LSHTM MSc in Medical Statistics 2022–23

THE ANALYSIS OF HIERARCHICAL AND OTHER DEPENDENT DATA

ASSIGNMENT

This assignment has two parts. You should attempt both parts. Part I carries 90% of the marks, and part II 10%.

1 Part I

The prothrombin data

The data were derived from a randomized clinical trial that compared an active and placebo treatment on patients affected by chronic liver disease, who had abnormally high prothrombin levels. To monitor their progress, prothrombin measurements were taken at 3, 6, 12 and 24 months. Prothrombin was also measured at baseline prior to treatment (baseline covariate). Prothrombin is a protein produced by your liver and is one of many factors in the blood that help it to clot appropriately.
The data are held in the file called `assignment_2023.dta`. You will find it in the usual folder on the U: drive, and on Moodle.

The variables are:

```
-----------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-----------------------------------------------------------------------------------------------
id              int     %8.0g                 Personal identifier
treat           byte    %8.0g      treat      Treatment
sex             byte    %8.0g      sex        Sex
time            float   %9.0g                 Time (months)
pro             float   %9.0g                 Prothrombin (sec)
-----------------------------------------------------------------------------------------------
```

Aims

(i) Model the observed prothrombin trajectories over time (**with time assumed continuous in both fixed effects and covariance structures**) and their between- and within-patient variation using mixed effects models. The primary aim is to understand how post-treatment trajectories differ between treatment arms and by sex.

(ii) Re-estimate your final model using GEEs. Compare the predicted population-average trajectories of patients treated with the active/placebo treatment that you obtained in (i) with the predictions obtained using the same explanatory variables but estimated using GEEs.

Tasks

The part of the assignment consists of 4 parts:

1. Data description (contributing to 18% of the final mark):
   Carefully describe the study and the available data, including the prothrombin trajectories and their distribution. Take particular care in describing any missing data patterns.

2. Methods (contributing to 27% of the final marks):
   Describe the methods you used to address aims (i) and (ii). In particular describe your selection steps for both variance and fixed effects structures in your final model, defining the final model algebraically stating all its assumptions. Also, describe what steps you have taken to estimate the population average trajectories using GEEs.

3. Results (contributing to 27%):
   Report all the relevant analyses that have led you to your results; take care in using appropriate tabular and/or graphical summaries of your results.

4. Models critique and results summary (contributing to 18%):
   Briefly summarise the essential findings of your analyses and discuss the main analytical issues you have encountered (including comments on data quality and plausibility of the assumptions).

2 Part II

This part of the assignment is designed to test your ability to apply the concepts you have learned in a novel research setting.

On Tuesday, 14th March at 9:30am UK time, we will listen to a lecture by a final year researcher from the University of Malawi. Following this we will have a live zoom-Q&A with the student. The live Q&A will be recorded, but this recording will only be available for a single viewing to online students who, by prior arrangement, have been unable to join the live session.

After listening to the lecture and taking part in the discussion (which you are encouraged to guide by asking questions), you should answer the following question in a paragraph of not more than half a page.

Question (contributing 10%)

Briefly summarise the principal research question, and the methods used to investigate it. In the context of this research, briefly describe the shared random effect model, the generalised estimating equation model and the latent variable model, highlighting any major differences. Finally, briefly summarise the research findings.

3 Format

- Address each part separately, as outlined above. Do not include an introduction.
- You can use any software you wish but you must declare which one (or ones) you have used.
- Do not include any parts of your computer output.
- Do not use Stata variables names; use their proper names.
- Label tables and figures carefully (or you will lose marks).
- There is no minimum or maximum number of figures and tables.
- You should use *Times New Roman* font of size at least 11pt; use page margins of at least 2cm; do not use multiple columns.
- IN TOTAL you should report your findings in no more than 4 pages, which should include figures and tables.

Reports that are 2% to 10% over length will have 1 grade point deducted, those that are more that 10% over length will not be marked and given an automatic zero

Your work should be submitted by 5pm on Tuesday 21st March 2022