used during exercise ECG: 1 = yes; 0 = no)
- prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
- nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
- pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
- diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)
- proto: exercise protocol 1 = Bruce, 2 = Kottus, 3 = McHenry, 4 = fast Balke, 5 = Balke, 6 = Noughton, 7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!), 8 = bike 125 kpa min/min, 9 = bike 100 kpa min/min, 10 = bike 75 kpa min/min, 11 = bike 50 kpa min/min, 12 = arm ergometer
- thaldur: duration of exercise test in minutes
- thaltime: time when ST measure depression was noted
- met: mets achieved
- thalach: maximum heart rate achieved
- thalrest: resting heart rate
- tpeakbps: peak exercise blood pressure (first of 2 parts)
- tpeakbpd: peak exercise blood pressure (second of 2 parts)

- dummy
- trestbpd: resting blood pressure
- exang: exercise induced angina (1 = yes; 0 = no)
- xhypo: (1 = yes; 0 = no)
- oldpeak = ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment - Value 1: upsloping - Value 2: flat - Value 3: downsloping
- rldv5: height at rest
- rldv5e: height at peak exercise
- ca: number of major vessels (0-3) colored by flourosopy
- restckm: irrelevant
- exerckm: irrelevant
- restef: rest raidonuclid (sp?) ejection fraction
- restwm: rest wall (sp?) motion abnormality, 0 = none, 1 = mild or moderate, 2 = moderate or severe, 3 = akinesis or dyskmem (sp?)
- exeref: exercise radinalid (sp?) ejection fraction
- exerwm: exercise wall (sp?) motion
- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
- thalsev: not used
- thalpul: not used
- earlobe: not used
- cmo: month of cardiac cath (sp?) (perhaps "call")
- cday: day of cardiac cath (sp?)
- cyr: year of cardiac cath (sp?)

- num: diagnosis of heart disease (angiographic disease status) - Value 0: ¡ 50% diameter narrowing - Value 1: ¿ 50% diameter narrowing (in any major vessel: the next 10 variables)
- lmt
- ladprox
- laddist
- diag
- cxmain
- ramus
- om1
- om2
- rcaprox
- rcadist
- lvx1: not used
- lvx2: not used
- lvx3: not used
- lvx4: not used
- lvf: not used
- cathef: not used
- junk: not used

Important note: I adjusted the datasets, to work with R. I hope all is working well. Make sure early in the project to make sure that R is reading the data correctly. Also, some variables are missing from every patient so you may consider deleting them to make sure you have only useful information left in the data.frame you use.

2 Research questions

You need to answer the following questions:

- Perform a test to see if for different exercises protocols (variable: "proto") there is equality of means for the vector of variables: (chol, thaldur, thaltime, met, thalach, thalrest, tpeakbps, tpeakbpd, trestbpd, oldpeak, rldv5, rldv5e)
- Use a multivariate regression model to estimate the coefficients for regressing (chol, thaldur, thaltime, met, thalach, thalrest, tpeakbps, tpeakbpd, trestbpd, oldpeak, rldv5, rldv5e) on variables proto, restecg, dig, prop, nitr, pro, diuretic
- Use a dimension reduction technique to see if you can identify which of all the variables affect the diagnosis of the heart disease the most (diagnosis is variable "num"). Exclude dummy variables and variables that are missing from every patient.
- Use discriminant analysis/classification tools to see if you can built a classification rule to distinguish patients based on their disease. Evaluate the models (if you build more than one). Compare with the most important variables you identified in dimension reduction question above.

3 General Project Guidelines

Imagine that you have applied for a job as a statistical consultant and you will get the job if you perform well on this report. Therefore make sure that the report is well presented.

3.1 Deliverables

You should submit

- a report (as .pdf file) by email,
- your code (as a .txt or .R file) by email
- cover page with the names and signatures of the students in the group and how the work was allocated between group members. This should be turned in by hand on the due date in class.

Your grade will be 75% on your report and 25% on your presentation. Cover page is important to show that you have done the work together and the split of work was mutually agreed. I will not mark a project without the cover page submitted.

Due date: April 29th, by 3pm.

3.2 Report

The project report should have 3 Chapters:

- Introduction and exploratory data analysis for the variables.
- Analysis to answer each research question (use subsections for each research question)
- Summary - Conclusions

There is no page limit (because you can write the same report using 30 pages with small figures or 60 pages with large figures), but I expect about 2-3 pages for each research question (some may take longer if the output is longer). Try not to do always the same thing so that it is not boring. For example you have to do plots for 5 variables. Don't do 5 histograms. Do some boxplots or scatterplots as well.

Your report grade will be: 50% on the appropriate analysis being used explained and tested, presentation of your findings, 25% neatness and organization in the report and finally 25%, creativity and initiative.

- All Figures and Tables should be numbered.
- Each Chapter should start in a new page.
- R output or R code should be boxed and numbered as well (output is needed but I do not encourage you to put too much code in there as the code will be sent in a different file).
- The code needs to have appropriate comments for me to understand.

3.3 Group work

This is a group work. Part of your evaluation will be on how you communicate as a team. I do not mind how you split the work (as long as you all agree with the split). Make sure you set clear tasks early and stay on schedule. Set important meetings well before hand, so that you know well ahead

what is the deadline. It is important that all members of the group stay on schedule with their tasks and discuss things between them on meetings.

If a group fails to communicate appropriately or if one person is not following the group schedule, it is important that you let me know as soon as possible in writing. If we can't resolve the issue, then the members of the group will not necessarily get the same grade on the project.