Advances in Applied Mathematics 应用数学进展, 2023, 12(6), 2781-2787
Published Online June 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.126279

Hans汉斯

带有交互项的广义部分函数型线性模型的应用

毛可敬，李颂萱，肖维维*

北方工业大学理学院，北京

收稿日期：2023年5月16日；录用日期：2023年6月9日；发布日期：2023年6月16日

摘　要

随着科技的发展，数据信息逐渐呈现出多元化的特点，传统数据分析已经不能再满足人们的需求，因此越来越多的学者开始关注函数型数据分析。目前，函数型数据分析被应用到医学、气象学、环境学、经济学等各个领域。本文针对预测变量是函数型和标量型的混合变量，且考虑函数型预测变量之间的交互作用的情况，提出了一个带有交互项的广义部分函数型线性模型，利用主成分分析法对函数型预测变量进行降维处理，再运用加权最小二乘法对未知参数迭代求解，最后将此模型应用于粮食产量的研究中。研究结果表明：除特定时期外，降水量和气温在一定程度上均会促进粮食产量的增加，农作物总播种面积、农用机械总动力、化肥使用量对粮食产量的增加同样具有促进作用。

关键词

函数型数据分析，交互项，主成分分析，粮食产量

Application of Generalized Partially Function Type Linear Models with Interaction Terms

Kejing Mao, Songxuan Li, Weiwei Xiao*

School of Science, North China University of Technology, Beijing

Received: May 16th, 2023; accepted: Jun. 9th, 2023; published: Jun. 16th, 2023

Abstract

With the development of technology, data information is gradually presenting more and more diversified characteristics, traditional data analysis can no longer meet people's needs, so more and more scholars are beginning to focus on functional data analysis. At present, functional data anal-

*通讯作者。

文章引用: 毛可敬, 李颂萱, 肖维维. 带有交互项的广义部分函数型线性模型的应用[J]. 应用数学进展, 2023, 12(6): 2781-2787. DOI: 10.12677/aam.2023.126279

ysis has applications in a wide range of fields such as medicine, meteorology, environmental science and economics. In this paper, a generalized partially functional linear model with interaction terms is proposed for the case where the predictor variables are a mixture of functional and scalar variables, and the interaction between the functional predictor variables is considered. The functional predictor variables are reduced in dimensionality using principal component analysis, and then the weighted least squares method is applied to iteratively solve for the unknown parameters. The results of the study show that, except for certain periods, precipitation and temperature contribute to the increase in grain yield to a certain extent, and that the total area sown, total power of agricultural machinery and fertiliser use also contribute to the increase in grain yield.

Keywords

Functional Data Analysis, Interaction Items, Principal Component Analysis, Grain Production

1. 引言

当今世界，人们收录的数据形式越来越多样化，传统数据分析已经不能满足人们的需求，所以我们运用函数型数据分析来研究这些数据间的关系。回归分析[1]、主成分分析[2]、方差分析[3]等统计方法已经很好地融入到了函数型数据分析中，但是仍然需要做进一步的探究。

Ramsay [4]明确指出了什么是函数型数据，并将经典数据分析技术扩展到函数型数据的问题中。Ramsay 和 Dalzell [5]对函数型数据进行分析与讨论，提出了函数型数据分析的方法，Ranmsay 和 Silverman [6]利用已有的函数型数据分析研究成果，总结了函数型线性模型的基本形式和估计算法，并举例说明了函数型数据的处理方法。Nelder & Wedderburn [7]首次引入广义线性模型，该模型通过一个连接函数建立了连续型响应变量和预测变量之间的相关关系，并分析了不同的离散型响应变量和预测变量之间的关系。James [1]提出了一种将广义线性模型扩展到预测变量是曲线或函数的方法，Müller 和 Stadtmüller [8]针对响应变量是标量，预测变量是随机函数的回归情况，提出了广义函数型线性回归模型。在进行实际数据分析时，预测变量之间往往会存在一定的关系，因此，许多研究者对带有交互项的回归模型展开了研究。Yang 等人[9]考虑深度谱图和温度时间序列之间的函数交互作用，提高了对密苏里河下游鲟鱼产卵率的预测，Usset [10]提出了一种能适应双向交互作用的具有标量响应变量和函数型预测变量的函数回归模型，Luo 和 Qi [11]、Fuchs [12]、H. Matsui [13]、Yifan Sun [14]等人在带有交互项的回归模型的研究上也有所成就。

粮食作为人们生活的必需品，是影响着人们的生活、生存与发展的重要问题。近年来由于极端天气的肆虐，粮食面临减产的危机，为了应对即将到来的粮食危机，采用科学有效的方法提高粮食产量迫在眉睫。因此本文提出了连接函数已知、预测变量是函数和向量的混合变量，且函数型预测变量之间具有交互效应的广义部分函数型线性模型，并应用此模型研究粮食产量的部分影响因素。

2. 模型及其估计

2.1. 模型简介

我们假设有响应变量 $Y_i, i=1,\cdots,n$，函数型预测变量 $X_{ij}\left(t_j\right), j=1,2$ 和标量型预测变量 $Z=\left(Z_1,Z_2,\cdots,Z_q\right)$，其中 $X_{ij}\left(t_j\right)\in L^2\left(T_j\right)$，连接函数 $g(\cdot)$ 已知，且二阶导连续。定义响应变量 Y_i 和预测变量 $X_{ij}\left(t_j\right)$、Z 有以

下关系：

$$Y_i = g\left(\alpha + \int_{T_1} X_{i1}(t_1)\beta_1(t_1)\mathrm{d}t_1 + \int_{T_2} X_{i2}(t_2)\beta_2(t_2)\mathrm{d}t_2 + \iint_{T_1T_2} X_{i1}(t_1)X_{i2}(t_2)\beta(t_1,t_2)\mathrm{d}t_1\mathrm{d}t_2 + Z^{\mathrm{T}}\gamma\right) + \varepsilon_i \quad (1)$$

其中 $\alpha \in \mathrm{R}$ 是截距，$\beta_1(t_1),\beta_2(t_2)$ 和 $\beta(t_1,t_2)$ 是两个函数型预测变量和交互项对应的回归函数，$\gamma = (\gamma_1,\gamma_2,\cdots,\gamma_q)^{\mathrm{T}}$ 是标量型预测变量 Z 对应的未知回归系数，且 ε_i 满足 $E(\varepsilon_i)=0$，$Var(\varepsilon_i)=\sigma^2$。

由于函数型数据具有无穷维的特点，因此我们首先运用主成分分析法对函数型数据进行降维，$X_{ij}(t_j)$ 使用 K-L 展开为

$$X_{ij}(t_j) = \sum_{k=1}^{\infty} \xi_{ijk}\phi_{jk}(t_j)$$

其中，ξ_{ijk} 为函数型主成分得分，$\phi_{jk}(t_j)$ 为函数型主成分基，且 $\int_{T_j}\phi_{jk}^2(t_j)\mathrm{d}t_j = 1$。

同理，回归系数函数 $\beta_j(t_j),\beta(t_1,t_2)$ 使用 K-L 展开分别为：

$$\beta_j(t_j) = \sum_{k=1}^{\infty}\beta_{jk}\phi_{jk}(t_j)$$

$$\beta(t_1,t_2) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} u_{kl}\phi_{1k}(t_1)\phi_{2l}(t_2)$$

将上述展开式带入模型(1)，预测变量在 p_j 处截断，且 p_j 随着 $n\to\infty$ 渐近增加，得到截断模型(2)

$$Y_i = g\left(\alpha + \sum_{k=1}^{p_1}\xi_{i1k}\beta_{1k} + \sum_{l=1}^{p_2}\xi_{i2l}\beta_{2l} + \sum_{k=1}^{K}\sum_{l=1}^{L}\rho_{ikl}u_{kl} + Z^{\mathrm{T}}\gamma\right) + \varepsilon_i,\ i = 1,2,\cdots,n \quad (2)$$

其中，$K = p_1, L = p_2$。

2.2. 参数估计

定义参数向量

$$\vartheta = \left(\beta_{11},\cdots,\beta_{1p_1},\beta_{21},\cdots,\beta_{2p_2},u_{11},\cdots,u_{1L},u_{21},\cdots,u_{2L},\cdots,u_{K1},\cdots,u_{KL},\gamma_0,\gamma_1,\cdots,\gamma_q\right)^{\mathrm{T}}$$

以及

$$\eta_i = \alpha + \sum_{k=1}^{p_1}\xi_{i1k}\beta_{1k} + \sum_{l=1}^{p_2}\xi_{i2l}\beta_{2l} + \sum_{k=1}^{K}\sum_{l=1}^{L}\rho_{ikl}u_{kl} + Z^{\mathrm{T}}\gamma$$

$$\mu_i = g(\eta_i)$$

$$\delta_i = \left(\xi_{i11},\cdots,\xi_{i1p_1},\xi_{i21},\cdots,\xi_{i2p_2},\rho_{i11},\cdots,\rho_{i1L},\rho_{i21},\cdots,\rho_{i2L},\cdots,\rho_{iK1},\cdots,\rho_{iKL},z_0,z_1,\cdots,z_q\right)^{\mathrm{T}}$$

利用最大似然估计，有

$$U(\vartheta) = \sum_{i=1}^{n}\frac{(Y_i - g(\eta_i))g'(\eta_i)}{\sigma^2(\mu_i)}\delta_i = 0 \quad (3)$$

引入矩阵

$$V = diag\left(\sigma^2(\mu_1),\cdots,\sigma^2(\mu_n)\right)$$

$$H = diag\left(g'(\eta_1),g'(\eta_2),\cdots,g'(\eta_n)\right)$$

$$A_0 = A_{n,q+1} = (z_{im})_{1\le i\le n,0\le m\le q}$$

$$A_j = A_{n,p_j} = \left(\xi_{ijr}\right)_{1\le i\le n,0\le r\le p_j,1\le j\le 2}$$

$$A_{12} = A_{n,KL} = \left(\rho_{ikl}\right)_{1\le i\le n,1\le k\le K,1\le l\le L}$$

$$A = A_{n,q+1+p_1+p_2+KL} = diag\left(A_1, A_2, A_{12}, A_0\right)$$

引入向量 $Y = \left(Y_1,\cdots,Y_n\right)^{\mathrm{T}}$， $\mu = \left(\mu_1,\cdots,\mu_n\right)^{\mathrm{T}}$， b_j,γ,u 则方程(3)可被写成

$$A^{\mathrm{T}}V^{-1}H\left(Y-\mu\right) = 0 \tag{4}$$

通过加权最小二乘法对(4)进行迭代求解即得 β_j,γ,u 的估计值

$$\tilde{\beta}_j = \left(A_j^{\mathrm{T}}IA_j\right)^{-1}A_j^{\mathrm{T}}Ig^{-1}\left(Y\right)$$

$$\tilde{\gamma} = \left(A_0^{\mathrm{T}}IA_0\right)^{-1}A_0^{\mathrm{T}}Ig^{-1}\left(Y\right)$$

$$\tilde{u} = \left(A_{12}^{\mathrm{T}}IA_{12}\right)^{-1}A_{12}^{\mathrm{T}}Ig^{-1}\left(Y\right)$$

其中， $I = V^{-1}H^2$。

3. 实例研究

本文使用的数据来源于中国环境监测总站和各地区统计公报中收集的 2020 年 1 月 1 日至 2020 年 12 月 31 日的北京、成都、包头、新乡等 58 个城市的降水量、气温、农作物总播种面积、农用机械总动力、化肥使用量和粮食产量等数据。

我们的目的是利用所提的模型研究降水量、气温、农作物总播种面积、农用机械总动力和化肥使用量对粮食产量的影响。其中，以 2020 年各城市的降水量和气温作为 2 个函数型预测变量，分别记为 X_1 和 X_2；以农作物总播种面积、农用机械总动力和化肥使用量作为3个标量型预测变量，分别记为 Z_1,Z_2 和 Z_3；以 2020 年各城市的粮食产量作为响应变量，记为 Y。我们首先对各城市的粮食产量数据进行预处理，规定当粮食产量大于 200 万吨时，该城市的粮食产量较高，用 1 来表示，反之，粮食产量较低，用 0 来表示。图 1 展示了部分城市 2020 年的降水量和气温的情况。

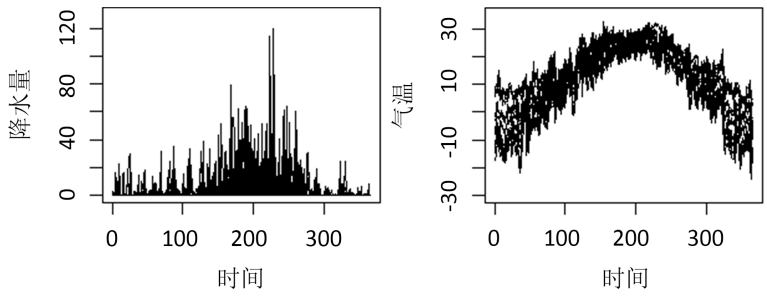

Figure 1. Precipitation and temperatures in selected cities
图 1. 部分城市的降水量和气温

将预处理的数据代入到模型中，回归系数 $\hat{\gamma}$ 的结果如表 1 所示，可以看出农作物总播种面积、农用机械总动力和化肥使用量与粮食产量之间呈正相关关系，其中农作物总播种面积对粮食产量的影响最为显著。即在一定条件下，农作物总播种面积、农用机械总动力和化肥使用量越多，粮食产量越多，沈一鸣[15]、祝正芳[16]等人的研究与我们得出的结果是一致的。农作物总播种面积是粮食生产的载体和基础，播种面积越大，资源投入越大，粮食产量越多。农业机械化的发展能够推动农业的生产发展，提高农业

生产效率[17]，因此农用机械总动力越高，粮食产量越多。而农用化肥中含有许多农作物生长所需的营养物质和微量元素，因此科学的增加化肥施用量能够促进农作物生长，从而增加粮食产量。

Table 1. Estimates of regression coefficients and their significance levels
表 1. 回归系数的估计值及其显著性水平

$\hat{\gamma}$	Estimate	Std. Error	t value	Pr(>|t|)
$\hat{\gamma}_{面积}$	0.27834	0.02800	9.939	9.67e−14***
$\hat{\gamma}_{动力}$	0.08564	0.04822	1.776	0.08146 .
$\hat{\gamma}_{化肥}$	2.68839	0.81361	3.304	0.00171**

回归系数函数 $\hat{\beta}_1(t_1)$ 和 $\hat{\beta}_2(t_2)$ 的结果见图 2，我们可以看出降水量和粮食产量具有明显的正相关性，即降水量越大，粮食产量越多。但是在雨季，降水量过大可能会影响农作物的呼吸作用，使农作物根系受到伤害，土壤养分流失，增加农作物病害的风险，从而造成粮食产量降低。对于气温和粮食产量的关系，我们可以从图中看出，除夏季外，气温与粮食产量总体上呈正相关关系。在一定范围内，温度升高能够加快农作物的生长，增加粮食产量，但是夏季温度过高则会对农作物的传粉、受精、灌浆等过程造成不良影响，导致农作物籽粒减产，粮食产量降低，这一结论在郭军伟[18]、樊廷海[19]等人的研究中得到了印证。图 3 和图 4 显示了 $\hat{\beta}(t_1,t_2)$ 在置信区间内的变化特征，当 $t_2\in[-30,-10]$ 时，$\hat{\beta}(t_1,t_2)$ 随着 t_1 的增大而减小，$t_2\in[-10,20]$ 时，$\hat{\beta}(t_1,t_2)$ 随着 t_1 的增大而增大，$t_2\in[20,30]$ 时，$\hat{\beta}(t_1,t_2)$ 随着 t_1 的增大而减小，说明降水量和气温具有一定的相互作用，共同影响着粮食产量。

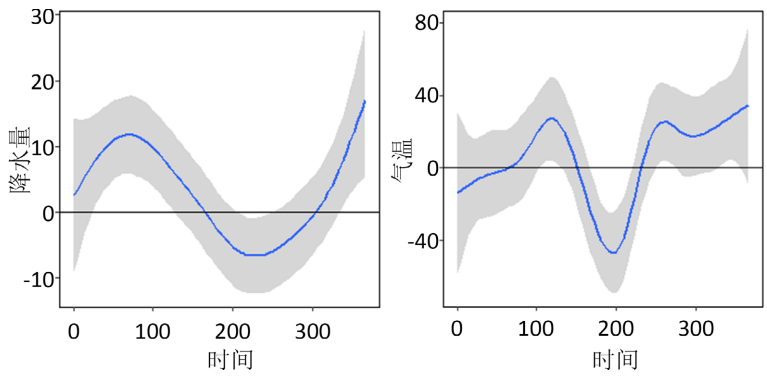

Figure 2. $\hat{\beta}(t)$ and its 95% confidence interval band
图 2. $\hat{\beta}(t)$ 及其 95%的置信区间带

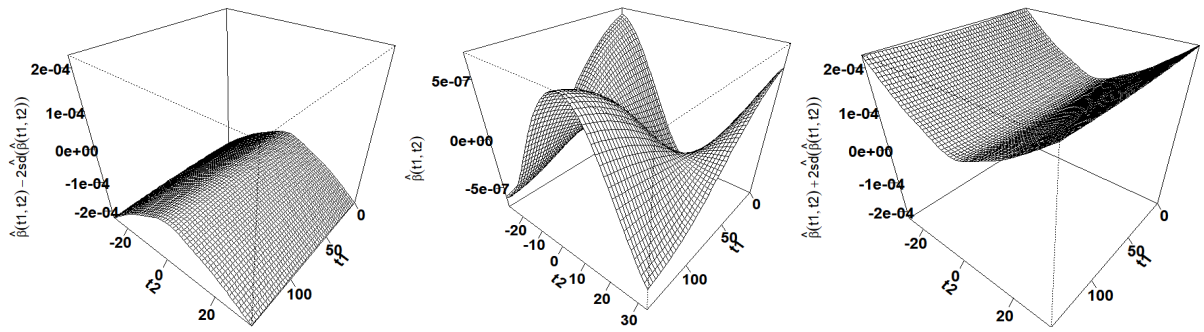

Figure 3. Visualisation of $\hat{\beta}(t_1,t_2)$ in three dimensions
图 3. $\hat{\beta}(t_1,t_2)$ 的可视化三维图

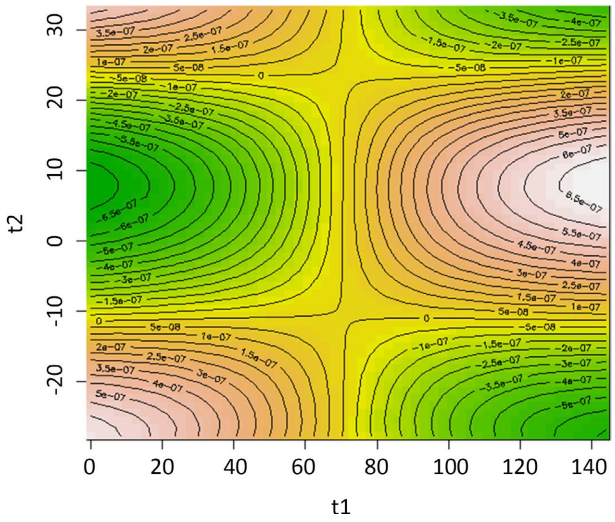

Figure 4. Contour map of $\hat{\beta}(t_1,t_2)$

图 4. $\hat{\beta}(t_1,t_2)$ 的等高线图

4. 结论

本文介绍了带有交互项的广义部分函数型线性模型，既考虑了函数型预测变量和标量型预测变量的影响，又考虑了函数型预测变量之间的交互作用，并运用模型来研究降水量、气温、农作物总播种面积、农用机械总动力、化肥使用量对粮食产量的影响。将数据代入到模型中，我们得出结论：降水量和气温对粮食产量的影响总体上呈正相关性，但是在雨季，降水量与粮食产量呈负相关性；在夏季，气温与粮食产量也呈现出负相关性。而农作物总播种面积、农用机械总动力、化肥使用量对粮食产量的影响均是正相关的。因此为了增加粮食产量，抵御粮食危机，我们应该保护耕地，积极处理现有的环境问题，防止土地荒漠化，优化农业生产模式，加大对农业科技的投入，用科学的方法合理种植农作物。

基金项目

北方工业大学毓杰人才项目，No. 107051360023XN075-04。

参考文献

[1] James, G.M. (2002) Generalized Linear Models with Functional Predictors. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **64**, 411-432. https://doi.org/10.1111/1467-9868.00342

[2] Locantore, N. and Marron, J.S. (1998) Robust Principal Component Analysis for Functional Data. *Test*, **8**, 1-73. https://doi.org/10.1007/BF02595862

[3] Brumback, B.A. and Rice, J.A. (1998) Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves. *Journal of the American Statistical Association*, **93**, 991-994. https://doi.org/10.2307/2669843

[4] Ramsay, J.O. (1982) When the Data Are Functions. *Psychometrika*, **47**, 379-396. https://doi.org/10.1007/BF02293704

[5] Ramsay, J.O. and Dalzell, C.J. (1991) Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society*, **53**, 539-572. https://doi.org/10.1111/j.2517-6161.1991.tb01844.x

[6] Ramsay, J.O. and Silverman, B.W. (1997) Principal Components Analysis for Functional Data. In: Ramsay, J.O. and Silverman, B.W., Eds., *Functional Data Analysis*, Springer, New York, 285-290. https://doi.org/10.1007/978-1-4757-7107-7_6

[7] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*: *Series A* (*General*), **135**, 370-384. https://doi.org/10.2307/2344614

[8] Müller, H.G. and Stadtmüller, U. (2005) Generalized Functional Linear Models. *The Annals of Statistics*, **33**, 774-805.

https://doi.org/10.1214/009053604000001156

[9] Yang, W.H., Wikle, C.K., Holan, S.H. and Wildhaber, M.L. (2013) Ecological Prediction with Nonlinear Multivariate Time-Frequency Functional Data Models. *Journal of Agricultural, Biological, and Environmental Statistics*, **18**, 450-474. https://doi.org/10.1007/s13253-013-0142-1

[10] Usset, J., Staicub, A.M. and Maity, A. (2016) Interaction Models for Functional Regression. *Computational Statistics and Data Analysis*, **94**, 317-329. https://doi.org/10.1016/j.csda.2015.08.020

[11] Luo, R.Y. and Qi, X. (2019) Interaction Model and Model Selection for Function-on-Function Regression. *Journal of Computational and Graphical Statistics*, **28**, 309-322. https://doi.org/10.1080/10618600.2018.1514310

[12] Fuchs, K., Scheipl, F. and Greven, S. (2015) Penalized Scalar-on-Functions Regression with Interaction Term. *Computational Statistics and Data Analysis*, **81**, 38-51. https://doi.org/10.1016/j.csda.2014.07.001

[13] Matsui, H. (2020) Quadratic Regression for Functional Response Models. *Econometrics and Statistics*, **13**, 125-136. https://doi.org/10.1016/j.ecosta.2018.12.003

[14] Sun, Y.F. and Wang, Q.H. (2020) Function-on-Function Quadratic Regression Models. *Computational Statistics and Data Analysis*, **142**, Article ID: 106814. https://doi.org/10.1016/j.csda.2019.106814

[15] 沈一鸣. 粮食产量预测模型研究与应用[D]: [硕士学位论文]. 武汉: 武汉轻工大学, 2022.

[16] 祝正芳. 我国粮食产量与播种面积、施肥量、降水量关系实证研究[J]. 中国市场, 2013(16): 85-86, 94.

[17] 方方. 京津冀地区农业生产效率的时空格局及收敛性研究[J]. 世界地理研究, 2019, 28(5): 130-140.

[18] 郭军伟, 吴志岐, 祁国梅. 温度升高对水稻生长及品质的影响[J]. 农业科技与信息, 2022(6): 22-25.

[19] 樊廷海. 温度对玉米生长发育及产量的影响[J]. 种子科技, 2022, 40(14): 17-19.