Journal of Statistical Planning and Inference 142 (2012) 2599–2606

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

A note on generalized functional linear model and its application

Qi Long

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

ARTICLE INFO

Article history:
Received 1 January 2011
Accepted 14 February 2012
Available online 1 April 2012

Keywords:
Generalized functional linear model
Measurement errors
Spline smoothing

ABSTRACT

Motivated by a biomarker study for colorectal neoplasia, we consider generalized functional linear models where the functional predictors are measured with errors at discrete design points. Assuming that the true functional predictor and the slope function are smooth, we investigate a two-step estimating procedure where both the true functional predictor and the slope function are estimated through spline smoothing. The operating characteristics of the proposed method are derived; the usefulness of the proposed method is illustrated by a simulation study as well as data analysis for the motivating colorectal neoplasia study.

1. Introduction

In biomedical studies, predictors are often measured from the same subjects repeatedly over time or a certain spatial structure and are therefore of functional nature. In particular, our work here is motivated by a colorectal neoplasia study, where the goal is to associate a subject's disease status with gene biomarkers whose expression levels in terms of protein contents are measured along the length of colon crypts, a microscopic structure in the human colon mucosa (Daniel et al., 2009). As Fig. 1 in Daniel et al. (2009) shows, the distribution of gene expression levels can be measured from the base to the apex of a semi-crypt, which forms a natural one dimensional spatial structure.

Studies like this can be naturally modeled using the generalized functional linear regression (GFLM, for short), where the dependence of a scalar outcome of interest, y, on a functional predictor, $x(\cdot)$ is characterized by a conditional density from the exponential family:

$$f(y|x) = \exp\{[y\eta(x) - b(\eta(x))]/a(\phi) + c(y,\phi)\}, \tag{1}$$

where

$$\eta(x) = \alpha_0 + \int_{\mathcal{T}} x(t)\beta_0(t)\,dt \tag{2}$$

is the natural parameter, ϕ is a nuisance parameter, and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specific smooth functions. In parallel to the classical generalized linear models (McCullagh and Nelder, 1989), α_0 and $\beta_0(\cdot)$ are referred to as the intercept and slope function, respectively, in GFLM. The goal is to estimate the intercept and slope function from n iid copies of the pair: $(x_i(\cdot), y_i)$, $i=1, \ldots, n$. Oftentimes, the task is further complicated by the lack of direct observations of $x_i(\cdot)$. Rather, one only has access to noisy observations of $x_i(\cdot)$ at discrete design points:

$$z_{ij} = x_i(t_{ij}) + \epsilon_{ij}, \quad j = 1, \ldots, m_i, \tag{3}$$

where the measurement error ϵ_{ij} is independent of the random function $x_i(\cdot)$.

E-mail address: qlong@emory.edu

0378-3758/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2012.02.027

Due to its wide applications, models with functional predictors have drawn much attention in recent years including functional linear models (FLM, for short) (Cardot et al., 2003, 2007; Li and Hsing, 2006; Yao et al., 2005; Crambes et al., 2009; Yuan and Cai, 2010) and GFLM (Cardot and Sarda, 2005; Muller and Stadtmüller, 2005). Most of existing approaches (Cardot et al., 2003; Cardot and Sarda, 2005; Muller and Stadtmüller, 2005; Yuan and Cai, 2010) assume that direct observations of $x(\cdot)$ are available, i.e., $x(\cdot)$ is fully observed without errors. This restriction is lifted only in several recent studies of the standard FLM (Cardot et al., 2007; Li and Hsing, 2006; Yao et al., 2005; Crambes et al., 2009) where $x(\cdot)$ is not fully observed and is measured with errors, and these models can be viewed as a special case of Model (1); most of these current work (Yao et al., 2005; Crambes et al., 2009) exploits the fact that there is a closed-form solution for estimating $\beta_0(t)$ in FLM when $x(\cdot)$ is fully observed without errors. Since such closed-form solution is not available for GFLM when $x(\cdot)$ is fully observed without errors, it is not trivial to extend these results developed for FLM to the more general GFLM and it is also unclear to what extent the existing results for FLM apply to GFLM.

To address our problem of interest, we adopt an approach similar to Li and Hsing (2006); specifically, we investigate a two-step estimating procedure, where both the functional predictor $x_i(\cdot)$ and the slope function $\beta_0(\cdot)$ are estimated through spline smoothing. We provide the details of the estimating procedure in Section 2 and study its operating characteristics in Section 3. In Section 4, we conduct a small simulation study to evaluate the finite sample performance, and illustrate the proposed approach using a colorectal cancer study. Finally, we make some conclusion remarks in Section 5. An outline of the proofs for the main theoretical results is given in Appendix.

2. Estimation

To fix idea, we assume that the true slope function $\beta_0(\cdot)$ belongs to the qth order periodic Sobolev space:

$$W^q_{per,2} = \{f : f, f^{(1)}, \ldots, f^{(q-1)} \text{ absolutely continuous}$$
$$\text{and} \quad g^{(k)}(0) = g^{(k)}(1) \text{ for } 0 \le k \le q-1,\ f^{(q)} \in L_2[0,1]\}.$$

Furthermore, we shall assume that the functional predictor $x(\cdot)$ belongs to the same functional space almost surely with $E(\|x^{[q]}\|^2) < \infty$. Throughout, we denote by $\|\cdot\|$ the usual L_2 norm, and by $\langle \cdot, \cdot \rangle$ the usual L_2 inner product. We note that the qth order Sobolev space:

$$W^q_2 = \{f : f, f^{(1)}, \ldots, f^{(q-1)} \text{ absolutely continuous},\ f^{(q)} \in L_2[0,1]\}$$

is the sum of two spaces, one is $W^q_{per,2}$ and the other is the space spanned by the first $q-1$ polynomial basis functions. Hence, when x and β_0 belong to W^q_2, our results also hold. Hence, this setting is suitable for most applications including the aforementioned colorectal cancer study, and is commonly adopted in the previous studies of functional linear regression (see, e.g., Li and Hsing, 2006).

To motivate our method, we consider first the situation where the functional predictor x_i's are available. It is evident that in this case, the negative log-likelihood can be expressed as

$$L(\alpha,\beta) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i\eta(x_i) - b(\eta(x_i))\} \tag{4}$$

up to terms not depending on α and β. The intercept and slope function can then be estimated through penalization:

$$(\hat{\alpha},\hat{\beta}) = \arg\min\left(L(\alpha,\beta) + \frac{\lambda}{2}J(\beta)\right), \tag{5}$$

where $\lambda \ge 0$ is a tuning parameter, and J is a penalty functional. In particular, we consider the following popular choice of the penalty functional:

$$J(\beta) = \int_0^1 (\beta^{[q]}(t))^2\, dt.$$

Proposition 1. *Suppose that for $L(\alpha,\beta)$ is continuous and convex with respect to its second argument. Then $\hat{\alpha}$ and $\hat{\beta}$ are uniquely defined if and only if $var(\int_T x(t)\, dt) > 0$.*

Proposition 1 can be readily proved and it indicates that this procedure indeed leads to valid estimates when $x_i(\cdot)$ is directly observed. In this case, the estimate of the slope function can be obtained by extending the method proposed in Yuan and Cai (2010), and is not the focus of this article.

We now consider the case when $x_i(\cdot)$ is not observable and only z_{ij}'s as given in (3) are observed. To use the procedure described above, we first construct an estimate of $x_i(\cdot)$, say, $x^*_i(\cdot)$. In particular, we propose to estimate x_i by means of penalized regression splines using the first $2K+1$ Fourier basis functions:

$$x^*_i(\cdot) = \arg\min\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\{z_{ij} - x(t_{ij})\}^2 + \frac{\lambda_i}{2}\int_0^1 (x^{[q]}(t))^2\, dt\right) \tag{6}$$