Computational Statistics and Data Analysis 73 (2014) 1–15

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Exploratory time varying lagged regression: Modeling association of cognitive and functional trajectories with expected clinic visits in older adults

Damla Şentürk [a,*], Samiran Ghosh [b,c], Danh V. Nguyen [d,e]

[a] Department of Biostatistics, University of California, Los Angeles, CA, USA
[b] Department of Family Medicine and Public Health Sciences, Wayne State University School of Medicine, MI, USA
[c] Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, MI, USA
[d] Department of Medicine, University of California, Irvine, CA, USA
[e] Institute for Clinical and Translation Science, University of California, Irvine, CA, USA

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form 9 April 2013
Accepted 1 November 2013
Available online 13 November 2013

Keywords:
Functional data analysis
Irregular, infrequent and unsynchronized longitudinal design
Lagged effects
Varying coefficient models
Transfer functions

ABSTRACT

Motivated by a longitudinal study on factors affecting the frequency of clinic visits of older adults, an exploratory time varying lagged regression analysis is proposed to relate a longitudinal response to multiple cross-sectional and longitudinal predictors from time varying lags. Regression relations are allowed to vary with time through smooth varying coefficient functions. The main goal of the proposal is to detect deviations from a concurrent varying coefficient model potentially in a subset of the longitudinal predictors with nonzero estimated lags. The proposed methodology is geared towards irregular and infrequent data where different longitudinal variables may be observed at different frequencies, possibly at unsynchronized time points and contaminated with additive measurement error. Furthermore, to cope with the curse of dimensionality which limits related current modeling approaches, a sequential model building procedure is proposed to explore and select the time varying lags of the longitudinal predictors. The estimation procedure is based on estimation of the moments of the predictor and response trajectories by pooling information from all subjects. The finite sample properties of the proposed estimation algorithm are studied under various lag structures and correlation levels among the predictor processes in simulation studies. Application to the clinic visits data show the effect of cognitive and functional impairment scores from varying lags on the frequency of the clinic visits throughout the study.

1. Introduction

Consider the standard varying coefficient model (VCM; Cleveland et al., 1991; Hastie and Tibshirani, 1993),

$$E\{Y(t)-\mu_Y(t)\}=\beta_1(t)\{X(t)-\mu_X(t)\} \tag{1}$$

where the regression function $\beta_1(t)$ is allowed to vary with time. The VCMs have been widely used in longitudinal data analysis in the past decade (e.g., see Fan and Zhang, 2000, 2008; Huang et al., 2002; Hoover et al., 1998; Wu and Chiang, 2000). When the time index is set to the duration of the longitudinal study, the regression function displays the varying

* Correspondence to: Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, CA 90095-1772, USA. Tel.: +1 310 825 5250; fax: +1 310 267 2113.
E-mail address: dsenturk@ucla.edu (D. Şentürk).

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.11.001

relation between the longitudinal response and the predictor throughout the study. Note that in (1) the regression relation is modeled only between the concurrent/contemporaneous times of the response and the predictor. However, in some applications it is of interest to predict or associate the response at the current time with lagged times of the predictor; e.g., a subset of previous predictor values. For example, Senturk and Mueller (2008) and Koru-Sengul et al. (2007) discovered lagged relations between acute phase protein levels and between a child's growth index and maternal cigarette smoking and alcohol use, respectively.

We propose a time varying lagged regression model to assess the association between predictors, including cognitive and functional impairment scores, with the frequency of the clinic visits of older adults aged 65–93. The approach focuses on exploratory modeling of lagged association between previous cognitive and functional impairment statuses with current clinic visits, by sequential conditional modeling where lags are chosen to optimize a normalized covariation between the response and the predictor processes. Such a modeling approach provides important information potentially useful for individual prognosis assessment as well as formulation of managed care strategies. Informative lagged predictors (e.g., based on routine visit assessments of cognitive and functional impairment) can be used as markers to monitor future activities, such as intervention adherence or health care services utilization, for instance. The data which motivates our model development consists of measurements taken annually on multiple health scores (cognitive and functional impairment scores) as well as the total number of clinic visits every three months for four years on 703 older adults. Challenges with the motivating data requires several modeling innovations; although useful in other contexts, existing methods are not directly applicable. The observations are from infrequent time points, where the response and predictors are not necessarily observed at concurrent times. Existing useful models such as the lagged VCM (Koru-Sengul et al., 2007; Senturk and Mueller, 2008) requires equidistant time grid for estimation. Also, although the recent approach of Mueller and Yang (2010) based on transfer functions, can handle irregular and infrequent data, it is not practical as the number of predictors increase. Therefore, several significant modeling challenges are addressed in the current work, including data sparsity, non-synchronicity of measurement times and the curse of dimensionality.

We first briefly review the aforementioned existing models that explore lagged effects. To study lagged predictor effects, Senturk and Mueller (2008) and Koru-Sengul et al. (2007) proposed lagged varying coefficient models

$$E\{Y(t) - \mu_Y(t)\} = \sum_{r=1}^{p} \beta_r(t)\{X(t-r) - \mu_X(t-r)\},$$

where a separate varying coefficient function explains the time dependent effect of the predictor from each lag $t - r$. In this model, an equidistant grid is assumed for the observation times, where r denotes size of the lag on this equidistant grid. Koru-Sengul et al. (2007) also proposed an imputation algorithm to fill occasional missing values in the equidistant grid, although this would be impossible for irregular data where subjects are observed at subject specific observation times. Mueller and Yang (2010) proposed the transfer functions

$$E\{Y(t) - \mu_Y(t)|X(s)\} = \beta(t,s)\{X(s) - \mu_X(s)\}, \tag{2}$$

for jointly Gaussian processes where the transfer function $\beta(t,s)$ reflects the effect of a lag ($s < t$) of the predictor process on the value of the response at the current time t. Unlike the prior proposals of lagged varying coefficient models, this general model can be estimated from irregular and infrequent data which may not be observed concurrently. However the dimension of the transfer function will increase with the number of predictors considered, hence the model is only feasible for a single predictor process in many applications.

In this work, we propose an *e*xploratory time *var*ying *lag*ged (EVarlag) regression model that addresses these challenges to analyze the aforementioned data. The main goal of the proposed model is to embed the classical VCM in a larger class of models to detect deviations from the concurrent nature of the classical VCM via estimated time varying lags. The EVarlag model (for a single predictor) is

$$E\{Y(t) - \mu_Y(t)\} = \beta_1(t)\{X(t-\Delta_t) - \mu_X(t-\Delta_t)\} \tag{3}$$

which relates the response process to time varying lags $t - \Delta_t, 0 < \Delta_t < t$, of the predictor process. Lagged associations are explored in (3) via estimation of the time dependent lag $t - \Delta_t$ by maximizing the absolute value of a normalized covariance criterion between lags of the predictor process and the response from time t. For homoskedastic predictor processes, the lag search corresponds to finding the lag of the predictor with the highest absolute correlation with the response. This also corresponds to choosing the path in the two-dimensional transfer function $\beta(t,s)$ with the highest absolute value as will be shown in Section 2.1. The classical VCM is a special case of (3) with concurrent relations, i.e. $\Delta_t = 0$. If a nonzero lagged relation is determined from the EVarlag model, this is informative for further investigation of the nature of the lag detected, including whether it is from a specific slice in time or a lagged time interval. Thus, follow-up analysis, such as functional linear models can be used to model the effects of longitudinal predictors from lagged intervals of time on the response (Senturk and Mueller, 2010; Malfait and Ramsay, 2003; Mueller and Zhang, 2005).

The proposed estimation algorithm is designed for irregular and infrequent data and does not require a common grid, similar to the estimation procedure for the transfer functions. Also, it can accommodate the response and predictor processes that may not be measured at the same frequency, hence at concurrent times, as encountered in the clinic visits data that will be analyzed in Section 3. Unlike the transfer function approach which is only feasible for a single longitudinal predictor