Chapter 15
Goodness-of-fit Tests for Functional Linear Models Based on Integrated Projections

Eduardo García-Portugués, Javier Álvarez-Liébana,
Gonzalo Álvarez-Pérez and Wenceslao González-Manteiga

Abstract Functional linear models are one of the most fundamental tools to assess the relation between two random variables of a functional or scalar nature. This contribution proposes a goodness-of-fit test for the functional linear model with functional response that neatly adapts to functional/scalar responses/predictors. In particular, the new goodness-of-fit test extends a previous proposal for scalar response. The test statistic is based on a convenient regularized estimator, is easy to compute, and is calibrated through an efficient bootstrap resampling. A graphical diagnostic tool, useful to visualize the deviations from the model, is introduced and illustrated with a novel data application. The R package `goffda` implements the proposed methods and allows for the reproducibility of the data application.

15.1 Functional Linear Models

15.1.1 Formulation

Given two separable Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$, we consider the regression setting with centered $\mathbb{H}_2$-valued response $\mathcal{Y}$ and centered $\mathbb{H}_1$-valued predictor $\mathcal{X}$:

Eduardo García-Portugués (✉)
Department of Statistics and UC3M-Santander Big Data Institute, Carlos III University of Madrid, Avda. Universidad 30, 28911 Leganés, Spain, e-mail: edgarcia@est-econ.uc3m.es

Javier Álvarez-Liébana
Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo, Spain, e-mail: alvarezljavier@uniovi.es

Gonzalo Álvarez-Pérez
Department of Physics, University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo, Spain, e-mail: gonzaloalvarez@uniovi.es

Wenceslao González-Manteiga
Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Rúa Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, Spain, e-mail: wenceslao.gonzalez@usc.es

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*, Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_15

$$\mathcal{Y} = m(\mathcal{X}) + \mathcal{E}, \tag{15.1}$$

where $m : x \in \mathbb{H}_1 \mapsto \mathbb{E}[\mathcal{Y}|\mathcal{X} = x] \in \mathbb{H}_2$ is the regression operator and the $\mathbb{H}_2$-valued error $\mathcal{E}$ is such that $\mathbb{E}[\mathcal{E}|\mathcal{X}] = 0$. When $\mathbb{H}_1 = L^2([a,b])$ and $\mathbb{H}_2 = L^2([c,d])$, the Functional Linear Model with Functional Response (FLMFR; see, e.g., [15, Chapter 16]) is the most well-known parametric instance of (15.1). If the regression operator is assumed to be Hilbert–Schmidt, m is parametrizable as

$$m_\beta(\mathcal{X}) = \int_a^b \beta(s,\cdot)\mathcal{X}(s)\,\mathrm{d}s =: \langle\langle \beta, \mathcal{X} \rangle\rangle, \tag{15.2}$$

for $\beta \in \mathbb{H}_1 \otimes \mathbb{H}_2 = L^2([a,b] \times [c,d])$ a square-integrable kernel. The present work considers this framework and is concerned with the goodness-of-fit of the family of $\mathbb{H}_2$-valued and $\mathbb{H}_1$-conditioned linear models

$$\mathcal{L} := \{\langle\langle \beta, \cdot \rangle\rangle : \beta \in \mathbb{H}_1 \otimes \mathbb{H}_2\}. \tag{15.3}$$

Any $\mathcal{X} \in \mathbb{H}_1$ and $\mathcal{Y}, \mathcal{E} \in \mathbb{H}_2$ can be represented in terms of orthonormal bases $\{\Psi_j\}_{j=1}^\infty$ and $\{\Phi_k\}_{k=1}^\infty$ as $\mathcal{X} = \sum_{j=1}^\infty x_j \Psi_j$, $\mathcal{Y} = \sum_{k=1}^\infty y_k \Phi_k$, and $\mathcal{E} = \sum_{k=1}^\infty e_k \Phi_k$, where $x_j = \langle \mathcal{X}, \Psi_j \rangle_{\mathbb{H}_1}$, $y_k = \langle \mathcal{Y}, \Phi_k \rangle_{\mathbb{H}_2}$, and $e_k = \langle \mathcal{E}, \Phi_k \rangle_{\mathbb{H}_2}$, $\forall j,k \geq 1$. Also, $\beta \in \mathbb{H}_1 \otimes \mathbb{H}_2$ can be expressed as

$$\beta = \sum_{j=1}^\infty \sum_{k=1}^\infty b_{jk}(\Psi_j \otimes \Phi_k), \quad b_{jk} = \left\langle \beta, \Psi_j \otimes \Phi_k \right\rangle_{\mathbb{H}_1 \otimes \mathbb{H}_2}, \quad \forall j,k \geq 1.$$

Therefore, the population version of the FLMFR based on (15.2) can be expressed as

$$y_k = \sum_{j=1}^\infty b_{jk} x_j + e_k, \ k \geq 1. \tag{15.4}$$

15.1.2 Model Estimation

The projection of (15.4) into the truncated bases $\{\Psi_j\}_{j=1}^p$ and $\{\Phi_k\}_{k=1}^q$ opens the way for the estimation of β given a centered sample $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^n$. Indeed, the truncated sample version of (15.4) is expressed as

$$\mathbf{Y}_q = \mathbf{X}_p \mathbf{B}_{p,q} + \mathbf{E}_q, \tag{15.5}$$

where $\mathbf{Y}_q$ and $\mathbf{E}_q$ are $n \times q$ matrices with the respective coefficients of $\{\mathcal{Y}_i\}_{i=1}^n$ and $\{\mathcal{E}_i\}_{i=1}^n$ on $\{\Phi_k\}_{k=1}^q$, $\mathbf{X}_p$ is the $n \times p$ matrix of coefficients of $\{\mathcal{X}_i\}_{i=1}^n$ on $\{\Psi_j\}_{j=1}^p$, and $\mathbf{B}_{p,q}$ is the $p \times q$ matrix of coefficients of β on $\{\Psi_j \otimes \Phi_k\}_{j,k=1}^{p,q}$.

Several estimators for β have been proposed; see, e.g., [16, 13, 5, 1, 14]. A popular estimation paradigm is Functional Principal Components Regression (FPCR; [15]), which considers the (empirical) Functional Principal Components (FPC) $\{\hat{\Psi}_j\}_{j=1}^p$