Introduction to MECfda Package

Heyang Ji

Abstract

Package MECfda provide functions that sovle a variaty of scalar-on-function linear regression models and estimation methods that correct the bias due to measurement error in such models.

Introduction

Scalar-on-function linear regression is an important problem in functional data analysis. Our aim is to develop an R package that can solve all kinds of scalar-on-function linear regression models and correct the bias due to measurement error in scalar-on-function linear regression models.

```
library(MECfda)
```

Scalar-on-function Linear Regression Models

The generic form of scalar-on-function linear regression model is

$$T(F_{Y_i|X_i,Z_i}) = \sum_{l=1}^{L} \int_{\Omega_l} \beta_l(t) X_{li}(t) dt + (1, Z_i^T)\gamma$$

where Y_i represents the scalar-valued response variable, $X_{li}(t)$ represents the function-valued covariates $(l = 1, \ldots, L)$, β_l and X_{li} $(t \in \Omega_l)$ are in $L^2(\Omega_l)$, $Z_i = (Z_{1i}, \ldots, Z_{qi})^T$ represents the scalar valued covariates, $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)^T$, $F_{Y_i|X_i,Z_i}$ represents the CDF of the Y_i given X_i and Z_i, $T(\cdot)$ is a statistical functional. In ordinary scalar-on-function generalized linear (mixed) models,

$$T(F_{Y_i|X_i,Z_i}) = g\left\{\int_{\mathbb{R}} y dF_{Y_i|X_i,Z_i}(y)\right\} = g(E[Y_i|X_i, Z_i]),$$

where $g(\cdot)$ is a link function.

In scalar-on-function quantile linear regression models,

$$T(F_{Y_i|X_i,Z_i}) = Q_{Y_i|X_i,Z_i}(\tau) = F^{-1}_{Y_i|X_i,Z_i}(\tau),$$

where $F^{-1}_{Y_i|X_i,Z_i}(\cdot)$ is the inverse of $F_{Y_i|X_i,Z_i}(\cdot)$, $\tau \in (0, 1)$.

Functional Data

Function-valued variables (functional variables) are usually recorded by the value of the function(s) at some certain (time) points in its domain. The data of a function variable is often in a form of a matrix $(x_{ij})_{n\times m}$, where $x_{ij} = f_i(t_j)$, represents the value of $f_i(t_j)$, each row represent an observation (subject), each column is corresponding to a measurement (time) point.

In MECfda package, we have a s4 class, "functional_variable" that represents the data of a functional variable in this matrix form. The class has a slot for matrix $(x_{ij})_{n\times m}$, and slots for the domain of the functional variable and (time) points that the functional variable is measured.

```
fv = functional_variable(
  X = matrix(rnorm(10*24),10,24),
  t_0 = 0,
  period = 1,
  t_points = (0:9)/10
)
dim(fv)
#>     subject time_points
#>          10          24
```

Basis Expansion

$\{\rho_k\}_{k=1}^{\infty}$ is a basis sequence of $L^2(\Omega)$. For an arbitrary function-valued parameter $\beta(\cdot) \in L^2(\Omega)$, there exist number sequence $\{c_k\}_{k=1}^{\infty}$ s.t.

$$\beta(t) = \sum_{k=1}^{\infty} c_k \rho_k(t)$$

and for a function-valued variable $X_i(t)$,

$$\int_{\Omega} \beta(t) X_i(t) dt = \int_{\Omega} X_i(t) \sum_{k=1}^{\infty} c_k \rho_k(t) dt = \sum_{k=1}^{\infty} c_k \int_{\Omega} \rho_k(t) X_i(t) dt$$

We do a truncation for the infinite basis sequence, then

$$\int_{\Omega} \beta(t) X_i(t) dt \approx \sum_{k=1}^{K} c_k \int_{\Omega} X_i(t) \rho_k(t) dt$$

For statistical models with part(s) in the form of $\int_{\Omega} \beta(t) X_i(t) dt$ we use $c_k \int_{\Omega} X_i(t) \rho_k(t) dt$ to approximate $\int_{\Omega} \beta(t) X_i(t) dt$ and treat $\int_{\Omega} \rho_k(t) X_i(t) dt$ as the variables, then the scalar-on-function linear models is converted to a ordinary scalar-on-scalar linear model, and problem of estimating $\beta(t)$ is converted to estimating c_k, $k = 1, \ldots p$.

Perform basis expansion methods to data of functional variable in matrix form as we mentioned is to compute $(b_{ik})_{n \times p}$, where $b_{ik} = \int_{\Omega} f(t) \rho_k(t) dt$.

Usually the domain Ω of a function-valued variable $\{X(t), t \in \Omega\}$ is an interval. And the commonly used basis sequence types includes Fourier basis, b-splines basis, and eigen function basis.

The pakcage MECfda provide method "fourier_basis_expansion" and "bspline_basis_expansion" for class "functional_variable" to do basis expansion using Fourier basis and b-spline basis respectively

```
data(MECfda.data.sim.0.0)
fv = MECfda.data.sim.0.0$FC[[1]]
BE.fs = fourier_basis_expansion(fv,5L)
BE.bs = bspline_basis_expansion(fv,5L,3L)
```

Fourier basis

The Fourier basis of $L^2([t_0, t_0 + T])$ consists of

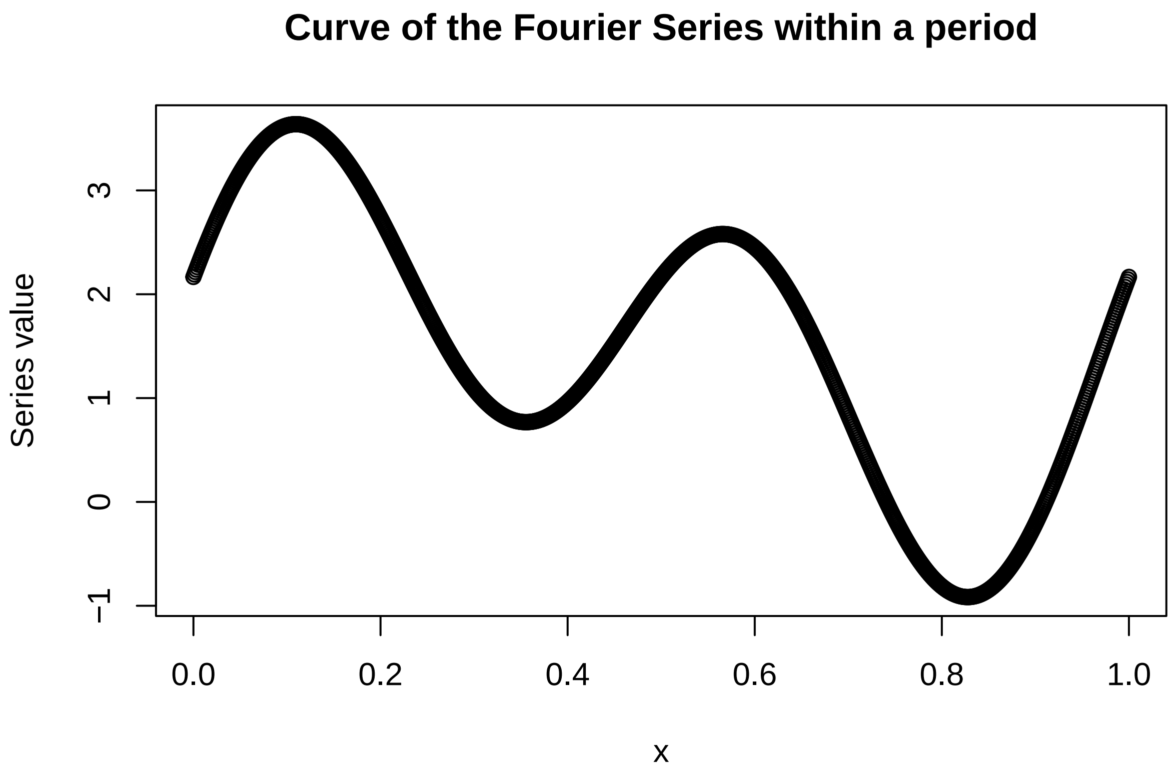
$$\frac{1}{2},\ \cos(\frac{2\pi}{T} k[x - t_0]),\ \sin(\frac{2\pi}{T} k[x - t_0])$$

where $k = 1, \ldots, \infty$.

In MECfda package, we have a s4 class, "Fourier_series" that represents the linear combination of Fourier basis functions

$$\frac{a_0}{2} + \sum_{k=1}^{p_a} a_k \cos\left(\frac{2\pi}{T} k(x - t_0)\right) + \sum_{k=1}^{p_b} b_k \sin\left(\frac{2\pi}{T} k(x - t_0)\right), \qquad x \in [t_0, t_0 + T].$$

```
fsc = Fourier_series(
  double_constant = 3,
  cos = c(0,2/3),
  sin = c(1,7/5),
  k_cos = 1:2,
  k_sin = 1:2,
  t_0 = 0,
  period = 1
)
plot(fsc)
```



The object "fsc" represents the summation

$$\frac{3}{2} + \frac{2}{3}\cos(2\pi \cdot 2x) + \sin(2\pi x) + \frac{7}{5}\sin(2\pi \cdot 2x).$$

B-splines basis

A b-spline basis $\{B_{i,p}(x)\}_{i=-p}^{k}$ on the interval $[t_0, t_{k+1}]$ is defined as

$$B_{i,0}(x) = \begin{cases} I_{(t_i, t_{i+1}]}(x), & i = 0, 1, \ldots, k \\ 0, & i < 0 \; or \; i > k \end{cases}$$

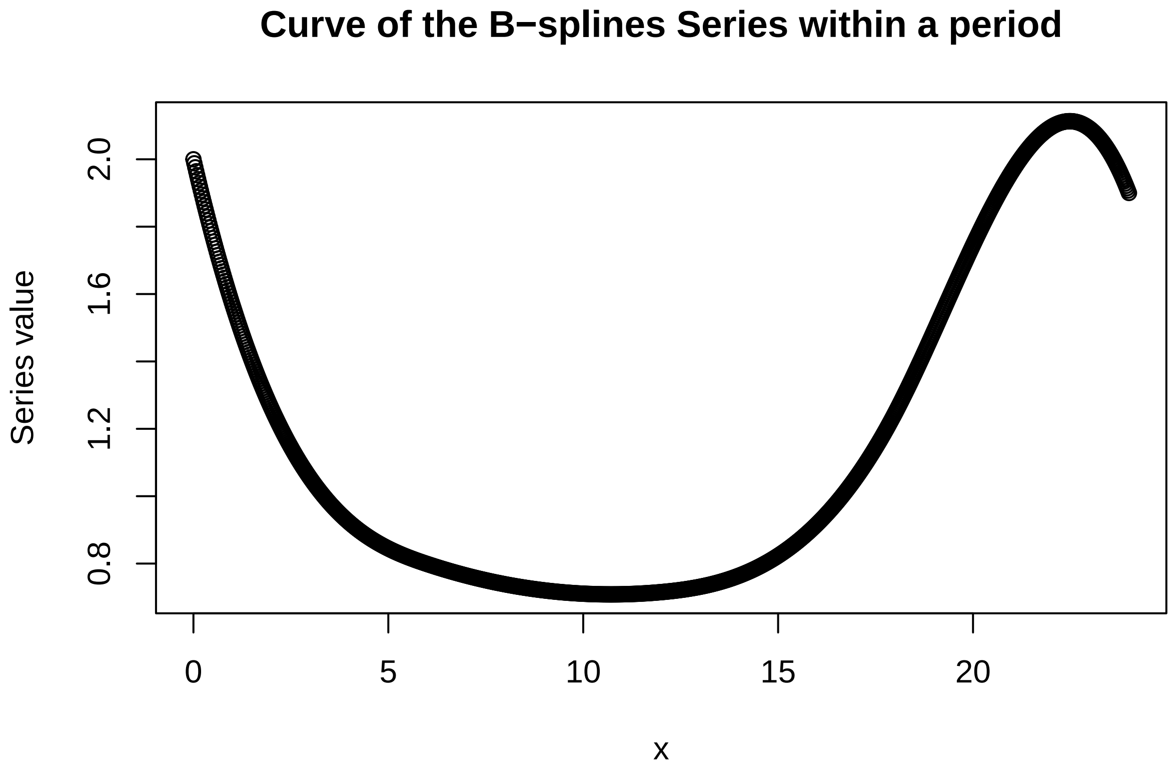
$$B_{i,r}(x) = \frac{x - t_i}{t_{i+r} - t_i} B_{i,r-1}(x) + \frac{t_{i+r+1} - x}{t_{i+r+1} - t_{i+1}} B_{i+1,r-1}(x)$$

For all the discontinuity points of $B_{i,r}$ $(r > 0)$ in the interval (t_0, t_k), let the value equals its limit, which means

$$B_{i,r}(x) = \lim_{t \to x} B_{i,r}(t).$$

In MECfda package, we have a s4 class, "bspline_basis" that represents a b-spline basis $\{B_{i,p}(x)\}_{i=-p}^{k}$ on the interval $[t_0, t_{k+1}]$. And we a s4 class, "bspline_series" that represents the summation $\sum_{i=0}^{k} b_i B_{i,p}(x)$.

```
bsb = bspline_basis(
  Boundary.knots = c(0,24),
  df             = 7,
  degree         = 3
)
bss = bspline_series(
  coef = c(2,1,3/4,2/3,7/8,5/2,19/10),
  bspline_basis = bsb
)
plot(bss)
```



The object "bsb" represents $\{B_{i,3}(x)\}_{i=-3}^{0}$, and object "bss" represents the

$$2B_{i,-3}(x) + B_{i,-2}(x) + \frac{3}{4}B_{i,-1}(x) + \frac{2}{3}B_{i,0}(x) + \frac{7}{8}B_{i,1}(x) + \frac{5}{2}B_{i,2}(x) + \frac{19}{10}B_{i,3}(x),$$

where $x \in [t_0, t_4]$ and $t_0 = 0$, $t_k = t_{k-1} + 6$,$k = 1, 2, 3, 4$.

Eigenfunction basis

Suppose $K(s,t) \in L^2(\Omega \times \Omega)$, $f(t) \in L^2(\Omega)$. Then K induces an linear operator $\mathcal{K}$ by

$$(\mathcal{K}f)(x) = \int_\Omega K(t,x)f(t)dt$$

If $\xi(\cdot) \in L^2(\Omega)$ s.t.

$$\mathcal{K}\xi = \lambda\xi$$

where $\lambda \in C$, then we call ξ a eigenfunction/eigenvector of $\mathcal{K}$ and λ a eigenvalue associated with ξ.

All the eigenfunctions of $\mathcal{K}$ make a basis of $L^2(\Omega)$. We call the basis induced by

$$K(s,t) = \mathrm{Cov}(X(t), X(s))$$

a functional principal component (FPCA) basis, where $\{X(t), t \in \Omega\}$ is a stochastic process.

Numerical Computation of Integrals

We use

$$\frac{1}{|T|}\sum_{t\in T} \rho_k(t)X_i(t)$$

to compute the integral

$$\int_\Omega \rho_k(t)X_i(t)dt$$

where T is the measurement (time) points of $X_i(t)$, $|T|$ represents the cardinal number of T.

Scalar-on-function Linear Regression in MECfda

fcRegression

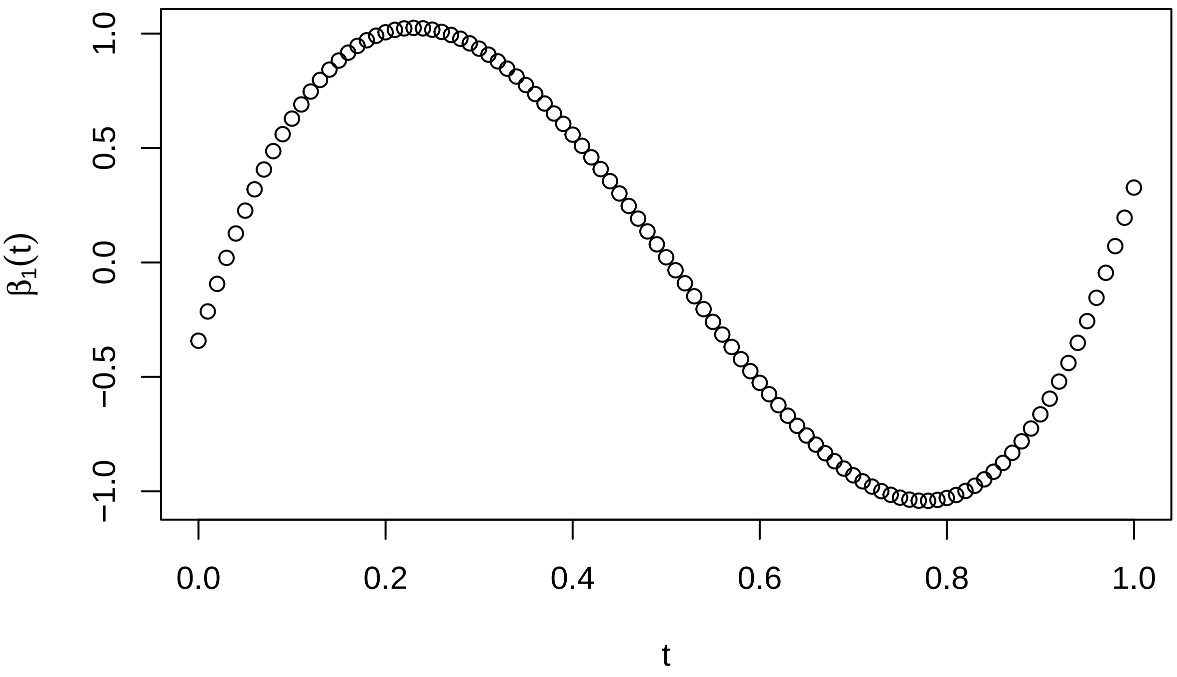
The MECfda package provides a function “fcRegression” to fit generalized linear mixed effect models, including ordinary linear model, generalized linear model with fixed and random effect, using basis expansion to discretize the function-valued covariates. The function “fcRegression” can solve a linear model in the following form:

$$g(E(Y_i|X_i, Z_i)) = \sum_{l=1}^{L} \int_{\Omega_l} \beta_l(t)X_{li}(t)dt + (1, Z_i^T)\gamma.$$

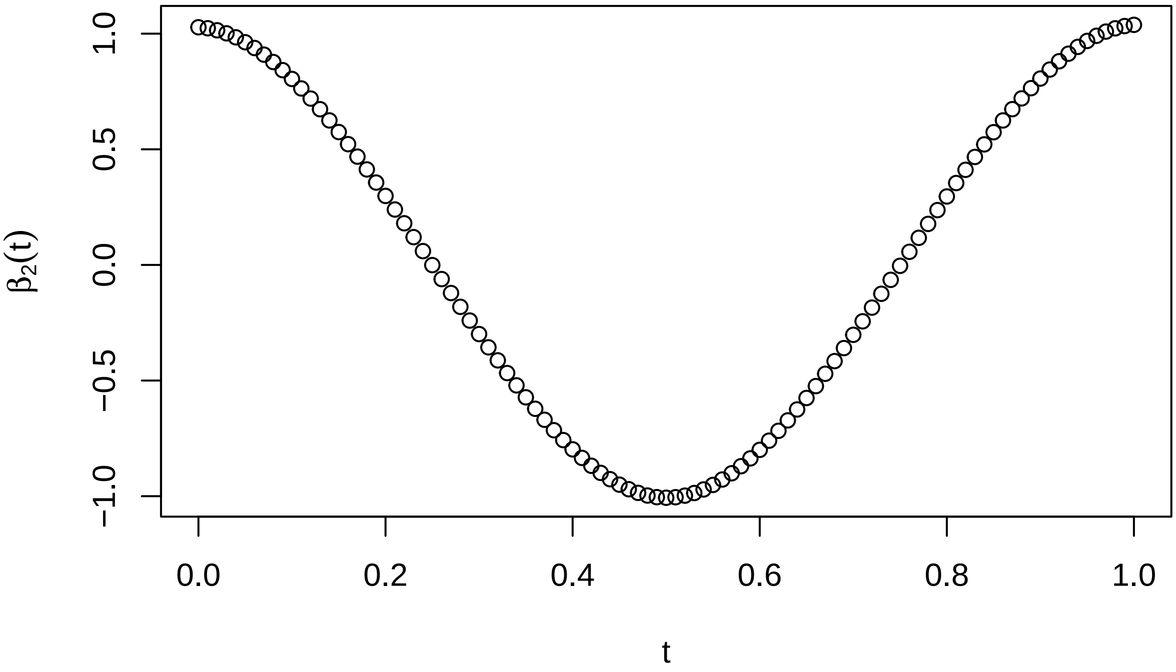
The funciton can allow one or multiple function-valued covariate(s) as fixed effect(s), and zero, one, or multiple scalar-valued covariate(s) as fixed or random effect(s). Response variable, function-valued covariate(s), and scalar-valued covariate(s) are input separately as three different arguements, Y, FC, and Z. The format of the input data can be very flexible. For response variable, the input format can be an atomic vector, a one-column matrix or data frame. Recommended form is a one-column data frame or matrix with column name, because in this case, the name of response variable is specified. For input data of function-valued covariate(s), a “functional_variable” object or a matrix or a data frame or a list of these object(s) can be accepted. When one “functional_variable” object or a matrix or a data frame is input as argument FC, there is only one function-valued covariate in the model. When list of these object(s) is input as argument FC, the model can have multiple function-valued covariates, each element of the list is correspondent to a function-valued covariate. For input data of scalar-valued covariates, a matrix, data frame, atomic vector, NULL or, not input can be accepted. When not assign input value for argument Z, there is no scalar-valued covariate in the model and argument formula.Z should also be NULL or not input. When an atomic vector is input as argument Z, there is only one scalar-valued in the model. And in this case, the name of the scalar-valued covariate is not specified. So even if there is only one scalar-valued covariate, a matrix or data frame with colname is recommended to be input as argument Z. The argument formula.Z is used to specify which part of the argument Z is used and how to treat the scalar-valued covariates, whether to use them as fixed effects or random effects. The argument "family" can specify the distribution type of response variable and link function to be used in the regression. Argument "basis.type" indicate the type of basis function to be used in basis expansion process. Available options are 'Fourier' and 'Bspline', represent Fourier basis and b-spline basis respectively. Argument "basis.order" indicates number of the function basis to be used. When using Fourier basis $\frac{1}{2}, \sin kt, \cos kt, k = 1, \ldots, p_f$, "basis.order" is the number p_f. When using b-splines basis $\{B_{i,p}(x)\}_{i=-p}^{k}$, "basis.order" is the number of splines, equal to $k + p + 1$. Argument "bs_degree" specify the degree of the piecewise polynomials of b-spline basis function if use b-splines basis. This argument is need only when using b-spline basis.

The function "fcRegression" returns an object of s3 class "fcRegression". It is a list that contains the information of the structure of statistical model, estimation results, input data, and the basis functions. We can use method "predict" to get predicted value from the model and use method "fc.beta" to get the value of function-valued linear coefficient parameters $\beta_l(t)$.

```
data(MECfda.data.sim.0.0)
res = fcRegression(FC = MECfda.data.sim.0.0$FC,
                   Y=MECfda.data.sim.0.0$Y,
                   Z=MECfda.data.sim.0.0$Z,
                   family = gaussian(link = "identity"),
                   basis.order = 5, basis.type = c('Bspline'),
                   formula.Z = ~ Z_1 + (1|Z_2))
t = (0:100)/100
plot(x = t, y = fc.beta(res,1,t), ylab = expression(beta[1](t)))
```



```
plot(x = t, y = fc.beta(res,2,t), ylab = expression(beta[2](t)))
```



```
data(MECfda.data.sim.1.0)
predict(object = res, newData.FC = MECfda.data.sim.1.0$FC,
        newData.Z = MECfda.data.sim.1.0$Z)
#>        1        2        3        4        5
#> 6.500129 5.690171 2.388979 5.441011 4.821000
```

fcQR

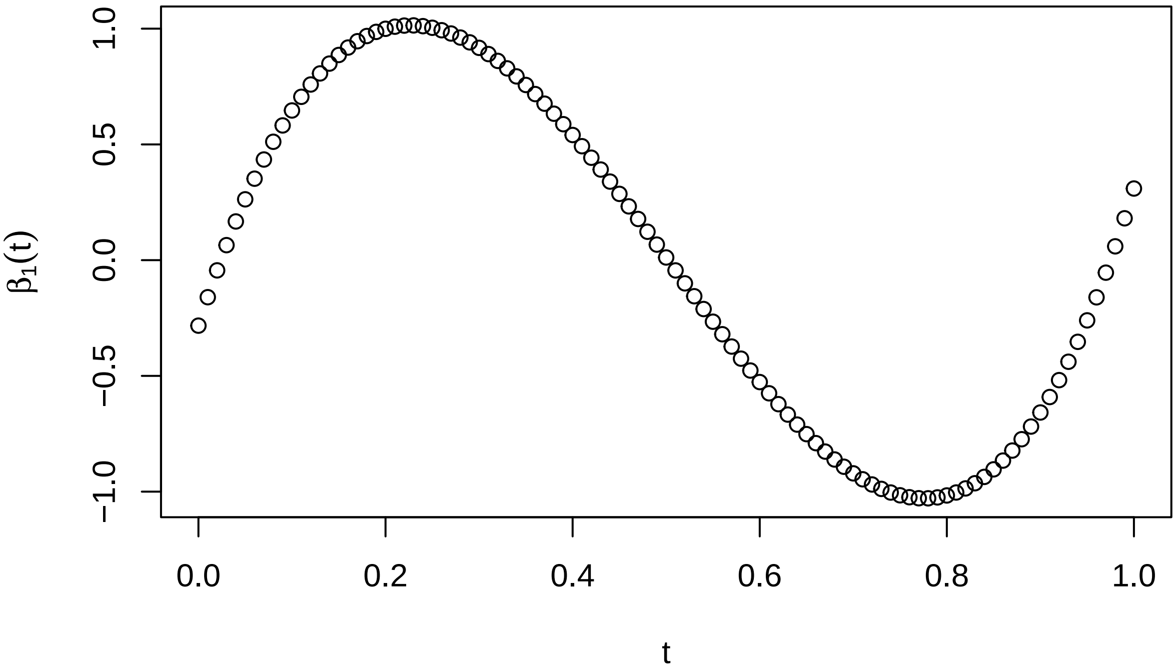
The MECfda package provides a function “fcQR” to fit quantile linear regression models. The method to deal with function-valued covariates is also discretization using basis expansion. The function “fcQR” can solve a linear model in the following form:

$$Q_{Y_i|X_i,Z_i}(\tau) = \sum_{l=1}^{L} \int_{\Omega_l} \beta_l(\tau, t) X_{li}(t)dt + (1, Z_i^T)\gamma.$$

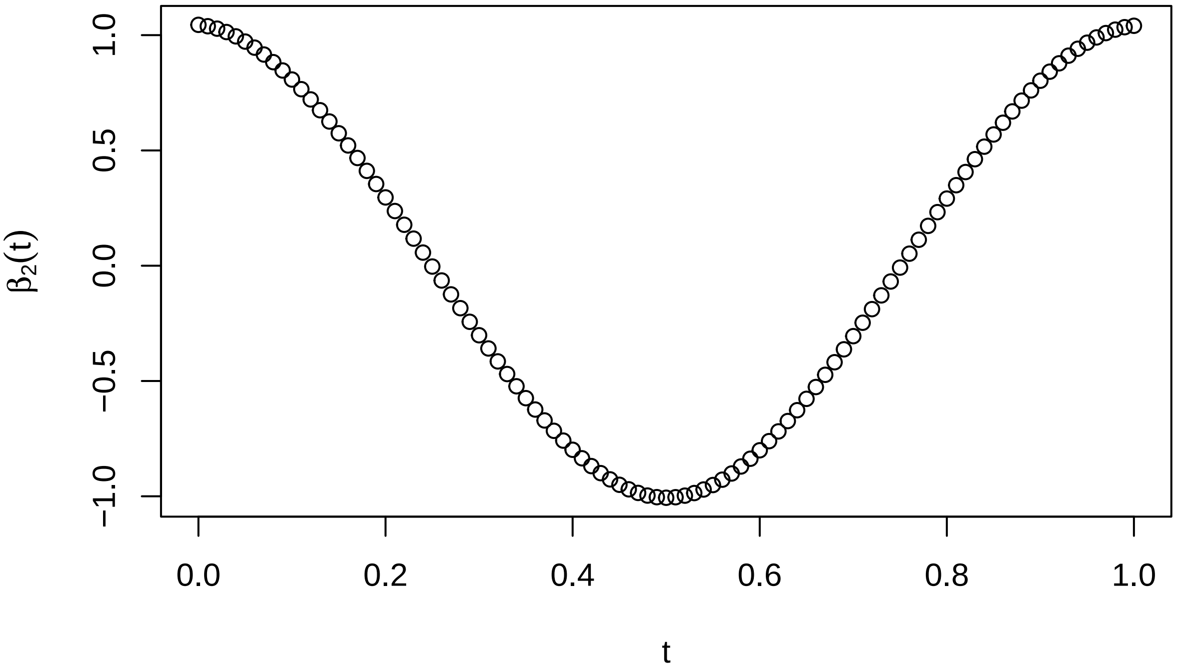
The funciton can allow one or multiple function-valued covariate(s), and zero, one, or multiple scalar-valued covariate(s). The way and rules to input the data is exact the same as function “fcRegression”. How to treat the scalar-valued covariates is also specified by the argument “formula.Z”, similar to in function “fcRegression”. The only difference is that there is no random effect in quantile linear regression model. The quantile τ is specified by the argument “tau”. The type and parameters of the basis function are also specified by argument “basis.type”, “basis.order”, and “bs_degree” as in function “fcRegression”.

The function “fcQR” returns an object of s3 class “fcQR”. It is a list that contains the information of the structure of statistical model, estimation results, input data, and the basis functions. We can use method “predict” to get predicted value from the model and use method “fc.beta” to get the value of function-valued linear coefficient parameters $\beta_l(t)$.

```
data(MECfda.data.sim.0.0)
res = fcQR(FC = MECfda.data.sim.0.0$FC,
           Y=MECfda.data.sim.0.0$Y,
           Z=MECfda.data.sim.0.0$Z,
           tau = 0.5,
           basis.order = 5, basis.type = c('Bspline'),
           formula.Z = ~ .)
t = (0:100)/100
plot(x = t, y = fc.beta(res,1,t), ylab = expression(beta[1](t)))
```



```
plot(x = t, y = fc.beta(res,2,t), ylab = expression(beta[2](t)))
```



```
data(MECfda.data.sim.1.0)
predict(object = res, newData.FC = MECfda.data.sim.1.0$FC,
        newData.Z = MECfda.data.sim.1.0$Z)
#>        1        2        3        4        5
#> 6.497759 5.682573 2.404464 5.440699 4.830085
```

Measurement Error Models

Data in real world often have measurement error, especially for function-valued data. And measurement error in data set may lead to bias in estimation. Package MECfda also provides functions of bias correction estimation method for some certain linear regression models with measurement error.

ME.fcRegression_MEM

Luan et. al. proposed a mixed effect model based bias correction estimation method for scalar-on-function generalied linear regression model with measurement error. (Luan et al. 2023) Their statistical model is as follow:

$$
\begin{aligned}
&g(E(Y_i|X_i,Z_i)) = \int_\Omega \beta(t)X_i(t)dt + (1,Z_i^T)\gamma \\
&h(E(W_{ij}(t)|X_i(t))) = X_i(t) \\
&X_i(t) = \mu_x(t) + \varepsilon_{xi}(t)
\end{aligned}
$$

where response variable Y_i and scalar-valued covariates Z_i are measured without error, function-valued covariate $X_i(t)$ is repeatedly measured with error as $W_{ij}(t)$. And they have additional assumption:

1. $Y_i|X_i,Z_i \sim EF(\cdot)$, EF refers to an exponential family distribution.

2. $g(\cdot)$ and $h(\cdot)$ are known monotone, twice continuously differentiable functions.
3. $Cov\{X_i(t), W_{ij}(t)\} \neq 0$,
4. $W_{ij}(t)|X_i(t) \sim EF(\cdot)$
5. $f_{Y_i|W_{ij}(t),X_i(t)}(\cdot) = f_{Y_i|X_i(t)}(\cdot)$, f refers to PDF.
6. $X_i(t) \sim GP\{\mu_x(t), \Sigma_{xx}\}$, GP refers to Gaussian process.

They proposed a mixed effect model based estimation method to correct the bias due to the measurement error. The Mixed-effect model (MEM) approach is a two-stage-based method that employs functional mixed-effects models. It allows us to delve into the nonlinear measurement error model, where the relationship between the true and observed measurements is not constrained to be linear, and the distribution assumption on the observed measurement is relaxed to encompass the exponential family rather than being limited to the Gaussian distribution. The MEM approach employs point-wise (UP_MEM) and multi-point-wise (MP_MEM) estimation procedures to avoid potential computational complexities caused by analyses of multi-level functional data and computations of potentially intractable and complex integrals.

Package MECfda provide a function "ME.fcRegression_MEM" to apply their bias correction estimation method.

ME.fcQR_IV.SIMEX

Tekwe et. al. proposed a simulation extrapolation (SIMEX) estimation method to correct the bias in scalar-on-function quantile linear regression due to measurement error using instrumental variable. (Tekwe et al. 2022) Their statistical model is as follow:

$$
\begin{aligned}
&Q_{Y_i|X_i,Z_i}(\tau) = \int_\Omega \beta(\tau,t) X_{li}(t)dt + (1, Z_i^T)\gamma(\tau) \\
&W_i(t) = X_i(t) + U_i(t) \\
&M_i(t) = \delta(t)X_i(t) + \eta_i(t)
\end{aligned}
$$

where response variable Y_i and scalar-valued covariates Z_i are measured without error, function-valued covariate $X_i(t)$ is measured with error as $W_i(t)$, and $M_i(t)$ is an measured instrumental variable. And they have additional assumption:

1. $Cov\{X_i(t), U_i(s)\} = 0$,
2. $Cov\{M_i(t), U_i(s)\} = 0$,
3. $E(W_i(t)|X_i(t)) = X_i(t)$
4. $U_i(t) \sim GP\{\prime, \Sigma_{uu}\}$, GP refers to Gaussian process.

for $\forall t, s \in [0,1]$ and $i = 1, \ldots, n$.

Their bias correction estimation method performs a two-stage strategy to correct the measurement error of a function-valued covariate and then fit a linear quantile regression model. In the first stage, an instrumental variable is used to estimate the covariance matrix associated with the measurement error. In the second stage, simulation extrapolation (SIMEX) is used to correct for measurement error in the function-valued covariate.

Package MECfda provide a function "ME.fcQR_IV.SIMEX" to apply their bias correction estimation method.

ME.fcQR_CLS

Zhang et. al. proposed a corrected loss score estimation method for scalar-on-function quantile linear regression to correct the bias due to measurement error. (Zhang et al. 2023) Their statistical model is as follow:

$$
\begin{aligned}
&Q_{Y_i|X_i,Z_i}(\tau) = \int_\Omega \beta(\tau,t) X_{li}(t)dt + (1, Z_i^T)\gamma(\tau) \\
&W_i(t) = X_i(t) + U_i(t)
\end{aligned}
$$

where response variable Y_i and scalar-valued covariates Z_i are measured without error, function-valued covariate $X_i(t)$ is measured with error as $W_i(t)$.

1. $E[U_i(t)] = 0$.
2. $Cov\{U_i(t), U_i(s)\} = \Sigma_U(s,t)$, where $\Sigma_U(s,t)$ is unknown.
3. $U_i(t)$ are i.i.d for different i.

Zhang et al. proposed a new corrected loss function for a partially functional linear quantile model with functional measurement error in this manuscript. They established a corrected quantile objective function of the observed measurement that is an unbiased estimator of the quantile objective function that would be obtained if the true measurements were available. The estimators of the regression parameters are obtained by optimizing the resulting corrected loss function. The resulting estimator of the regression parameters is shown to be consistent.

Package MECfda provide a function "ME.fcQR_CLS" to apply their bias correction estimation method.

ME.fcLR_IV

Tekwe et. al. proposed a bias correction estimation method for scalar-on-function linear regression model with measurement error using instrumental variable. (Tekwe et al. 2019) Their statistical model is as follow:

$$\begin{aligned} Y_i &= \int_0^1 \beta(t) X_i(t) dt + \varepsilon_i \\ W_i(t) &= X_i(t) + U_i(t) \\ M_i(t) &= \delta X_i(t) + \eta_i(t) \end{aligned}$$

where response variable Y_i and are measured without error, function-valued covariate $X_i(t)$ is measured with error as $W_i(t)$, and $M_i(t)$ is an measured instrumental variable. And they have additional assumption:

1. $E\varepsilon_i(t) = 0$,
2. $EU_i(t) = 0$,
3. $E\eta_i(t) = 0$,
4. $Cov\{X_i(t), \varepsilon_i\} = 0$,
5. $Cov\{M_i(t), \varepsilon_i\} = 0$,
6. $Cov\{M_i(t), U_i(s)\} = 0$,

for $\forall t, s \in [0,1]$ and $i = 1, \ldots, n$.

Package MECfda provide a function "ME.fcLR_IV" to apply their bias correction estimation method.

References

Luan, Yuanyuan, Roger S Zoh, Erjia Cui, Xue Lan, Sneha Jadhav, and Carmen D Tekwe. 2023. "Scalable Regression Calibration Approaches to Correcting Measurement Error in Multi-Level Generalized Functional Linear Regression Models with Heteroscedastic Measurement Errors." *arXiv Preprint arXiv:2305.12624*.

Tekwe, Carmen D, Mengli Zhang, Raymond J Carroll, Yuanyuan Luan, Lan Xue, Roger S Zoh, Stephen J Carter, David B Allison, and Marco Geraci. 2022. "Estimation of Sparse Functional Quantile Regression with Measurement Error: A SIMEX Approach." *Biostatistics* 23 (4): 1218–41.

Tekwe, Carmen D, Roger S Zoh, Miao Yang, Raymond J Carroll, Gilson Honvoh, David B Allison, Mark Benden, and Lan Xue. 2019. "Instrumental Variable Approach to Estimating the Scalar-on-Function Regression Model with Measurement Error with Application to Energy Expenditure Assessment in Childhood Obesity." *Statistics in Medicine* 38 (20): 3764–81.

Zhang, Mengli, Lan Xue, Carmen D Tekwe, Yang Bai, and Annie Qu. 2023. "PARTIALLY FUNCTIONAL LINEAR QUANTILE REGRESSION WITH MEASUREMENT ERRORS." *Statistica Sinica* 33: 2257–80.