Analysis of Nonnormal Longitudinal Data with Generalized Linear Models

Ludwig Fahrmeir, University of Regensburg

This paper surveys models and methods for time series and panel data based on extensions of generalized linear models as a unifying tool. Though the linear normal case is covered within this frame, focus is on nonnormal, in particular discrete and categorical data. Models are classified according to their type of parameters, which may be constant, time-varying or cross-varying.

1. Introduction

Longitudinal or panel data are repeated observations on a set of variables for each unit of a population of sample size n. We assume that observations are made at discrete times \( s = 1, 2, \ldots \) and that the set of variables consists of a variable of primary interest, the response variable \( y \), and a vector \( x \) of covariates. Thus, the data are given by \((y_{is}, x_{is})\), \(i = 1, \ldots, n, \ s = 1, \ldots, t_i\), where \( i \) denotes the unit and \( s \) the time period. These data may be viewed as a cross-section of individual time series \( i, \ i = 1, \ldots, n \), with possibly different length \( t_i \).

For \( n = 1 \), i.e. if only one unit is observed, panel data reduce to time series data \((y_{s}, x_{s})\), \(s = 1, \ldots, t\).

Most of the work on longitudinal data analysis has been devoted to models for normally distributed responses, see e.g. Hsiao (1986) for a recent survey on panel data, while models for nonnormal time series and panel data have received attention only recently. In this survey
we consider models and methods for nonnormal longitudinal data, using appropriate extensions of generalized linear models (e.g. McCullagh and Nelder, 1983, Fahrmeir and Kredler, 1984) as a unifying tool for responses which may be continuous, discrete or categorical.

Response variables are assumed to be (conditionally) distributed according to an exponential family or, more generally, to a quasi-likelihood. Covariates may be constant over time or individuals, and they may be deterministic or random. In the latter case, analysis is conditional on observed covariates, and therefore no distributional assumptions on covariates are necessary. Individual responses \( y_{is} \) are related to present and past covariates and past responses by individual generalized linear models with individual parameters \( \beta_{is} \) (Section 2). We distinguish four types of individual parameters:

(i) \( \beta_{is} = \beta \) for all \( i,s \), i.e. parameters are constant over time and over units; (ii) \( \beta_{is} = \beta_s \) for all \( i \), i.e. parameters are constant over units but time-varying, (iii) \( \beta_{is} = \beta_i \) for all \( s \), i.e. parameters are cross-varying but constant in time, (iv) parameters are cross- and time-varying.

For time series (Section 3) only (i) and (ii) are meaningful. For panel data (Section 4), (i), (ii) may be suitable for a homogeneous (sub-)population, while (iii) and (iv) take into account heterogeneity explicitly. Application of the models to discrete duration data is sketched in Section 5.

Finally, we mention work on longitudinal data analysis which is not covered by the framework of this survey. Jacobs and Lewis (1983) develop D(iscrete) ARMA processes, Lawrance and Lewis (1985) discuss time series models with exponential marginal distributions. In another approach binary time series are assumed to be generated by