An Approach for PLS Regression Modeling of Functional Data

[1]Shengshuai Wang [2]Jie Wang [1]Huiwen Wang [3]Gilbert Saporta

[1] Beijing University of Aeronautics and Astronautics, Beijing 100191, China
E-mail: wss2002buaa@sina.com
[2] Dagong Global Credit Rating Co., Ltd., Beijing, China
E-mail: wangjie@dagongcredit.com
[3] Cedric,Conservatoire National des Arts et Métiers, Paris, France
E-mail: gilbert.saporta@cnam.fr

Abstract: Partial Least Squares (PLS) approach is employed for linear regression modeling when both the dependent variables and independent variables are functional data (curves). After the introduction of the constant-style mean, variance and the correlative coefficient of functional data, an approach for PLS regression modeling of functional data is proposed to overcome the multicollinearity existing in the independent variables set. An empirical study of the functional regression modeling shows that the proposed approach provides a tool for building regression model on functional data under the condition of multicollinearity. The empirical study conclusion, which is coincident with the wildly accepted economic theory, indicates that the Compensation of Employees is the most important variable that contributes to the Total Retail Sales of Consumer Goods in China, while the Government Revenue and Income of Enterprises are less important.

Keywords: PLS regression, Functional data, Multicollinearity

1 Introduction

Functional data was firstly proposed by Jim. O. Ramsay in 1991, the feature which distinguishes functional data from other data is that each cell of the data sheet is a function. Functional linear model has received many concerns recently. West, Harrison developed a dynamic generalized functional linear model in 1989. Furthermore, Hastie and Tibshirani (1993) proposed varying-coefficient model, in which the variables and the regression coefficients are functions with arguments time t. Müller and Stadtmüller (2005) also investigated generalized functional linear model. Li, Aragon, Shedden et al. (2003) offered an approach that combines the methods of concurrent functional linear model, principal components analysis and the varying-coefficient model. Escabias, Aguilera et al. (2004), James (2002) and Cardot and Sarda (2004) adapted the generalized linear model to the presence of a functional predictor variable. In the purpose of dimensional reduction, Escabias (2004) introduced principal components analysis to functional linear model. Preda and Saporta (2007) introduced PLS to the classification of functional data. Huiwen Wang and Jie Wang (2008) proposed a constant coefficient linear regression modeling method for functional data.

The multicollinearity is a common problem in the multiple linear regression analysis. As a result, the model will be inaccurate and unstable when we build model based on Ordinary Least Squares (OLS) principle under the condition of multicollinearity. Variable Screening, Ridge Regression and Principal Component Regression are

common methods adopted to overcome the multicollinearity. Partial Least Squares Regression proposed by S.Wold and C.Albano (1983) is a novel and effective method to solve this problem.

As functional data is transformed from ordinary discrete data, so multicollinearity also exists in functional data. Huiwen Wang and Jie Wang (2008) proposed the conceptions and algorithms, which deal with the problem of multicollinearity in the multiple linear regression of functional data with Gram-Schmidt orthogonal transformation. In this paper, PLS is introduced to the multiple linear regression analysis of functional data to solve the multicollinearity existing in functional data.

2 Multicollinearity in Functional Data

To investigate the multicollinearity in functional data, correlation coefficient between two functional variables should be defined. The important concepts, including the inner product of functional data, constant-style mean, variance, covariance, correlation coefficient of functional data, are presented as follows:

In a functional data space, for functions $x(t) \in L^2[a,b]$, $y(t) \in L^2[a,b]$, the inner product of $x(t)$, $y(t)$ is defined as:

$$\langle x(t), y(t) \rangle = \int_a^b x(t) \cdot y(t) dt ; \tag{1}$$

For the sake of simplicity and without confusion, formula (1) is simply noted as:

$$\langle x(t), y(t) \rangle = \int x(t) \cdot y(t) dt ;$$

In addition, for $\forall t \in [a,b]$,there is $\mathbf{I}(t) \equiv 1$,and $\mathbf{I}(t)$ is called unit function on the interval $[a,b]$.

Based on the inner product defined above, for functional variables $x_j = \left(x_{1j}(t), \cdots, x_{nj}(t)\right)^T$, $x_k = \left(x_{1k}(t), \cdots, x_{nk}(t)\right)^T$ with n observations $x_{ij}(t), x_{ik}(t) \in L^2[a,b]$, $i = 1,2,\cdots,n$,the following related definitions can be deduced.

(1) $\overline{x}_j \in \mathbf{R}$,as the constant mean of x_j,can be defined as:

$$\int \overline{x}_j \cdot \mathbf{I}(t) dt = \frac{1}{n} \sum_{i=1}^{n} \int x_{ij}(t) dt ; \tag{2}$$

(2) $s_j^2 \in \mathbf{R}$, as the constant variance of x_j,can be defined as:

$$\int s_j^2 \cdot \mathbf{I}(t) dt = \frac{1}{n} \sum_{i=1}^{n} \int [x_{ij}(t) - \overline{x}_j \cdot \mathbf{I}(t)]^2 dt ; \tag{3}$$

(3) $s_{jk} \in \mathbf{R}$, as the constant covariance of x_j, x_k,can be defined as:

$$\int s_{jk} \cdot \mathbf{I}(t) dt = \frac{1}{n} \sum_{i=1}^{n} \int [x_{ij}(t) - \overline{x}_j \cdot \mathbf{I}(t)][x_{ik}(t) - \overline{x}_k \cdot \mathbf{I}(t)] dt ; \tag{4}$$

Combining formula (2), (3) and (4), the correlation coefficient of x_j, x_k can be defined as:

$$r\left(x_j, x_k\right) = \frac{s_{jk}}{s_j \cdot s_k} ; \tag{5}$$

If the correlation between functional variables x_j and x_k is high, we can find that $x_{ij}(t)$ and $x_{ik}(t)$ almost change in the same pattern through n observations. In the constant coefficient linear regression model of functional data, if there are high correlation coefficients in some variables of the independent variables set $x_1, x_2, \cdots, x_p$,we infer that multicollinearity exists in the independent variables. The same as ordinary discrete data modelling, when OLS method is adopted under this condition, the precision of regression coefficients will be degraded and the stability of model will be deteriorated.

3 PLS Regression Modeling of Functional Data

In functional data sheet, each data cell is a function. Suppose that there are q functional dependent variables $Y=\left[y_1,y_2,\cdots,y_q\right]$, and p functional independent variables $X=\left[x_1,x_2,\cdots,x_p\right]$. The sample size is n, $y_k=(y_{1k}(t),y_{2k}(t),\cdots,y_{nk}(t))^T$, $k=1,2,\cdots,q$; $x_j=\left(x_{1j}(t),x_{2j}(t),\cdots,x_{nj}(t)\right)^T$, $j=1,2,\cdots,p$; and $y_{ik}(t)$, $x_{ij}(t)\in L^2[a,b]$, $i=1,2,\cdots,n$.

The functional data above can be standardized, according to formula (2) and (3). Take independent variables for example, the formula for standardizing functional variables is as following:

$$\tilde{x}_{ij}(t)=\frac{x_{ij}(t)-\overline{x}_j\cdot\mathbf{I}(t)}{s_j},\quad i=1,2,\cdots,n,\ j=1,2,\cdots,p \tag{6}$$

Note that, the following formulas can be inferred according to formula (6) :

$$\overline{\tilde{x}}_j=\frac{1}{n(b-a)}\sum_{i=1}^{n}\int\tilde{x}_{ij}(t)dt=0;$$

$$\tilde{s}_j^2=\frac{1}{n(b-a)}\sum_{i=1}^{n}\int[\tilde{x}_{ij}(t)-0\cdot\mathbf{I}(t)]^2dt=1;$$

Similarly, dependent variables can also be standardized.

The standardized independent variables and dependent variables can be noted as:

$$E_0=\left[\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_p\right],\quad F_0=\left[\tilde{y}_1,\tilde{y}_2,\cdots,\tilde{y}_q\right];$$

According to formula (1), the inner product of two functional variables can be defined as:

$$\left\langle\tilde{x}_j,\tilde{y}_k\right\rangle=\sum_{i=1}^{n}\int\tilde{x}_{ij}(t)\cdot\tilde{y}_{ik}(t)dt; \tag{7}$$

Hence, the computational method of $E_0^TF_0$ can be given as:

$$E_0^TF_0=\begin{bmatrix}\sum_{i=1}^{n}\int\tilde{x}_{i1}(t)\cdot\tilde{y}_{i1}(t)dt & \sum_{i=1}^{n}\int\tilde{x}_{i1}(t)\cdot\tilde{y}_{i2}(t)dt & \cdots & \sum_{i=1}^{n}\int\tilde{x}_{i1}(t)\cdot\tilde{y}_{iq}(t)dt\\ \sum_{i=1}^{n}\int\tilde{x}_{i2}(t)\cdot\tilde{y}_{i1}(t)dt & \sum_{i=1}^{n}\int\tilde{x}_{i2}(t)\cdot\tilde{y}_{i2}(t)dt & \cdots & \sum_{i=1}^{n}\int\tilde{x}_{i2}(t)\cdot\tilde{y}_{iq}(t)dt\\ \vdots & \vdots & & \vdots\\ \sum_{i=1}^{n}\int\tilde{x}_{ip}(t)\cdot\tilde{y}_{i1}(t)dt & \sum_{i=1}^{n}\int\tilde{x}_{ip}(t)\cdot\tilde{y}_{i2}(t)dt & \cdots & \sum_{i=1}^{n}\int\tilde{x}_{ip}(t)\cdot\tilde{y}_{iq}(t)dt\end{bmatrix}_{p\times q} \tag{8}$$

Obviously, $E_0^TF_0\in\mathbf{R}^{p\times q}$ is a $p\times q$ scalar matrix.

Then we can derive components from dependent and independent variable sets respectively.

In the first step, calculate the matrix $E_0^TF_0F_0^TE_0=\left(E_0^TF_0\right)\cdot\left(E_0^TF_0\right)^T$ according to formula (8). The eigenvector corresponding to the maximum eigenvalue of the matrix $E_0^TF_0F_0^TE_0{}^T$ is $w_1=\left(w_{11},w_{12},\cdots,w_{1p}\right)^T\in\mathbf{R}^{p\times1}$. Thus, we derive the first PLS component $z_1=\left(z_{11}(t),\cdots,z_{n1}(t)\right)^T$ as:

$$z_1=E_0w_1=w_{11}\tilde{x}_1+w_{12}\tilde{x}_2+\cdots+w_{1p}\tilde{x}_p$$

Regress E_0 and F_0 with z_1 respectively by constant coefficient linear regression models of functional data, we can obtain:

$$E_0=z_1p_1^T+E_1$$

$$F_0=z_1r_1^T+F_1$$

Here, $p_1 \in \mathbf{R}^{p\times 1}$, $r_1 \in \mathbf{R}^{q\times 1}$ are regression coefficient vectors, that is,

$$p_1 = \frac{E_0^T z_1}{z_1^T z_1}, \quad r_1 = \frac{F_0^T z_1}{z_1^T z_1}$$

And that:

$$E_0^T z_1 = \left(\sum_{i=1}^{n} \int \tilde{x}_{i1}(t) \cdot z_{i1}(t) dt, \cdots, \sum_{i=1}^{n} \int \tilde{x}_{ip}(t) \cdot z_{i1}(t) dt \right)^T$$

$$F_0^T z_1 = \left(\sum_{i=1}^{n} \int \tilde{y}_{i1}(t) \cdot z_{i1}(t) dt, \cdots, \sum_{i=1}^{n} \int \tilde{y}_{iq}(t) \cdot z_{i1}(t) dt \right)^T$$

$$z_1^T z_1 = \sum_{i=1}^{n} \int z_{i1}(t) \cdot z_{i1}(t) dt$$

In the second step, substitute E_0, F_0 with E_1, F_1, and then repeat step one. When the number of PLS components l is finally determined by the method of Cross Validation, the iteration terminates.

After deriving components $z_1, z_2, \cdots, z_l$, the regression model will be:

$$\hat{F}_0 = z_1 r_1^T + z_2 r_2^T + \cdots + z_l r_l^T \tag{9}$$

Formula (9) can be transformed into the regression model with the original dependent variables y_k, $k = 1, 2, \cdots, q$ and original independent variables $x_1, x_2, \cdots, x_p$.

In the practical calculation, Discrete Fourier Transform (DFT) can be applied to formula (1) and formula (7).

4 Empirical Study

In the related theories of national income, the distribution pattern of national income refers to how the national income distributes among government, enterprises and residents. This paper attempts to establish a model based on functional data, and analyze the impacts of the three styles of income on Chinese consumer demand, and also validate the effectiveness of the proposed approach.

This paper adopted Total Retail Sales of Consumer Goods, Compensation of Employees, Government Revenue, and Income of Enterprises of 31 provinces and areas of China, during 1993 to 2007, and the price factor has been deducted, as shown in Figure 1.

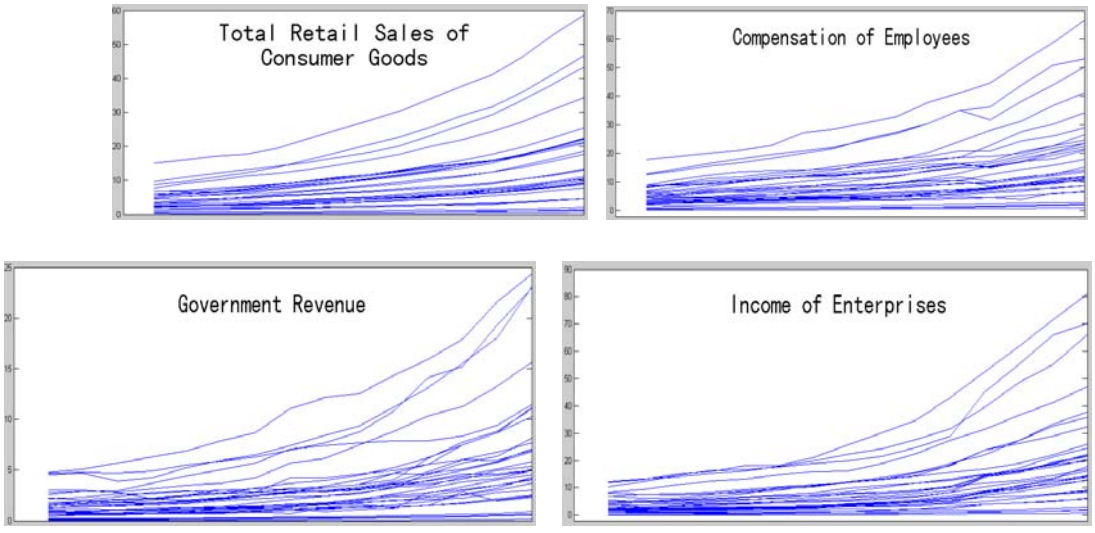

Figure 1 Curves of Total Retail Sales of Consumer Goods and distribution of national income of 31 provinces and areas of China (1993-2007)

Total Retail Sales of Consumer Goods (Con) is adopted as dependent variable, Compensation of Employees

(Emp), Government Revenue (Gov) and Income of Enterprises (Ent) are independent variables.

Firstly, we employ DFT to obtain functional data from the original discrete data.

The correlation coefficient matrix is obtained by formula (5), as shown in Table 1.

Table 1 Correlation coefficient matrix of independent variables

	Emp	Gov	Ent
Emp	1	0.9207	0.9665
Gov		1	0.9507
Ent			1

Table 1 shows that the correlation coefficients of independent variables are as high as 0.9 or above, thus OLS method is unsuitable.

The regression model obtained by the approach proposed in this paper is:

$$\mathrm{Con}(t) = 0.0121 \cdot \mathbf{I}(t) + 0.4688 \cdot \mathrm{Emp}(t) + 0.1183 \cdot \mathrm{Gov}(t) + 0.4007 \cdot \mathrm{Emp}(t) \tag{10}$$

Two components are extracted by cross validation, and the extracted components are used in PLS regression model. Figure 2 is the scatter plot of the observed values and the predicted values. Figure 2 indicates that the points of the observed values and the predicted values are distributed on a diagonal line, so the degree of fitting is good.

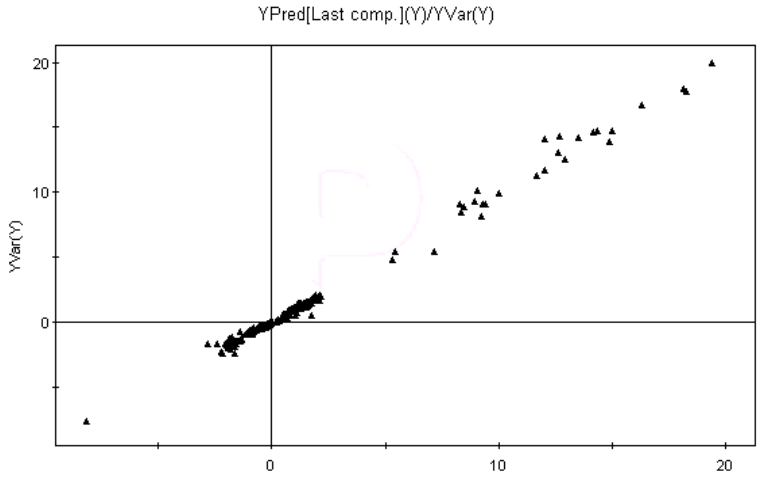

Figure 2 Scatter plot for the observed values and the predicted values

Formula (10) shows that, in the initial distribution of national income of China, Compensation of Employees contributes the most to Total Retail Sales of Consumer Goods, the Income of Enterprises contributes less, and the Government Revenue contributes the least. According to classical consumption theory, marginal consumption rate of labour income is higher than marginal consumption rate of capital revenue. Regression model (10) indicates that the increase of resident income is the most important factor affecting the growth of consumer demand.

In conclusion, in the initial distribution of national income, Compensation of Employees, Government Revenue and Income of Enterprises contribute to the growth of consumer demand. However, Compensation of Employees plays the most important role. According to this conclution, the structure of national income distribution should be adjusted, so as to promote the consumption of residents, and realize the harmonious growth of national economy.

5 Conclusions

This paper discussed the conception of multicollinearity existing in functional data, defined correlation between

two functional variables, and proposed an approach for Functional PLS regression modeling. An empirical study shows that the proposed method is effective. In the empirical study, the regression model of the initial distribution of national income and consumer demand indicates that the Compensation of Employees is the most important factor that contributes to the consumer demand, and the Government Revenue and Income of Enterprises contribute less. This conclusion is coincident with classical consumption theory.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC), under grant numbers: 70771004, 70531010 and 70521001.

Reference

[1] Bastien P., Esposito Vinzi E., Tenenhaus M. (2005) PLS generalised linear regression, *Computational Statistics and Data Analysis*, 48, 17-46.

[2] Cardot H., Ferraty F., Sarda P. (1999) Functional linear model, *Statistics & Probability Letters*, 45, 11-22.

[3] 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 (3-4), 365-384.

[4] Fan Y.F. (1994) *Analysis of Procedure and Distribution of National Income*, China Renmin University Press, Beijing.

[5] Hastie T., Tibshirani R. (1993) Varying-coefficient models, *Journal of the Royal Statistical Society*, 55, 757-796.

[6] James G.M. (2002) Generalized linear models with functional predictors, *Journal of the Royal Statistical Society*, 64, 411-432.

[7] Li K.C., Aragon Y., Shedden K., Thomas Agnan C. (2003) Dimension reduction for multivariate response data, *Journal of the American Statistical Association*, 98, 99-109.

[8] Müller H.-G., Stadtmüller U. (2005) Generalized functional linear models, *The Annals of Statistics*, 33 (2), 774-805.

[9] Preda C., Saporta G., Leveder C. (2007) PLS classification of functional data, *Computational Statistics*, 22, 223-235.

[10] Ramsy J.O., Dalzell C.J. (1991) Some tools for functional data analysis (with discussion), *Journal of the Royal Statistical Society*, 53, 539-572.

[11] Ramsay J.O., Silverman B.W. (2005) *Functional Data Analysis-Second Edition*, Springer Science+Business Media, New York.

[12] Wang H.W., Chen M.L., Saporta G. (2008) Gram-Schmidt regression and application in cutting tool abrasion prediction, *Journal of Beijing University of Aeronautics and Astronautics*, 34 (6), 729-733.

[13] Wang H.W., Wu Z.B., Meng J. (2006) *Partial Least-Squares Regression-Linear and Nonlinear Methods*, National Defense Industry Press, Beijing.

[14] West M., Harrison P.J. (1989) *Bayesian Forecasting and Dynamic Models*, Springer, New York.

[15] Wold S., Martens H., Wold H. (1983) The multivariate calibration problem in chemistry solved by the PLS method. Ruhe A, Kågström B (Eds), *Proc. Conf. Matrix Pencils, Lectures Notes in Mathematics*, Springer-Verlag, Heidelberg.

[16] Yang T.Y. (2001) *Income Distribution and Effective Demand*, Economic Science Press, Beijing.