Data analysis with R (PUHR11103) - Assignment instructions

2024

Value: 100% of overall course mark

Date handed out: Friday 2nd February 2024

Date to be submitted: Monday 19th February 2024 – 2pm (UK time)

Maximum Length: 2500 words (text, excluding R code and outputs), 25 pages (excluding the cover page)

Dataset

We will be exploring a subset of the data from a prospective cohort study on residents of the town of Framingham, Massachusetts. The study looks at the risk factors for coronary heart disease (CHD).The dataset provides the values of 15 variables for 4,238 patients. There are 14 risk factors, for which the information was obtained at the first examination cycle, and one outcome variable, which is whether the participant has ever had any type of coronary heart disease by the end of the study.

The dataset is provided in the comma-separated values format file **framingham_2024.csv** in the Learn page. The variables are described in the table below.

Variable	Description
sex	Sex: 1=male, 2=female
age	Age of the participant at exam (years)
currentSmoker	Current smoker: 1=Yes, 0=No
cigsPerDay	Number of cigarettes that the respondent smoked on average in one day
BPMeds	Use of Anti-hypertensive medication at exam: 1=Yes, 0=No
prevalentStroke	Respondent had a stroke previously: 1=Yes, 0=No
prevalentHyp	Prevalent Hypertensive. Subject was defined as hypertensive if treated or if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg: 1=Yes, 0=No
diabetes	Diabetic according to criteria of first exam treated or first exam with casual glucose of 200 mg/dL or more: 1=Yes, 0=No
totChol	Serum total cholesterol level (mg/dL)
sysBP	Systolic blood pressure (mean of last two of three measurements) (mm Hg)
diaBP	Diastolic blood pressure (mean of last two of three measurements) (mm Hg)
BMI	BMI (kg/m^2)
heartRate	Heart rate (Ventricular rate) (BPM)
glucose	Casual serum glucose level (mg/dL)
anyCHD	Outcome: Occurrence of any Coronary Heart Disease (Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease: 1=Yes, 0=No

Project goals

Use these data to investigate the following questions, and write a report summarising your findings by using the R Markdown template provided:

1. We want to study the association between the occurrence of any coronary heart disease and two risk factors:

    a. What is the relative risk of occurrence of any coronary heart disease given the smoking status at the first examination?

    b. How is occurrence of any coronary heart disease associated with age category? Hint: The American Heart Association (AHA) reports that the incidence of cardiovascular disease (CVD) in US men and women is different between age groups 40–59 years, 60–79 years, and above the age of 80.

2. We want to study the association between risk factors:

    a. Is systolic blood pressure associated with smoking status?

    b. Is prevalence of hypertension associated with smoking status?

3. How much do the following risk factors contribute to explaining occurrence of coronary heart disease? Are these contributions different between men and women? - total cholesterol, smoking status, age category, systolic blood pressure, BMI, and prevalence of hypertension

Approach to Reporting

You should follow the sections of the template.

- Data preparation
    - Prepare your data for analysis (e.g. create a working dataset, recode variables as factors, change factor levels, etc.).
- Exploratory Data Analysis
    - Explain how you are exploring the dataset or relevant variables, and why you are using the different graphs and tables. Explore the dataset informed by the project goals (e.g. create relevant plots, summary statistics, contingency tables). Comment on any patterns you see in the exploratory data analysis.
- Data analysis plan
    - Describe how you plan to answer the questions based on the variable types. Define your null and alternative hypotheses and set your significance level.
- Assumptions
    - Explain which assumptions for which methods you are checking. You can refer to outputs from the Exploratory data analysis step if they are relevant. Check the assumptions for the methods described in your Data analysis plan. Comment on whether the assumptions are met, and if not, what your alternative is for the analysis. **We do not expect you to check the assumptions for logistic regression as this is not covered in the course**.
- Carry out analysis
    - Perform the analyses mentioned above. If necessary, explain any extra step taken during the analysis.
- Interpretation
    - Explain what each analysis tells you about the variables of interest, and how they answer the questions from the project goals. Remember to make use of all the information available (plots, tables, p-values, confidence intervals, odds-ratio, relative risk, etc.).

- Conclusion
  - Write a short paragraph to summarise the conclusions from this analysis.

If you cannot use RMarkdown, it is possible to write the report as a Word document, following the structure explained above. The R code itself will not be assessed for style, we will only assess its output to see if it answers the question, but it is useful for us to see the R code to give feedback.

The percentage of marks allocated for each section is noted below in square brackets:

- Data preparation [**5%**]
- Exploratory Data Analysis [**20%**]
- Data analysis plan [**15%**]
- Assumptions [**15%**]
- Carry out analysis [**25%**]
- Interpretation [**15%**]
- Conclusion [**5%**]

Submission instructions:

1. Your project should be submitted as a single PDF document. **Your name should not appear anywhere in your assignment.** Please also check your R code for path that may contain your name or student number and only use paths within an R project. If you do not know your exam number, please email lcourage@exseed.ed.ac.uk or consult the **My Personal Details** section of MyEd. The document should have a cover sheet which states the following information only:
   - the assignment title
   - your word count (excluding any R code or output)
   - your exam number.
2. Assignments must be submitted no later than the specified date/time. Marks will be deducted for assignments which are handed in late at a rate of 5% per day (except where there are acceptable extenuating circumstances and the student has submitted a special circumstances form in advance). This applies for up to five days, after which a mark of zero will be given.
3. Marks will be deducted for assignments over the stated length. An assignment which is up to 49 words over the word limit will incur no penalty. After that, 1% will be deducted for each 50 words over the limit, up to a maximum of 10%. Longer reports will be awarded zero. **Note**: Data (R code or output) should be excluded from the word count. It won't be necessary to use the full word limit – a shorter formal word count is likely and acceptable.
4. Please submit your assignment electronically to the TurnItIn drop-box in the **Assessment** folder. Use a filename of format: B123456.pdf
5. Plagiarism will be treated extremely seriously.
6. Assignments will be graded in line with the University of Edinburgh postgraduate common marking scheme:

Grade	Percentage	Description
A1	90 - 100	An excellent performance, worthy of a distinction
A2	80 - 89	
A3	70 - 79	
B	60 - 69	A very good performance
C	50 - 59	A good performance
D	40 - 49	A satisfactory performance for a diploma or certificate but inadequate for a masters
E	30 - 39	A marginal fail
F	20 - 29	A clear fail
G	10 - 19	A bad fail
H	0 - 9	

Remember: Named assignments and assignments without an exam number will not be marked.