Análisis de datos funcionales - Taller 4

Valeria Bejarano - Camilo Avellaneda

Contents

(1) Resuelva el ejercicio 4.2 del libro de Kokoszka and Reimherr [2017].
(2) Usando el dataset TECATOR:

- Formule un modelo de regresión entre la cantidad de grasa (variable de respuesta) y las curvas espectrométricas (variable regresora) que caracterizan las muestras de carne dadas.
- Encuentre y analice las estimaciones del parámetro Beta usando las tres propuestas presentadas en el libro de Kokoszka and Reimherr [2017].

Con el objetivo de complementar los resultados, se describen de manera breve las tres metodologías propuestas en Kokoszka and Reimherr [2017] para el caso donde se tienen variables funcionales como regresoras y escalares en la respuesta. Estas se presentan a continuación. De esta manera, la ecuación de regresión que relaciona una variable funcional y una variable de tipo escalar se muestra en la ecuación (1).

$$Y_i = \alpha + \int \beta(t)X(t)dt + \epsilon_i \tag{1}$$

- Estimación a partir de funciones base

En esta parte lo que se realiza es reescribir el término $\beta(t)$, como se muestra en la ecuación (2), donde B_k corresponde a una función base.

$$\beta(t) = \sum_{k=1}^{K} c_k B_k(t) \tag{2}$$

De esta forma, reemplazando la ecuación (2) en la ecuación (1), se tiene lo que se muestra en la ecuación (3), en donde se observa que a partir de reescribiendo todo en términos de funciones base, el problema se reduce a la estimación de los parámetros de un modelo de regresión múltiple de la manera usual.

$$\int \beta(t)X_i(t)dt = \sum_{k=1}^{K} c_k \int B_k(t)X_i(t)dt =: \sum_{k=1}^{K} x_{ik}c_k \tag{3}$$

- **Estimación a partir de penalizaciones por rugosidad:**

La idea principal de esta propuesta es considerar un parámetro de penalización por curvaturas. La función a optimizar en este planteamiento se muestra en la ecuación (4). La estimación de $\int \beta(t)X(t)dt$ se realiza de manera análoga a la metodología presentada en la metodología de estimación sin penalización, mientras que la estimación de λ se realiza mediante validación cruzada generalizada.

$$P_\lambda(\alpha, \beta) = \sum_{i=1}^{N} \left[Y_i - \alpha - \int \beta(t)X_i(t)df \right]^2 + \lambda \int [(L\beta)(t)]^2 dt \tag{4}$$

- **Regresión a partir de los componentes principales funcionales:**

donde $\hat{\xi}_{ij} = \int [X_i(t) - \hat{\mu}(t)]\, \hat{v}_j(t)$

$$X(t) \approx \mu(t) + \sum_{j=1}^{p} \hat{\xi}_{ij} \hat{v}_j(t) \tag{5}$$

$$\begin{aligned} Y_i &= \alpha + \int \beta(t) \left(mu(t) + \sum_{j=1}^{p} \hat{\xi}_{ij} \hat{v}_j(t) \right) dt + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^{p} \hat{\xi}_{ij} \beta_j + \epsilon_i. \end{aligned} \tag{6}$$

Se observa en la ecuación (6) que los parámetros a estimar son los ξ_{ij} a partir de los métodos de regresión tradicionales.

En Kokoszka and Reimherr [2017] se presenta una breve descripción de cómo implementar estas metodologías a partir de la librería *refund* [Goldsmith et al., 2020] y en este documento se realizan pasos similares, únicamente que se aplican a el dataset de tecator [Febrero-Bande and Oviedo de la Fuente, 2012]. De esta forma, la variable respuesta Y_i es la cantidad de grasa, mientras que las curvas espectrométricas son las funciones aleatorias explicativas en este caso.

La estimación del modelo mediante los componentes principales funcionales, modelo sin penalización y el que considera la penalización por rugosidad se presentan a continuación. En el segundo, el argumento $fx = TRUE$ significa que no se penaliza. En los casos 2 y 3 el número de funciones base se obtuvo mediante la función *k.check* del paquete *mgcv* [Wood, 2003].

```
fit.fpcr = pfr(Y ~ fpc(absorp,k=50))
fit.lin = pfr(Y ~ lf(absorp, bs = "ps", k = k_optimo, fx = TRUE))
fit.pfr = pfr(Y ~ lf(absorp, bs = "ps", k = k_optimo))
```

Para resumir el comportamiento de las estimaciones de $\beta(t)$, el Gráfico 1 muestra a $\hat{\beta}(t)$, de acuerdo a los diferentes métodos de estimación descritos. Por otro lado, Gráfico 2 muestra lo mismo en un rango más reducido de valores en el eje de las abcisas, ya que en la primer figura se observa como uno de los casos toma valores elevados, lo cual hace parecer que la regresión por componentes principales arroja una función casi constante como resultado, mientras que en la segunda figura se observa cómo la función oscila, solo que menos que los otros métodos de estimación, tanto el que penaliza como el que no lo hace. Una manera de realizar la interpretación de la estimación de los $\beta(t)$ es similar a diferentes ponderaciones asignadas a cada $X(t)$. Por ende, como la estimación realizada sin penalización obtiene valores mayores en valores absoluto en los extremos del intervalo, esto quiere decir que en ese caso se asigna una mayor importancia a los valores de $X(t)$ con t cercanos a 0 o a 100. Como en el caso de FPCA se observa una función que no oscila en la misma proporción que las otras, se podría pensar en que las ponderaciones son más homogéneas.

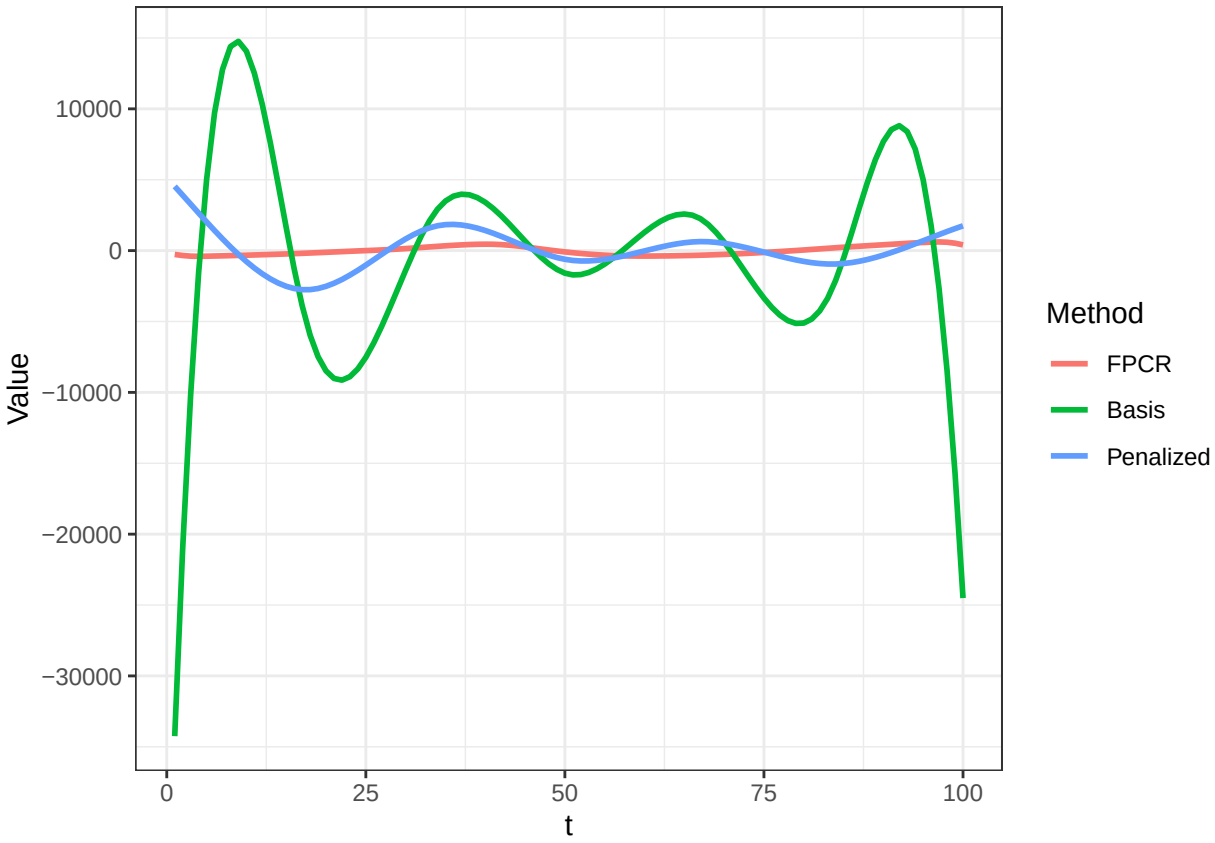

Gráfico 1: Funciones de regresión estimadas de acuerdo a las diferentes metodologías de estimación.

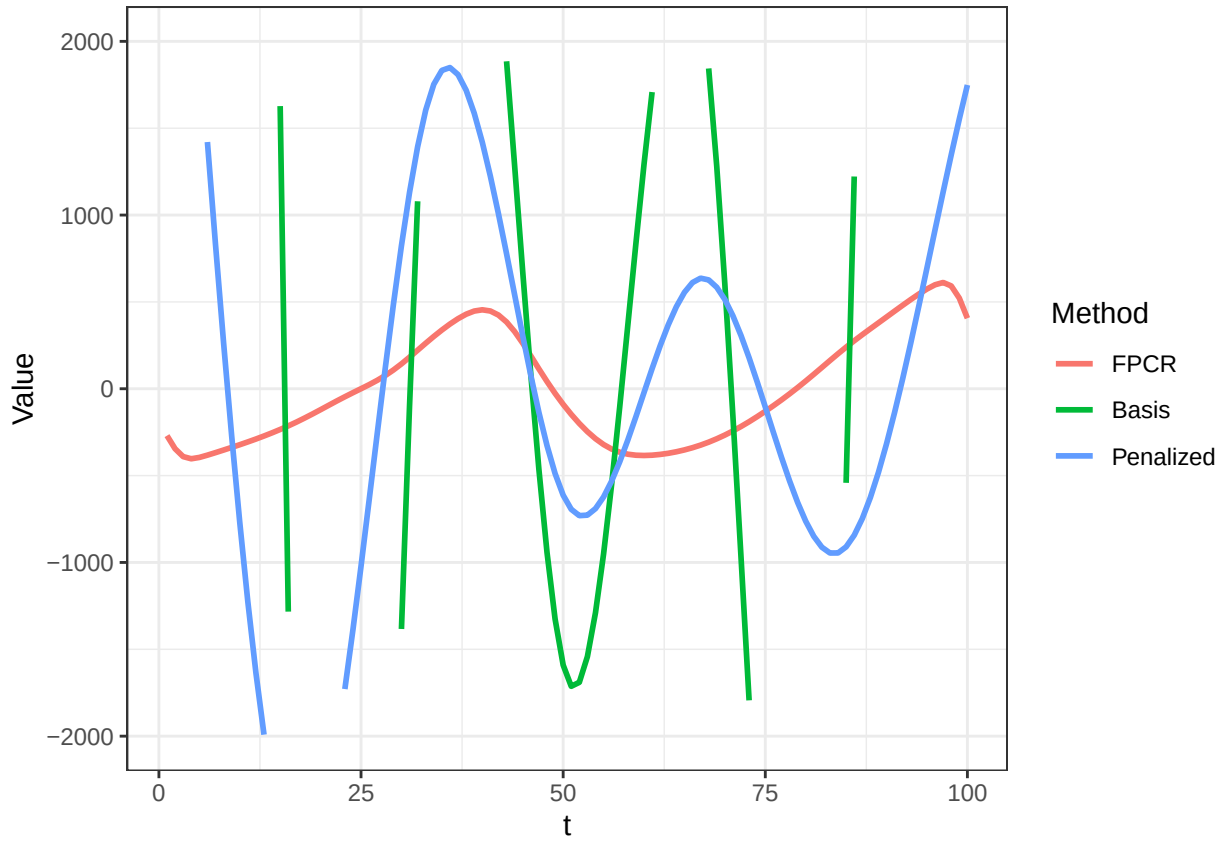

Gráfico 2: Funciones de regresión estimadas de acuerdo a las diferentes metodologías de estimación en un rango reducido de valores en el eje de las abcisas.

El Gráfico 3 muestra las estimaciones y los valores observados de Y_i de acuerdo a los diferentes métodos. Allí se observa que el caso que utiliza los FPCA arroja residuales mayores que los otros. Para corroborar lo que se establece anteriormente, se calculan los cuadrados medios del error (CME) para cada caso. La Tabla 1 muestra los CME, en donde se observa que cuando se utiliza los FPCA es el de peor rendimiento, mientras que el que menor CME tiene es el método que no penaliza por rugosidad.

Tipo	MSE
FPCR_predict	55.545627
Basis_predict	7.351828
Penalized_predict	8.965630

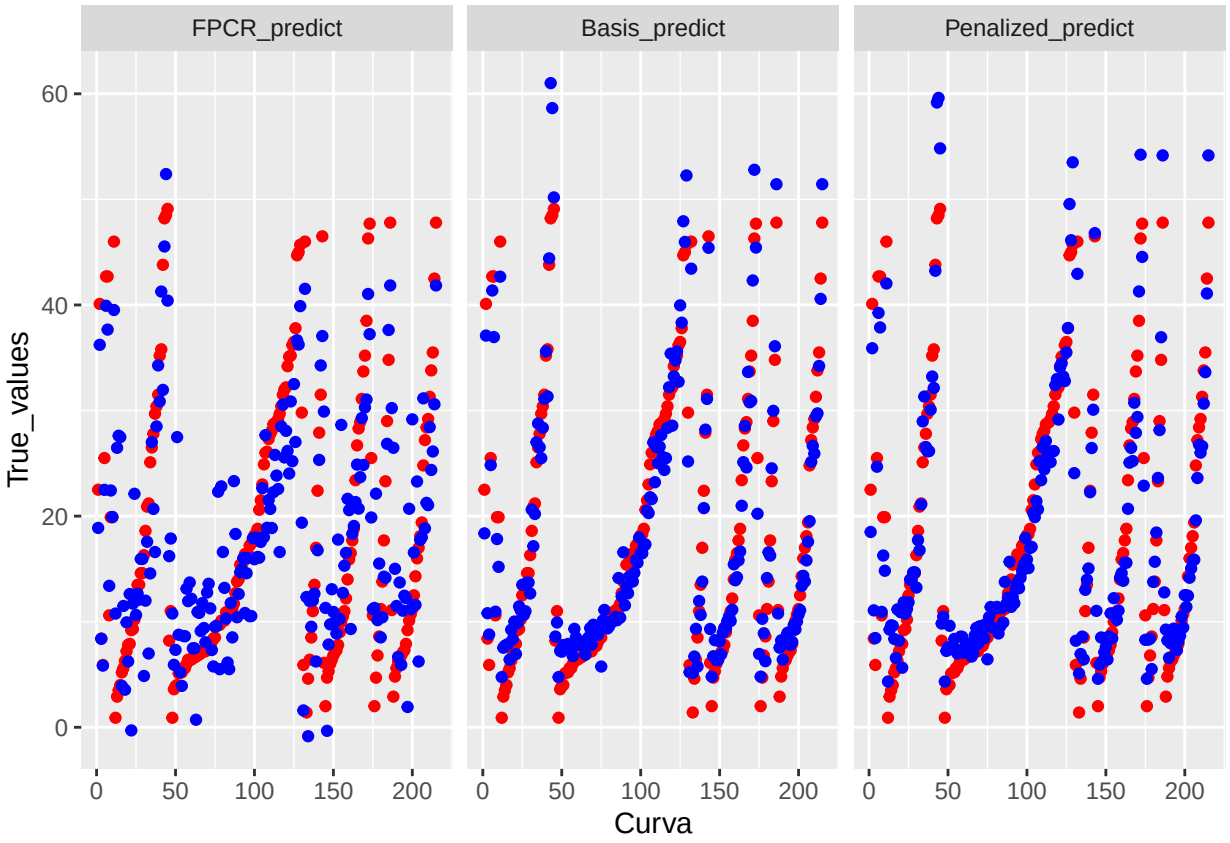

Gráfico 3: Valores observados de la variable respuesta junto con sus respectivas estimaciones de acuerdo a las diferentes metodologías de estimación.

Tabla 1: Cuadrados medio del error de acuerdo a la metodología de estimación de la función de regresión.

- Encuentre e interprete el coeficiente de determinación.

Para complementar esta revisión, la Tabla 2 muestra los valores de R^2 calculados para cada caso. La formula utilizada para el cálculo de los R^2 se muestra en la ecuación (7), donde SSE es la suma de cuadrados del error y SST es la suma de cuadrados totales. Los resultados y conclusiones en comparación de los tres métodos son los mismos que se obtuvieron con los CME. La interpretación de los R^2 es análoga al de los modelos de regresión usuales, donde dichos valores representan los porcentajes de variación de la variable respuesta explicados por la regresión.

$$R^2 = 1 - \frac{SSE}{SST} \tag{7}$$

- Formule un modelo de regresión adicionando la información suministrada por la primera derivada de las curvas espectrométricas. Encuentre las estimaciones de los parámetros y el coeciente de determinación. Compare estos resultados con los resutlados anteriores.

Tipo	SSE	SST	R2
FPCR_predict	11942.310	34735.44	0.6561924
Basis_predict	1580.643	34735.44	0.9544948
Penalized_predict	1927.610	34735.44	0.9445060

Tipo	MSE
FPCR_predict	6.468830
Basis_predict	5.959644
Penalized_predict	6.337808

Tabla 2: Sumas de cuadrado y coeficientes de determinación de acuerdo a la metodología de estimación de la función de regresión.

Se repitió el mismo procedimiento, de manera análoga al caso que se describió previamente, solamente que en este punto se incluyó una variable funcional adicional como explicativa dentro de la formula. El Gráfico 6 muestra las estimaciones obtenidas por este nuevo modelo a junto con los respectivos valores observados de acuerdo a los tres métodos de estimación. Nuevamente se observa que el modelo que peor rendimiento tiene en términos de proximidad entre estimaciones y valores observados es el que utiliza los FPCA, lo cual también se observa en las tablas 3 y 4, donde se tienen los CME y los coeficientes de determinación. Sin embargo, la mejora en la metodología de FPCA incluyendo la derivada es notoria.

Tabla 3: Cuadrados medio del error de acuerdo a la metodología de estimación de la función de regresión en el caso en que se adiciona una variable funcional correspondiente a la derivada de las funciones iniciales.

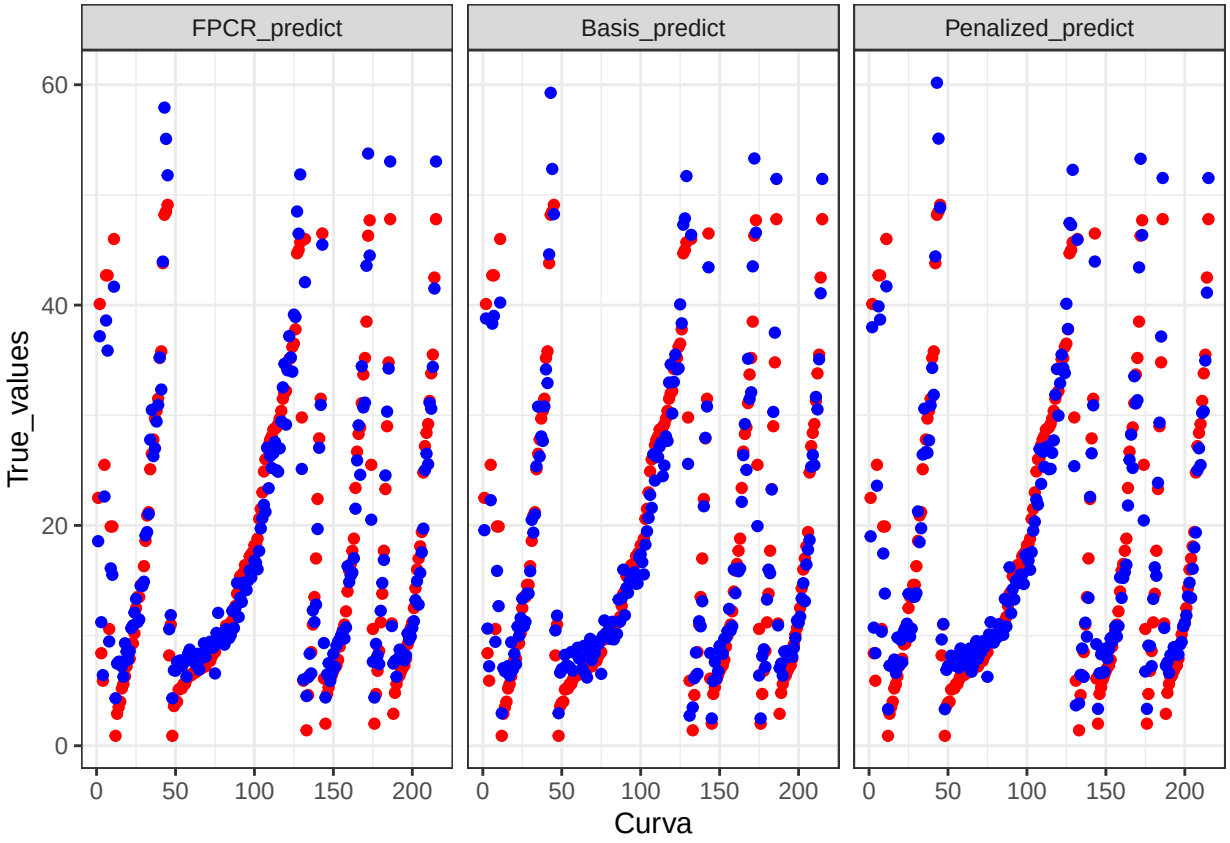

Gráfico 6: Valores observados de la variable respuesta junto con sus respectivas estimaciones de acuerdo a las diferentes metodologías de estimación en el caso en que se adiciona una variable funcional correspondiente a

Tipo	SSE	SST	R2
FPCR_predict	1390.799	34735.44	0.9599602
Basis_predict	1281.323	34735.44	0.9631119
Penalized_predict	1362.629	34735.44	0.9607712

la derivada de las funciones iniciales.

Tabla 4: Sumas de cuadrado y coeficientes de determinación de acuerdo a la metodología de estimación de la función de regresión en el caso en que se adiciona una variable funcional correspondiente a la derivada de las funciones iniciales.

References

Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012. URL http://www.jstatsoft.org/v51/i04/.

Jeff Goldsmith, Fabian Scheipl, Lei Huang, Julia Wrobel, Chongzhi Di, Jonathan Gellar, Jaroslaw Harezlak, Mathew W. McLean, Bruce Swihart, Luo Xiao, Ciprian Crainiceanu, and Philip T. Reiss. *refund: Regression with Functional Data*, 2020. URL https://CRAN.R-project.org/package=refund. R package version 0.1-23.

Piotr Kokoszka and Matthew Reimherr. *Introduction to functional data analysis*. CRC press, 2017.

S. N. Wood. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114, 2003.