MATH97124 - Applied Statistics

Autumn 2020 - Assessed Coursework

Deadline - 12.00 Noon Friday 11th December 2020

The final report must be typed up and should be a properly structured document in PDF. Where appropriate, phrase your answer using proper sentences. Your report should be uploaded to blackboard (see inside the Coursework (Coursework 3) folder) by the deadline stated above.

Once the report is uploaded there is no option for re-uploading so you should **upload your final version only**. If possible, avoid last minute uploads as the system can crash if it simultaneously receives too many requests.

When reporting values in your report, round figures to 4 decimal points — using the `round` command in `R`. There is a **page limit of 10 pages (excluding Appendices)**.

Any material beyond the page limit will not be marked. Do not use small fonts i.e. less than 11pt or less than the normal font size in LaTeX.

Quality of presentation and conciseness of your answers will count towards your mark. Ensure that your answers are well written, organised and are in the form of properly written sentences that include your full statistical reasoning.

Any figures or tables should be included in your report for a good reason, and this reason should be described in your report – all tables and figures should be referenced in the main body of your text. All axes should be appropriately labelled.

You are strongly discouraged from excessively using code in place of written sentences and mathematics. Include long pieces of code in appendices if necessary.

Note that you may run into trouble if you copy and paste code from a PDF document into `R`. To use code from a PDF document, type it yourself into `R`.

Please remember to always use your own words to explain concepts, models, methods etc and to cite all sources and provide appropriate references.

This Coursework counts for 15% of your mark in Applied Statistics. Questions 1-2 constitute 20% of the marks, 3-4 60%, 5 constitutes 10% and the remaining percentage is for clarity and presentation. Remember to always show your justification and reasoning for your answer rather than just stating an answer - as marks will be allocated for justification and reasoning.

This coursework will use data which can be accessed at:
`http://wwwf.imperial.ac.uk/~nadams/AppliedStatistics2020/your-data.csv`
where you should replace "your-data" with "CID1234567", where 1234567 is your CID *with the leading zero removed.*

The data set that you will find via the above link was acquired in a psychophysical study concerned with the sensitivity of the human eye to colour and its relationship with age. Light can be mapped to a three-dimensional space (the colour space) that is useful for representing the relationships between colours. The experimental procedure measured the frequency at which a subject was able to detect a change in colour, that is the *threshold* detection level. The procedure also measured the direction of the change *in colour space*. Other variables include age, gender and IQ, with slightly different procedures deployed to handle subjects of different age where necessary.

The variables in the dataset are:

- `Thresh` – the response variable – a measure of the frequency level required to detect a change of chromatic (colour) stimulus.
- `Age` – age in years of the experimental subject.
- `Axis` – a categorical variable indicating in which of three `directions` (in colour space) the change occurred. The names of the levels follow the names of the axes of the colour space: `Protan`, `Duetan` and `Tritan`.
- `IQ` – The intelligence quotient (IQ) of the subject, measured in various ways for different ages, and normalised.
- Gender – The gender of the subject.

The coursework requires you to find a linear model that best describes the relationship between the response variable and the predictor variables, and allows for valid inference of model parameters.

Q1) Conduct an exploratory data analysis (EDA) of the data set, noting important features, relationships and issues.

Q2) Missing values are a perennial problem in applied statistics. For the purpose of this coursework, we will consider two crude methods:

- Listwise deletion: any case (observation, subject) with missing values is deleted from the data set.
- (simple) Mean imputation: missing values for a variable are replaced by the mean value for that variable.

Research these two methods and, using your own words, briefly report on their strengths and weaknesses. Select **one** method to produce a data set that will be the subject of the remaining questions.

Q3) Fit a normal linear model to the data, performing a range of model selection and diagnostics procedures. Report the key steps, with justifications for the choices. Specify the final selected normal linear model.

Q4) The response variable `Thresh` is known to be strictly positive in the range of the data and based upon the optical physics of the problem it is possible to motivate use of the Gamma GLM for modelling for this problem. Research the Gamma GLM and, in your own words, provide a brief summary of its characteristics (including its mathematical specification) and discuss its appropriateness for the data set here. Fit two GLMs, as follows:

- A Gamma GLM with identity link function, where the linear predictor has an intercept, and terms for `Axis`, `Age` and the reciprocal of `Age`.
- A Gamma GLM with identity link function, where the linear predictor has an intercept, and only interaction terms between `Axis` and `Age` and `Axis` and reciprocal of `Age`.

Compare these two models using appropriate procedures, and make a recommendation about which GLM model you favour.

Q5) From your final selected normal linear model in Q3 and your recommended GLM model in Q4, make a final recommendation of which model you can recommend and why. Highlight any concerns that arise in your recommendation.