scientific reports

OPEN

Generalized partially functional linear model

Weiwei Xiao[1], Yixuan Wang[1✉] & Haiyan Liu[2,3]

In this paper, a generalized partially functional linear regression model is proposed and the asymptotic property of the proposed estimated coefficients in the model is established. Extensive simulation experiment results are consistent with the theoretical result. Finally, two application examples of the model are given. One is sleep quality study where we studied the effects of heart rate, percentage of sleep time on total sleep in bed, wake after sleep onset and number of wakening during the night on sleep quality in 22 healthy people. The other one is mortality rate where we studied the effects of air quality index, temperature, relative humidity, GDP per capita and the number of beds per thousand people on the mortality rate across 80 major cities in China.

Nowadays, more and more data, in various fields such as medicine, economics, biology and computer, are recorded in the form of curves or images. For these high-dimensional data, traditional multiple regression analysis is insufficient. Ramsay [23] proposed the concept of functional data. Functional data can be thought of as a real valued function defined on a compact interval. In other words, the data sampled can be represented as a function by interpolation or fitting. Indeed, functional data is different from scalar data because of its infinite dimension. Ramsay and Dalzell[24] proposed functional data analysis. Since then, functional data analysis has been very popular among researchers. Many statistical methods such as principal component analysis, functional regression, and cluster analysis have been developed and widely used. For details on functional data analysis, see monographs Ramasy and Sliverman[25], Ferraty and Vieu[10], Horvath and Kokoszka[14], Hsing and Eubank[15], and so on.

With respect to functional regression, functional and scalar data can respectively appear in the regression model as predictors and responses. And estimation of regression coefficients is important in functional regression model. The case that response is scalar and predictors are functional is well-developed. For examples, Cardot et al.[5] studied two estimation methods of principal component estimation and penalty spline, Cai and Hall[3] used a principal components approach to estimate regression coefficients, Kou and Liu[19] used a wavelet basis to estimate regression coefficients, and so on. The case where the response is scalar with functional and scalar predictors can be referred to Shin[31] which used the least squares method to estimate regression parameters, Shin and Lee[32] and its regression parameter estimation is based on functional principal components regression (FPCR) and the alternative is functional ridge regression (FRR) based on Tikhonov regularization, Kong et al.[18] characterized the effects of regularization on the resulting estimators, Wang et al.[33] used a two-step estimation method based on functional effective dimension reduction, slice inverse regression and kernel estimation to give the estimation of partial functional linear models, and gave the convergence rate of the estimation, and so on.

Generalized functional regression model has attracted more and more researchers' interest. Müller[21] develop generalized functional regression but only include one functional predictor and proposed a functional estimating equation which is maximizing a functional quasi-likelihood. Goldsmith et al.[11] proposed the generalized linear mixed model and compared likelihood based and Bayesian estimation. However the response in their simulation study and application is continuous, and the corresponding asymptotic property of estimated parameters is not obtained. Crainiceanu et al.[8] introduced generalized multilevel functional linear models. They proposed and compared two methods for inference: a two-stage frequentist approach and a joint Bayesian analysis. They studied the generalized functional linear model in which the predictive variables are both functional and scalar and the response variables are scalar. However, they do not give corresponding results for the generalized functional linear model with binary or Poisson response variables. Ieva and Paganoni[16] proposed a generalized functional linear regression model for binary response and applied the model to the ECG signals. They studied the generalized functional linear model with binary response variables, but they did not give the corresponding asymptotic property of estimated parameters, and the predictive variables only have functional data in their application. For general response such as binary or Poisson the theory is not well-developed. In this paper, We

[1]Department of Mathematics, North China University of Technology, Beijing 100144, China. [2]Department of statistics, University of Leeds, Leeds LS2 9JT, UK. [3]The Alan Turing Institute, London NW1 2DB, UK. ✉email: 1426044457@qq.com

establish generalized partially functional regression model which can deal with general response and multiple scalar and functional predictors and we develop quasi-likelihood method to estimate the regression coefficients.

Mental health problem reduces quality of life, especially during this pandemic, and sleep quality can result in severe mental health problem. Therefore study of sleep quality has received extensive attention from researchers, see Zhang[36], Kadoya[17] and so on. Our work is motivated by a sleep quality study aiming to investigate the relationship between sleep quality which is binary and covariates which are functional and scalar. The Multilevel Monitoring of Activity and Sleep in Healthy people (MMASH) dataset provides 24 hours of continuous beat-to-beat heart data, sleep quality, physical activity and psychological characteristics (i.e., anxiety status, stress events and emotions) for 22 healthy participants. Moreover, saliva bio-markers (i.e. cortisol and melatonin) and activity log were also provided in this dataset. In MMASH database, we have curves like heart rate, inter-beat intervals; we have scalar data like wake after sleep onset, number of awakening during the night, hormones concentrations in the saliva and so on, and we have binary response i.e. good or bad sleep quality. Our work is also motivated by another study in public health, i.e. to investigate the impact of air quality index (AQI), temperature, relative humidity (RH), GDP per capita and the number of beds per thousand people on the mortality rate across different cities in China. In particular, we are interested in studying the effect of AQI on mortality. AQI reflects the degree of air pollution and it is judged by the concentration of pollutants in the air, where the main pollutants in the air include PM2.5, PM10, carbon monoxide, nitrogen dioxide, sulfur dioxide, ozone and so on.

The paper is organized as follows. The generalized partially functional linear model is proposed in Section "Generalized partially functional linear regression". The estimation of the regression coefficients within the generalized functional linear model is discussed in Section "Estimation of β and γ". In Section "Asymptotic inference", asymptotic normality of estimators is derived where the needed appropriate metrics are given. Simulation results are reported in Section "Simulation study". Two real data examples, the sleep quality study in healthy people and the mortality rate study in major Chinese cities, are given in Section "Application". We conclude in Section "Conclusion". Proofs and other supplementary materials are followed in Appendix.

Generalized partially functional linear regression

The data we observe for the i-th subject are $\{(X_{i1}(t_1), t_1 \in T_1), (X_{i2}(t_2), t_2 \in T_2), \ldots, (X_{id}(t_d), t_d \in T_d), Z_i, Y_i\}, i = 1 \ldots n$. We assume that these data are independent identically distributed (i.i.d.) copies of $(X_1, \ldots, X_d, Z, Y)$. For $j = 1, \ldots, d$, the functional predictor $X_{ij}(t_j)$ is a random curve which is observed for subject i and corresponds to a square integrable stochastic process on a real interval T_j, i.e. $X_{ij}(t_j) \in L^2(T_j)$. And the scalar predictor vector $Z = (Z_1, Z_2, \ldots Z_q)$ is a q dimensional random vector. The dependent variable or response Y_i is a real-valued random variable which may be continuous or discrete (e.g. binary, count etc.).

We assume there is a known link function $g(\cdot)$ which is a monotone and twice continuously differentiable function with bounded derivatives and is thus invertible. We further assume there is a variance function $\sigma^2(\cdot)$ which is defined on the range of the link function and is strictly positive.

We assume the following relation between Y_i and $(X_{i1}, \ldots, X_{id}, Z_i)$

$$Y_i = g\left(\alpha + \sum_{j=1}^{d} \int_{T_j} X_{ij}(t_j)\beta_j(t_j)dt_j + Z_i\gamma\right) + \varepsilon_i, \tag{1}$$

where $\alpha \in \mathbb{R}$ is the intercept, $\beta_j(\cdot) \in L^2(T_j)$ is the regression function corresponding to functional predictor X_j, and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^T$ is the regression coefficient vector corresponding to the scalar predictor vector Z. Furthermore, ε is random variable from exponential family that satisfies $\mathrm{E}[\varepsilon \mid X_j(t_j), Z] = 0$ with $\mathrm{E}[\varepsilon^2] = \mathrm{Var}[\varepsilon_i] = \mathrm{Var}[Y] = \sigma^2(\mathrm{E}[Y])$. Therefore, the generalized partially functional linear model is determined by parameter coefficient function $\beta_j(\cdot)$, parameter coefficient vector γ, link function $g(\cdot)$ and variance function $\sigma^2(\cdot)$.

Define linear predictors η,

$$\eta = \alpha + \sum_{j=1}^{d} \int_{T_j} X_j(t_j)\beta_j(t_j)dt_j + Z\gamma. \tag{2}$$

We have conditional mean $\mathrm{E}[Y \mid X_j(t_j), Z] = \mu = g(\eta)$ and $\mathrm{Var}(Y \mid X_j(t_j), Z) = \sigma^2(\mu) = \tilde{\sigma}^2(\eta)$ for a function $\tilde{\sigma}^2(\eta) = \sigma^2(g(\eta))$.

For simplicity, assume that both functional predictors $X(t)$ and scalar predictors Z are centralized, i.e. $\mathrm{E}[X_j(t_j)] = 0, j = 1, 2, \ldots, d$ and $\mathrm{E}[Z_l] = 0, l = 1, 2, \ldots, q$. In order to solve the problem of infinite dimension of functional data, we adopt functional principal component analysis method to reduce dimension. For functional predictor $X_{ij}(t_j)$, by Karhunen-Loeve (KL) expansion and Mercer's theorem, it can be expanded as:

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

where ξ_{ijk} is called functional principal component score and the functions $\phi_{jk}(\cdot), k = 1, 2, \ldots$ are called functional principal components obtained via the covariance structure of $X_{ij}(t_j)$. Notice that $\phi_{jk}(\cdot), k = 1, 2, \ldots$ forms an orthonormal basis for the function space $L^2(T_j)$ and clearly $\int_{T_j} \phi_{jk}^2(t_j)dt_j = 1$. Then parameter coefficients $\beta_j(t_j)$ can be expanded as

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

with $\sum b_{jk}^2 < \infty$.

Therefore, after plugging the above two expansions in to (1), we have:

$$Y_i = g\left(\alpha + \sum_{j=1}^{d}\sum_{k=1}^{p_j}\xi_{ijk}b_{jk} + Z_i\gamma\right) + \varepsilon_i, i = 1, 2, \ldots, n. \tag{3}$$

Notice that in (3), in order to solve the difficulty caused by the infinite dimension of the functional predictors, we truncate the predictors at p_j, and the dimension p_j increases asymptotically with $n \to \infty$.

Estimation of β and γ

Denote parameter vector

$$\boldsymbol{\theta} = (b_{11}, b_{12}, \ldots, b_{1p_j}, b_{21}, b_{22}, \ldots, b_{2p_j}, \ldots, b_{d1}, b_{d2}, \ldots, b_{dp_j}, \gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_q)^T.$$

Then the maximum likelihood estimator of $\boldsymbol{\theta}$ can be obtained by solving the score equation:

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)g'(\eta_i)}{\sigma^2(\mu_i)}\boldsymbol{\omega}_i = 0, \tag{4}$$

where $\eta_i = \alpha + \sum_{j=1}^{d}\sum_{k=1}^{p_j} b_{jk}\xi_{ijk} + Z_i\gamma, \mu_i = g(\eta_i), i = 1, 2, \ldots, n$ and

$$\boldsymbol{\omega}_i = (\xi_{i11}, \xi_{i12}, \ldots, \xi_{i1p_1}, \xi_{i21}, \xi_{i22}, \ldots, \xi_{i2p_2}, \ldots, \xi_{id1}, \xi_{id2}, \ldots, \xi_{idp_d}, z_{i0}, z_{i1}, \ldots, z_{iq})^T.$$

We denote the MLE

$$\hat{\boldsymbol{\theta}} = \left(\hat{b}_{11}, \hat{b}_{12}, \ldots, \hat{b}_{1p_j}, \hat{b}_{21}, \hat{b}_{22}, \ldots, \hat{b}_{2p_j}, \ldots, \hat{b}_{d1}, \hat{b}_{d2}, \ldots, \hat{b}_{dp_j}, \hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_q\right)^T$$

and therefore $\hat{\boldsymbol{b}}_j = (\hat{b}_{j1}, \hat{b}_{j2}, \ldots, \hat{b}_{jp_j})^T, j = 1, \ldots, d, \hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_q)^T, \hat{\alpha} = \hat{\gamma}_0$ are the estimators of $\boldsymbol{b}_j, \boldsymbol{\gamma}, \alpha$ respectively.

We introduce the following matrices:

$$D_0 = D_{n,q+1} = \left(\frac{g'(\eta_i)z_{il}}{\sigma(\mu_i)}\right)_{1\leq i\leq n, 0\leq l\leq q},$$

$$D_j = D_{n,p_j} = \left(\frac{g'(\eta_i)\xi_{ijk}}{\sigma(\mu_i)}\right)_{1\leq i\leq n, 1\leq k\leq p_j}, j = 1, 2, \ldots, d,$$

$$D = D_{n,q+1+\sum_j^d p_j} = \text{diag}(D_1, D_2, \ldots, D_d, D_0),$$

$$V = \text{diag}(\sigma^2(\mu_1), \sigma^2(\mu_2), \ldots, \sigma^2(\mu_n)).$$

and vectors $Y = (Y_1, Y_2, \ldots, Y_n)^T$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$. Then the score Eq. (4) can be rewritten as

$$D^T V^{-1/2}(Y - \mu) = 0.$$

This equation is usually solved iteratively by the method of integrated weighted least squares. Under our basic conditions in Section "Generalized partially functional linear regression", $\frac{1}{n}\text{E}(D^T D)$ is a fixed positive definite matrix.

Asymptotic inference

Generalized auto-covariance operator. Given an L^2 integrable kernel function $R(s,t): T^2 \to R$, define the linear integral operator $A_R: L^2(ds) \to L^2(dt)$ on the Hilbert space $L^2(ds)$ for $f \in L^2(dt)$ by

$$\left(A_R f\right)(t) = \int f(s)R(s,t)ds.$$

An operator A_R is compact self-adjoint Hilbert-Schmidt operator if

$$\int |R(s,t)|^2 dsdt < \infty,$$

and can then be diagonalized.

Integral operator of special interest is the auto-covariance operator A_K of K with kernel

$$K(s,t) = \text{cov}(X(s), X(t)) = \text{E}[X(s)X(t)],$$

and the generalized auto-covariance operator A_G with kernel

3

$$G(s,t) = \mathrm{E}\left[\frac{g'(\eta)^2}{\sigma^2(\mu)}X(s)X(t)\right]$$

where $\lambda_{j,k_1 k_2}$ is a non-increasing sequence of eigenvalues.

Hilbert-Schmidt operators A_R generate a metric in L^2,

$$d_R^2(f,g) = \int (f(t) - g(t))(A_R(f-g))(t)\mathrm{d}t$$

$$= \iint (f(s) - g(s))(f(t) - g(t))R(s,t)\mathrm{d}s\mathrm{d}t.$$

for $f,g \in L^2(ds)$, and given an arbitrary orthonormal basis $\{\phi_j, j = 1, 2, \ldots\}$, the Hilbert-Schmidt kernels R can be expressed as

$$R(s,t) = \sum_{k,l} r_{kl}\phi_k(s)\phi_l(t)$$

for suitable coefficients $\{r_{kl}, k, l = 1, 2, \ldots\}$.

Thus, for each $X_j(t_j)$, $j = 1, \ldots, d$, the correlated Hilbert-Schmidt operator of A_G produces a metric in L^2:

$$d_G^2\left(\hat{\beta}_j, \beta_j\right) = \iint \left(\hat{\beta}_j(s) - \beta_j(s)\right)\left(\hat{\beta}_j(t) - \beta_j(t)\right)G(s,t)\mathrm{d}s\mathrm{d}t$$

$$= \iint \left(\hat{\beta}_j(s) - \beta_j(s)\right)\left(\hat{\beta}_j(t) - \beta_j(t)\right)\mathrm{E}\left[\frac{g'(\eta)^2}{\sigma^2(\mu)}X_j(s)X_j(t)\right]\mathrm{d}s\mathrm{d}t$$

$$= \left(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j\right)^T \Lambda_j\left(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j\right), j = 1, 2, \ldots, d,$$

where $\hat{\boldsymbol{b}}_j = (\hat{b}_{j1}, \hat{b}_{j2}, \ldots)^T$ is the estimator of $\boldsymbol{b}_j = (b_{j1}, b_{j2}, \ldots)^T$, $\Lambda_j = (\lambda_{j,k_1 k_2})_{1 \leq k_1, k_2 \leq \infty} = \left(\mathrm{E}\left[\frac{g'(\eta)^2}{\sigma^2(\mu)}\xi_{jk_1}\xi_{jk_2}\right]\right)_{1 \leq k_1, k_2 \leq \infty}$ is a symmetric and positive definite matrix with elements to be the eigenvalues of generalized auto-covariance operator A_G.

For the sequence of p_j truncated model (3), we have

$$d_G^2\left(\hat{\beta}_j, \beta_j\right) = \left(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j\right)^T \widetilde{\Lambda}_j\left(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_j\right) + \sum_{k_1, k_2 = p_j+1}^{\infty} \lambda_{j,k_1 k_2}\bar{\boldsymbol{b}}_j^2, j = 1, 2, \ldots, d,$$

where $\hat{\boldsymbol{b}}_j = (\hat{b}_{j1}, \hat{b}_{j2}, \ldots, \hat{b}_{jp_j})^T$ is the estimator of $\boldsymbol{b}_j = (b_{j1}, b_{j2}, \ldots, b_{jp_j})^T$, and here,

$$\widetilde{\Lambda}_j = (\lambda_{j,k_1 k_2})_{1 \leq k_1, k_2 \leq p_j},$$

$$\lambda_{j,k_1 k_2} = \mathrm{E}\left[\frac{g'(\eta)^2}{\sigma^2(\mu)}\xi_{jk_1}\xi_{jk_2}\right],$$

$$\widetilde{\Lambda}_j^{-1} = (\zeta_{j,k_1 k_2})_{1 \leq k_1, k_2 \leq p_j}.$$

We note that $\widetilde{\Lambda}_j = \frac{1}{n}\mathrm{E}(D_j^T D_j)$ is a symmetric and positive definite matrix and that the inverse matrix $\widetilde{\Lambda}_j^{-1}$ exists. And $\bar{\boldsymbol{b}}_j = \left(b_{j(p_j+1)}, b_{j(p_j+2)}, \ldots\right)^T$.

We note that higher-order oscillations associated with property $\sum_{k_1, k_2 = p_j+1}^{\infty} \lambda_{j,k_1 k_2}\bar{\boldsymbol{b}}_j^2, j = 1, 2, \ldots, d$ contribute to the L^2 norm of the parameter functions $\beta_j(t_j)$, relative to the oscillations of processes $X_j(t_j)$, i.e.,

$$\sum_{k_1, k_2 = p_j+1}^{\infty} \lambda_{j,k_1 k_2}\bar{\boldsymbol{b}}_j^2 = o\left(\frac{\sqrt{p_j}}{n}\right),$$

The specific proof follows by combining the results of corollary 4.1 of Müller (2005).

To derive the asymptotic property of $d_G^2\left(\hat{\beta}_j, \beta_j\right)$, in addition to the basic assumptions in Section "Generalized partially functional linear regression" and usual conditions on variance and link functions, we require some technical conditions which restrict the growth of p_j and the independence of X_j and Z. The basic model assumption is as follows:

(A1) The link function g is monotone, invertible and has two continuous bounded derivatives with $\|g'(\cdot)\| \leq c$, $\|g''(\cdot)\| \leq c$ for a constant $c \geq 0$. The variance function $\sigma^2(\cdot)$ has a continuous bounded derivative and there exists a $\delta > 0$ such that $\sigma^2(\cdot) > \delta$.

(A2) The fourth moment of X_j is finite, i.e. $\mathrm{E}\left[\int_T \{X_j(t)\}^4 \mathrm{d}t\right] < \infty$.

(A3) $Z = (Z_1, \ldots, Z_q)$ and $(X_1(t), X_2(t), \ldots, X_d(t))$ are independent of each other.

(A4) The number of truncated terms p_j in the sequence of approximating p_j truncated model (3) satisfies $p_j \to \infty$ and $p_j n^{-\frac{1}{4}} \to 0$ as $n \to \infty$.

Remark 1 In the assumption (A1), because the link function g we defined is continuous, its first derivative is bounded. And its second derivative is bounded in order to make sure that the Hessian matrix is a meaningful existence when we prove corresponding asymptotic property. In the assumptions (A2) and (A3), the point is to simplify our proof, and we do not have to worry about the interactions between the predictors. As shown in the assumption (A4), the truncation p_j goes to infinity, but in order for lemmas to hold and for the rate of convergence to be even faster, we have to $p_j n^{-\frac{1}{4}} \to 0$ to control the rate at which p_j goes to infinity. Since d and q represent finite number of predictors, then $(q + 1 + \sum_{j=1}^{d} p_j) n^{-\frac{1}{4}} \to 0$.

Asymptotic convergence of $\beta_j(t)$ and γ. **Lemma 1** As $n \to \infty$, assumptions (A1)–(A4)

$$\left\| \sqrt{n}(\hat{\theta} - \theta) - \left(\frac{D_j^T D_j}{n} \right)^{-1} \frac{U(\theta)}{\sqrt{n}} \right\|_2 = o(1).$$

Remark 2 Lemma 1 essentially implies $\sqrt{n}(\hat{\theta} - \theta) \sim \left(\frac{D_j^T D_j}{n} \right)^{-1} \frac{U(\theta)}{\sqrt{n}}$.

Lemma 2 *Under the assumptions (A1)–(A4), we have that*

$$\left\| \Psi_{nj}^2 - I_{nj} \right\|_2^2 = O\left(\frac{1}{p_j} \right).$$

where Ψ_{nj} are $p_j \times p_j$ matrices

$$\Psi_{nj} = \widetilde{\Lambda}_j^{\frac{1}{2}} \left(\frac{D_j^T D_j}{n} \right)^{-1} \widetilde{\Lambda}_j^{\frac{1}{2}}$$

and I_{nj} are $p_j \times p_j$ diagonal matrices.

Theorem 1 *If the basic conditions in Section "Generalized partially functional linear regression" and assumptions (A1)–(A4) are satisfied, then*

$$\begin{pmatrix} \frac{nd_G^2(\hat{\beta}_1, \beta_1) - p_1}{\sqrt{2p_1}} \\ \frac{nd_G^2(\hat{\beta}_2, \beta_2) - p_2}{\sqrt{2p_2}} \\ \vdots \\ \frac{nd_G^2(\hat{\beta}_d, \beta_d) - p_d}{\sqrt{2p_d}} \\ \sqrt{nv_0}(\gamma_0 - \hat{\gamma}_0) \\ \sqrt{nv_1}(\gamma 1 - \hat{\gamma}_1) \\ \vdots \\ \sqrt{nv_q}(\gamma_d - \hat{\gamma}_d) \end{pmatrix} \xrightarrow{d} N(0, I_a)$$

where $v_l = \mathrm{E}\left[\frac{g'(\eta_i)^2}{\sigma^2(\mu_i)} z_{il}^2 \right]$, $l = 0, 1, \ldots, q$ and I_a is a $\left(q + 1 + \sum_{j=1}^{d} p_j \right) \times \left(q + 1 + \sum_{j=1}^{d} p_j \right)$ identity matrix.

Remark 3 For $j = 1, 2, \ldots, d$, each $nd_G^2\left(\hat{\beta}_j, \beta_j \right)$ obeys a asymptotically normal distribution with the mean of p_j and the variance of $2p_j$. For $l = 0, 1, \ldots, q$, each $\sqrt{n}(\gamma_l - \hat{\gamma}_l)$ obeys a asymptotically normal distribution with the mean of 0 and the variance of $\mathrm{E}\left[\frac{g'(\eta_i)^2}{\sigma^2(\mu_i)} z_{il}^2 \right]^{-1}$. Because we assume that $(X_{i1}(t_1), X_{i2}(t_2), \ldots, X_{id}(t_d), z_{i0}, z_{i1}, \ldots, z_{iq})$ are independent of each other, the variance term in the jointly asymptotically normal distribution is a diagonal matrix.

Corollary 1 *Denote the eigenvectors and eigenvalues of the matrix $\tilde{\Lambda}_j$ by*

$$(e_{j,1}, \lambda_{j,1}), (e_{j,2}, \lambda_{j,2}), \ldots, (e_{j,p_j}, \lambda_{j,p_j}),$$

and let

$$\boldsymbol{e}_{j,k} = (e_{j,k1}, e_{j,k2}, \ldots, e_{j,kp_j})^T,$$
$$\omega_{j,k}(t) = \sum_{l=1}^{p_j} \phi_{j,l}(t) e_{j,k}, k = 1, 2, \ldots, p_j.$$

where $\{\phi_{j,l}(t), l = 1, 2, \ldots\}$ is an orthonormal basis, which is mentioned in Section "Generalized partially functional linear regression".

Then for large n and p_j an approximate $(1 - \alpha)$ simultaneous confidence band is given by

$$\hat{\beta}_j(t) \pm \sqrt{c(\alpha) \sum_{k=1}^{p_j} \frac{\omega_{j,k}(t)^2}{\lambda_{j,k}}}.$$

where $c(\alpha) = [p_j + \sqrt{2p_j}\Phi(1 - \alpha)]/n$. And when $\alpha = 0.05$, then $\Phi(1 - \alpha) = 1.96$.

Simulation study

We consider the case with two functional predictors and three scalar predictors and the response is binary.

For the functional predictors $X_{i1}(t_1)$, $t_1 \in [0, 1]$, and $X_{i2}(t_2)$, $t_2 \in [-1, 1]$, $i = 1, ..., n$, we first generate ξ_{ijk}, $j = 1, 2$ which satisfy

$$\xi_{i1k} \sim N(0, \lambda_{1k}), \ k = 1, 2, 3, \ \text{with } \lambda_{11} = 6, \lambda_{12} = 4, \lambda_{13} = 2$$

and

$$\xi_{i2k} \sim N(0, \lambda_{2k}), \ k = 1, 2, 3, \ \text{with } \lambda_{21} = 6, \lambda_{22} = 4, \lambda_{23} = 2, \lambda_{24} = \frac{1}{2}, \lambda_{25} = \frac{1}{4}.$$

Defining orthonormal basis $\phi_{1k}(t_1)$, $t_1 \in [0, 1]$ and $\phi_{2k}(t_2)$, $t_2 \in [-1, 1]$, which satisfy

$$\phi_{1k}(t_1) = \begin{cases} \sqrt{2}\cos(k\pi t_1), k = 1, 2, 3 \\ 0, k \geqslant 4. \end{cases}$$

$$\phi_{2k}(t_2) = \begin{cases} \sqrt{\frac{(4k-3)}{2}}t_2^{2(k-1)}, k = 1, 2, 3, 4, 5 \\ 0, k \geqslant 6. \end{cases}$$

Therefore $X_{i1}(t_1)$, $t_1 \in [0, 1]$, and $X_{i2}(t_2)$, $t_2 \in [-1, 1]$, $i = 1, ..., n$ satisfy

$$X_{i1}(t_1) = \sum_{k=1}^{3} \xi_{i1k}\phi_{1k}(t_1),$$

$$X_{i2}(t_2) = \sum_{k=1}^{5} \xi_{i2k}\phi_{2k}(t_2).$$

where ϕ_{jk}, $j = 1, 2$ are principal component base. Figure 1 plots part of the $X_1(t_1)$ and $X_2(t_2)$.

We simulate scalar predictors as follows: $z_1 \sim N(0, 1)$, $z_2 \sim N(0, 3)$ and $z_3 \sim B(1, 0.5)$.

We assume the true regression coefficients are generated as

$$\gamma = (2, 3, 5)^T,$$

$$\beta_1(t_1) = \sum_{k=1}^{3} b_{1k}\phi_{1k}(t_1),$$

$$\beta_2(t_2) = \sum_{k=1}^{5} b_{2k}\phi_{2k}(t_2),$$

where $b_{1k} = 2k^2$, $b_{2k} = k^2$.

Therefore the binary response is generated as follows: We define

$$p(X, Z) = g\left(\sum_{j=1}^{2} \int_T X_j(t_j)\beta_j(t_j)\mathrm{d}t_j + Z\gamma\right),$$

and choose logit link

$$g(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

Then we generate response

$$Y(X, Z) \sim \text{Bernoulli}(p(X, Z), 1)$$

as pseudo-Bernoulli random variable sequence with probability $p(X, Z)$, when $p > 0, Y = 1$, otherwise $Y = 0$.

Therefore, we have a sample

$$\{(X_{i1}(t_1), t_1 \in [0, 1]), (X_{i2}(t_2), t_2 \in [-1, 1]), Z_i, Y_i\}, \ i = 1 \ldots n,$$

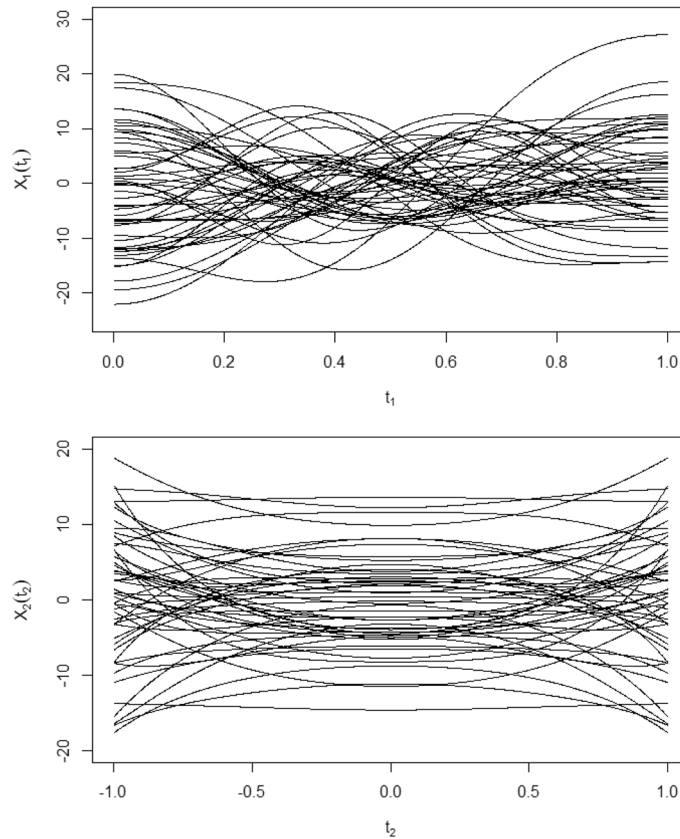**Figure 1.** Part of the $X_1(t_1)$ and $X_2(t_2)$.

n	McFadden	r2ML
50	0.736	0.976
500	0.816	0.984
1000	0.925	0.996

Table 1. Statistics and statistical criteria under different sample sizes.

where n is the sample size.

In the simulation study, in order to study the asymptotic property of our estimators we choose separately $n = 50, \ 500, \ 1000$, and the number of functional principal components that explain 90% of cumulative variation contribution are $p_1 = 1, \ 2, \ 2$, $p_2 = 2, \ 3, \ 4$. We do 100 simulations.

Table 1 shows the values of statistics McFadden's pseudo R^2 (McFadden) and Maximum likelihood pseudo R^2 (r2ML) for different n. The simulation results show that the performance of the model gets better and better with the increase of sample size. This is consistent with our expectations.

Then, in order to study the performance of $\hat{\beta}_1(t_1)$ and $\hat{\beta}_2(t_2)$, we show the 95% confidence band under different sample sizes of the estimator $\hat{\beta}_1(t_1)$ and $\hat{\beta}_2(t_2)$ in Fig. 2. The red curves are true $\beta_1(t_1)$ and $\beta_2(t_2)$, the black curves are the corresponding estimations $\hat{\beta}_1(t_1)$ and $\hat{\beta}_2(t_2)$, the gray parts are the 95% confidence band. As we can see, as the sample size increases, the confidence band becomes narrower and narrower and the estimate (red line) gets closer to the theoretical true value (black line). Therefore, the simulation clearly indicates that the larger the sample size, the closer the estimated and true values.

In order to verify Theorem 1, we calculate the mean and variance of $nd_G^2\left(\hat{\beta}_j, \beta_j\right)$, $j = 1, 2$ and $\sqrt{n}(\gamma_l - \hat{\gamma}_l)$, $l = 1, 2, 3$. To be specific, the theoretical values of the mean and variance of $nd_G^2\left(\hat{\beta}_j, \beta_j\right)$ are 3, 6 and 5, 10 respectively for $j = 1, 2$. The theoretical values of the mean of $\sqrt{n}(\gamma_l - \hat{\gamma}_l)$, $l = 1, 2, 3$ are 0. Table 2 shows the sample mean and variance of $nd_G^2\left(\hat{\beta}_j, \beta_j\right)$, $j = 1, 2$ and $\sqrt{n}(\gamma_l - \hat{\gamma}_l)$, $l = 1, 2, 3$. From Table 2 we can see that the sample mean and sample variance of $nd_G^2\left(\hat{\beta}_j, \beta_j\right)$ tend to 3, 6 for $j = 1$ and 5, 10 for $j = 2$ and the sample mean of $\sqrt{n}(\gamma_l - \hat{\gamma}_l)$, $l = 1, 2, 3$ tend to 0 as n increases which verifies Theorem 1. The simulation results show that the estimation of regression coefficient functions gets better and better as the sample size increases.
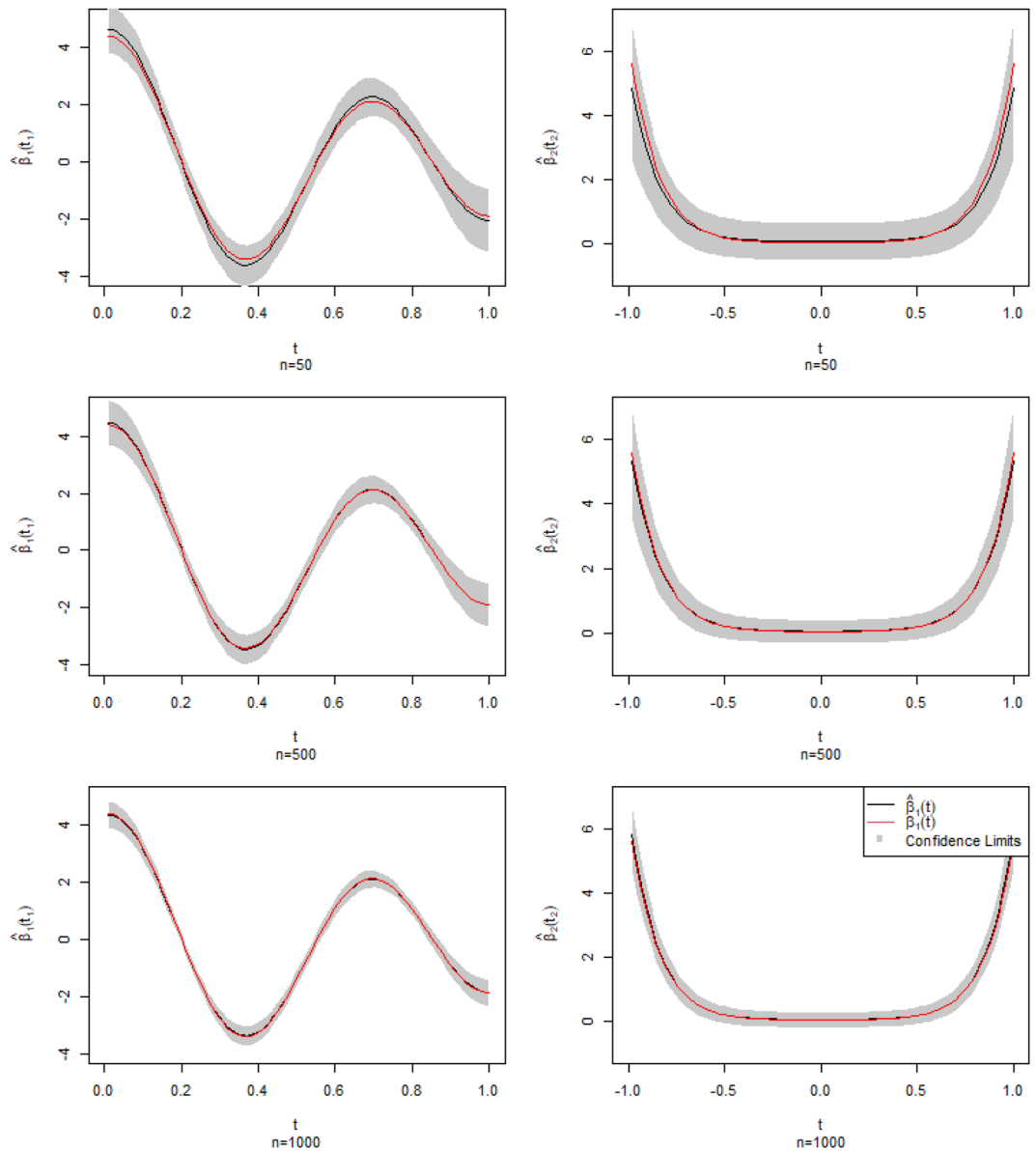
Figure 2. 95% confidence band for the estimator $\hat{\beta}_1(t_1)$ and $\hat{\beta}_2(t_2)$ under different sample sizes. The red curves are theoretical $\beta_1(t_1)$ and $\beta_2(t_2)$, the black curves are the corresponding estimation and it's the result of a simulation, the gray parts are the 95% confidence band.

Table 3 shows the estimated values and corresponding standard deviations of the $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)^T$ under different sample sizes. The simulation results show that as the sample size increases, the standard deviation becomes smaller and smaller, and $\hat{\gamma}$ tends to the theoretical value, where the theoretical values of $\gamma = (\gamma_1, \gamma_2, \gamma_3)^T$ are 2,3,5.

We use GCV to demonstrate the predictive accuracy of the estimators. When $n = 50, 500, 1000$, the corresponding GCV values are 0.104, 0.005, 0.001 which imply that as the sample size increases the prediction becomes more accurate. For details of GCV, see Roozbeh et al.[27], Roozbeh et al.[28], Roozbeh[29], Amini and Roozbeh[1] for details.

Application

Sleep quality. Data on activity and sleep of healthy person from PhysioNet Databases were analyzed in this section. The data were collected and provided by BioBeats (biobeats.com)[34] in collaboration with researchers from the University of Pisa. Data were recorded on 22 healthy males in various aspects of their daily lives, such as cardiovascular responses, psychological perception, sleep quality, and exercise information.

We investigated the effects of hourly heart rate (HR), percentage of sleep time on total sleep in bed (Efficiency), wake after sleep onset (WASO) and number of awakenings during the night (Number) on sleep quality. The volunteers' Pittsburgh Sleep Quality Questionnaire index was used as response, which follows a Bernoulli distribution where good sleep quality is represented by 1 and bad sleep quality is represented by 0. In our study,

	n	Mean	Variance
$nd_G^2\left(\hat{\beta}_1, \beta_1\right)$	50	4.15	7.95
	500	3.31	6.75
	1000	3.03	6.08
$nd_G^2\left(\hat{\beta}_2, \beta_2\right)$	50	6.06	13.42
	500	5.51	11.49
	1000	5.06	10.63
$\sqrt{n}(\gamma_1 - \hat{\gamma}_1)$	50	0.19	1.26
	500	0.16	1.11
	1000	0.09	1.08
$\sqrt{n}(\gamma_2 - \hat{\gamma}_2)$	50	0.07	0.15
	500	0.04	0.13
	1000	0.03	0.11
$\sqrt{n}(\gamma_3 - \hat{\gamma}_3)$	50	0.10	5.01
	500	0.05	4.02
	1000	0.04	3.93

Table 2. The mean and variance of $nd_G^2\left(\hat{\beta}_j, \beta_j\right)$ and $\sqrt{n}(\gamma_l - \hat{\gamma}_l)$ in Theorem 1.

n	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$
50	1.84 (0.16)	2.97 (0.06)	4.89 (0.34)
500	2.06 (0.04)	2.99 (0.02)	4.95 (0.09)
1000	2.01 (0.03)	2.99 (0.01)	5.03 (0.05)

Table 3. The estimated values and corresponding standard deviations in brackets of the estimator $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)^T$.

	Estimate	SE	t value	Pr(> |t|)	
$\hat{\gamma}_{Efficiency}$	0.106	0.213	0.497	0.063	–
$\hat{\gamma}_{WASO}$	0.059	0.063	0.938	0.037	–
$\hat{\gamma}_{Number}$	− 0.143	0.007	− 0.07	0.006	**
r2ML = 0.917					

Table 4. Parameter coefficient estimation and significance levels.

7 men had a bad sleep quality and 15 had a good sleep quality. Efficiency, WASO and Number were used as scalar predictors. The volunteers' hourly heart rate was used as a functional predictor. We chose $g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$ as link function. We chose the number of functional principal components that explain 80% of cumulative variation contribution and $p = 5$.

By substituting the data into our model, the functional parameter coefficients $\hat{\beta}$ and the non-functional parameter coefficients $\hat{\gamma}$ can be obtained. Moreover, with respect to the prediction accuracy, the GCV is obtained: is 0.358. The results of the estimation are shown in Table 4 and Fig. 3.

From the estimated non-functional parameter coefficient vector $\hat{\gamma}$ shown in Table 4, we can see that Efficiency and WASO are positively correlated with sleep quality. In other words, the greater the proportion of sleep time or the longer the first waking time after falling asleep, the better the corresponding sleep quality. On the contrary, Number is negatively correlated with sleep quality, that is, the more awakenings during the night, the poorer sleep quality.

Figure 3 shows the estimated regression coefficient function $\hat{\beta}(t)$ and its 95% confidence band. We can tentatively conclude that heart rate is positively correlated with sleep quality between 8:00 and 11:00. On the contrast, heart rate is mainly negatively correlated with sleep quality at other time. So, proper exercise and hard work to raise your heart rate in the morning will help improve the quality of sleep at night. While, some people may think that doing some exercise at night will improve their sleep quality because they are tired, but this is exactly the opposite. Sajjadieh (2020) used the method of time domain spectral analysis and obtained that heart rate has a negative correlation with sleep quality. But we give more insights on the dynamic and temporal behaviour of the influence of heart rate on sleep quality.
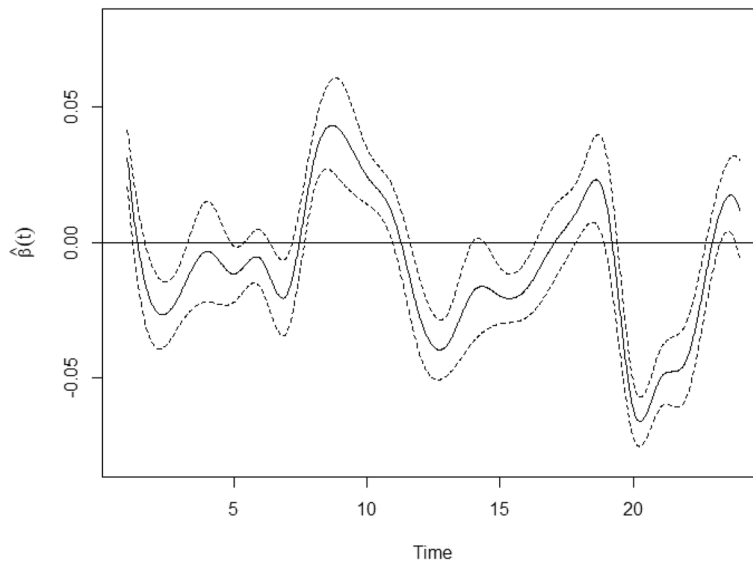
Figure 3. Regression coefficient function $\hat{\beta}(t)$ and its 95% confidence band.

Mortality. We have a dataset which includes daily temperature, daily relative humidity (RH), daily air quality index (AQI), GDP per capita, the number of beds per thousand people and mortality in 80 cities in China collected during 2019. A main goal of the study was to investigate the impact of temperature, RH, AQI, GDP per capita and the number of beds per thousand people on the mortality rate. We apply the proposed generalized partially functional linear model to the data for different cities.

We have three functional predictors: daily AQI, daily temperature and daily RH from 1 January 2019 to 31 December 2019. The two scalar predictors are GDP per capita and the number of beds per thousand people. The response is the mortality rate for each city in 2019, which is documented in the statistical bulletin of each city. Mortality follows a Bernoulli distribution where the mortality rate greater than 6‰ is considered high and represented by 1, otherwise the mortality rate is low and the mortality rate is represented by 0. In our study, there are 30 cities with high mortality and 50 with low mortality. Figure 4 shows daily AQI, temperature and RH in some cities in 2019.

We chose $g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$ as link function. We chose the number of functional principal components that explain 75% of cumulative variation contribution, i.e. $p_{\text{AQI}} = 11$, $p_{\text{Temp}} = 4$, $p_{\text{RH}} = 10$ as standard orthogonal basis. By substituting the data into our model, the regression coefficients of functional predictors $\hat{\beta}$ and regression coefficients of scalar predictors $\hat{\gamma}$ are shown in Table 5 and Fig. 5. Moreover, the prediction accuracy is shown by the GCV with value 0.037.

From the estimated regression coefficient vector $\hat{\gamma}$ shown in Table 5, we can see that GDP per capita and the number of beds per thousand people are negatively correlated with mortality rate. In other words, the greater GDP per capita or the number of beds per thousand people, the lower the mortality rate.

Figure 5 shows the estimated regression coefficient functions $\hat{\beta}(t)$ and their 95% confidence band. From Fig. 5, we can conclude that AQI is positively correlated with mortality in winter, early spring and fall. In other words, the higher the AQI, i.e. the worse the air conditions, the higher the death rate. Kong[18] studied the effect of PM2.5 on mortality in US cities from 1 April 2000 to 31 August 2000. Unsurprisingly, in our study, the effect of AQI on mortality from April to August is concise with their results. For the effect of temperature on mortality, we can conclude that temperature is negatively correlated with mortality from March to May (in spring), and is positively correlated with mortality for the rest of the year. For the effect of RH on mortality, we can conclude that RH is positively correlated with mortality from August to November (in autumn), and is negatively correlated with mortality for the rest of the year. The effect of temperature and RH on mortality is very easy to use in traditional Chinese medicine (TCM) theory. In TCM, Yin and Yang are emphasized, and there is a saying that spring is born, summer is growing, autumn is converging and winter is stored. In spring, the gradually high temperature indicates that the climate is normal and Yang occurs normally. But if temperature is low in spring then Yang occurs abnormally which results in health problems. In autumn, Yang is converging, and both temperature and RH should gradually decrease. If the RH is still high, it shows that Yang cannot dive into the internal storage, floating over the outside, excessive consumption of human healthy.

Conclusion

In this paper, we propose a generalized functional linear regression model with multiple scalar and functional predictors. We develop maximum likelihood estimators for the regression coefficients. For the functional predictors, we adopt the method of functional principal component analysis to reduce their dimensions. We then propose the generalized auto-covariance operator, based on which an appropriate measure quantifies the difference between the estimators and their true values is established. The asymptotic joint distribution of estimated
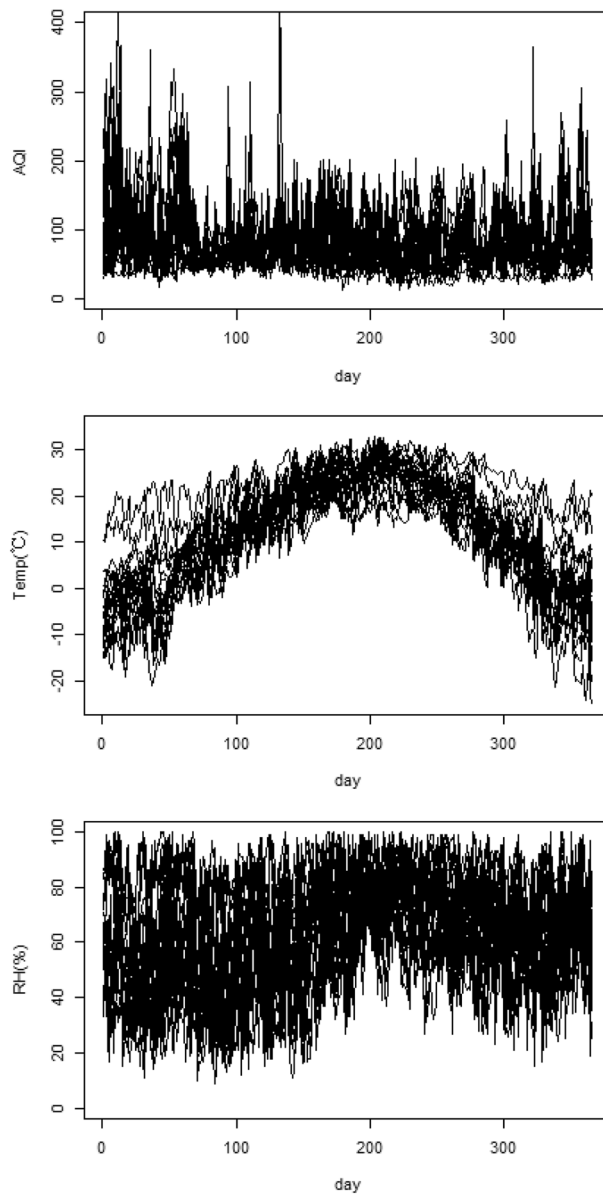
Figure 4. Daily AQI, temperature and RH.

|	Estimate	SE	t value	Pr(> |t|)	
$\hat{\gamma}_{\text{GDP per capita}}$	− 3.181e−06	1.659e−06	− 2.165	0.035	*
$\hat{\gamma}_{\text{Number of beds}}$	− 3.048e−01	4.718e−02	− 1.917	0.061	−

Table 5. Regression coefficient estimation and significance levels.

regression functions is proved. For the scalar predictors, we establish a distance between the estimated value and the true value, and prove the asymptotic property of the estimated regression coefficients. The model is applied to two examples: sleep quality study and mortality rate study, and the research results clearly show that the predictors in this model explain the responses well and reveal the influence of predictors on response.

In sleep quality research, we find that Efficiency and WASO are positively correlated with sleep quality, and Number is negatively correlated with sleep quality. We also find that heart rate is positively correlated with sleep quality between 8:00 and 11:00. On the contrast, heart rate is mainly negatively correlated with sleep quality at other time. In mortality rate research, we conclude that GDP per capita and the number of beds per thousand people are negatively correlated with mortality rate. AQI is positively correlated with mortality in winter, early
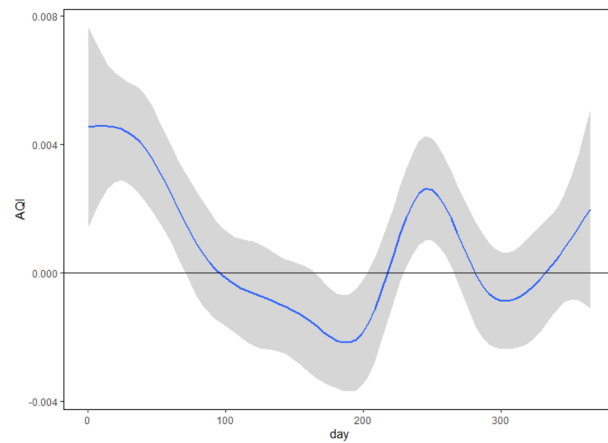
(a) AQI
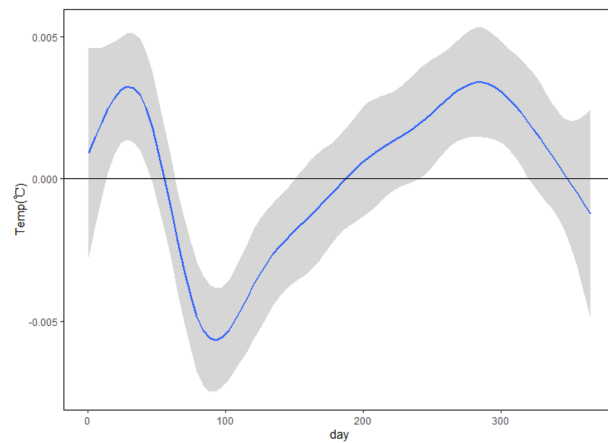
(b) Temperature
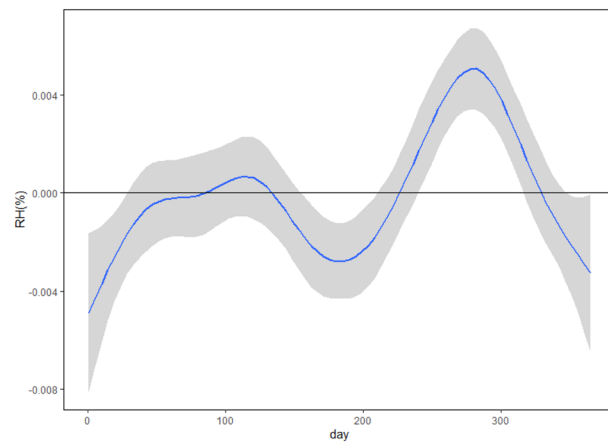
(c) Relative humidity

Figure 5. Regression coefficient function $\hat{\beta}(t)$ and its 95% confidence band.

spring and fall. Temperature is negatively correlated with mortality from March to May (in spring). Relative humidity is positively correlated with mortality from August to November (in autumn).

The application of generalized partially functional linear model has been further extended, which lays a foundation for further research on the generalized partial-function linear model of unknown link function and variance function, predictors with interactions and variable selection with high-dimensional predictors.

References

1. Amini, M. & Roozbeh, M. Optimal partial ridge estimation in restricted semiparametric regression models. *J. Multivar. Anal.* **136**, 26–40 (2015).
2. Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R. & Kupfer, D. J. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res.* **28**(2), 193–213 (1989).
3. Cai, T. T. & Hall, P. Prediction in functional linear regression. *Ann. Stat.* **34**(5), 2159–2179 (2006).
4. Capra, W. B. & Müller, H. G. An accelerated time model for response curves. *Am. Stat. Assoc.* **92**(437), 72–83 (1997).
5. Cardot, H., Ferraty, F. & Sarda, P. Spline estimators for the functional linear model. *Stat. Sinica* **13**(3), 571–591 (2003).
6. Chiou, J. M. & Müller, H. G. Quasi-likelihood regression with unknown link and variance functions. *Am. Stat. Assoc.* **93**(444), 1376–1387 (1998).
7. Conway, J. B. *A Course in Functional Analysis* 2nd edn. (Springer, 1990).
8. Crainiceanu, C., María, Staicu Ana & Di, Chong-Zhi. Generalized multilevel functional regression. *J. Am. Stat. Assoc.* **104**(488), 1550–1561 (2009).
9. Delaigle, A. & Hall, P. Methodology and theory for partial least squares applied to functional data. *Ann. Stat.* **40**(1), 322–352 (2012).
10. Ferraty, F. & Vieu, P. *Nonparametric Functional Data Analysis* (Springer, 2006).
11. Goldsmith, J., Crainiceanu, C. M., Caffo, B. & Reich, D. Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **61**(3), 453–469 (2012).
12. Hall, P. & Horowitz, J. M. Methodology and convergence rates for functional linear regression. *Ann. Stat.* **35**(1), 70–91 (2007).
13. Hall, P. & Yang, Y. J. Ordering and selecting components in multivariate or functional data linear prediction. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **72**(1), 93–110 (2010).
14. Horvath, L. & Kokoszka, P. *Inference for Functional Data with Applications* (Springer, 2012).
15. Hsing, T. & Eubank, R. L. *Theoretical Foundations of Functional Data Analysis with an Introduction to Linear Operators* (Wiley, 2015).
16. Ieva, F. & Paganoni, A. M. Risk prediction for myocardial infarction via generalized functional regression models. *Stat Methods Med. Res.* **25**(4), 1648–1660 (2016).
17. Kadoya, M. *et al.* Serum macro TSH level is associated with sleep quality in patients with cardiovascular risks-HSCAA study. *Sci. Rep.* **7**, 44387 (2017).
18. Kong, D. H., Xue, K. L., Yao, F. & Zhang, H. H. Partially functional linear regression in high dimensions. *Biometrika* **139**(10), 147–159 (2016).
19. Kou, J & Liu, Y. Wavelet regression estimations with strong mixing data. *Stat. Methods Appl.* **27**, 667–688 (2018).
20. McCullagh, P. Quasi-likelihood functions. *Ann. Stat.* **11**(1), 59–67 (1983).
21. Müller, H. G. & StadtMüller, U. Generalized functional linear models. *Ann. Stat.* **33**(2), 774–805 (2005).
22. Nelder, J. A. & Wedderbun, R. W. M. Generalized linear models. *J. R. Stat. Soc. Ser. A (General)* **135**(3), 370–384 (1972).
23. Ramsay, J. O. When the data are functions. *Psychometriha* **47**(4), 379–396 (1982).
24. Ramsay, J. O. & Dalzell, C. J. Some tools for functional data analysis. *J. R. Stat. Soc. Ser. B (Methodological)* **53**(3), 539–572 (1991).
25. Ramsay, J. O. & Silverman, B. W. *Functional Data Analysis* 2nd edn. (Springer, 2005).
26. Rice, J. A. & Silverman, B. W. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Stat. Soc. Ser. B (Methodological)* **53**(1), 233–243 (1991).
27. Roozbeh, M., Hesamian, G. & Akbari, M. G. Ridge estimation in semi-parametric regression models under the stochastic restriction and correlated elliptically contoured errors. *J. Comput. Appl. Math.* **378**, 112940 (2020).
28. Roozbeh, M., Arashi, M. & Hamzah, N. A. Generalized cross-validation for simultaneous optimization of tuning parameters in ridge regression. *Iran. J. Sci. Technol. Trans. A Sci.* **44**, 473–485 (2020).
29. Roozbeh, M. Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion. *Comput. Stat. Data Anal.* **117**, 45–61 (2018).
30. Sajjadieh, A. *et al.* The association of sleep duration and quality with heart rate variability and blood pressure. *Tanaffos* **19**(2), 135–143 (2020).
31. Shin, H. Partial functional linear regression. *J. Stat. Plan. Inference* **139**(10), 3405–3418 (2009).
32. Shin, H. & Lee, M. H. On prediction rate in partial functional linear regression. *J. Multivar. Anal.* **103**(1), 93–106 (2012).
33. Wang, G. C., Feng, X. N. & Chen, M. Functional partial linear single-index model. *Scand. J. stat. Theory Appl.* **43**(1), 261–274 (2015).
34. World Health Organization. World health statistic 2019: monitoring health for the SDGs, sustainable development goals, 2019.
35. Yao, F., Müller, H. G. & Wang, J. L. Functional linear regression analysis for longitudinal data. *Ann. Stat.* **33**(6), 2873–2903 (2005).
36. Zhang, Y. S. *et al.* Prevalence and socio-demographic correlates of poor sleep quality among older adults in Hebei province, China. *Sci. Rep.* **10**, 12266 (2020).

Acknowledgements

Author contributions

Statement: 1. all methods were carried out in accordance with relevant guidelines and regulations. 2. all experimental protocols were approved by a named institutional and licensing committee. W.X. and H.L. conceived the idea. Y.W. did the statistical inference, analysed the data, carried out the simulation and wrote the first version of the manuscript. All authors contributed to developing the statistical models and inference, designing simulations, reviewing the literature, writing, reading and revising this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-02896-7.

Correspondence and requests for materials should be addressed to Y.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.