MATH5824 Assessed Practical

Robert G Aykroyd

3/13/23

Practicalities

Released on Tuesday 14th March 2023. Practical sessions are planned in the week 20-24 March where you can work on the assessment while I am present to help – check your timetable for the time and location. Note that it is not currently clear if the potential UCU strike action planned over the next few days will affect these Practical Sessions. Please check the module webpage on Minerva for the latest information.

Your task is to fit a suitable model to a data set to investigate how coronary heart disease depends on two explanatory variables (blood pressure and cholesterol) and write a report describing your analysis and conclusion. Some hints are included below to help you carry out the analysis.

Please follow these instructions when writing your report:

- Your report must start with a short summary explaining your conclusions and interpreting your model in language suitable for a nonspecialist.
- Your report must be no longer than eight pages with a *standard* layout. Put your name, student number, report title, and date of submission at the top of page 1. Do not include a separate title page. Anything on page nine or later will not be marked.
- You must include a completed academic integrity form with your submission but that does not count towards the page limit.
- Since space is limited, focus your discussions on the essentials and think about what is most important to include in your report. Do not include R code or output in the body of your report, but include your code as a separate appendix (this will not count towards the eight-page limit).
- Submit your completed practical report using the links in the *Learning Resources / Assessed Practical* folder in Minerva by **12 noon on Friday 31st March**. Late work will be penalized by 5% of the available marks for each calendar day, or part day, it is late.

- The marking will be done in Gradescope and Turnitin will check for plagiarism. The deadline applies to submission to both Gradescope and Turnitin. That is your submission is only considered on-time if you submit the report to both of Gradescope and Turnitin on-time. If one of them is late, or not submitted at all, then it will considered as a late submission, or as a non-submission.

Background

The table below gives the number (Y) of men diagnosed as having coronary heart disease (CHD) in an American study of 1325 men. The serum cholesterol level (CHOL) and blood pressure (BP) were recorded for each man. Cholesterol values are represented as rows and blood pressure as columns. The variables BP and CHOL are each reported in one of four categories, giving a 4×4 cross-classified table in which each cell contains the number Y of men with CHD and the number M of men examined. For example, of the 118 men with Normal BP and Normal CHOL, 3 were diagnosed as having coronary heart disease.

Table 1: Results from a study into male coronary heart disease

CHOL	BP Normal	BP Elevated	BP Stage I	BP Stage II
Normal	3/118	4/122	3/52	4/25
Medium risk	3/89	3/102	1/41	3/24
High risk	9/126	12/219	7/73	7/48
Dangerous	8/75	11/112	11/56	12/43

These data can be obtained from the `chd.txt` file which is available online at:

- rgaykroyd.github.io/MATH3823/Datasets/*filename.ext* where *filename.ext* is the stated filename with extension.

Note that you may need to explicitly declare as factors any qualitative variables, but you are not expected to declare any as ordered.

The task

Fit a suitable model to this dataset to explain how CHD depends on the two explanatory variables. Your final model should balance goodness of fit and model complexity. That is, fit the data well but with relatively few parameters – a parsimonious model.

It may be useful to bear in mind the following considerations when carrying out your analysis:

- using appropriate plots in an exploratory analysis,
- using analysis of deviance tables and residual plots to help choose a suitable model,

- combining factor levels when there is little difference between them,
- replacing a qualitative factor by a single quantitative variable taking simple values.

This is not meant to be an exhaustive list and so you may want to also follow some of your own ideas.

R hints:

- If `x` is a vector in R, then `(x==3.4)` is a logical vector of the same length as `x`, which equals `TRUE` if `x=3.4` and `FALSE` otherwise. Other logical vectors include `(x>3.4)`, `(x>=3.4)`, `(x<3.4)`, `(x<=3.4)`, and `(x!=3.4)` (for not equal).
- When used in arithmetic expressions, logical vectors assume the values 1 when TRUE, and 0 when FALSE.
- For example, suppose `x` is a factor taking the values `x=c(1,2,3,4,5)`, and suppose we wish to combine levels 3 and 4 together into a new factor with new levels 1, 2, 3, and 4. This can be achieved by defining a vector `xnew`

```
xnew=-1*(x==1)+2*(x==2)+3*(x==3)+ 3*(x==4)+4*(x==5)
new = as.factor(xnew) # check your result
```

  The second line makes `xnew` into a factor.
- If a factor `x` takes levels 1:n (consecutive integers starting at 1), it can be turned back into a quantitative variable (i.e. a vector) by the command `x = as.numeric(x)`
- If `a, x, y` are vectors of the same length, a plot of `y`*vs*.`x` labelled by the values of `a` can be produced by the command `plot(x,y, pch=as.character(a))`

Additional level 5 component to the practical

Data were collected from 25 car engines. For each engine, two variables were recorded:

- `size`: size of engine, and
- `wear`: an index of wear.

The data file `engine.txt` is available online at:

- rgaykroyd.github.io/MATH5824/Datasets/*filename.ext* where *filename.ext* is the stated filename with extension.

Your task is to plot wear against size and fit some cubic smoothing splines with different values of the smoothing parameter λ. You should discuss how the fit changes with the smoothing parameter in terms of visual behaviour and comment briefly on how the fit behaves as the smoothing parameter tends to zero and infinity. Use your judgement to suggest a range of suitable values for the smoothing parameter and suggest which single value you think is best.

When investigating different smoothing parameters, it may be helpful to vary λ by a factor of 10 between successive fits.

The written version should be about **two or three pages**, plus your **R** code included as an appendix. Please write this up as a separate report from your Level 3 report, and submit it separately on Minerva by the same deadline as the Level 3 part, **12 noon on Friday 31st March** (there will be separate links to submit this report).