Assignment 1
02429 Analysis of correlated data: Mixed linear models
DTU Compute, Autumn 2024

Preamble

Perform a statistical analysis using R (see details below), and present your analysis and conclusions in a written report. The report should be submitted via DTU learn no later than Thursday 26 September 2019 at 8:00 hrs.

The report and all R code must be written by groups of three students. The R code must be included as an appendix in the report and should be neat, readable, and commented. The report may be no longer than 5 pages excluding the front page, table of contents, figures, references, and appendices. All figures and tables included in the report must be referred to and discussed in the text and should be numbered and have an informative legend.

There should be no R code (or raw R output) in the main body of the report other than in the R code appendix. Some of the intermediate statistical analysis, in raw R output if you wish, could be presented in a second appendix. Some additional comments:

- The report should not contain excessive material. The overall presentation of your work including its coherence and the relevance of included material is assessed as well - not just the correctness of the statistical analysis.
- Pay attention to the teaching material and get inspired by e-Note 3 for the report writing.
- Remember that much of Statistics is about quantifying and describing variability and uncertainty: When presenting, e.g., parameter estimates and Least-Squares means (LS-means), remember to include confidence intervals.
- Think of yourself as a consulting statistician when writing your report. You should include enough detail and explanation in the body of your report:
    a) to allow a statistician with no knowledge of R to replicate your analysis and validate your results using other statistical software *and*
    b) for someone with little or no knowledge of Statistics to understand the overall purpose and conclusions of your report in the context of the data.

1 Exercise I. Potato plants

Data

In an experiment with potato plants, 5 potato varieties were grown in 5 different environments, and the yield was recorded. The purpose of the experiment was to compare the 5 varieties in terms of their yield.

The variable GEN indicates the genotype (variety) of the potato plant (A, B, C, D or E), ENV states the environment (1, 2, 3, 4 or 5) and YLD is the yield measured in tons per hectare. The data is available in the file assignment1.csv.

Statistical analysis and report writing

Part 1

Using R, perform a statistical analysis in which you investigate the effect of genotype on the yield. The statistical inference should not only be valid for the 5 environments in our experiment (which we assume were chosen at random from a larger population of possible environments).

Your report should contain the following elements.

i) Frontpage

ii) Table of contents

iii) Introduction. Include a description of the data

iv) Descriptive/exploratory analysis of the data

v) *Some sections of your choice*

vi) Discussion and conclusions

vii) Appendices

The description of the data should include a short description of the experiment and present all relevant variables together with their type (factor or numeric) and values (factor levels or range observed in the data). It should be indicated whether each factor is balanced or unbalanced, considered fixed or random in the analysis, and which factors are nested or crossed.

Present the mathematical expression of the models. Include no more than a sufficient number of significant digits when reporting numerical results.

The descriptive/exploratory analysis of the data should include relevant plots, tables, and summary measures.

The *sections of your choice* should (when applicable) include.

- descriptions of relevant statistical models (including their assumptions) using appropriate mathematical notation and accompanying text.
- documentation of hypothesis tests performed in connection with model reduction.
- parameter estimates, documentation of post-hoc analyses/tests of particular interest, and illustrations related to the final model.

The discussion and conclusions section should be used to briefly reflect on the results of the analysis in the context of the experiment and the data.

Part 2

Perform a short statistical analysis in which you investigate the effect of genotype and environment on the yield. The statistical inference should now only be valid for the 5 environments in our experiment (i.e. you can consider the environment effect as fixed this time).

What happens? Compare the results from the two analyses.
Hint: Look at the confidence intervals. Why did they change? and does it make sense? When and why should the environments be considered random?

You can add this short analysis/discussion as a separate section after your conclusion from the first analysis.

2 Exercise II. Simulation study

In class we have seen a simulation example of a simple random effects model $y_{ij} = \mu + a_i + \epsilon_{ij}$, $j = 1, \ldots, n_i$, $i = 1 \ldots, G$. Simulate data from the model, this time with groups of different sizes and a number of groups of your own choice. Consider the following scenarios:

a) Two different total sample sizes $N = \sum_i n_i$: small/large.

b) For three intraclass correlation values: 0, .45, .95.

Report in a tabular form, the estimated mean value and its CI, the estimated intraclass correlation, and the corresponding population/theoretical parameter values. Include some plots of the data. More than one plot in a figure for easy comparison. From the Table of results, make some observations.