https://doi.org/10.53465/ER.2644-7185.2024.1.22-31

FUNCTIONAL LINEAR REGRESSION: A CASE STUDY FROM THE FOOD INDUSTRY[1]

PETER KNÍŽAT[2]

Abstract: *A problem of analyzing functions in statistics is a very recent phenomenon that has been extensively investigated. The functionality in observed data, that is, when data are generated through a process naturally described as functional, occurs in many areas of sciences. For example, time series in financial engineering, imagining records in medicine, or spectrometric wavelengths in chemometrics can be considered as a discrete approximation of a continuous set of mathematical objects. In the initial step, observed real-valued data are transposed into functional data by smoothing or interpolation technique. These functional objects can then be directly used in the framework of statistical regressions. The aim of this paper is to show an estimation of the functional linear regression, where the observed covariate is functional, and its corresponding response is scalar. In the empirical study, we utilize spectrometric data with an objective to predict the fat content in a meat sample given the spectrum of absorbances, which are recorded through the infrared analyzer, by using the estimated functional linear regression. The functional regression coefficient is estimated through the basis spline expansion, for which we evaluate a different number of basis splines and propose the best predictor estimator.*

Keywords: *functional data, basis splines, functional linear regression, spectrometric data.*

JEL Classification: C13, C49, C63

[1] The paper was presented on conference EDAMBA 2023.

[2] Peter Knížat, University of Economics in Bratislava, Slovak Republic, e-mail: peter.knizat@euba.sk, https://orcid.org/0000-0001-5100-1319.

1 Introduction

Many natural phenomena allow to record real-valued data over a very fine grid that results in the collection of very large data sets of observations. The real-valued data samples in such high-dimensional datasets can be transformed into a collection of mathematical objects that are observable on an infinite continuum.

The first step in the functional data analysis is to choose a data representation through a smoothing or an interpolation process. Assuming that a matrix $\mathbf{X_{ij}}$, where each sample $\mathbf{i, i = 1, \dots, n,}$ is observed for variables $\mathbf{j, j = 1, ..., p}$, as a set of real-valued vectors, the first task is to convert these vectors into functions $\boldsymbol{\chi_i(t)}$ for each $\mathbf{i}$ that are computable for any argument value $\mathbf{t}$. If discrete values are assumed to be errorless, then it is an interpolation process, but if they have some observational error that needs removing, then the conversion from discrete data to functions may involve smoothing. The common technique to impose functions $\boldsymbol{\chi_i(t)}$ on observed data is a spline basis approximation. A comprehensive study of data smoothing using splines is provided in De Boor (2001).

Once data is transposed into a functional form, we can proceed to explore the variability in functional objects. The pioneering research using functional data in the framework of the linear regression is carried out by Cardot, Ferraty and Sarda (1999), where the authors propose an adjusted computational strategy for the least squares minimization.

In the paper Knížat (2022), the author shows an application of the functional analysis of variance (ANOVA). In the general form of functional regression, the response variable is of a functional form and the matrix of covariates is coded zero and one that corresponds to each treatment category of observed curves. In its empirical analysis, the spectrometric data set is adapted to an experimental design that is proposed by the author.

In the paper Ferraty and Vieu (2002), the authors propose a functional non-parametric regression to predict the fat content in meat samples using the spectrometric data set. They conclude that the proposed model provides satisfactory prediction results.

The state-of-the-art literature on functional data analysis, which also include the practical application in the R software, is provided in Ramsay and Silverman (2005) and Ramsay, Hooker and Graves (2009).

The main objective of this paper is to show an estimation of the functional linear regression, where the observed covariate is functional, and its corresponding response is scalar. In the empirical study, we utilize spectrometric data with an objective to predict the fat content in a meat sample given the spectrum of absorbances, which are recorded through the infrared analyzer, by using the estimated functional linear regression. The functional regression coefficient is estimated through the basis spline expansion, for which we evaluate a different number of basis splines and propose the best predictor estimator based on the mean squared error.

The paper is organized as follows. Section 2 outlines a theoretical framework for fitting basis splines into observed data, to transform real data into functional data, and for ordinary least squares when using functions in the estimator. Section 3 shows the application of functional linear regression on the spectrometric data set. It evaluates its prediction accuracy when using a different number of basis splines in the expansion of the functional regression parameter. Conclusion summarizes the results and outlines further research possibilities.

2 Theoretical Framework

The notion of the classic form of the linear model can be extended to the functional context. The linear model can be functional in the following way: either response or covariate, or both, can be of the functional form.
This chapter shows the theoretical framework for estimating a functional linear regression where the response is a real-valued vector given the functional covariate. The computational methodology defines basis functions within the least squares criteria, with a way of thinking very similar to the classic regression approach. The main difference is that the regression coefficients now become the regression coefficient function, **β(t)**, observable on a continuous domain **t.**

2. 1 B-spline expansion

Let us assume that we observe a matrix of real numbers $\mathbf{X}_{\mathbf{ij}}$ as specified in Section 1. It follows that by assuming that an observed data set comes from a functional process, the need dictates transforming the observed real data into functional objects. A function is fitted into a matrix of real numbers $\mathbf{X}_{\mathbf{ij}}$ across each sample observation **i**, and, therefore, a **p**-dimensional space is mapped into an infinite, or functional, space. A notation for such a functional covariate is $\chi_i(t)$, where **i = 1, …, n** refers to the number of sample curves.

The basis spline, also called B-spline, expansion can be used to individually construct a continuous and sufficiently differentiable functions $\chi_i(t)\colon \mathbf{R} \rightarrow \mathbf{R}$. The functions $\chi(t) = \chi_i(t)$ can be expressed in terms of B-spline expansion as De Boor (2001):

$$\chi(t) = \sum_{k=1}^{K} C_k B_k(t, \tau_l). \tag{1}$$

The spline curves, $B_k\ (t,\tau_l\ ),$ are piecewise polynomials of order m that are automatically tied together at the breakpoint, or knot, sequence τ_l, **l = 1, …, L – 1,** where **k = 1, …, K** refers to the number of basis splines. The $\mathbf{C}_k$ is a matrix of parameters to be estimated for each observed sample curve **i**, with a number of parameters **K.**

The code for working with B-splines is available in a wide range of programming languages, including R, S-PLUS and MATLAB®. The interested reader should consult De Boor (2001) for more theoretical details related to basis spline expansions.

The application of the B-spline interpolation to real data can be done through the familiar technique of fitting statistical models to data by minimizing the sum of squared errors, or least squares estimation, which leads to an estimation of parameters $\mathbf{C}_k$. The number of basis splines is defined by the user that drives a degree of smoothing. In our case, we use a sufficiently large number of basis splines such that the B-spline provides an interpolation of the original data that guarantees no loss of information.

It follows that after defining a least square criterion, setting it equal to zero, and solving its derivative, the estimated $\hat{\mathbf{C}}_k$ are defined as:

$$\hat{C}_k = (B_k{}'B_k)^{-1}B_k{}'X \tag{2}$$

where the superscript ' denotes a transpose. The functional covariate $\chi(t)$ in Eq. (1) can be re-expressed as:

$$\chi(t) = B_k{}'\hat{C}_k. \tag{3}$$

The functions $\chi(t)$ are observable on a continuous domain **t**, and can be discretized at any values on **t.**

2.2 Functional linear regression

Further statistical analysis normally comprises of identifying a structure and covariability in the response variable, **y = $\mathbf{y_i}$,** given the functional covariate, **χ(t) = $\boldsymbol{\chi_i}$(t).** They can be placed within the following functional form:

$$\boldsymbol{y} = f\big(\boldsymbol{\chi}(\boldsymbol{t})\big) + \boldsymbol{\varepsilon} \tag{4}$$

where ε is a real-valued vector of residuals **i = 1, …, n** and the unknown regression function **f(χ(t))** can be of parametric or nonparametric form.

A functional parametric model can be described by the estimated functional parameter(s) that determine the model structure. In the functional nonparametric model, the estimated structure of the model is determined by the specified semi-metric measure that is based on the kernel density function.

In the further analysis, we consider the functional parametric regression that takes the following form Ramsay and Silverman (2005):

$$\boldsymbol{y} = \beta_0 + \int \boldsymbol{\chi}(\boldsymbol{t})\beta(\boldsymbol{t})dt + \boldsymbol{\varepsilon}. \tag{5}$$

Eq. (5) is a functional extension of the linear regression where the usual summation of **Xβ** is replaced by the integration over a continuous index **t,** where **β(t)** is a regression coefficient function observable on a continuous domain **t**. The real-valued vector ε are random errors, which are assumed to be independent and identically distributed **ε~N(0,$\boldsymbol{\sigma^2}$)**, and $\boldsymbol{\beta_0}$ is a point-wise intercept of the regression function.

To express Eq. (5) in terms of basis splines, the regression coefficient function **β(t)** can be decomposed as:

$$\beta(\boldsymbol{t}) = \sum_{k=1}^{K_\beta} b_k\theta_k(\boldsymbol{t}) \tag{6}$$

where $\boldsymbol{\theta_k}$ ***(t)*** is a vector of basis splines of length $\mathbf{K_\beta}$, with a corresponding vector of coefficients $\mathbf{b_k}$**.** The functional covariate **χ(t)** is expressed as in Eq. (1). Noting that for the simplification purpose, we omit the notation for knots $\boldsymbol{\tau_l}$**.**

It follows that Eq. (5) can be rewritten as Ramsay and Silverman (2005):

$$\boldsymbol{y} = \beta_0 + \int \boldsymbol{C}_k \boldsymbol{B}_k(\boldsymbol{t})\theta_k(\boldsymbol{t})' b_k dt + \boldsymbol{\varepsilon}. \tag{7}$$

In Eq. (7), a number of basis splines $\boldsymbol{\theta_k}$ ***(t)*** in the expansion of the functional regression parameter $\boldsymbol{\beta}$***(t)*** is defined by the user, with the corresponding unknown parameters b_k estimated using the functional least squares method.

Now, we define a matrix $\boldsymbol{J_{B\theta}}$ of length $\boldsymbol{K}\times\boldsymbol{K_\beta}$,

$\boldsymbol{J_{B\theta}}$***=∫***$\boldsymbol{B_k}$ ***(t)***$\boldsymbol{\theta_k}$***(t)'dt.*** To further simplify the notation, the vector **u** of length $\boldsymbol{(K_\beta + 1)}$ and the matrix $\boldsymbol{Z}$ of length $\boldsymbol{N\times(K_\beta+1)}$ are defined as:

$$\boldsymbol{u} = (\beta_0, b_1, \dots, b_K)' \qquad \boldsymbol{Z} = [1\boldsymbol{C}_k\boldsymbol{J}_{B\theta}] \tag{8}$$

where 1 denotes a vector of ones with length K_β used in the estimation of β_0. Eq. (7) can now be re-expressed as:

$$\boldsymbol{y} = \int \boldsymbol{Zu} + \boldsymbol{\varepsilon}. \tag{9}$$

It follows that the least squares criterion (LSC) can be written as:

$$LSC(\beta_0, b_1, \dots, b_K) = \int \boldsymbol{y} - \boldsymbol{Zu}. \tag{10}$$

The least squares estimate of the augmented parameters vector **u**, which minimizes Eq. (10), is the solution of the equation:

$$\hat{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}. \tag{11}$$

A detailed derivation of Eqs. (7) up to (11) is provided in Ramsay and Silverman (2005).

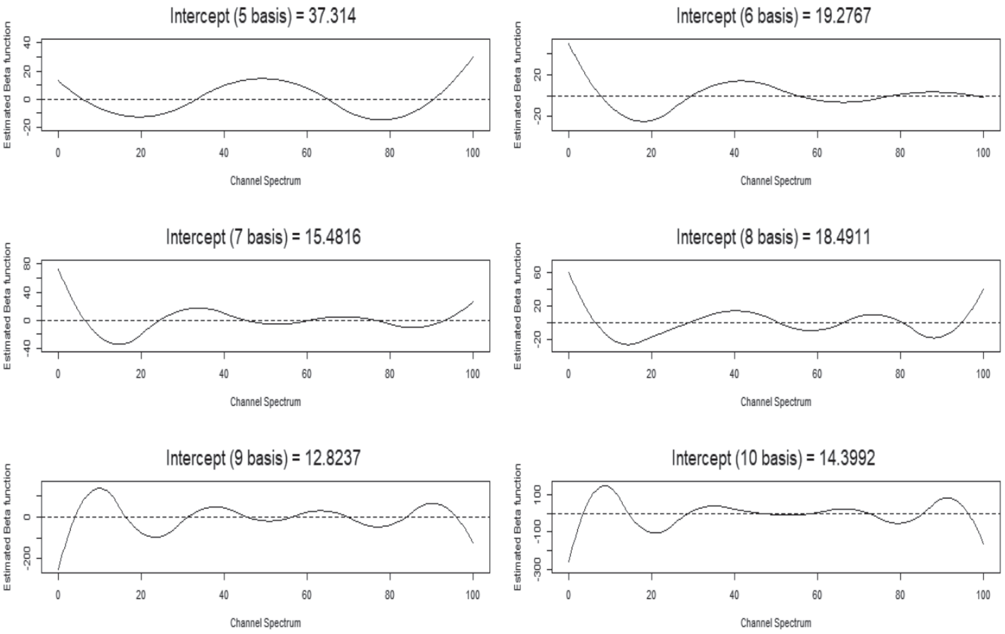
3 Results

In the empirical analysis, we utilize data that originate from the investigation of quality control in the food industry, which are referred to as a spectrometric data set. Detailed description of the spectrometric (tecator) data set can be found at http://lib.stat.cmu.edu/datasets/tecator.

The spectrometric data set was also used in the paper Knížat (2022), where the author shows the application of the functional analysis of variance. The transformation of observed data, which are measured as 100 channel spectra of absorbances by the infrared analyzer for each meat sample, into functional objects is the same as in this paper, i.e., we use 30 basis splines of order 3 to generate 215 sample curves $\boldsymbol{\chi_i(t)}$ as in Eq. (1). The fat content for each meat sample is a real-valued response $\boldsymbol{y_i}$.

Figure 1 shows the estimated functional regression parameter $\boldsymbol{\hat{\beta}(t)}$, including the point-wise intercept $\boldsymbol{\hat{\beta}_0}$, that is measured on the scale from 0 to 100 as the original observed absorbances on 100 channel spectra. For the expansion of $\boldsymbol{\hat{\beta}(t)}$ as in Eq. (6), we use a different number of basis splines to evaluate its impact.

Fig. 1: Estimated $\boldsymbol{\hat{\beta}(t)}$ coefficient functions, with 5, 6, 7, 8, 9 and 10 basis splines used in its expansion



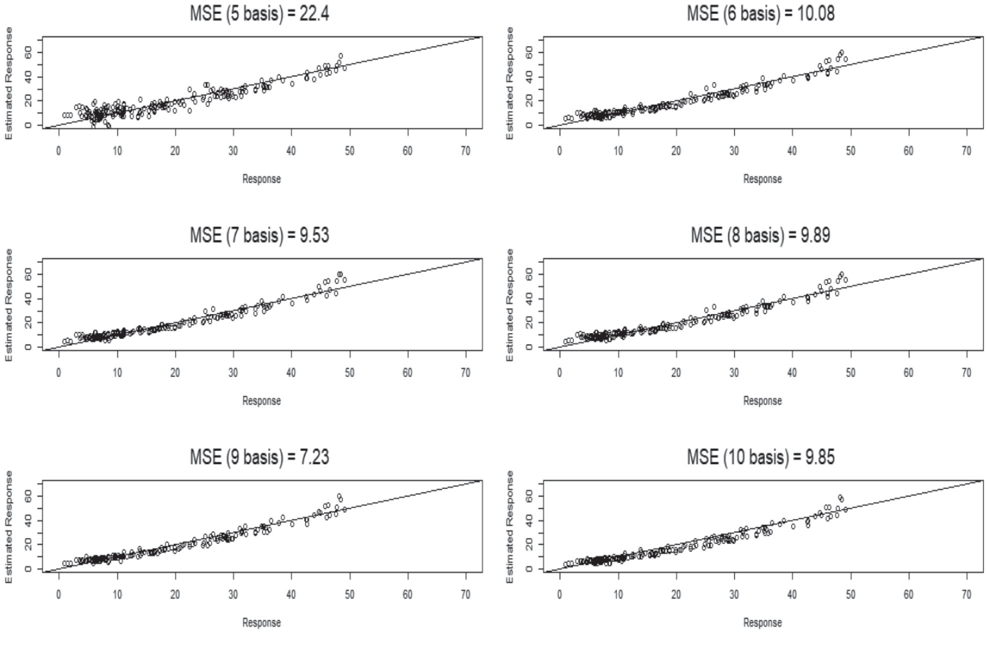
Source: author's calculations

Figure 1 shows that both shape and magnitude of the estimated $\boldsymbol{\hat{\beta}(t)}$ coefficient function change by varying the number of basis splines in its expansion.

However, it is not straightforward to interpret $\hat{\beta}(t)$. We can deduce that it places more emphasis, shown as departures from zero axis, on approximately 10th, 40th, and 85th channel spectra, which might indicate that these absorbances play a more pivotal role in the estimation of the fat content in meat samples. Noting that the emphasis changes when using a different number of basis splines. Moreover, the estimated intercept $\hat{\beta}_0$ increases by decreasing the number of basis splines.

Figure 2 shows the estimated versus observed fat content values, including the mean squared error[3] for each $\hat{\beta}(t)$ expansion.

Fig. 2: The estimated versus observed fat content value, with 5, 6, 7, 8, 9 and 10 basis splines used in the $\hat{\beta}(t)$ expansion



Source: author's calculations

Figure 2 shows that the best prediction fit is obtained when using 9 basis splines in the $\hat{\beta}(t)$ expansion, which is confirmed by its corresponding mean squared error.

[3] The mean squared error is calculated as a mean of squared differences between observed and predicted responses divided by the number of observations.

4 Conclusions

The main objective of this paper is the application of the theoretical framework of functional linear regression. In the first part, we present a computational method of the least squares criterion that is used when the covariate is functional. In the empirical analysis, we use a spectrometric data set to evaluate the prediction ability of the functional linear model. We show that the number of basis splines in the expansion of the functional regression coefficient plays a pivotal role in the fitted model’s prediction accuracy. The best prediction for the fat content in meat samples is achieved when nine basis splines are used in the $\hat{\beta}(t)$ expansion. Moreover, it is not straightforward to interpret the estimated functional regression coefficient as in its classical counterpart.

Further research is required to evaluate diagnostics of the fitted model and its assumptions. Moreover, a nonparametric functional regression could be used for the response prediction and its results compared to its parametric counterpart. A comparison can also be made with classic multivariate statistical regression models, similar to Ahn (2022).

Acknowledgements

This work was supported by the Grant Agency of the Slovak Republic VEGA 1/0211/21 "Econometric Analysis of Macroeconomic Impacts of Pandemics in the World with Emphasis on the Development of EU Economies and especially the Slovak Economy" and VEGA 1/0047/23 „The importance of spatial spillover effects in the context of the EU's greener and carbon-free Europe priority“.

REFERENCES

[1] Ahn, K. (2022). Comparative study between functional data analysis and multivariate data analysis for functional data. *Journal of the Korean Data And Information Science Society,* 33(5), 817 – 827. https://doi.org/10.7465/jkdi.2022.33.5.817

[2] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* 3rd edition. New York: Wiley.

[3] Cardot, H., Ferraty, F., & Sarda, P. (1999). Functional linear model. *Statistics &*

Probability Letters, 45(1), 11 ─ 22. https://doi.org/10.1016/S0167-7152(99)00036-X

[4] De Boor, C. (2001). *A Practical Guide to Splines.* Revised Edition. New York: Springer.

[5] Ferraty, F., & Romain, Y. (2011). *The Oxford Handbook of Functional Analysis.* Oxford University Press.

[6] Ferraty, F., & Vieu, P. (2002). The Functional Nonparametric Model and Application to Spectrometric Data. *Computational Statistics,* 17(4), 545 – 564. https://doi.org/10.1007/s001800200126

[7] Knížat, P. (2022). Functional Analysis of Variance: Case Study From Food Industry (in Slovak). In: *Mezinárodní vědecký seminář. Nové trendy v ekonometrii a operačním výzkumu: mezinárodní vědecký seminář,* 30. listopad - 2. prosinec, Praha: Vydavateľstvo EKONÓM, 71 ─ 77.

[8] Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis.* Second Edition. Springer Series in Statistics, Springer. https://doi.org/10.1007/b98888

[9] Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional Data Analysis with R and MATLAB.* Springer. https://doi.org/10.1007/978-0-387-98185-7

[10] Tecator. (2023). Tecator data set. Database. Available at: http://lib.stat.cmu.edu/datasets/tecator