M3S2/M4S2 - Coursework Spring 2020

Background

The data come from an experiment to study the effect of an artificial stimulant on heart rate, and how this varies for individuals with different health profiles. The stimulant is known to target the same receptors as caffeine, and so the clinicians who designed the study aimed to study a group that contains as many coffee drinkers as non-coffee drinkers. Your sample should reflect this, although the balance may not be exact because of patients who dropped out.

The study was conducted over two days. On the first day, the resting heart rate of each experimental subject was taken. On the second day, each subject was given a pill containing 100mg of the stimulant. Their heart rate was measured 30 mins later.

Data

You have access to a personalised dataset on Blackboard, stored in the object `dat`. It contains the following variables, for each subject

`rest_pulse` - heart rate in beats per minute on day 1.

`stimulated_pulse` - the heart rate in beats per minute on day 2, after administration of the stimulant.

`bmi` - body mass index, in kg/m^2.

`coffee_reg` - whether or not the subject regularly consumes coffee. Binary indicator (1 indicates a consumer).

Modelling

The clinical team began with a model in which the stimulated heart rate is explained by the resting heart rate and bmi.

```
fit0<-lm(stimulated_pulse~rest_pulse + bmi)
```

In a brief consultation with a statistician, the clinicians were advised instead to use the difference between the stimulated and resting heart rates as a response variable, and fit a Gamma GLM using the inverse link function, with BMI as a covariate. They are unfamiliar with GLMs, and so have turned to you for help.

Workflow

1. Download the data from blackboard and read it in to R.
2. Carry out an exploratory analysis of the data, producing plots and summaries.
3. Fit the clinicians' initial model, and use summaries and diagnostic plots to evaluate the model fit.
4. Fit the model suggested by the statistician, and evaluate the model fit.
5. Fit your own models and evaluate the quality of fit. (You can stick with the gamma family, but try different linear predictors.)
6. Write a report suitable for another statistician to read: your report should include

- A clear statement of your aims.
- A description of the models you have considered, and the assumptions they require.
- A detailed explanation of the numerical scheme used to fit the models, including code where appropriate.

- Evaluation of how well the models fit the data.
- Estimates of the model parameters, together with confidence intervals, for the most reasonable model.
- Discussion of any limitations of the model and the experimental design.
- Conclusions in the context of the problem.

7. Append to your report a standalone section that would be accessible to the team who originally collected the data, who do not have much statistical training. You should explain any difficulties with their original model, and clearly describe the relationship (if any) between the variables provided.

Your report should not exceed 5 sides of A4. It should be 10 point font or larger, with normal margins. You should also submit carefully structured and commented R code, as a separate file.

Note that while you may use inbuilt routines in R, such as `glm`, you are expected to include in your report a full, working numerical algorithm to fit the specific GLM considered here, with a method for choosing a sensible starting point.

Marking rubric

Your reports will be marked out of 5 in each of the following categories

1. **Accuracy**: are statistical tools used correctly, and are the conclusions drawn reasonable?
2. **Reproducibility**: could your analysis be repeated according to the summary you have provided? Where decisions are made, e.g. in removing suspect data points, are these justified and their consequences considered?
3. **Completeness**: are the main statistical and modelling issues addressed?
4. **Accessibility**: are the conclusions described clearly, in plain language, for the clinicians who collected the data?

Submission

Please submit your report on blackboard. The deadline is **4:00 PM Monday 9th March**

Mastery Material: for M4S2 students only

1. Use simulation to assess the suitability of the commonly used asymptotic approximations for the sampling distribution of the maximum likelihood esimators and scaled deviance in this problem.
2. Find data on the relationship between BMI and resting pulse from another study. Evaluate how well the conclusions of this study would generalize.

Marks will be awarded according to the same criteria considered above. This section will be scaled to be 20% of the total.