Essentials of Mixed and Longitudinal modeling - 2/2024

Roula Tsonaka, Biomedical Data Sciences, Medical Statistics, LUMC

Case study: Progression of joint destruction in rheumatoid arthritis

1 Instructions

The aim of this assignment is for you to show that: (1) you have sufficiently comprehended the basics of the course, (2) you are able to analyse on your own a dataset and (3) you are able to correctly interpret the analysis results.

You are expected to prepare a reportpresenting the analysis for two longitudinal outcomes, i.e., a continuous and a binary one. The case study report should not be read as an annotated R output. It should comprise:

- an *Introduction section* where the study, the research questions and structure of the report are discussed (10 - 15 lines),
- a *Descriptive Analysis section* for each longitudinal outcome, where the available data are presented using summary statistics tables and plots and your conclusions/observations are clearly stated (1 - 2 pages),
- a *Statistical Analysis section* for each longitudinal outcome, where you present the final model including its assumptions (e.g., the final model is a linear mixed effects model where in the mean part we used the main effects of time and group and their interaction, for the within subject correlation we used random intercepts and for the error terms we assumed they are normally distributed, homogeneous and independent of the random effects). You will also present the analysis results with the statistical tests you used. Specifically you are expected to present the estimated parameters with standard errors or confidence intervals, p-values and answer to the specific questions made (2 - 3 pages),
- You should include an *Appendix* to discuss the procedure and the statistical techniques (hypothesis tests and residual analyses) you used to decide the structure of this final model. Be selective on the tables and plots you will show. You are expected to deliver the .Rmd file which reproduces your analysis (2 - 3 pages for each outcome).

Your report will be graded as follows: 10% will be given on the general presentation i.e., structure, typos, grammatical mistakes and clarity; 20% on descriptive analysis for both outcomes and 70% on the statistical analysis of both outcomes including the Appendix.

On the front of the report there should be a declaration that all students in the group have contributed sufficiently to the study and that the work (analysis, R programs and text) is original and not copied from elsewhere, signed by all students.

2 Introduction

The data set considered in this assignment comes from a longitudinal study, in which the goal is to investigate how the joint destruction evolves over time in rheumatoid arthritis patients and how it is related to the gender, age and genetic background of the patients. Joint destruction is quantified using the Sharp-van der Heijde scoring method on X-rays on hands and feet of the patients. The Sharp index takes values from 0 to 448 and the higher the score the higher the disease activity is. In addition to the Sharp index, the disease severity is also evaluated by the clinicians as low or high based on a combination of clinical parameters.

In this study 500 patients have been included between 1990 and 2006, and X-rays were taken at baseline and yearly thereafter during 7-years. For these patients several variables are recorded including: the date of inclusion in the study, the dates of each X-ray, the value of the Sharp index on each visit, their sex and age at inclusion in the study. In addition, various genetic variants have been genotyped, but here we will concentrate on the effect of one single nucletide polymorphism (SNP) on the disease progression.

A complication in this study is that not all planned measurements have been recorded for several reasons e.g., loss to follow-up, relocation, remission, etc. Another complication is that the treatment strategies have changed through the years. Thus it has been suggested by the treating clinicians to take into account in the analysis the inclusion period of each patient, i.e., a patient who joined before "1996-01-01" belongs to period A, while a patient who joined after "1996-01-01" belongs to period B.

3 Data file

- DiseaseActivity.csv
- Variables
  - id: patient indicator.
  - sex: sex indicator (1: males, 2: females).
  - Age: age of the patient at inclusion.
  - SNP: single nucleotide polymorphism (0: no variant alleles, 1: variant allele, 2: variant alleles).
  - Period: inclusion period.
  - Visit: Visit number.
  - SHS: Sharp index at each of the 8 visits, respectively.
  - Severity: Disease severity (0: low, 1: high) at each of the 8 visits, respectively.

4 Questions: Sharp index

1. Present summary statistics for the sample at hand i.e., number of males/females, age distribution, etc.

2. Describe the longitudinal outcome using summary statistics and graphical techniques to explore the mean structure, the variance structure and the correlation structure. Summarize your conclusions e.g., do you see non-linearities, differences between the groups, etc.

3. The researchers are interested to study how the mean Sharp scores evolve over time and how this evolution is related to the gender, age and SNP of the patients while taking into account the inclusion period. Build an appropriate linear mixed effects model to answer this question. Describe in the Appendix all the steps you have performed to reach your final model (including residuals analysis). Based on the final model, present your conclusions using effect estimates with standard errors or confidence intervals and p-values. You may use plots with this information to communicate your findings.

5 Questions: Disease Severity

1. Describe the longitudinal outcome using summary statistics and graphical techniques to explore the mean structure. Summarize your conclusions e.g., do you see non-linearities, differences between the groups, etc.

2. The researchers are interested to study how the probability of high disease severity evolves over time and how this evolution is related to the gender, age and SNP of the patients while taking into account the inclusion period. Build an appropriate mixed effects logistic regression model to answer this question. Describe in the Appendix all the steps you have performed to reach your final model. Based on the final model, present your conclusions using effect estimates with standard errors or confidence intervals and p-values. You may use plots with this information to communicate your findings.