X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

Implementation of functional GLM and GAM models in R

Manuel Oviedo y Manuel Febrero Bande

Universidade de Santiago de Compostela

ABSTRACT

The generalized functional linear model (GFLM) is used to estimate the relationship between a scalar response and functional covariates. The theoretical framework of these models is growing, but there are few programs available for use. Therefore, the aim of this paper is to provide the information necessary to use such models in practice. In addition, we also implemented the extension of the generalized additive model (GAM) to the functional case. Both methods (GFLM and GFSAM) are available in **fda.usc** package that provides an integrated framework for the treatment of functional data.

KEY WORDS: Additive Model, Functional Data Analysis, Functional Regression, Linear Model.

1. INTRODUCTION

There are many theoretical and applied studies available literature that extends the generalized linear model (McCullagh and Nelder, 1989) to the case in which the predictor is functional and the response is scalar. However the software to use these procedures is scarce. **fda** package is a basic reference work on the programming environment **R** with functional data, but all the techniques included are restricted to the space of $\mathcal{L}_2$ functions. Other software can use the functional generalized linear models (GFLM), such as **PACE** in **Matlab** or the work of Crainiceanu, 2010 in **Winbugs**. However, the package **fda.usc** provides a framework for functional data analysis broader than previous one by integrating the functional nonparametric methods implemented by Ferraty and Vieu 2006 and to complement and expand some of the functions **fda** package such as those shown in this work.

In the same way as for the GLM models extend generalized additive models (Hastie and Tibshirani, 1990) in which the predictor can be a nonlinear smooth function, such as a spline. This paper has been implemented in R the functional version of GLM and GAM model using basis representation of the the functional covariates. Finally, we use practical example in the case of a binary response using the functional logistic regression.

2. Generalized Functional regression models (GFLM)

In several applications for instance when the response is binary the functional linear model (FLM) may be too restrictive. One natural extension of this model is the generalized functional linear regression model (GFLM) Müller, 2005 which allows various types of the response and its expected value is related to this linear predictor via a link function. For example, with this approach in the case of a count data or binary variable would have the functional poisson or binomial regression, respectively.

In the GLM framework it is generally assumed that $y_i|x_i$ can be chosen within the set of distributions belonging to the exponential family with probability density function:

$$f(\theta, \phi, y) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{1}$$

where ϕ represents a scale (or dispersion) parameter and θ is the canonical parameter of distribution. The functions $a()$, $b()$ and $c()$ are known and differ for the distinct Y distributions, e.g., the normal, binomial or poisson distribution.

The estimation of the model parameters should be carried out by maximizing the likelihood function. The log-likelihood is:

$$l(\theta, \phi, y) = logf(\theta, \phi, y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \tag{2}$$

The model is specified as follows:

$$E[y|X] = b'(\theta) = \mu$$

$$Var[y|X] = b''(\theta)a(\phi) = V(\mu)\phi$$

$$g(\mu) = \left(\int_T X\beta + dt\right) + Z\beta$$

where μ is the expected value of response, $g()$ is the link function that specified the dependence between μ and the regressors, $V[\mu]$ is the conditional variance.

In **R** some of the principal distribution are specified together with the link function, see table 1.

The GFLM model is given by:

$$y_i = g^{-1}\left(\alpha + \beta_1 Z_i^1 + \cdots + \beta_p Z_i^p + \int_{T_1} X_i^1(t)\beta_1(t)dt + \cdots + \int_{T_f} X_i^q(t)\beta_q(t)dt\right) + \epsilon_i \tag{3}$$

where $Z = \left[Z^1, \cdots, Z^p\right]$ are the non functional covariates, $X(t) = \left[X^1(t_1), \cdots, X^q(t_q)\right]$ are the functional covariates and ϵ_i are random errors with mean zero and finite variance σ^2.

And the model 3 is estimated by the expression:

Distribution	ψ	$E(\mu)$	$V(\mu)$	Canonical link; in **R**
Binomial/n	1/n	μ	$\mu(1-\mu)$	$log(\mu(1-\mu))$; **logit**
Poisson	1	μ	μ	$log(\mu)$; **log**
Negative Binomial	1	$log\left(\frac{\mu}{1+1/\phi}\right)$	$\mu+\frac{\mu^2}{\phi}$	$log\,(\mu(\phi+\mu))$; **log**
Normal	σ^2	μ	1	μ; **identity**
Gamma	$1/\upsilon$	$-1/\upsilon$	μ^2	μ^{-1}; **inverse**

Table 1: Principal distributions used in GLMs.

$$\hat{y} = g^{-1}\left(\tilde{\mathbf{X}}\beta\right) = g^{-1}\left(\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}})^{-\mathbf{1}}\tilde{\mathbf{X}}^{\mathbf{T}}\right)y = g^{-1}\left(\mathbf{H}\right)y$$

where the first columns of $\tilde{X}$ are the p non-functional covariates Z and the following columns are the q scores. This scores can be done by: (i) basis expansion of class “fd” (see Ramsay and Silverman, 2005):

$$\tilde{X} = \left[Z^1,\cdots,Z^p,(\mathbf{C^1})^{\mathbf{T}}\psi(\mathbf{t_1})\phi^{\mathbf{T}}(\mathbf{t_1}),\cdots,(\mathbf{C^q})^{\mathbf{T}}\psi(\mathbf{t_q})\phi^{\mathbf{T}}(\mathbf{t_q})\right]$$

(ii) functional principal components basis s_j^i:

$$\tilde{X} = \left[Z^1,\cdots,Z^p,\left\{s_1^1,\ldots,s_{k_1}^1\right\},\cdots,\left\{s_1^q,\ldots,s_{k_q}^q\right\}\right]$$

Maximum likelihood estimates of β can be obtained via iteratively weighted least squares (IWLS) algorithm. For a more complete description (see McCullagh, 1989).

3. Generalized functional spectral additive regression models (GFSAM)

Regression models are those techniques for modeling and analyzing the relationship between a dependent variable and one or more independent variables. When one of the variables have a functional nature, we have functional regression models. The previous section was devoted to on the functional predictor is assumed to be linear. In GAM framework the response is a smooth function f_j^i of the functional scores $\tilde{X}$ of the predictor process (see Müller, 2005), (iii) functional expansion basis or principal componentes:

$$\tilde{X} = \left[f_1(Z^1),\cdots,f_p(Z^p),\left\{f_1^1(s_1^1),\ldots,f_{k_1}^1(s_{k_1}^1)\right\},\cdots,\left\{f_1^q(s_1^q),\ldots,f_{k_q}^q(s_{k_q}^q)\right\}\right]$$

4. Example of Functional Binomial Regression Model

In this section we focus on the particular case in which the response is binary, this model is also called functional logistic regression (FLR), (see Escabias, 2005). The functional logistic regression model the probability, π_i , of the occurrence of an event, $Y_i = 1$, rather than the event $Y_i = 0$, conditional on a vector of functional covariate $X_i(t)$ is expressed as:

$$y_i = \pi_i + \epsilon_i\ , i = 1,\ldots,n$$

where π_i is the expectation of Y given $X_i(t)$ that will be modeled as:

$$\pi_i = P[Y = 1|x_i(t) : t \in T] \quad , i = 1, \ldots, n$$

For logistic the canonical link is: **logit**: (a) $g(\pi) = log(\pi/(1-\pi))$. Other link function are also used: (b) **probit**: $g(\pi) = \phi^{-1}(\mu_i)$, where ϕ is the normal cumulative distribution function. (c) the complementary log-log, **cloglog**: $g(\pi) = log(log(1-\pi))$ and (d) the **cauchit**, $g(u) = tan(\pi(u - 1/2))$.

Below, we show how to apply the FLR model with binary response (dichotomized fat content, 1 for **fat**> 15, 0 otherwise) in the Tecator dataset. The following code we uses a training sample (first 129 curves) of the second derivative of absorbance curves $X.d2$.

```
ind<-1:129
Fat.bin<-ifelse(tecator$y$Fat<15,0,1)
X.d2<-fdata.deriv(tecator$absorp,nderiv=2)
dataf=data.frame(tecator$y[ind,],Fat.bin[ind])
ldata=list("df"=dataf,"X.d2"=X.d2[ind])
basis.x=list("X.d2"=create.pc.basis(absorp[ind],1))
f1<-Fat.bin ~ X.d2
```

For illustration, the fitted object returned (**res.glm2**) can be used in other functions of the *"glm"* class such as: **summary()**.

```
R> summary(res.glm)
Call: glm(formula = pf)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0955  -0.7910  -0.1840   0.7318   2.2264
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.03986    0.22156   -0.18    0.857
X.d2.PC1    88.65403   16.00282    5.54 3.03e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 178.76  on 128  degrees of freedom
Residual deviance: 123.79  on 127  degrees of freedom
AIC: 127.79
Number of Fisher Scoring iterations: 5
```

And the call for additive model:

```
R> f2<-Fat.bin ~ s(X.d2)
R> res.gsam=fregre.gsam(f2,ldata,family=binomial,basis.x=basis.x)
R> summary(res.gsam)
```

```
Family: binomial
Link function: logit
Formula: "Fat.bin~+s(X.d2.PC1,k=-1)"
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1292     0.3143  -0.411    0.681
Approximate significance of smooth terms:
              edf Ref.df Chi.sq  p-value
s(X.d2.PC1) 4.444  5.501  35.17 2.39e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
R-sq.(adj) =  0.398   Deviance explained = 35.2%
UBRE score = -0.017579  Scale est. = 1         n = 129
```

If new data is observed, the response can be predicted by:

```
R> pred.glm <- predict.fregre.glm(res.glm, list("absorp"=absorp[-ind]))
R> pred.gsam<-predict.fregre.gsam(res.gsam,list("absorp"=absorp[-ind]))
```

Logistic regression method can be used for a binary classification variable. For GFLM, the prediction of dichotomized fat content is correct in 80.6% of cases in training sample (first 129 data) and 80.2% in test sample (last 86 data). In order to test the results we have repeated 200 times changing the data in the sample (length 129) and summarized into the table 2 the percentage of good classification of the binary response. We repeat the estimation process by different link and basis functions and predict new response values. We uses the 200 fitted models (training sample of length 129) and new curves (test sample of length 86) to predict the binary response, see table 2. This example does not appreciate differences between the 4 link functions used (table 2), perhaps the cloglog link function have lower percentages of good classification. GSAM model is equal or slightly better than GFLM model when using the basis of the first principal component and is worse when using a bspline basis.

		Mean		Median	
Basis	Link/Model	GFLM	GFSAM	GFLM	GFSAM
1st PC	**logit**	79	79	78	79
1st PC	**probit**	78	79	78	79
1st PC	**cloglog**	76	79	76	79
1st PC	**cauchit**	80	80	80	80
5 bspline	**logit**	91	90	91	91
5 bspline	**probit**	91	90	91	91
5 bspline	**cloglog**	90	90	91	91
5 bspline	**cauchit**	91	90	91	91

Table 2: Percentage of good classification for test sample.

5. CONCLUSION

In this work we have implemented two advanced models for functional regression. The generalized functional linear functional model (GFLM) and generalized functional additive spectral model (GFSAM) have been incorporated into the **fda.usc** package in R. The implementation was done by mimicking the format of similar functions in multivariate environment for the better usability of the user.

ACKNOWLEDMENTS

This work was supported by grants MTM2008-03010 from the Ministerio de Ciencia e Innovación, 10MDS207015PR from the Xunta de Galicia and GI-1914 MODESTYA-Modelización estadística y aplicaciones.

BIBLIOGRAPHY

Cardot H, Ferraty F, Sarda P (1999). Functional Linear Model. Statist. Probab. Lett., 45(1), 1122.
Cardot H, Ferraty F, Sarda P (2003). Spline Estimators for the Functional Linear Model. Statistica Sinica,, 13, 571591.
Crainiceanu CM, Goldsmith AJ (2010). Bayesian functional data analysis using win-bugs Journal of Statistical Soft, 32(11).
Escabias M, Aguilera AM, Valderrama, MJ(2005). Modeling environmental data by functional principal component logistic regression. Environmetrics, 16(1), 95107.
McCullagh P and Nelder JA (1989) Generalized Linear Models. Second ed. London: Chapman and Hall.
Müller HG and StadtMüller U. (2005). Generalized functional linear models. Ann. Statist., 33, 774805.
Mller, H.G., Yao, F. (2008). Functional additive models. Journal of the American Statistical Association 103, 426-437
Ramsay JO. and Silverman BW (2005). Functional Data Analysis. Springer Series in Statistics, 2nd edition. Springer-Velag, New York.