Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus

- Ana M. Aguilera, Manuel Escabias, Mariano J. Valderrama
- Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus
- *Computational Statistics & Data Analysis, Volume 53, Issue 1, 2008, Pages 151-163*
- DOI: https://doi.org/10.1016/j.csda.2008.07.001

Volume 169
May 2022
ISSN 0167-9473
ELSEVIER
COMPUTATIONAL STATISTICS & DATA ANALYSIS
An Official Publication of
iasc
CMStatistics
Available online at www.sciencedirect.com
ScienceDirect

Computational Statistics and Data Analysis 53 (2008) 151–163

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus

Ana M. Aguilera [*], Manuel Escabias, Mariano J. Valderrama

Department of Statistics and O.R., University of Granada, Spain

ARTICLE INFO

Article history:
Received 23 October 2007
Received in revised form 2 July 2008
Accepted 3 July 2008
Available online 6 July 2008

ABSTRACT

The relationship between time evolution of stress and flares in Systemic Lupus Erythematosus patients has recently been studied. Daily stress data can be considered as observations of a single variable for a subject, carried out repeatedly at different time points (functional data). In this study, we propose a functional logistic regression model with the aim of predicting the probability of lupus flare (binary response variable) from a functional predictor variable (stress level). This method differs from the classical approach, in which longitudinal data are considered as observations of different correlated variables. The estimation of this functional model may be inaccurate due to multicollinearity, and so a principal component based solution is proposed. In addition, a new interpretation is made of the parameter function of the model, which enables the relationship between the response and the predictor variables to be evaluated. Finally, the results provided by different logit approaches (functional and longitudinal) are compared, using a sample of Lupus patients.

1. Introduction

Systemic lupus erythematosus (SLE) is a chronic disease that affects the autoimmune system and is characterized by spontaneous exacerbations, or flares, that markedly decrease a patient's quality of life and can sometimes cause death. The diagnosis of flares is based on the presence of certain clinical characteristics, some of which can only be assessed by performing a blood test (see Petri et al. (1991)). A functional model for explaining flare dynamics based on differential equations has recently been developed (Ramsay et al., 2007). The association between the time evolution of stress and flares in SLE patients has been studied by many authors (see for example Pawlak et al. (2003)). It would be very helpful to control flares before they occur, and so in this paper a methodology is proposed that may help predict the probability of a flare some time before it occurs, by controlling the daily stress level suffered by the patient.

Functional models describe practical situations in which the data involved are sample curves that are observed only at a finite set of time points. Examples of functional data analysis (FDA) methodologies include functional principal component analysis (FPCA), which is used to reduce the dimension in a stochastic process, functional linear regression (used to model a scalar response variable from a functional covariate) and functional canonical correlation analysis, which is applied to investigate different modes of variability in two sets of curves. A very good review of functional methods and their applications in a diverse range of subject areas can be consulted in Ramsay and Silverman (2002, 2005) and Ferraty and Vieu (2006). With respect to functional regression models, different methods, in which the response and predictor variables are functional, have been proposed (Zeger and Diggle, 1994; Yao et al., 2005), while studies have also been made of functional generalized linear models with different types of response and link functions when only the predictor is functional

[*] Corresponding address: Universidad de Granada, Departamento de Estadística e I.O., Facultad de Ciencias, Campus Fuentenueva s/n, 18071-Granada, Spain. Tel.: +34 958246306; fax: +34 958243267.
E-mail addresses: aaguiler@ugr.es (A.M. Aguilera), escabias@ugr.es (M. Escabias), valderra@ugr.es (M.J. Valderrama).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.07.001

(Ratcliffe et al., 2002; Escabias et al., 2005; James, 2002). On the other hand, principal component prediction (PCP) models, used to forecast a functional variable in a future period from its recent past, have been studied in depth (Valderrama et al., 2000). Several reviews of the current state of development of statistical methods for analyzing functional data and new trends in this active field of statistical research have recently been published in special issues on FDA in leading statistical journals. See, for example, Davidian et al. (2004) in the special issue *Emerging Issues in Longitudinal and Functional Data Analysis* of the journal Statistica Sinica, (Valderrama, 2007) in *Modelling Functional Data in Practice* in the journal Computational Statistics and (González-Manteiga and View, 2007) in *Statistics for Functional Data* in the journal Computational Statistics and Data Analysis.

In order to predict flares in SLE patients from the historical evolution of their daily stress level, the functional logistic regression (FLR) model is proposed. This model has been designed to predict a time-independent binary response variable (occurrence of a flare) from the time evolution of a functional predictor variable (the daily stress level). The FLR model has been studied in several recent papers (James, 2002; Ratcliffe et al., 2002; Escabias et al., 2004, 2005; Müller and StadtMüller, 2005).

Time-dependent data involved in functional models are well known in several fields, such as in medicine, as longitudinal data. These data consist of observations carried out on a single variable repeatedly of the same subject at different time points. Longitudinal data have previously been used to predict a response variable by considering them as observations of different correlated variables and by including the dependence framework in a classical multiple regression model (see Diggle et al. (2002) for a good review of longitudinal data analysis and (Liang and Zeger, 1986) for the particular case of generalized linear models with longitudinal data). As a result of the considerable development of functional models, a different point of view can now be used to predict a response from longitudinal data by considering them as observations of sample curves. A comparison of the two perspectives and methods in functional data analysis and longitudinal data analysis is provided in the article by Rice (2004). In the present paper, these two points of view (longitudinal and functional data analysis) are compared to predict the risk of flares in SLE patients by using both multiple and functional logistic regression approaches.

In order to introduce longitudinal data into a functional model, the true functional form of the underlying curves has to be reconstructed. The most commonly used method consists of considering that the curves belong to a finite space generated by a basis of functions. Many types of basis have been described in the literature, depending on the nature of the sample paths. If the curves are expected to be smooth with a sinusoidal aspect, the basis most often used is that of trigonometric functions (Aguilera et al., 1995), while B-spline functions (Escabias et al., 2004) or wavelet functions (Ocañna et al., 1998) are employed to explain the local behaviour of sample curves. Monotone piecewise cubic interpolation of the sample paths has been proposed to approximate the mean of a doubly stochastic Poisson process (Bouzas et al., 2006). Diverse methods have been proposed for computing the basis coefficients of sample curves from their discrete time observations (longitudinal data). Unobserved basis coefficients estimated by using the EM algorithm have been considered by James (2002). Quasi-natural cubic spline interpolation on longitudinal data observed without error has been developed by Escabias et al. (2005). Many of the papers on FDA assume that the functional variable is observed at the same time points on each subject and that the measurements grid is sufficiently fine for this purpose. Functional nonparametric statistical methods based on functional local weighting kernel techniques could be used to smooth sample curves from this type of discretized data (Ferraty and Vieu, 2006). This kind of balanced data set is not commonly found, however, due to the absence of certain data or the impossibility of observing a variable on a certain subject at a specific time point. In order to approximate the basis coefficients when longitudinal data are missing (SLE, for example), the use of least squares approximation for each individual functional observation (stress level) is now proposed.

Most papers on FLR models focus on predicting the response variable; for example, the probability of a high risk birth outcome is predicted from periodically stimulated foetal heart rate tracings (Ratcliffe et al., 2002). In addition to estimating the probability of lupus flares, there is special interest in interpreting the relationship between lupus flares and stress levels, in terms of obtaining an accurate estimation of the parameter function of a FLR model. A new method to interpret the parameter function is now proposed, based on evaluating the change in the odds of success (lupus flare) for a general functional increment of the functional predictor (stress level) in the observation interval.

The approximated solution most commonly used for estimating a functional model consists in transforming it into a multiple model after approximating sample paths and parameter functions in terms of a set of basic functions. In the case of functional logistic regression, this model fits well and provides good predictions but inaccurate estimations of the parameter function because of the high correlation between longitudinal observations (multicollinearity) (Escabias et al., 2004). Different approaches have been developed to solve this problem; principal component regression and principal covariate regression are often employed to forecast a response variable from highly correlated predictors (see Heij et al. (2007) for a comparison of the forecast accuracy of these two methods). In a recent paper, a functional PLS logit regression (FPLSLR) model that can be seen as a generalization of the functional PLS regression model (proposed by Preda and Saporta (2005)) was proposed (Escabias et al., 2007). On the basis of several simulation studies, it has been concluded that the accuracy of the estimations provided by FPLSLR is similar to that of the functional principal component logit regression (FPCLR) model developed by Escabias et al. (2004), and better than that of alternative discrimination models such as the classification and regression tree procedure (CART), functional discriminant analysis (FDA), multivariate partial least-squares regression (MPLSR), penalized discriminant analysis (PDA) and nonparametric curve discrimination (NPCD) (see Ferraty and Vieu (2003) for a detailed explanation).

A functional PCA based approach is proposed to overcome this problem in the case of SLE data. The model is based on using as covariates of the logit model a reduced set of principal components associated to the design matrix of the multiple model obtained after basis expansion of the sample curves and the parameter function. An alternative nonparametric estimation procedure of this functional logit model could be developed by applying nonparametric smoothing procedures to estimate the mean and covariance functions of sample curves, and then the corresponding principal components (Rice and Silverman, 1991). In this context, generalized functional linear models that include functional binary regression models for longitudinal data have recently been generalized to the extreme case of sparse longitudinal data (few irregularly spaced observations), by using a new nonparametric FPCA estimation procedure (Yao et al., 2005; Müller, 2005).

This article is organized in four sections. The first is an introduction describing the objectives of the paper and the present state of art concerning the FLR model. A classical multiple logistic regression treatment of longitudinal data and a principal component (PC) based solution to solve the multicollinearity problem is developed in Section 2. An estimation of the FLR model from missing longitudinal data, a PC based solution to the multicollinearity problem and a new method for interpreting the association between response and predictor variables are introduced in Section 3. These methodologies are then applied in Section 4 to estimate the probability of lupus flare and to interpret its relationship with stress level.

2. Multiple logistic regression for longitudinal data

The logistic regression model is the most commonly used method to explain a binary response variable from a set of related covariates. In order to use longitudinal data to predict a binary response Y by the multiple logistic regression model, it is assumed that the observations of a functional variable, $X(t)$, at different time points, $t_1, \ldots, t_p$, are observations of different variables, $X(t_1), \ldots, X(t_p)$ (Diggle et al., 2002). One of the main disadvantages of this multiple approach with respect to the functional one is that the functional variable has to be observed at the same time points for each individual, and so incomplete variables (missing data) are usually removed.

Let $\left\{\left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)' = \left(x_i\left(t_1\right), x_i\left(t_2\right), \ldots, x_i\left(t_p\right)\right)', i = 1, \ldots, n\right\}$ be the observations at p different time points of a sample of a functional variable, and let $(y_1, y_2, \ldots, y_n)'$ be the vector of observations of an associated binary random variable. Then the multiple logistic regression model is defined as $y_i = \pi_i + \varepsilon_i$, where the probability of success is modelled as

$$\pi_i = P\left[Y = 1|(x_{i1}, x_{i2}, \ldots, x_{ip})\right] = \frac{\exp\left\{\alpha + \sum_{j=1}^{p} x_{ij}\beta_j\right\}}{1 + \exp\left\{\alpha + \sum_{j=1}^{p} x_{ij}\beta_j\right\}}, \tag{1}$$

ε_i are centered independent random errors with unequal variances $\pi_i\left(1 - \pi_i\right)$ and $\alpha, \beta_1, \ldots, \beta_p$ are the parameters to be estimated. This model is usually expressed in matrix linear form in terms of the logit transformations $l_i = \ln\left[\pi_i / \left(1 - \pi_i\right)\right]$ as $L = \mathbf{1}\alpha + X\beta$, with $L = (l_1, \ldots, l_n)'$ being the vector of logit transformations, $\mathbf{1} = (1, \ldots, 1)'$ the n-length vector of ones, X the $n \times p$ matrix that has as rows each longitudinal sample observation and $\beta = \left(\beta_1, \ldots, \beta_p\right)'$ the vector of parameters.

In the longitudinal case, the β_j parameter is associated to the variable $X(t_j)$ (jth observation of the longitudinal variable) and can be interpreted as the additive change produced in the logit transformation when the longitudinal variable is increased by one unit in t_j and remains constant in the rest (see Hosmer and Lemeshow (2000)). Equivalently, $\exp\left\{\beta_j\right\}$ is the multiplicative change in the odds of success ($Y = 1$) produced when an additive change of one unit is produced in the longitudinal variable at the time t_j.

The estimation of the parameters of this model is not very accurate when there is strong dependence between the explicative variables (multicollinearity) (Ryan, 1997). This fact impedes the correct interpretation of such parameters in terms of odds ratios. Nevertheless, the model may fit well if accurate goodness of fit measures are applied. The usual goodness of fit measures for logistic regression are the area under the ROC curve and correct classification rates (CCR) for the response. A CCR is defined as the rate of correctly classified individuals taking into account that individuals are correctly classified when their estimated probabilities agree with their response observations, that is, after setting a cut-off point, individuals are classified as $Y = 1$ if their estimated probabilities are greater than the cut-off point and $Y = 0$ otherwise. On the other hand, it is known that standard goodness-of-fit tests (residual deviance and Pearson chi-square statistics) behave unsatisfactorily when the data contain only a small number of observations for each pattern of covariate values (extreme sparseness). This is the case of the SLE data analyzed in this paper, where there is only one observation for each different stress curve. The Hosmer–Lemeshow (H&L) test is frequently recommended to solve this problem. This is a chi-squared goodness of fit test that is computed from a new grouping of the observations, to avoid sparseness, and which depends on the estimated probabilities being usually grouped in deciles. See Hosmer and Lemeshow (2000) for details about the definitions of such goodness of fit statistics. Alternative non-standard tests that also perform well with sparse data (Farrington test and the information matrix test) have been compared with the (H&L) test on simulated data (Kuss, 2002).

When the covariates of the logit model come from the observations of a longitudinal variable at different time points, there is necessarily multicollinearity between the explicative variables being considered. Different solutions have been proposed to avoid problems related to multicollinearity, and the dependence framework between covariates in the model could be involved in this (Liang and Zeger, 1986). Unlike these solutions, the use of a reduced set of principal components (PCs) of the longitudinal data as covariates of the logistic model is considered in this paper. Thus, the original parameters are reconstructed in terms of those estimated by the resulting principal component model.

2.1. Principal component estimation

The principal components (PCs) of a set of variables are defined as the centred uncorrelated variables with maximum variance, obtained as linear spans of the original variables whose coefficients are the eigenvectors of the sample covariance matrix. Let $Z = XV$ be the matrix of PCs of the X matrix with V being the matrix of eigenvectors of the covariance matrix of X. The logistic model can be expressed in matrix form in terms of all PCs as $L = 1\alpha + X\beta = 1\alpha + Z\gamma$, where $\beta = V\gamma$. The principal component logistic regression (PCLR) model is then defined in terms of a compact set of s PCs by $L_{(s)} = 1\alpha_{(s)} + Z_{(s)}\gamma_{(s)}$, where $Z_{(s)}$ is the matrix whose columns are s selected PCs and $V_{(s)}$ is its associated matrix of eigenvectors. An estimation of the original β parameters can then be obtained through the estimation of this model as $\widehat{\beta}_{(s)} = V_{(s)}\widehat{\gamma}_{(s)}$. This estimation is more accurate than that provided by the logit model without PCs (Aguilera et al., 2006).

Different criteria have been employed to select the PCs to be retained in principal component regression. The most commonly used method consists in discarding the PCs with the lowest variances (eigenvalues of the sample covariance matrix). However, in many cases the last PCs may be more explicative of the response than the first ones. A criterion based on the partial correlation coefficient between the PCs and the explicative variables was proposed in Foucart (2000). The inclusion of variables in a principal component linear regression model in the order given by a stepwise method was considered by Aucott et al. (2000). In addition, the number of PCs to be retained in the model is an important aspect to take into account.

It has been shown by simulation that including PCs in the PCLR model in the order given by a stepwise method based on conditional likelihood ratio tests provides more accurate estimated parameters than in the order given by the explained variance (Aguilera et al., 2006). Moreover, it has also been shown that after adding the next PC (in the stepwise order) to the model that provided the best possible estimation of the parameters (minimum mean squared error), the estimated variance of the estimated parameters increases very noticeably. Therefore, in order to obtain the most accurate estimation of the β parameters, PCs are included in the logistic model according to their statistical significance given by a stepwise method based on the conditional likelihood ratio test, and the best PCLR model is chosen as the one with a number of PCs previous to a significant increase in the estimated variance of the estimated parameters given by $\mathrm{Var}_{(s)} = \mathrm{Var}\left[\widehat{\beta}_{(s)}\right] = V_{(s)}\left(Z'_{(s)}\widehat{W}_{(s)}Z_{(s)}\right)^{-1}V'_{(s)}$, with $\widehat{W}_{(s)}$ being the diagonal matrix of elements $\widehat{\pi}_{i(s)}\left(1 - \widehat{\pi}_{i(s)}\right)$, and $\widehat{\pi}_{i(s)}$ being the probabilities estimated by the corresponding PCLR model.

3. Functional logistic regression

This section presents the theoretical tools of the functional point of view for longitudinal data, and provides an introduction to the FLR model, a PC based solution to the multicollinearity problem and a new procedure for interpreting the change in the odds of success from the estimated parameter function. The main advantage of this functional approach is that it takes into account the distances between the observed data, and so the functional predictor can be observed at different and unequally spaced time points for each sample individual. The first step is to summarize different procedures for approximating the true functional form of observations from longitudinal data.

3.1. Reconstructing the functional form of missing longitudinal data

Let $\{x_{i1}, x_{i2}, \ldots, x_{im_i}, i = 1, \ldots, n\}$ be observations of a single variable taken repeatedly on the ith subject of a size n sample at m_i time points (longitudinal data). These are seen by FDA methods as observations of a set of n curves at different time points. That is, for the ith subject $x_{ik} \approx x_i(t_{ik})$, $(k = 1, \ldots, m_i)$, with $\{x_i(t), t \in T\}$ being its associated sample curve. As it is impossible to observe each curve continuously in time, it is necessary to reconstruct its true functional form from the corresponding discrete-time observations.

In most studies published on this question, it is assumed that sample curves belong to a finite dimension space generated by a basis of functions $\left\{\phi_1(t), \ldots, \phi_p(t)\right\}$, such that each curve can be expressed in terms of the basis as $x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t)$.

When discrete-time observations are assumed to be measured without error, $x_{ik} = x_i(t_{ik})$, an interpolation method to estimate the basis coefficients can be used. This is the case of quasi-natural cubic spline interpolation used by Escabias et al. (2005) for estimating the parameter function of a functional logistic model. On the other hand, if a degree of error is assumed to exist in the observations $x_{ik} = x_i(t_{ik}) + \varepsilon_{ik}$, least squares approximation is usually used for estimating the basis coefficients for a specific curve as $a_i = \left(a_{i1}, \ldots, a_{ip}\right)' = \left(\Phi_i'\Phi_i\right)^{-1}\Phi_i'x_i$, with $\Phi_i = \left(\phi_j(t_{ik})\right)_{m_i \times p}$ and $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{im_i}\right)'$ (Escabias et al., 2004).

Note that the functional variable could be recorded at different time points for each individual (missing longitudinal data). In the proposed approach, the basis coefficients are estimated by using least squares approximation on the same basis for each patient from the days he/she answers the stress test. In addition, two different types of basis, cubic B-splines and Fourier basis, are used.

3.2. Principal component estimation

Firstly, the FLR model is formulated as a generalization of the logistic regression model and estimated after considering the sample curves and parameter function expressed in terms of a basis. Secondly, an accurate estimation of the parameter function is obtained by using a functional principal component logistic regression (FPCLR) model.

Let $x_1(t), \ldots, x_n(t)$ be a random sample of curves of a functional variable $\{X(t) : t \in T\}$, and let $y_1, y_2, \ldots, y_n$ be the random observations of the binary response variable Y associated with the curves. Then, the FLR model is given by $y_i = \pi_i + \varepsilon_i$ $(i = 1, \ldots, n)$ with ε_i being zero mean independent random errors with variance $\pi_i(1-\pi_i)$, and the probability of response $Y = 1$ for a specific curve $x_i(t)$ modelled as

$$P\{Y = 1 | X(t) = x_i(t)\} = \pi_i = \frac{\exp\left\{\alpha + \int_T x_i(t)\beta(t)\mathrm{d}t\right\}}{1 + \exp\left\{\alpha + \int_T x_i(t)\beta(t)\mathrm{d}t\right\}},$$

with α being a real parameter and $\beta(t)$ a parameter function that has to be estimated. Equivalently, in terms of the logit transformations, $l_i = \ln\left[\pi_i/\left(1-\pi_i\right)\right]$, the model can be seen as a functional generalized linear model (James, 2002) given by

$$l_i = \alpha + \int_T x_i(t)\beta(t)\mathrm{d}t, \quad i = 1, \ldots, n. \tag{2}$$

As in the linear case (see Ramsay and Silverman (2005)) it is impossible to obtain a direct estimation of model (2). Therefore, what is usually done is to fit a related multiple model derived from the reconstruction of the sample curves from the longitudinal data, making the assumption that the parameter function and the sample curves belong to the same space generated by a basis $\left\{\phi_1(t), \ldots, \phi_p(t)\right\}$. Then, they can be expressed as

$$\beta(t) = \sum_{k=1}^{p} \beta_k \phi_k(t), x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t), \quad i = 1, \ldots, n, \tag{3}$$

so that the functional model given by Eq. (2) is equivalent to the following multiple logit model given in matrix form by

$$L = \mathbf{1}\alpha + A\Psi\beta, \tag{4}$$

with $L = (l_1, \ldots, l_n)'$, A the matrix that has the basis coefficients of the sample curves as rows, $\Psi = \left(\psi_{jk}\right)_{p\times p}$ the one that has the L^2-usual inner products between the basis functions as entries, $\left(\psi_{jk} = \int_T \phi_j(t)\phi_k(t)\mathrm{d}t\right)$, and $\beta = (\beta_1, \ldots \beta_p)'$ the vector of the parameter function basis coefficients.

Observe that the parameter function $\beta(t)$ could belong to a different space from that of the sample functions spanned by a different basis $\{\varphi_1(t), \ldots, \varphi_m(t)\}$. In this case $\beta(t) = \sum_{k=1}^{m} \beta_k\varphi_k(t)$, so that in Eq. (4) Ψ would be a $p \times m$ matrix whose entries are the L^2-usual inner products between the functions of the two different bases, $\left(\psi_{jk} = \int_T \phi_j(t)\varphi_k(t)\mathrm{d}t\right)$, and $\beta = (\beta_1, \ldots \beta_m)'$.

The last multiple model has a high degree of multicollinearity between the columns of its design matrix (see Escabias et al. (2004)). As stated above, this dependence provides inaccurate estimations of the parameters of the multiple model and hence of the parameter function too. Estimation of these parameters is improved by using a limited set of PCs of the design matrix $X = A\Psi$ (FPCLR model).

It can be shown that multiple PCA of $A\Psi$ is equivalent to functional PCA of the sample curves with respect to an inner product different from the usual one in $L^2(T)$ (Ocaña et al., 2007). A detailed study of the equivalences between FPCA with a given inner product and FPCA with a given well-suited inner product can be found in Ocaña et al. (1999). We now obtain the following PC decomposition of the original sample curves $x_i(t) = \bar{x}(t) + \sum_{r=1}^{p} z_{ir}f_r(t)$, where z_r are the PCs of the $A\Psi$ matrix (columns of Z matrix) and $f_r(t)$ are the eigenfunctions given in terms of the basis functions as $f_r(t) = \sum_{j=1}^{p} f_{jr}\phi_j(t)$, $(r = 1, \ldots, p)$, where the basis coefficients are obtained in matrix form as $F = \left(f_{jr}\right)_{p\times p} = \Psi^{-1}V$ from the eigenvector matrix V of the sample covariance matrix of $A\Psi$.

Finally, let $\widehat{\beta}_{(s)} = \left(\widehat{\beta}_{1(s)}, \ldots, \widehat{\beta}_{p(s)}\right)'$ be the most accurate estimation of the parameters of the multiple model (4) obtained from the PCLR model by following the PCs selection criterion described in the previous section, based on the stepwise method. Then, the most accurate estimation of the parameter function is $\widehat{\beta}_{(s)}(t) = \sum_{k=1}^{p} \widehat{\beta}_{k(s)}\phi_k(t)$.

3.3. Interpreting the parameter function

Interpretation of the parameter function is important because it may help to assess the relationship between the response and the functional predictor variable. Several attempts have recently been made to interpret the parameter function of the FLR model. An interpretation of the parameter function asserting that high absolute values of the parameter function indicate times with a large influence on the response whereas small values represent times with little influence has been given in James (2002).

An interpretation of the parameter function $\beta(t)$ is now proposed in terms of odds ratios from which the change in the odds of flares in SLE patients can be assessed.

Let l_i^* be the logit transformation corresponding to a curve $x_i^*(t)$ obtained by increasing the curve $x_i(t)$ in accordance with a function $g(t)$ in an interval of amplitude h $(h > 0)$, denoted by $[t_0, t_0 + h]$

$$x_i^*(t) = \begin{cases} x_i(t) + g(t) & \text{if } t \in [t_0, t_0 + h] \\ x_i(t) & \text{if } t \notin [t_0, t_0 + h]. \end{cases}$$

Then, the difference between the two logits is

$$l_i^* - l_i = \int_{t_0}^{t_0+h} \beta(t) g(t) \mathrm{d}t. \tag{5}$$

So, the integral of the parameter function multiplied by $g(t)$ is the change in the logit transformation provided by an increment of the curve $x_i(t)$ according to $g(t)$ in the interval $[t_0, t_0 + h]$. The difference between the logit transformations is the logarithm of the odds ratio

$$\theta\left(\Delta x_i(t) = g(t) : t \in [t_0, t_0 + h]\right) := \frac{\frac{\pi_i^*}{1-\pi_i^*}}{\frac{\pi_i}{1-\pi_i}},$$

so that, the exponential of the integral (5) is the multiplicative change in the odds of success $(Y = 1)$ provided by a change of the curve $x_i(t)$ according to $g(t)$ in the cited interval.

A particular case of this situation is that of a constant increment $(g(t) = K > 0)$ in a sample curve at a specific interval, such that the change in the odds of success is the integral of the parameter function multiplied by K. James (2002) shows that each parameter of the multiple linear model in terms of all functional PCs is the additive change obtained in the response variable when a functional observation changes according to the corresponding eigenfunction. In fact, from the PC decomposition of the original sample curves, it can be deduced that if the rth PC is increased by one unit, then the functional observation is increased according to the associated rth eigenfunction $f_r(t)$. In the logistic model, this interpretation can be seen as a particular case of the one given by (5) when the pattern of variation is given by an eigenfunction across the entire domain, that is, $g(t) = f_r(t)$ $t \in T$. Then,

$$\int_T f_r(t)\beta(t)\mathrm{d}t = \int_T \left(\sum_{j=1}^{p} f_{jr}\phi_j(t)\right)\left(\sum_{k=1}^{p} \beta_k \phi_k(t)\right) \mathrm{d}t = F_r' \Psi \beta = V_r' \beta = \gamma_r,$$

with F_r' and V_r' being the rth columns of matrices F and V, respectively. Then the integral of the parameter function multiplied by the rth eigenfunction is the rth parameter of the model in terms of all PCs, and its exponential represents the change in the odds of response $Y = 1$ for an individual whose sample curve changes according to the rth eigenfunction. In the particular case of the FPCLR model in terms of s PCs, $\exp(\gamma_{r(s)})$ is the multiplicative change in the odds of success for an increment of the functional variable according to its associated eigenfunction $f_r(t)$.

4. Modelling lupus flares from daily stress level

This section discusses the performance of different logistic models in predicting the risk of flares in SLE patients from their daily stress level. First, a classical multiple logistic regression model with longitudinal data is examined, followed by various functional approaches, whose results are compared and interpreted.

With the aim of predicting the risk of flares in lupus patients, analysis was made of real data provided by the Autoimmune Diseases Section of the Internal Medicine Department of the *Virgen de las Nieves* hospital (Granada, Spain) in a highly ambitious project to study many aspects related to lupus patients, one of which is stress. Of particular relevance to the present study were 44 SLE patients, who were asked to respond to different stress tests over a period of 18 days. The stress level was evaluated by carrying out different tests (the Self-Efficacy Scale of Sherer and Adams (1983), among others). The responses were analyzed by a team of psychologists from the Department of Personality, Testing and Psychological Treatment of the University of Granada, who estimated a daily stress level for each individual. After these 18 days, the occurrence of flare was tested according to Petri et al. (1991), and two patients were found to have suffered a flare. This ratio is in agreement with the established prevalence of flares in Lupus patients, i.e. 0.65 flares per year (see Petri et al. (1991)). As a result, the observation of the binary response variable was given the value of one if a flare had taken place and zero otherwise.

Table 1
Parameters estimated by the multiple logit model (second column) and the PCLR model with the second PC as covariate (third column)

Param	$\widehat{\beta}_j$	$\widehat{\beta}_{(1)j}$
α	−38.13	−3.92
β_1	−5.17	−0.05
β_2	2.17	−0.03
β_3	5.24	−0.03
β_4	−1.23	−0.03
β_5	−0.79	−0.02
β_6	5.62	0.004
β_7	3.66	0.01
β_8	−2.62	0.007
β_9	−2.72	−0.02
β_{10}	4.60	−0.002
β_{11}	4.40	0.009
β_{12}	−2.19	0.01
β_{13}	−4.91	0.02
β_{14}	−2.76	0.02
β_{15}	−7.64	0.04
β_{16}	3.86	0.05

Table 2
Goodness of fit measures for different PCLR models (PCs entered by stepwise selection) and for the multiple logit model (with all the PCs)

s	PCs	CCR1	CCR2	Var$_{(s)}$	G^2(df)	H&L(df)	p
1	2	97.73	90.91	1.17E+00	1.08E+1(43)	7.44(8)	0.49
2	2,1	97.73	84.09	1.56E+02	6.52E+0(42)	2.27(8)	0.97
3	2,1,13	100.00	100.00	3.73E+11	2.49E−9(41)	1.01E−10(1)	1.00
...	...	...	...	...	...	...	...
16	1,2,...,16	100.00	100.00	2.69E+10	4.63E−9(28)	7.88E−10(4)	1.00

Correct classification rates with 0.5 and the proportion of flares in the sample as cut-off points (CCR1, CCR2, respectively). Variances of estimated parameters, G^2 and H&L statistics, degrees of freedom (df) and p-values (p).

4.1. The multiple approach

The first problem encountered in using the classical logistic regression model is that this model needs the stress level to be observed at the same time points for all the 44 patients. This was not the case in the real data analyzed because several patients did not provide observations on the third and twelfth days. These time observations were removed for all the 44 patients, and so 16 unevenly spaced measurements of the daily stress level were available for each patient before the blood test.

First, a multiple logit model with 16 explicative variables was fitted (one for each day of observation) and its estimated parameters obtained (second column of Table 1), together with the usual goodness of fit measures (last row of Table 2). The logit model was a good choice for these data, as is clear from the usual H&L goodness-of-fit test (H&L $=$ 7.88E−10 with p-value 0.49) and CCR $=$ 100%. Nevertheless, the variance of the estimated parameters was very high (2.69E+10), which implies that this estimation was not very accurate.

Secondly, a PCLR model was fitted, to improve the estimation of the parameters of this logit model. The first PC of the design matrix explained 42.4% of the total variability, and eleven PCs were needed to explain at least 90%. The PCs were entered in the PCLR model in the order given by the stepwise method based on the conditional likelihood ratio test. From Table 2 it can be seen that only the second PC was included in the optimum model because the following one (the first PC) dramatically increased the estimated variance, which rose from 1.17 to 156. Therefore, the initial multiple model with 16 covariates was reduced to a simple one in terms of the second PC. Moreover, this simple model fitted well (CCR1 $=$ 97.73%, CCR2 $=$ 90.91%) and was a good choice for the data (H&L $=$ 7.44 with p-value $=$ 0.97). The β parameters estimated by this model can also be seen in Table 1 (third column). The parameters estimated by the PCLR model with only the second PC as covariate are very different from those estimated by the initial multiple model, which is equivalent to a multiple model that has all the PCs as covariates. This is due to multicollinearity and the main consequence is that the interpretation of the relationship between lupus flares and stress may be very different and even the complete opposite, depending on the model used.

When a multiple model with longitudinal data is used, the exponential of each single parameter can be interpreted as the multiplicative change in the odds of flare provided by a unit change in the stress level on a specific day. For example, if the stress level increases by one unit two days before the blood test (see $\widehat{\beta}_{15}$ and $\widehat{\beta}_{(1)15}$ from Table 1), the odds of flare are multiplied by $e^{-7.64} = 4.8\text{E}{-4}$ if the multiple model with all the PCs is used (the probability of lupus flare decreases), and by $e^{0.04} = 1.04$ if the model with only the second PC is used (the probability of flare increases). Nevertheless, this interpretation is not very significant because the stress level usually changes over a period of time and not in a single day.

The results obtained show the importance of accurately estimating the parameters of the logistic model in the presence of multicollinearity. The use of PCs improves the estimation of the parameters of the multiple logistic model with longitudinal data. Nevertheless, the use of a multiple model for longitudinal data made it necessary to remove some time observations and to lose some information provided by them. There are many real situations in which the removal of data might mean that few data are available with which to fit the model.

4.2. The functional approach

The first step taken was to obtain the functional form of each stress level sample curve by least squares approximation from the longitudinal observations. Then, the FLR model was fitted on a different basis and the most frequently used goodness of fit measures and the estimated parameter function obtained. These estimations were not very accurate due to multicollinearity (moreover, the interpretations of the parameter function were inconsistent). Finally, a FPCLR model was fitted to solve this problem and the best possible parameter function estimation obtained. The results with two different bases of functions (Fourier and B-splines basis), the improvement in the estimation and the interpretation of the parameter function can be seen below.

In relation to the basis considered, an important factor is the number p of basis functions to be used. According to the literature, the most frequently used criterion for selecting the dimension of the Fourier basis in functional regression models is that of cross-validation (Ratcliffe et al., 2002). In the present case, a heuristic criterion was assumed, consisting in choosing the lowest number of basic functions that provided minimal changes in the most accurate estimated parameter function, this was found to be $p = 6$. In the B-spline basis case, the number of basis functions depends on the number of definition knots and their allocation, and the approach adopted was to make a heuristic selection among the observation knots given by the knots (0, 3, 5, 6, 10, 12, 14, 17), which provided a basis of cubic B-spline functions of dimension 10. Many articles have addressed the question of knot selection, for example Zhou and Shen (2001).

Some of the sample curves approximated by least squares with these two different bases, the curve of mean stress level among observations, given by $\bar{x}(t) = n^{-1}\left(\sum_{i=1}^{n} x_i(t)\right)$, for individuals with $y = 0$ and $y = 1$ can be seen in Fig. 1 together with the observed daily stress level (points) on the individual curves. In all cases, the solid line curve corresponds to the curve approximated by using cubic B-spline functions, while the broken line curve corresponds to the Fourier basis. This figure shows that the cubic B-spline basis provided better approximations than the Fourier basis, especially at the boundaries of the observation time period. In addition, the mean functions reveal a similar behaviour of the two bases, with little difference at the boundaries. Therefore, it is concluded that in the case of stress level, where irregular trends are commonly found, it is better to use the B-spline basis than the Fourier basis because cubic B-splines give a better approximation of the sample curve local behaviour.

After approximating the sample curves, the functional logit model (2) was fitted by using the usual multiple approximation given in Eq. (4). The parameter functions estimated with the two described bases are shown in Fig. 3. In order to compare the multiple and functional models, the natural cubic spline interpolation of the estimated parameters of the multiple approach was obtained (dotted line function). Clearly, there are large differences between the estimated parameter functions provided by the three approaches. Although the functional models fitted well with high CCR and low H&L statistics with both bases (see Table 3), the variances of the estimated parameter functions (Table 3) were extremely large (6.12E+12 for the Fourier basis and 2.23E+12 for the cubic B-spline basis). This is indicative of inaccuracy in the estimated parameter functions, which may be due to multicollinearity.

In order to improve these estimations, the FPCLR model was fitted. The first PC alone accounted for 63.17% of the total variance with the Fourier basis and 94.76% with the cubic B-splines basis, while four PCs were needed to account for at least 90% of the variability with the Fourier Basis. As stated in Section 2.1, on many occasions the most explicative principal components may not be the best predictors for the response. This is one example, and when PCs are entered in the functional model by a stepwise selection based on the conditional likelihood ratio test, the only significant PC for the response is the third one, with each of the approximating bases. The percentage of variance accounted for by the third principal component was 1.06% and 9.22% in the B-spline and Fourier basis cases, respectively. Fig. 2 illustrates the predictive ability, where the S-shaped models with the 3rd PC (B-spline basis) suggests that it is a better predictor than the first PC in spite of its lower explicative power. The corresponding curves for the Fourier basis case were very similar.

The FPCLR parameter function estimation, for both cases, is shown in Fig. 3. It can be seen that the estimated parameter functions are very similar except at the beginning of the domain. However, the cubic interpolation with the parameters estimated by the PCLR model is very different, revealing the improvement obtained in the estimation of the parameter function of the FLR model by using a small number of PCs (in the present case, only the third PC) versus not using them, and the improvement, moreover, of using the functional model against the classical multiple treatment of longitudinal data. Note, too, that although the shape of the estimated function when PCs are not used in the Fourier basis case does not differ greatly from that obtained by the optimum model with the third PC, the scale is very different, which causes large differences in the integrals of the parameter functions and their interpretation in terms of odds ratios.

Table 3 and Fig. 4 show the goodness of fit measures for the FPCLR models. The model with only the 3rd PC provides high CCR rates with both bases and similar estimated variances of the estimated parameter function, with 0.9892 and 1.074 for the Fourier and B-spline bases, respectively, which differs greatly from the variances estimated without using PCs. The p-values of the H&L test corroborate the goodness-of-fit of the models (p-value = 0.14 for the B-spline basis and p-value = 0.28

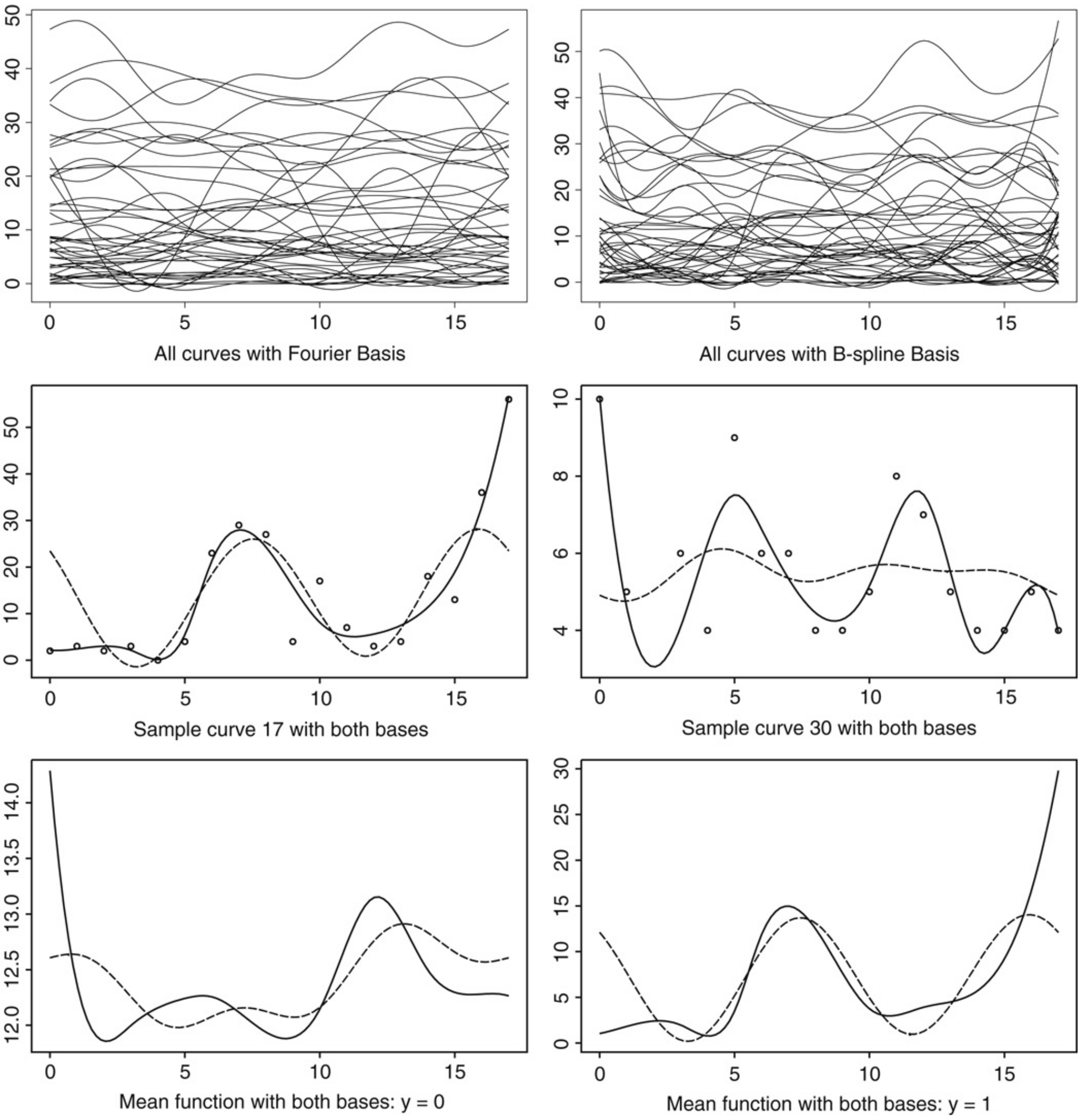

Fig. 1. Least squares approximations of stress level curves obtained with Fourier (broken line) and B-spline (solid line) bases and the discrete-time observations (points).

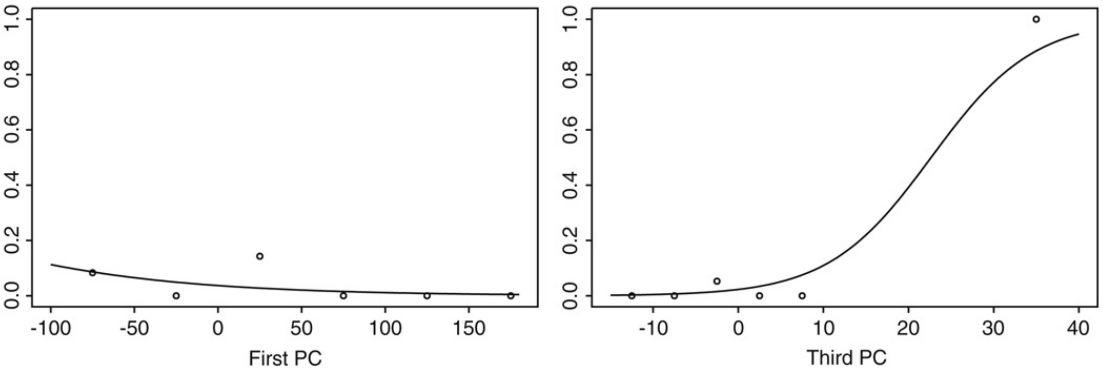

Fig. 2. Scatter plots of grouped data: average proportion of flares per class against class midpoints next to the fitted logit curve.

for the Fourier basis). The areas under ROC curves for the models with the 3rd PC as response and the models with all PCs are shown in Fig. 4, together with the associated ROC curves. Thus, these models provide a good fit and may provide a more accurate estimation of the parameter function.

In order to overcome the possible overestimate of the CCR as a measure of the goodness of fit, a bootstrap based study was developed, by using 100 training samples and 100 test samples of the response and the 3rd functional PC with the

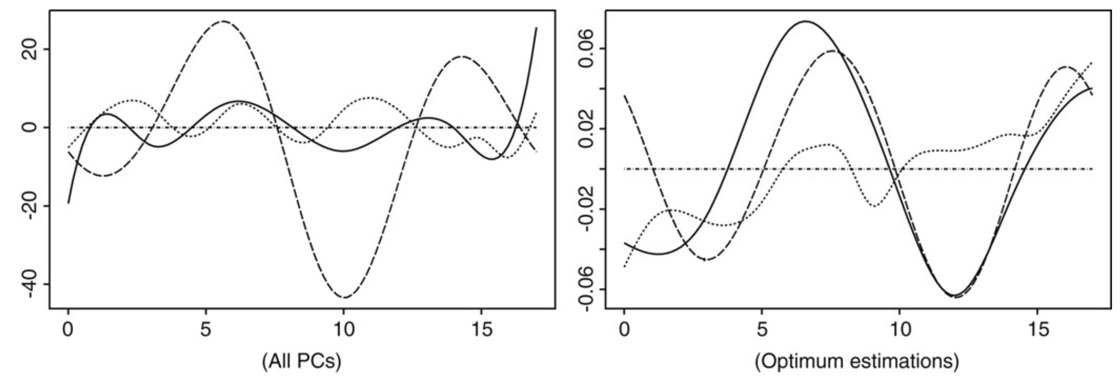

Fig. 3. Estimated parameter functions for functional models with B-spline basis (solid line) and Fourier basis (broken line), and the interpolation of the parameters estimated by the multiple model (dotted line).

Table 3
Goodness of fit measures for different FPCLR models: with all the PCs and with the 3rd

Model	CCR1	CCR2	Var	G^2(df)	p	H&L(df)	p
B-spline basis							
All PCs	100.00	100.00	2.23E+12	3.12E−9(34)	1	3.13E−10(3)	1.00
With 3rd PC	97.73	84.09	1.074	10.39(42)	1	12.19(8)	0.14
Fourier basis							
All PCs	100.00	100.00	6.12E+012	4.28E−9(38)	1	1.3E−11(1)	1.00
With 3rd PC	97.73	93.18	0.9892	10.39(42)	1	9.747904(8)	0.28

Correct classification rates with 0.5 and the proportion of flares in the sample as cut-off points (CCR1, CCR2, respectively). Variances of estimated parameter functions, G^2 and H&L statistics, degrees of freedom (df) and p-values (p).

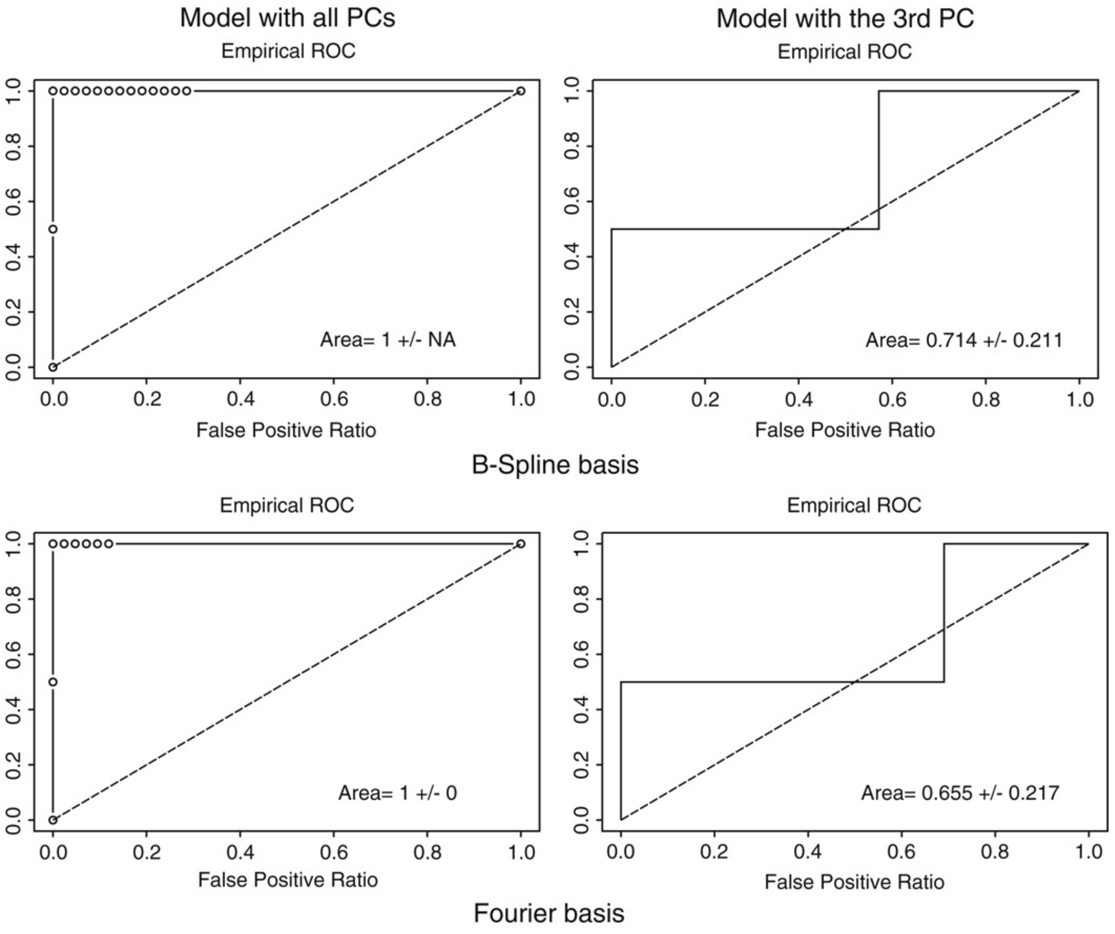

Fig. 4. ROC curves and areas under the curves.

Table 4
Bootstrap study for the CCR

	Training sample		Test sample	
	Mean	St. dev	Mean	St. dev
CCR1	97.54545	2.183272	96.97727	2.817141
CCR2	70.72727	30.51797	70.09091	29.46273

B-spline basis. The logit model was fitted with each training sample and the probabilities estimated over the test samples. The mean and standard deviation of the CCRs (over 100 samples) in each type of sample and using two different cut-off points, 0.5 (CCR1) and the proportion of flares in the sample (CCR2), are shown in Table 4. Although the CCR are somewhat slower in the test sample, from using the proportion of flares in the sample, the results show the models to have a reasonably good prediction ability.

Taking into account the similarity between the estimated parameter functions provided by the FPCLR model with the two bases, henceforth the interpretation of the relationship between lupus flares and stress level is analyzed on the basis of that given by B-splines.

First, looking at the shape of the estimated parameter function, it can be observed that people with a high stress level around the maxima of that function ($t = 6.6$ and $t = 17$) have a higher probability of suffering a lupus flare, while an absolute high level around the minima ($t = 1.2$ and $t = 12$) would reduce this probability. As a result, it can be concluded that the consequences of high stress level on a lupus flare have a lag of approximately five days.

Thus, it is possible to derive an interpretation based on the estimated parameter $\widehat{\gamma}_3 = 0.16$ (with both bases) of the optimum FPCLR model that has only the third PC as covariate. In this sense, if the stress level of an individual increases according to the third eigenfunction, that is, when the third PC increases by one unit, the odds of lupus flare will be multiplied by $\exp\{\widehat{\gamma}_3\} = \exp\{0.16\} = 1.17$. This means that if the stress level increases by five times the third eigenfunction, then the odds of lupus flares are doubled. However, if all functional PCs are used as in James (2002), the estimation of the third parameter is $\widehat{\gamma}_3 = 15.61$ for the Fourier basis and $\widehat{\gamma}_3 = 3.20$ for the B-spline basis. This means that if the stress level of an individual changes according to the third eigenfunction, the odds of lupus flare are multiplied by $\exp\{3.2\} = 24.53$ in the B-spline case and by $\exp\{15.6\} = 6.0\text{E}+6$ in the Fourier case. These results contradict those provided by the FPCLR model and would provide an erroneous interpretation because of the high dependence between longitudinal data.

5. Conclusions

This article examines a novel approach to functional logistic regression, the estimation of which constitutes a considerable improvement on that provided by the classical multiple logistic regression model, for the longitudinal data studied. This improvement in the estimation corroborates previous studies in which it has been shown that FDA approaches, such as functional PCA or functional regression, are superior to their multivariate counterparts in the case of unbalanced data (irregularly spaced grid of measurements that may change from one unit to another one) (Castro et al., 1986; Aguilera et al., 1999). Daily observations of the stress curves, considered as predictors of lupus flares, are an example of missing data that lead to irregularly spaced designs because some individuals do not respond to the stress test every day. Multivariate methods for analyzing correlated data do not take into account this unequal distance between observation times, which may be the cause of their poorer results in parameter estimation.

When longitudinal data of stress levels are used to obtain the functional form of the observations by least squares approximation, it is better to use the B-spline than the Fourier basis because the functions obtained provide a better fit to the points observed. In the functional logistic regression model, it is better to estimate by using a compact set of functional PCs included in the model in the order given by the stepwise method based on the conditional likelihood ratio test, because this greatly reduces the dimension of the problem and estimates the parameter function more accurately. This accurate estimation makes it possible to interpret the parameter function in terms of odds ratios that generalizes others discussed in the literature. In the application with lupus patients, it is concluded that the consequences of high stress levels on a lupus flare have a lag of approximately five days, and that the odds of lupus flare are doubled when the stress level increases by five times the third eigenfunction shown in Fig. 5. However, these findings should be considered with caution because we have analyzed a small sample with a very small proportion of lupus flares. This is justified because lupus is a rare sickness and to have a lupus flare is a rare event among lupus patients. It is well documented in the statistics literature that logistic regression can sharply underestimated the probability of rare events. Methods for computing probability and parameters estimates that correct problems due to small samples of rare events have been discussed by King and Zeng (2001). Any case, these problems will be innocuous in some applications and it is not usual among applied researchers to correct for the underestimation of event probabilities. In addition, the problem of rare events data affects more to the estimation of the probabilities than the one of the parameters and the main objective of our proposal is to provide an accurate estimation of the parameter function.

On the other hand, the functional methodology applied in this paper to estimate the relationship between lupus flares and stress level is not specific for lupus sickness. In fact, this is a general methodology for estimating functional logit models that can be applicable to estimate any binary response variable in terms of discrete time observations of a related

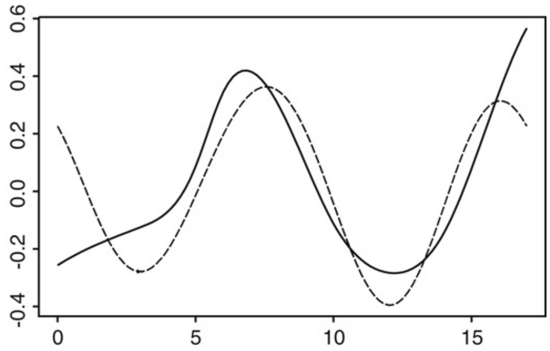

Fig. 5. 3rd eigenfunction: B-spline basis (solid line) and Fourier basis (broken line).

functional predictor. What differs in each real data application is the approximation technique used for reconstructing the true functional form of sample curves. If we consider, as in the stress level case, that time observations have been observed with error, least square approximation in terms of basis functions has proven to be effective. In other case, interpolation on the observed data could be more appropriate. As discussed in the introduction section, the basis must also be chosen depending on the nature and smoothness of the functional predictor sample curves.

Acknowledgements

This research has been funded by project P06-FQM-01470 from *"Consejería de Innovación, Ciencia y Empresa. Junta de Andalucía, Spain"* and project MTM2007-63793 from *Dirección General de Investigación, Ministerio de Educación y Ciencia, Spain.* The authors would like to thank professor Miguel Pérez (Department of Personality, Testing and Psychological Treatment of the University of Granada (Spain)) for providing us with the data. The authors also thank Professor J. Ramsay at McGill University for his helpful comments.

References

Aguilera, A.M., Escabias, M., Valderrama, M.J., 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Computational Statistics and Data Analysis 50 (8), 1905–1924.

Aguilera, A.M., Gutiérrez, R., Valderrama, M.J., 1995. Computational approaches to estimation in the principal component analysis of a stochastic process. Applied Stochastic Models and Data Analysis 11 (4), 279–299.

Aguilera, A.M., Ocaña, F.A., Valderrama, M.J., 1999. Forecasting with unequally spaced data by a functional principal component approach. Test 8 (1), 233–254.

Aucott, L.S., Garthwaite, P.H., Curral, J., 2000. Regression methods for high dimensional multicollinear data. Communications in Statistics: Computation and Simulation 29 (4), 1021–1037.

Bouzas, P.R., Valderrama, M.J., Aguilera, A.M., Ruiz-Fuentes, N., 2006. Modelling the mean of a doubly stochastic Poisson process by functional data analysis. Computational Statistics and Data Analysis 50 (10), 2655–2667.

Castro, P.E., Lawton, W.H., Sylvestre, E.A., 1986. Principal modes of variation for processes with continuous sample curves. Technometrics 28 (4), 329–337.

Davidian, M., Lin, X., Wang, J.-L., 2004. Emerging issues in longitudinal and functional data analysis with discussion. Statistica Sinica 14 (3), 613–630.

Diggle, P.J., Heagerty, P.J., Liang, K.Y., Zeger, S.L., 2002. Analysis of Longitudinal Data. Oxford, Oxford.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2004. Principal component estimation of functional logistic regression: Discussion of two different approaches. Journal of Nonparametric Statistics 16, 365–384.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2005. Modelling environmental data by functional principal component logistic regression. Environmetrics 16 (1), 95–107.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2007. Functional PLS logit regression model. Computational Statistics and Data Analysis 51 (10), 4891–4902.

Ferraty, F., Vieu, P., 2003. Curves discrimination: A nonparametric functional approach. Computational Statistics and Data Analysis 44, 161–173.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis. Springer-Verlag, New York.

Foucart, T., 2000. A decision rule for discarding principal components in regression. Journal of Statistical Planning and Inference 89, 187–195.

González-Manteiga, W., View, P., 2007. Statistics for functional data. Computational Statistics and Data Analysis 51 (10), 4788–4792.

Heij, C., Groenen, P.J.F., Dijk, D., 2007. Forecast comparison of principal component regression and principal covariate regression. Computational Statistics and Data Analysis 51 (7), 3612–3625.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, second edition. Wiley, New York.

James, G.M., 2002. Generalized linear models with functional predictors. Journal of the Royal Statistical Society, Series B 64 (3), 411–432.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political Analysis 9 (2), 137–163.

Kuss, O., 2002. Global goodness-of-fit tests in logistic regression with sparse data. Statistics in Medicine 21, 3789–3801.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.

Müller, H.G., 2005. Functional modelling and classification of longitudinal data. Scandinavian Journal of Statistics 32, 223–240.

Müller, H.G., StadtMüller, U., 2005. Generalized functional linear models. The Annals of Statistics 33 (2), 774–805.

Ocaña, F.A., Aguilera, A.M., Valenzuela, O., 1998. A wavelet approach to functional principal component analysis. In: Payne, R., Green, P. (Eds.), Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp. 413–418.

Ocaña, F.A., Aguilera, A.M., Valderrama, M.J., 1999. Functional principal components analysis by choice of norm. Journal of Multivariate Analysis 71 (2), 262–276.

Ocaña, F.A., Aguilera, A.M., Escabias, M., 2007. Computational considerations in functional principal component analysis. Computational Statistics 22 (3), 449–465.

Pawlak, R., Witte, T., Heiken, H., Hundt, M., Schubert, J., Wisse, B., Dischoff-Renken, A., Gerber, K., Licht, B., Goebel, M.U., Heijnen, C.J., Schmidt, R.E., Schedlowski, M., 2003. Flares in patients with systemic lupus erythematosus are associated with daily psychological stress. Psychotherapy and Psychosomatics 72, 159–166.

Petri, M., Genovese, M., Engle, E., Hochberg, M., 1991. Definition, incidence, and clinical description of flare in systemic lupus erythematosus. A prospective cohort study. Arthritis & Rheumatism 34, 937–944.

Preda, C., Saporta, G., 2005. PLS regression on a stochastic process. Computational Statistics and Data Analysis 48 (1), 149–158.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. Springer-Verlag, New York.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second edition. Springer-Verlag, New York.

Ramsay, J.O., Hooker, G., Cao, J., Campbell, D., 2007. Parameter estimation for differential equations: A generalized smoothing approach (with discussion). Journal of the Royal Statistical Society, Series B 69, 741–796.

Ratcliffe, S.J., Leader, L.R., Heller, G.Z., 2002. Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. Statistics in medicine 21 (8), 1115–1127.

Rice, J., 2004. Functional and longitudinal data analysis: Perspectives on smoothing. Statistica Sinica 14, 631–647.

Rice, J., Silverman, B., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society, Ser. B 53, 233–243.

Ryan, T.P., 1997. Modern Regression Methods. Wiley, New York.

Sherer, M., Adams, C.H., 1983. Construct validation of the self-efficacy scale. Psychological Reports 53 (3), 899–902.

Valderrama, M.J., Aguilera, A.M., Ocaña, F.A., 2000. Predicción Dinámica Mediante Análisis de Datos Funcionales. La Muralla-Hespérides, Madrid.

Valderrama, M.J., 2007. An overview to modelling functional data. Computational Statistics 22 (3), 331–334.

Yao, F., Müller, H.-G., Wang, J.-L., 2005. Functional data analysis for sparse longitudinal data. Journal of American Statistical Association 100, 577–590.

Yao, F., Müller, H.-G., Wang, J.-L., 2005. Functional linear regression analysis for longitudinal data. The Annals of Statistics 33 (6), 2873–2903.

Zeger, S.L., Diggle, P.J., 1994. Semiparametric models for longitudinal data with applications to CD4 cell numbers in HIV seroconverters. Biometrics 50, 689–699.

Zhou, S., Shen, X., 2001. Spatially adaptative regression splines and accurate knot selection. Journal of the American Statistical Association 96, 247–259.