MATH5824M Assessed Practical - Problem Description

Key dates

Released: Tuesday 12th March 2024
Supervised Sessions: Tuesday 19th / Wednesday 20th March 2024
Deadline: 2pm on Tuesday 23rd April 2024

Please note that if you require this information in a different format, then contact me: r.g. aykroyd@leeds.ac.uk

Important information

This R computer practical is included in the module assessment with submitted written reports counting a total of 20% of the overall module grade. This description was released on Tuesday 12th March 2024 and there are supervised computing sessions in the week 18-22 March where you can work on the assessment while I am present to help – check your timetable for the time and location. Your written report on this practical must be submitted by 2pm on Tuesday 23rd April 2024.

Please note that for students of MATH5824M, the practical is made-up of two parts: Problem 1 on generalised linear models and Problem 2 on spline modelling. Each has a set of instructions with separate reports which are to be submitted separately – details are below.

Problem 1 - Generalised Linear Models

Please follow these instructions when writing your report:

- Your report must start with a short summary explaining your conclusions and interpreting the models in language suitable for a nonspecialist.
- Your report should be written carefully and written in an academic style. Illustrate your ideas with appropriately chosen graphs and other R output. When reporting numerical values, but not for intermediate calculations, you should use appropriate rounding.
- 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, and selected output, as a separate appendix (this will not count towards the eight-page limit).
- When reporting numerical values, but not for intermediate calculations, you should use appropriate rounding.

Please follow these instructions when submitting your report:

- Submit your completed practical report using the links in the *Assessed Practical* folder in Minerva by the deadline of **2pm on Tuesday 23rd April**. 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.

Please follow these instructions when accessing your individual data set:

- The information below refers to a number $\texttt{N} \in (00, 01, \ldots, 99)$. This number represents the last two digits of your student ID. For example, if your student ID is 200123456, then $\texttt{N} = 56$ whereas if your ID is 200678900, then $\texttt{N} = 00$ – note that this is *double zero*. For these two student ID, the file names mentioned below would be *adelaide-56.csv* and *adelaide-00.csv* respectively.

Please follow these instructions when submitting key numerical answers for pre-deadline feedback (optional):

- Please note that answers submitted in this way are for feedback only and are not part of the assessed grade.
- To receive feedback, you should submit the key numerical answers as requested on the Microsoft Form accessed using the link "*Microsoft Form for submission of key numerical answers*" in the "*Assessed Practical*" folder in Minerva. Answers submitted before 2pm on Friday 22nd March will be graded automatically with feedback sent by email.
- When reporting your answers, but not for intermediate calculations, you should round values to 2 decimal places if RStudio returns more decimal places. As examples, a calculation from R of 5 should be returned as 5; 5.1 should be entered as 5.1; 5.13 should be entered as 5.13; but 5.138 should be entered as 5.14, etc. The RStudio command `round` can be used, for example `round(5.138, 2)`.

Description of the problem

This problem is based on the example given in Dobson and Barnett (2008), pp. 145-146, but individual data sets have been created. The data set contains historical information from the University of Adelaide, Australia, on the numbers of graduates, from 1938 to 1947, surviving for 50 years, as a function of the following explanatory variables:

- `year` – the year of graduation
- `faculty` – "M" for Medicine, "A" for Arts, "S" for Science and "E" for Engineering
- `sex` – "F" for Female and "M" for Male

with two response variables:

- `survive` – the number of graduates surviving for 50 years, and
- `total` – the total number of graduates with a particular combination of explanatory variables.

Note that you may need to explicitly declare as factors any qualitative variables.

Tasks you need to perform

1. Read into R the data stored in *adalaide-*`N`*.csv*, as defined above, from the usual data file location. For example using the command: `read.csv("http://rgaykroyd.github.io/MATH3823/Datasets/`*adalaide-*`N`*.csv*`")`.

   Then, perform an initial exploration of the data using appropriately chosen graphs.

   Enter the number of rows of data in the data file into the online form.

   Enter the total number of graduates into the online form.

2. Fit an appropriate model for which the variable `survive` depends on `year+faculty+sex`.

   To begin with, ignoring questions of statistical significance, explain what each parameter represents (that is, how do they affect the chances of survival).

   Then, use the fitted model to predict the probability of survival for each of the following combinations:
   `sex=M, year=1941, faculty=M`
   `sex=F, year=1938, faculty=E`

   You might have noted, however, that there are no women recorded doing Engineering! So what does the second fitted probability mean?

   Enter the estimated intercept parameter value into the online form.

   Enter the predicted probability of survival when sex=M, year=1941, faculty=M into the online form.

   Enter the p-value of the year term into the online form.

   Enter the std deviation of the deviance residuals into the online form.

3. For the model fitted in point 2, now consider statistical significance.

   Investigate whether, or not, a simpler model would be appropriate to describe the data. For example by removing variables or by combining factor levels. Also, consider more complicated models (for example, involving two-way interactions). If you do include any interactions, you should count the degrees of freedom carefully.

   For the model describing survival in terms of faculty and sex (without interaction), enter the value of the intercept parameter into the online form.

   For the model describing survival in terms of faculty, sex and their interaction, enter the p-value of the interaction term into the online form.

   It may be useful to bear in mind the following considerations when carrying out your 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, and replacing a quantitative variable by a multi-level factor to allow modelling of a non-linear relationship.

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

4. Next analyze the male and female data separately. Consider the models describing the probability of survival in terms of `faculty` only.

   For the model describing survival of women, enter the p-value of the faculty term into the online form.

For the model describing survival of men, enter the value of the intercept parameter into the online form.

Compare the output from these models with the model fitted to the original data and where the probability of survival depends on `faculty*sex`. Do you get any differences in interpretation? [This question is motivated by the fact that women did not do the full range of subjects.]

R hints:

- If `y` is a vector representing a quantitative variable, then `y.F = as.factor(y)` creates a corresponding factor where every unique value in `y` is a level.
- If `x` is a vector representing a qualitative factor with level *"a"*, *"b"*,*"c"*, *"d"*, say, then `x=="a"` is a logical vector of the same length as `x`, which equals `TRUE` if the element of `x` equals *"a"*, and `FALSE` otherwise.
- Values in the qualitative factor `x` can be *replaced* using, say, `x[x=="a"]="b"`. It is usually a good idea to only perform such manipulations on a copy of the variable.
- If `z` is a vector representing a quantitative variable taking the values `z=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 `z.new`

```
z.new = 1*(z==1)+2*(z==2)+3*(z==3)+ 3*(z==4)+4*(z==5)
z.new = as.factor(z.new) # 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 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))`

End of Problem 1 - Problem Description

Problem 2 - Spline modelling

Please follow these instructions when writing your report:

- Your report must start with a short summary explaining your conclusions and interpreting the models in language suitable for a nonspecialist.
- Your report should be written carefully and written in an academic style. Illustrate your ideas with appropriately chosen graphs and other R output. When reporting numerical values, but not for intermediate calculations, you should use appropriate rounding.
- Your report must be no longer than four 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 five 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 in the body of your report, but include your code and extra R output, as a separate appendix (this will not count towards the three-page limit).

Please follow these instructions when submitting your report:

- Submit your completed practical report using the links in the *Assessed Practical* folder in Minerva by the deadline of **2pm on Tuesday 23rd April**. 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.

Please follow these instructions when accessing your individual data set:

- The information below refers to a number $\texttt{N} \in (00, 01, \ldots, 99)$. This number represents the last two digits of your student ID. For example, if your student ID is 200123456, then $\texttt{N} = 56$ whereas if your ID is 200678900, then $\texttt{N} = 00$ – note that this is *double zero*. For these two student ID, the file names mentioned below would be *engine-56.csv* and *engine-00.csv* respectively.

Please follow these instructions when submitting key numerical answers for pre-deadline feedback (optional):

- Please note that answers submitted in this way are for feedback only and are not part of the assessed grade.
- To receive feedback, you should submit the key numerical answers as requested on the Microsoft Form accessed using the link "*Microsoft Form for submission of key numerical answers for MATH5824M*" in the "*Assessed Practical*" folder in Minerva. Answers submitted before 2pm on Friday 22nd March will be graded automatically with feedback sent by email.
- When reporting your answers, but not for intermediate calculations, you should round values to 2 decimal places if RStudio returns more decimal places. As examples, a calculation from R of 5 should be returned as 5; 5.1 should be entered as 5.1; 5.13 should be entered as 5.13; but 5.138 should be entered as 5.14, etc. The RStudio command `round` can be used, for example `round(5.138, 2)`.

Description of the problem

To investigate the relationship between a car's engine size and engine wear data were collected from car engines with two variables recorded:

- `size`: size of engine (L), and
- `wear`: an index of wear (higher values indicate more wear).

The aim is to investigate the relationship between engine wear and size using cubic smoothing splines.

Tasks you need to perform

1. Read into R the data stored in *engine*-`N`.*csv*, as defined above, from the usual data file location. For example using the command: `read.csv("http://rgaykroyd.github.io/MATH3823/Datasets/`*engine*-`N`.*csv*`")`.

   Then, perform an initial exploration of the data using appropriately chosen graphs.

   Enter the number of car engines recorded into the online form.

   Enter the mean `wear` into the online form.

2. Investigate the relationship between engine wear and size by fitting some cubic smoothing splines with different values of the smoothing parameter λ.

   For the spline with $\lambda = 0.01$, enter the value of the standard deviation of the residuals into the online form.

   You should discuss how the fit changes with the smoothing parameter in terms of visual behaviour, residuals and comment briefly on how the fit behaves as the smoothing parameter tends to zero and infinity. Discuss if either of these extreme cases produce an adequate description of the data.

3. 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.

   For your choice of best single value of the smoothing parameter, enter $\log_{10}\lambda$ into the online form.

4. Use the fitted splines to predict the wear for a car with engine size of 1.0 L and a car of 2.6 L. How do these predictions vary as λ is changed?

   For the spline with $\lambda = 0.01$, enter the predicted wear for a car of size 2.6 L into the online form.

5. Briefly, discuss a method for the automatic choice of λ. Investigate how the value of the corresponding criterion varies as λ is changed.

   For the smoothing parameter value which minimizes the generalised cross validation criterion, enter the value of $\log_{10}\lambda$ into the online form.

6. Suppose that we are now told that the relationship between wear and size must be monotonic. Find the smallest value of λ which leads to a monotonic fitted spline function.

   For the smoothing parameter value which leads to a monotonic fitted spline, enter the value of $\log_{10}\lambda$ into the online form.

R hints:

- For the output from a spline fit, for example using `my.fit = smooth.splint(wear, size, lambda=0.01)`, then the corresponding generalized cross validation criterion can be obtained using `my.fit$crit`.
- `log10(x)` gives the base 10 logarithm of `x`.

End of Problem 2 - Problem Description

End of Assessed Practical