Inferences from binary longitudinal data

In the previous chapters we focused on modeling the means over time from a continuous response vector. In clinical trials we often encounter cases, where our response is however not continuous, but rather discrete. Discrete data can stem from either count data, such that values are taken in (a subset) of the natural numbers, or ordinal data, where values represent distinct categories, or binary data. In the latter case only values 0 and 1 are taken and represent the presence or absence of a clinical status, such as alive or dead at time X, hospitalized or not hospitalized at time X or response or non-response on a specific scale at time X.

Additionally, we can rarely assume a linear relationship between the discrete response variable and the independent variables.

In this case we use *generalized linear models* for the analysis of discrete longitudinal data. Generalized linear models (GLMs) apply a suitable link function to deal with the nonlinearity problem and necessitate the choice of an appropriate distributional assumption on the errors.

GLMs differ from the linear models for continuous data in terms of their interpretation. Modeling approaches include marginal, random effects, and conditional models.

Marginal Models

A marginal model for binary data can be written as

$$logit\left(\,\mathrm{E}[\,Y_{i,j}\,|\,X\,]\,\right) = x'_{i,j}\,\beta\,,$$

where $Y_{i,j}$ is the binary response of subject i at visit j, X is a set of predictor variables and $x_{i,j}$ a covariate vector. Furthermore, the link function $logit(.)$ is used. It is defined as

$$logit(p) := \ln\left(\frac{p}{1-p}\right), \quad p \in (0,1).$$

The model is called a marginal model, as it uses the marginal distribution of an outcome vector Y given a set of predictor variables X. Models of this type are usually handled via Generalized Estimating Equations (GEEs). In R, the packages `gee` and `geepack` can be used.

Random Effect Models

We can add a vector of random effects to the predictor variables in X to obtain

$$logit\left(\,\mathrm{E}[\,Y_{i,j}\,|\,X, b_i\,]\,\right) = x'_{i,j}\,\beta \,+\, z'_{i,j} b_i\,.$$

This model is often referred to as a generalized linear mixed-effects model. In the random effects model all fixed effects estimates are conditional upon the random effects vector b_i, hence its inclusion on the left-hand side of the equation. In R, the package `glmm` can be used to fit random effects models for binary data.

Conditional Models

Conditional models are models of the type

$$logit\left(\,\mathrm{E}[\,Y_{i,j}\,|\,Y_{i,1}, \ldots, Y_{i,j-1}, x_{i,j}\,]\,\right) = x'_{i,j}\,\beta \,+\, \alpha\, Y_{i,j-1}.$$

Here no random effects are added, but expectations are based on earlier observations. As stated by (Mallinckrodt and Lipkovich 2016) "In clinical trials, interest is often on an overall, or population average treatment effect, not on a treatment effect associated with specific outcome histories". This limits the applicability of conditional models in the clinical trials context.

Further reading

For an in-depth discussion, refer to Chapter 10 in (Mallinckrodt and Lipkovich 2016) or Part III in (Fitzmaurice 2011).

Fitzmaurice, Laird, G. M. 2011. *Applied Longitudinal Analysis*. Vol. 2. USA: New York, Wiley. https://doi.org/10.1002/9781119513469.

Mallinckrodt, Craig, and Ilya Lipkovich. 2016. *Analyzing Longitudinal Clinical Trial Data*. Chapman; Hall/CRC. https://doi.org/10.1201/9781315186634.