STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:1552–1566
Published online 30 June 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2609

Functional regression analysis using an *F* test for longitudinal data with large numbers of repeated measures

Xiaowei Yang[1,2,*,†], Qing Shen[3], Hongquan Xu[4] and Steven Shoptaw[5]

[1]*Department of Public Health Sciences, Division of Biostatistics, University of California, Davis, CA 95616, U.S.A.*
[2]*BayesSoft Inc., 2221 Caravaggio Drive, Davis, CA 95616, U.S.A.*
[3]*Edmunds.com Inc., 2401 Colorado Ave., Suite 250, Santa Monica, CA 90404, U.S.A.*
[4]*Department of Statistics, University of California, Los Angeles, CA 90095-1554, U.S.A.*
[5]*UCLA-Integrated Substance Abuse Programs, 11075 Santa Monica Blvd, Suite 200, Los Angeles, CA 90025, U.S.A.*

SUMMARY

Longitudinal data sets from certain fields of biomedical research often consist of several variables repeatedly measured on each subject yielding a large number of observations. This characteristic complicates the use of traditional longitudinal modelling strategies, which were primarily developed for studies with a relatively small number of repeated measures per subject. An innovative way to model such 'wide' data is to apply functional regression analysis, an emerging statistical approach in which observations of the same subject are viewed as a sample from a functional space. Shen and Faraway introduced an *F* test for linear models with functional responses. This paper illustrates how to apply this *F* test and functional regression analysis to the setting of longitudinal data. A smoking cessation study for methadone-maintained tobacco smokers is analysed for demonstration. In estimating the treatment effects, the functional regression analysis provides meaningful clinical interpretations, and the functional *F* test provides consistent results supported by a mixed-effects linear regression model. A simulation study is also conducted under the condition of the smoking data to investigate the statistical power for the *F* test, Wilks' likelihood ratio test, and the linear mixed-effects model using AIC. Copyright 2006 John Wiley & Sons, Ltd.

KEY WORDS: functional *F* test; functional data analysis; functional regression analysis; longitudinal data analysis

*Correspondence to: Xiaowei Yang, Department of Public Health Sciences, Division of Biostatistics, Med Sci 1-C, University of California, Davis, CA 95616, U.S.A.
†E-mail: xdyang@ucdavis.edu

Contract/grant sponsor: National Institute of Drug Abuse; contract/grant numbers: N44 DA35513, R03 DA016721 and P50 DA 18185

Received 1 August 2005
Accepted 28 April 2006

1. INTRODUCTION

In biomedical research with longitudinal studies, subjects are repeatedly measured for a set of characteristics so that time-varying relationships between the responses and explanatory variables of interest can be modelled, e.g. growth trajectory and disease progression [1]. In certain fields of study, such as substance abuse, environmental, and public health research, repeated measures are sometimes collected at high frequencies over long periods of time. For example, in a 12-week smoking cessation study, carbon monoxide levels were collected three times weekly on each methadone-maintained tobacco smoker [2]. To analyse such longitudinal data with large-scale time grids, it may be unsatisfactory to apply traditional longitudinal modelling strategies (e.g. mixed-effects models, marginal models, and transition models) [3], which are mainly developed for data with a relatively small number of repeated measures per subject [4]. More advanced models such as nonlinear mixed effects models with smoothing schemes (e.g. kernel or spline methods) can be used [5], but the computation cost is considerable and the clinical interpretation is vague. Other multivariate-observation approaches, such as hierarchical models, latent variable models, and structure equation models, sometimes involve many parameters with unverifiable assumptions [6–8]. As of yet, there are not many alternatives that successfully address the unique problems presented by data collected in longitudinal studies with high dimensionality. This paper evaluates a recently developed method of functional data analysis for this purpose.

In the emerging statistical research field, functional data analysis refers to a collection of strategies for analysing functional data sets, such as curves, images, or shapes [9]. To a study observing seated automobile drivers' body motion patterns [10, 11], and to a study of urinary metabolites and a progesterone data set [12], several strategies of functional regression analysis have been applied.

Until very recently, functional data analysis and longitudinal data analysis have been viewed as distinct enterprises [13]. In the 2004 emerging issues of *Statistica Sinica* [4], it is seen that endeavour has been made to reconciling the two lines of methodology. For longitudinal data with dense time grids, one could conceive within-subject repeated measures as discrete samples from a functional curve over the studied time interval. A curve for each subject's response can be obtained via various smoothing techniques in connecting the discrete data points [14] and these individual subject response curves can be tested using functional data analysis. The approach to using functional data analysis provides an alternative with innovative insights to the practice of longitudinal data analysis. Unlike the long-form of representing longitudinal data in some computer procedures (e.g. PROC Mixed in SAS), where within-subject repeated measures are concatenated into one long vector, functional regression analysis does not change the original rectangular form of the data structure, which looks more natural to data analysts. With time-dependent coefficients, functional regression analysis captures the time-varying exposure–response relationship, thus providing a simpler data structure with intuitive interpretations. A time series plot of the estimated coefficient function vividly reveals how the effect of a predictor can change along the time axis. Most importantly, functional regression analysis could draw more robust conclusions as it has features similar to nonparametric methods, requiring fewer assumptions on the intra-subject error correlation and mean structures for the studied population [11, 15].

2. FUNCTIONAL LINEAR REGRESSION MODELS

A longitudinal study, usually collects continuous repeated measures, $\{y_i(t_{ij});\ \ i=1,\ldots,n,\ j=1,\ldots,m\}$, on a time grid, $\{t_1,\ldots,t_m\}$, that is either exactly or approximately the same for

DOI: 10.1002/sim

all n subjects. One may restrict that the same number of repeated measures be collected on each subject. Ideally, these repeated measures can be viewed as discrete samples from a continuous response curve, $y_i(t)$. In this setting, a functional linear regression model has the form of

$$y_i(t) = x_i^{\mathrm{T}}\beta(t) + \varepsilon_i(t)$$

where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is a vector of fixed covariates or predictor variables, $\beta(t) = (\beta_1(t), \ldots, \beta_p(t))^{\mathrm{T}}$ is a vector of coefficient functions, and $\varepsilon_i(t)$ is an error function of Gaussian process with mean zero and unknown covariance function $r(s,t) = \mathrm{cov}(\varepsilon_i(s), \varepsilon_i(t))$. Since $\beta(t)$ is a function of time, this model is sometimes referred as *varying-coefficient* regression model [16]. In more general settings, x_i may be also time-varying, although we only deal with the case of time-independent covariates in this paper. It is also assumed that $\varepsilon_i(t)$ and $\varepsilon_k(t)$ are independent of each other when $i \neq k$ (i.e., observations on different subjects are independent of each other).

The coefficient function $\beta(t)$ can be estimated by the least squares method, which leads to

$$\hat{\beta}(t) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y(t)$$

where $X = (x_1, \ldots, x_n)^{\mathrm{T}}$ is the model matrix and $Y(t) = (y_1(t), \ldots, y_n(t))^{\mathrm{T}}$ is the vector of response functions. The predicted (or fitted) responses are $\hat{y}_i(t) = x_i^{\mathrm{T}}\hat{\beta}(t)$ and the residuals are $\hat{\varepsilon}_i(t) = y_i(t) - \hat{y}_i(t)$. The residual sum of squares is $\mathrm{rss} = \sum_{i=1}^{n} \int (y_i(t) - \hat{y}_i(t))^2 \, \mathrm{d}t$.

In reality, only a finite number of measures (i.e. $y_i(t_{ij})$'s) exist for the ith response curve (i.e. $y_i(t)$). To apply functional regression analysis to discrete observational data, Shen and Faraway [11] recommended analysing the un-smoothed raw data directly over a common grid of time for different subjects. For a data set with unbalanced design, one may reconstruct the response curve from the observed data points to get estimates of $y_i(t)$ over a common grid $\{t_j;\ \ j = 1, \ldots, m\}$ via proper smoothing techniques, e.g. model-based cross-validation methods [14], kernel-based or spline-based nonparametric regression methods [17], and robust methods such as LOWESS [18]. The choice of different smoothing techniques usually has little impact on the analysis if there are plentiful underlying response curves (i.e. $y_i(t)$'s) with fairly smooth functional forms [11].

2.1. A functional F test for hypothesis testing and model selection

An important inference problem is to compare two nested linear models, ω and Ω, where $\dim(\omega) = q$, $\dim(\Omega) = p$, and model ω results from a linear restriction on the parameters of model Ω. There are relatively few satisfactory solutions available in the statistical literature to this situation. A naive approach is to examine the point-wise F statistics on each time point for testing $\beta(t)$. This method carries a serious problem with multiple-comparison and if Bonferroni correction were applied to the significance level, power would be significantly compromised considering that repeated measures are often strongly correlated. Ramsay and Silverman [9] and Faraway [10] proposed permutation- and bootstrap-based tests, which require intensive computation. As pointed out by Faraway [10], traditional multivariate test statistics such as Wilks' lambda likelihood ratio [19] are inappropriate due to the influence of unimportant variation directions.

To overcome these issues, Shen and Faraway [11] proposed a functional F test. Define

$$F = \frac{(\mathrm{rss}_\omega - \mathrm{rss}_\Omega)/(p-q)}{\mathrm{rss}_\Omega/(n-p)}$$

where rss_ω and rss_Ω are residual sum of squares under models ω and Ω, respectively. The null distribution of this statistic is $((n-p)/(p-q))\sum_{k=1}^{\infty} r_k\chi^2_{(p-q)}/\sum_{k=1}^{\infty} r_k\chi^2_{(n-p)}$, where $r_1 \geqslant r_2 \geqslant \cdots \geqslant 0$

DOI: 10.1002/sim

are eigenvalues of the covariance function $r(s,t)$ and all the χ^2 random variables are independent of each other. This null distribution can be effectively approximated by an ordinary F distribution with degrees of freedom $\mathrm{df}_1 = \lambda(p-q)$ and $\mathrm{df}_2 = \lambda(n-p)$, where $\lambda = \left(\sum_{k=1}^{\infty} r_k\right)^2 / \sum_{k=1}^{\infty} r_k^2$ is the *degrees-of-freedom-adjustment-factor*.

In practice, when repeated measures are observed on an evenly spaced time grid $\{t_1, \ldots, t_m\}$, we should replace the integration with summation, compute $\mathrm{rss} = \sum_{i=1}^{n} \sum_{k=1}^{m} (y_i(t_k) - \hat{y}_i(t_k))^2 / m$ and estimate the degrees-of-freedom-adjustment-factor by $\mathrm{trace}(E)^2/\mathrm{trace}(E^2)$, where $E = \hat{\Sigma}^{\Omega}$ is the empirical covariance matrix computed from the alternative model.

It is important to note that the functional F test works well even when the grid size m is larger than the sample size n, while most multivariate test statistics [20, 21] would fail. Other important work addressing the functional testing problem was provided by Fan and Lin [22], Eubank [23], and Abramovich *et al*. [24], but they only considered ANOVA-type models and their test statistics were formed by orthogonal (Fourier or Wavelets) expansion coefficients of response curves. Eubank [23] proved that among different ways of combining the coefficients into a test statistic, the L^2 norm, a simple sum of the squared coefficients, is asymptotically equivalent to the uniformly most powerful test when the grid size m goes to infinity. This result provides important evidence that the functional F-test statistic, which uses L^2 norm of the residual curves, is not only computationally cheaper but also more powerful than other methods.

Model selection is an important issue in regression analysis. Stepwise model selection requires an easy way of calibrating the p-value of a predictor in the full model, i.e. to test the null hypothesis '$H_{0j} : \beta_j(t) = 0$ for $j = 1, \ldots, p$' against the full model hypothesis '$H_1 : Y(t) = X\beta(t) + \varepsilon(t)$'. To test these hypotheses, one can fit each null model H_{0j} separately for $j = 1, \ldots, p$, and then use functional F statistics $F_j = (\mathrm{rss}_{0j} - \mathrm{rss}_1)/(\mathrm{rss}_1/(n-p))$ to make a decision on accepting or rejecting the null model. As shown by Shen and Faraway [11], it is indeed unnecessary to fit all the p null models, because F_j can be derived from quantities obtained directly from the fitting of the full model H_1, i.e.

$$F_j = \frac{(n-p)\int \hat{\beta}_j^2(t)\,\mathrm{d}t}{(X^{\mathrm{T}}X)^{-1}_{jj}\,\mathrm{rss}_1}$$

where $(X^{\mathrm{T}}X)^{-1}_{jj}$ denotes the jth diagonal element of $(X^{\mathrm{T}}X)^{-1}$, $\hat{\beta}_j(t)$ is the estimate of $\beta_j(t)$, and rss_1 is the residual sum of squares under the full model H_1. In practice, the operation of integration is replaced by that of summation. The null distribution of the functional F statistic F_j can be approximated by an ordinary F distribution with degrees of freedom $\mathrm{df}_1 = \lambda$ and $\mathrm{df}_2 = \lambda(n-p)$, where λ is the degrees-of-freedom-adjustment-factor.

2.2. Diagnostic check

It is important to identify outliers and highly influential curves (subjects) since including them in the analysis may give misleading results. As in the context of traditional linear regression for scalar responses, we define jackknife residuals and Cook's distances for functional regression. Let $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ be the hat matrix and define leverage h_{ii} as the diagonal entry of H. Define studentized residual as

$$S_i = \frac{\sqrt{\int \hat{\varepsilon}_i^2(t)\,\mathrm{d}t}}{\sqrt{(1-h_{ii})\mathrm{rss}/(n-p)}}$$

DOI: 10.1002/sim

and jackknife residual as

$$J_i = \frac{\sqrt{\int \hat{\varepsilon}^2_{(i)}(t)\,\mathrm{d}t}}{\sqrt{[1 + x_i^{\mathrm{T}}(X_{(i)}^{\mathrm{T}} X_{(i)})^{-1} x_i][\mathrm{rss}_{(i)}/(n-p-1)]}}$$

where $X_{(i)}$ is the X matrix with the ith row deleted, $\hat{\varepsilon}^2_{(i)}(t)$ is the ith residual from the model without the ith curve, and $\mathrm{rss}_{(i)}$ is the residual sum of squares from the model without the ith curve. Define Cook's distance as

$$D_i = \frac{\int (\hat{\beta}_{(i)}(t) - \hat{\beta}(t))^{\mathrm{T}} (X^{\mathrm{T}} X)(\hat{\beta}_{(i)}(t) - \hat{\beta}(t))\,\mathrm{d}t}{\mathrm{rss}} \cdot \frac{n-p}{p}$$

where $\hat{\beta}_{(i)}(t)$ is the estimate of $\beta(t)$ computed without the ith curve.

Shen and Xu [25] showed that jackknife residuals and Cook's distances can be computed directly from the studentized residuals and leverages as follows:

$$J_i = S_i \sqrt{\frac{n-p-1}{n-p-S_i^2}} \quad \text{and} \quad D_i = \frac{S_i^2}{p} \cdot \frac{h_{ii}}{1-h_{ii}}$$

These formulas provide efficient computations by avoiding fitting n regression models with each curve deleted. Shen and Xu [25] also showed that J_i^2 has a functional F distribution, which can be approximated by an ordinary F distribution with degrees of freedom $\mathrm{df}_1 = \lambda$ and $\mathrm{df}_2 = \lambda(n-p-1)$ if the ith curve is not an outlier. Thus, we can use the jackknife residual and F test to formally detect outliers.

3. APPLICATION TO A SMOKING CESSATION CLINICAL TRIAL

3.1. Background of the study, data exploration, and preliminary analysis

A 12-week clinical trial was performed to evaluate relapse prevention (RP) and contingency management (CM) as smoking cessation therapies for methadone-maintained tobacco smokers [2]. A total of 174 subjects were randomly assigned to one of four treatment conditions (Control; RP-only; CM-only; RP+CM). All subjects received nicotine replacement therapy in addition to their assignment to behavioural therapies: RP and/or CM. The repeated measures of most interest in this study were breath samples collected three times per week (i.e. $m = 36$), which were analysed for carbon monoxide levels (parts per million) to indicate recent tobacco smoking abstinence. The observed carbon monoxide levels in log-scale and their mean profiles for each group are depicted in Figure 1. The plots are sometimes called spaghetti plots where the light shaded background trajectories depicts the connected carbon monoxide levels for each subject. It is seen that the mean levels remain fairly stable across time for each group, while large variances are notable between subjects. This suggests that subject-related random effects are necessary to describe the heterogeneity among the smokers. Participants' age (Age), baseline carbon monoxide levels (BaseCO), and numbers of nicotine patches (Patches) were recorded as other predictors along with treatment conditions.

DOI: 10.1002/sim

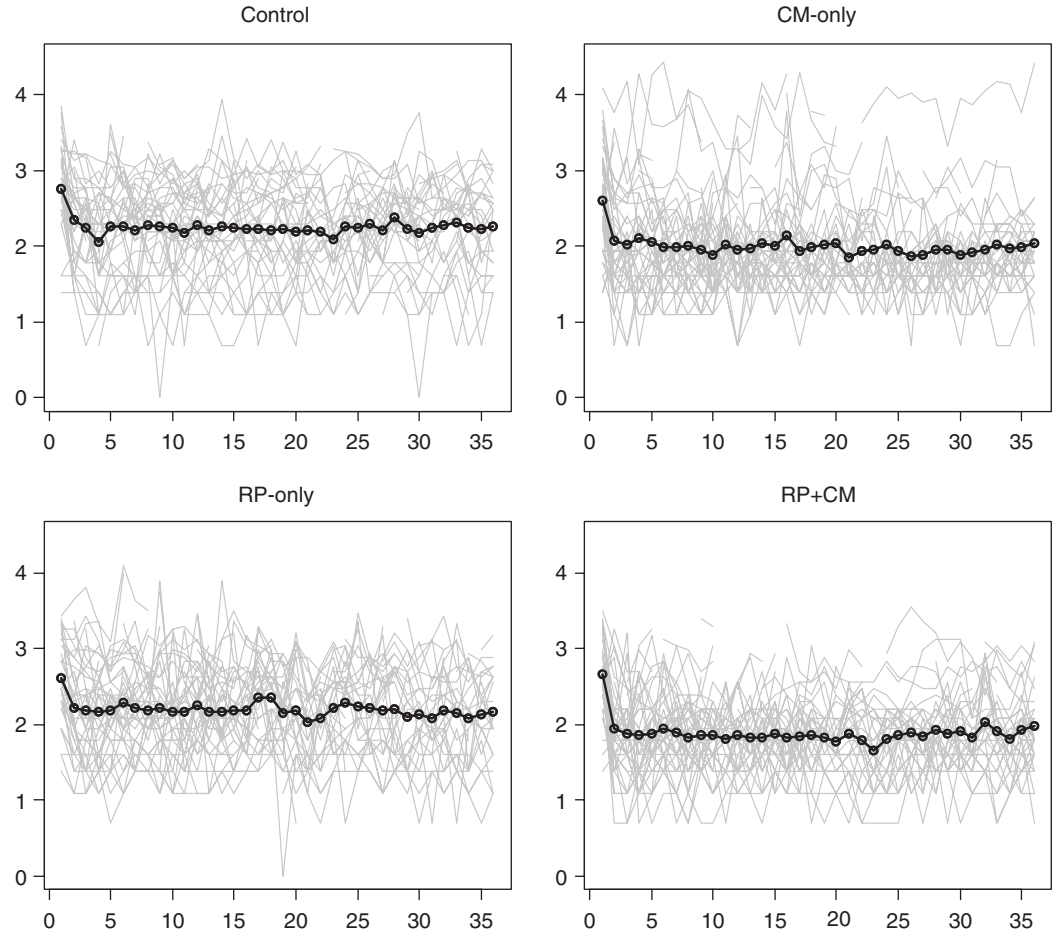

Figure 1. Mean levels of the carbon monoxide across the treatment groups. For each plot, the y-axis indicates $\log(1+y)$ transform of the original level of carbon monoxide (p.p.m.), the x-axis indicates number of clinic visit for study participants $(1, \ldots, 36)$. Both individual profiles and the mean profile are plotted for each of the four treatment conditions: Control, RP-only, CM-only, and RP+CM (RP, relapse prevention; CM, contingency management).

For significance testing, an insufficient approach was first applied to compare the carbon monoxide levels across treatment conditions on any given time point using the naive point-wise method. As depicted by Figure 2, at eight points significantly different carbon monoxide levels were indicated by the point-wise ANOVA with p-values smaller than 0.001. Because of the problem of *multiple comparison* [26], a significance level of 0.001 was used instead of the usual level of 0.05. Although this method provides some useful insights for exploratory purposes, it is relatively limited in making inferences on the overall treatment efficacy, because there is no simple way of combining these multiple p-values. Moreover, the point-wise ANOVA ignored the patterns showing that the average carbon monoxide levels were almost consistently lower for the treatment conditions involving CM.

DOI: 10.1002/sim

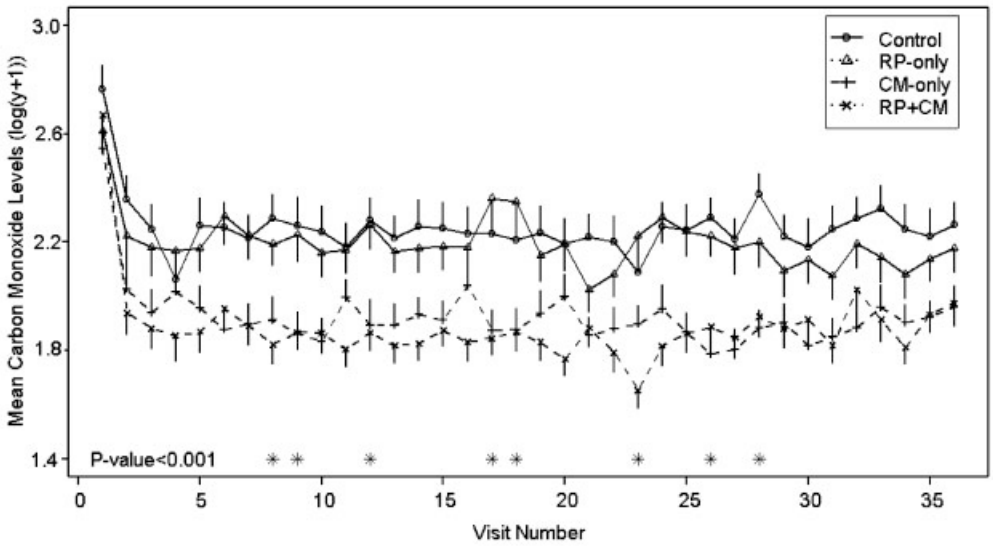

Figure 2. The average and standard deviation (SD) curves for the log-scaled carbon monoxide levels. On this plot, the four mean curves of the log-scaled carbon monoxide levels and the corresponding point-wise standard errors are drawn for each of the four treatment conditions: Control, RP-only, CM-only, and RP+CM (RP, relapse prevention; CM, contingency management). Vertical bars indicate the estimated standard errors of average carbon monoxide levels. The stars ('*') over the x-axis mark the time points (i.e. visit numbers) where the carbon monoxide levels are significantly different indicated by a point-wise ANOVA (p-value<0.001). y-axis indicates values of carbon monoxide levels after $\log(1+y)$ transform. x-axis represents number of clinic visit for study participants $(1, \ldots, 36)$.

In the original data, about 20% of the carbon monoxide levels were missing due to either occasional omission or premature withdrawal. To solve this problem, the method of multiple imputation [27] was applied. After the logarithmic transformation, repeated carbon monoxide levels for each participant could be viewed as multivariate normally distributed (i.e. $y_i \sim \mathrm{N}(\mu, \Sigma)$). Specifying a normal prior distribution for the mean vector (i.e. $\mu|\Sigma \sim \mathrm{N}(\mu_0, \tau^{-1}\Sigma)$) and an inverted Wishart distribution for the covariance matrix (i.e. $\Sigma \sim W^{-1}(r, \Lambda)$), we conducted multiple imputation using an R package named *norm* which implemented the iterative algorithm called data augmentation [28]. This algorithm consists of two steps per iteration. In the *imputation step*, for each person, we drew imputations of missing values conditionally on the observed values using a conditional normal distribution with parameters drawn in the previous iteration. In the *proposing step*, new parameters (μ, Σ) were proposed, given the complete data with current imputed values. Since no prior information was available, Jeffery's invariance principal was used to derive the non-informative form for the normal-inverse-Wishart prior distribution, i.e. $p(\mu, \Sigma) \propto |\Sigma|^{-(m+0.5)}$. The EM algorithm, a sub-function of the *norm* package, was first run to obtain the maximum likelihood estimates (i.e. $\hat{\mu}$, $\hat{\Sigma}$) as the starting point to initiate the data augmentation procedure. Various diagnostic tools suggested that the procedure converged within 200 iterations. Continuing the procedure with 2000 additional iterates, one set of imputed missing values was recorded after each 500 iterates, yielding totally four complete data sets.

DOI: 10.1002/sim

3.2. Functional regression analysis

For each of the above imputed data sets, a functional regression model, including all the interesting predictors, was fitted using the method of least squares estimation,

$$y(t) = \beta_0(t) + \mathrm{CM} \cdot \beta_1(t) + \mathrm{RP} \cdot \beta_2(t) + \mathrm{CM} * \mathrm{RP} \cdot \beta_3(t)$$
$$+ \mathrm{BaseCO} \cdot \beta_4(t) + \mathrm{Age} \cdot \beta_5(t) + \mathrm{Patches} \cdot \beta_6(t) + \varepsilon(t)$$

where CM = 1 (or 0) indicates whether a subject received CM (or not), RP = 1 (or 0) indicates whether a subject received RP (or not), and CM ∗ RP is an interaction term. In this coding scheme, the control group was coded as 'CM = 0 and RP = 0', and the RP+CM groups was coded by 'CM = 1 and RP = 1'. Since there was little difference between the four imputed data sets, the estimated coefficient functions were plotted in Figure 3 for the first imputed data set. Note that these functions are point-wise estimations and not smoothed. For the purpose of interpretation, one may consider smoothing the estimates. However, our purpose is mainly on model selection and the functional F test does not involve smoothing; therefore, we present the unsmoothed point-wise estimates. The fitted coefficient functions of RP and Age are close to the zero function, indicating that the treatment effect of the RP and the age effect are negligible. Further, the interaction term CM ∗ RP is not significant, indicating that CM does not interact with RP. Regression coefficient functions for CM and Patches are negative-valued throughout, suggesting favourable effects of CM and nicotine patch replacement. By contrast, the positive-valued coefficient function of the baseline carbon monoxide level implied that the higher the baseline carbon monoxide level, the more difficult to achieve tobacco abstinence.

The functional F-test statistics and their p-values of each predictor in this model are listed in Table I. For all four complete data sets, only the terms, CM, BaseCO, and Patches look significant using significance level $\alpha = 0.05$. After removing insignificant terms (RP, CM ∗ RP, and Age), the reduced model was fitted to the imputed data sets. The functional F-test statistics and their p-values for the remaining terms are listed in Table II. As expected, all predictors were significant at $\alpha = 0.01$ level this time. Since all the four data sets consistently supported the same results, we accept this three-predictor functional regression model as the final model to make inferences:

$$y(t) = \beta_0(t) + \mathrm{CM} \cdot \beta_1(t) + \mathrm{BaseCO} \cdot \beta_2(t) + \mathrm{Patches} \cdot \beta_3(t) + \varepsilon(t)$$

where the subscript indicating subjects is again suppressed. The fitting of this model indicated that, after adjusting out the effects of baseline levels (BaseCO) and number of nicotine patches applied (Patches), CM turned out to be significantly effective in helping this specific group of smokers achieve tobacco abstinence during treatment.

To check diagnostics for the above-selected model, jackknife residuals and Cook's distances for all the imputed data sets were computed. The charts of these statistics from the first imputed data set are shown in Figure 4. The jackknife residuals for the participants numbered 92 and 93 are bigger than the critical value (with Bonferoni adjustment) of the functional F distribution at significance level of $\alpha = 0.05$. Therefore, these two smokers may be declared as outliers. The record associated with the subject numbered 92 is also a highly influential point according to the Cook's distance. Checking the original records, both points with unusually high values for most of the observations were noted. After excluding these two 'outliers', we re-analysed the data using the above models and found consistent results.

DOI: 10.1002/sim

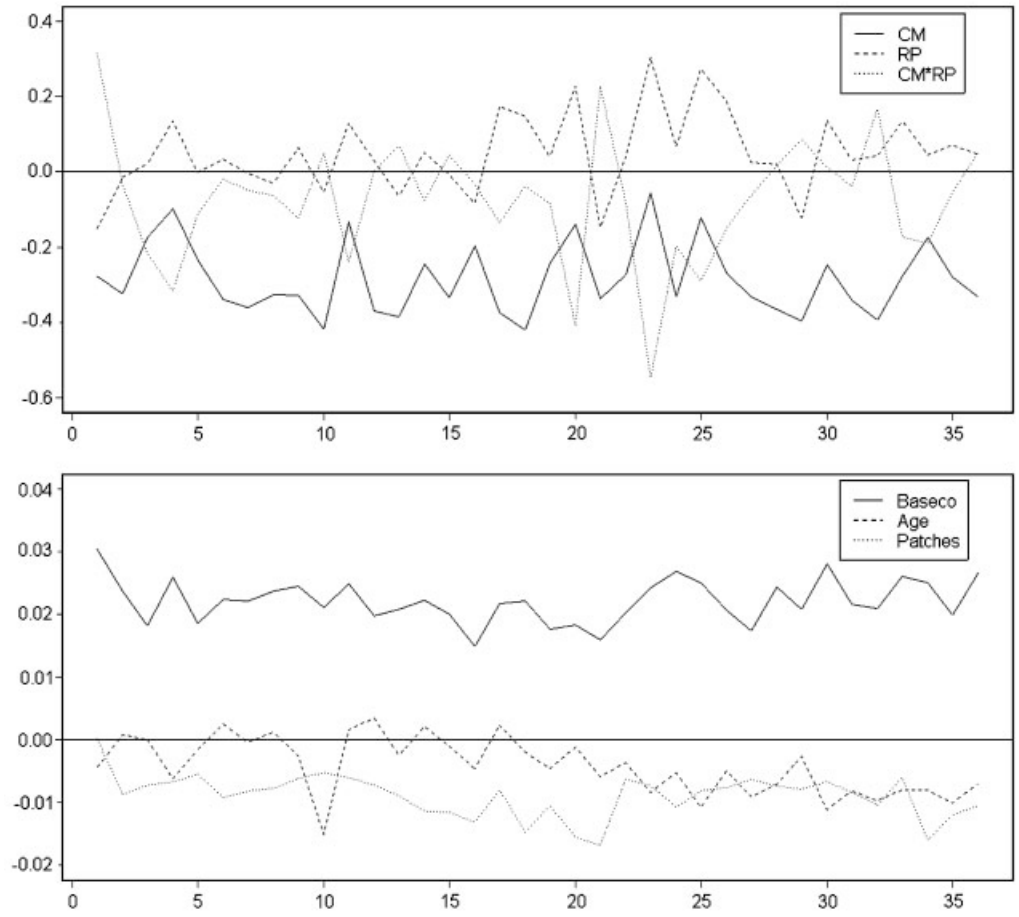

Figure 3. Estimated regression coefficient functions in functional regression analysis for the first imputed data set. The top panel shows the regression coefficient functions corresponding to effects of CM treatment, RP treatment and their interaction (CM $*$ RP); the bottom panel depicts the regression coefficient functions corresponding to baseline carbon monoxide level (BaseCO), smoker' age (Age), and number of nicotine patches a smoker has received during the study (Patches). y-axis indicates values of regression coefficients and x-axis indicates number of clinic visit for each smoker $(1, \ldots, 36)$.

Table I. Observed functional F test statistics (and p-values) for each covariate.

Data set	Intercept	CM	RP	CM $*$ RP	BaseCO	Age	Patches
Impute 1	98.9(*)	5.98(*)	0.89(0.45)	1.11(0.34)	24.8(*)	1.25(0.29)	24.9(*)
Impute 2	98.0(*)	4.89(*)	0.78(0.51)	1.17(0.32)	24.2(*)	1.83(0.14)	21.6(*)
Impute 3	104.5(*)	5.99(*)	0.71(0.54)	1.07(0.36)	26.1(*)	1.18(0.32)	30.9(*)
Impute 4	96.2(*)	5.01(*)	0.91(0.43)	1.29(0.28)	25.7(*)	1.05(0.37)	24.8(*)

* p-values are smaller than 0.01.

DOI: 10.1002/sim

Table II. Functional F-test statistics for each covariate in the final functional regression model.

Data set	Intercept	CM	BaseCO	Patches
Impute 1	254.6	14.71	25.1	27.3
Impute 2	239.3	13.75	24.7	24.0
Impute 3	272.0	15.35	26.6	33.4
Impute 4	250.7	14.04	26.3	26.8

All p-values are smaller than 0.01.

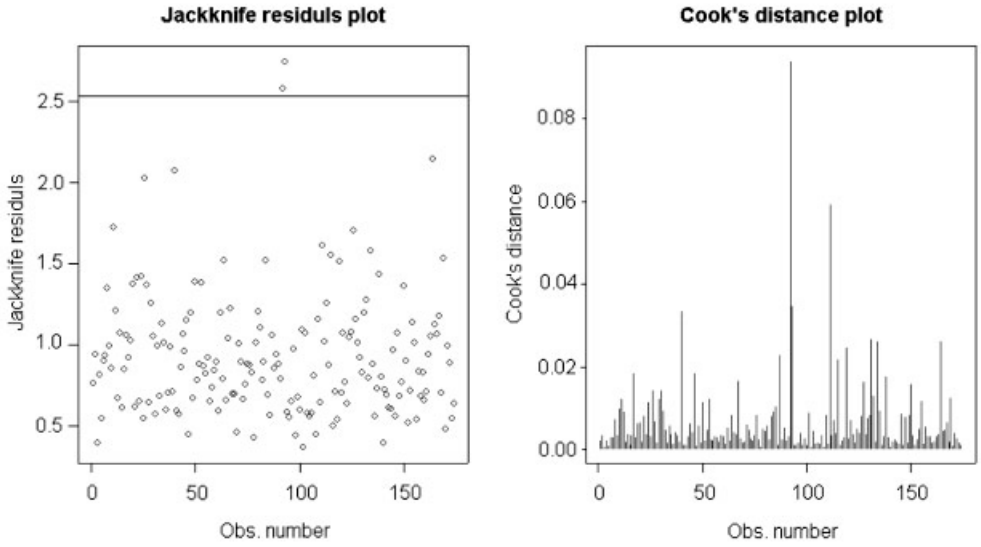

Figure 4. Diagnostics for the first imputed data set. The left panel draws jackknife residuals and the right panel depicts Cook's distances calculated from the functional regression model including three predictors: CM, Baseco, and Patches. In both plots, the x-axis corresponds to the labels of the 174 participants in the study. The y-axis corresponds to either the values of jackknife residuals or Cook's distances. The horizontal line on the jackknife residuals plot shows the critical value (with Bonferoni adjustment) of the functional F distribution at significance level of $\alpha = 0.05$. Two subjects (numbered 92 and 93) have jackknife residuals noticeable high and one subject (numbered 92) associates with the highest Cook's distance.

3.3. A random-intercept model

We also analysed the four complete data sets after imputation by a linear mixed effects model with random intercept to model heterogeneities across subjects:

$$y_{ij} = \beta_0 + \mathrm{CM} \cdot \beta_1 + \mathrm{RP} \cdot \beta_2 + \mathrm{CM} * \mathrm{RP} \cdot \beta_3 + \mathrm{BaseCO}_i \cdot \beta_4 + \mathrm{Age}_i \cdot \beta_5 + \mathrm{Patches}_i \cdot \beta_6 + u_i + \varepsilon_{ij}$$

DOI: 10.1002/sim

where y_{ij} stands for the jth carbon monoxide level of the ith smoker, CM, RP, RP$*$CM, BaseCO, Age, and Patches are fixed effects that are common for all observations on the same subject, $u_i \sim \mathrm{N}(0, \sigma_u^2)$ is the random intercept effect explaining the heterogeneity across subjects, and ε_{ij}'s are identically independently distributed normal random errors. Consistent conclusions were observed by fitting this linear mixed effects model: CM (p-value<0.01) is significant while RP and CM$*$RP are not. Additionally, Age (p-value$=0.43$) is not significant while BaseCO (p-value<0.01) and Patches (p-value<0.01) are significant.

3.4. Summary

As seen in this example, the scalar linear mixed effects model and the functional regression model differ in at least two ways. First, in the mixed effects model the fixed effects (i.e. β) are time independent, while in the functional regression model the effects (i.e. $\beta(t)$) are functions over time. Second, the random-intercept model implicitly assumes a compound symmetry error correlation structure within each smoker, while the functional regression model does not assume any specific forms on the intra-subject correlation structure. The time series plots of the estimated coefficient functions in Figure 3 for the functional regression model provide richer information with intuitive clinical interpretation than the point estimates of parameters in the mixed effects model. For example, there appears to be a slightly increasing negative effect of Patches over time. Although functional regression analysis and scalar linear mixed effects models supported no strong overall age effect, a negative influence of age on carbon monoxide levels (higher ages associate with lower carbon monoxide levels) was noticed starting from the eighth week. It appeared that older smokers stayed longer in the study, and the longer they stayed, the more likelihood they were to achieve smoking abstinence as measured using carbon monoxide levels. By applying both longitudinal and functional data analysis to the same set of data, the overall time-averaged treatment efficacy and the dynamic time-changing effects of treatment can be jointly targeted so that we may obtain multi-facet enhanced understanding of the studied phenomena.

4. SIMULATION STUDY

To further evaluate the performance of the functional F test, simulation studies were conducted under similar conditions to the smoking cessation data. Response carbon monoxide levels were simulated as the weighted average of predicted levels from the full functional regression model (Ω, i.e., the one with six predictors) and from the reduced model (ω, i.e., the one with three predictors) plus random errors from two covariance structures: compound symmetric (CS) and autoregressive type 1 (AR(1)) [1]. For the CS covariance structure, the correlation coefficient between y_{ij} and y_{ik} is the same (i.e., ρ for $j \neq k$), whereas for the AR(1) case, the correlation strength depends on the distance between the two observations (i.e., $\rho^{|j-k|}$). For both covariance structures, the residual variance was set as $\mathrm{Var}(y_{ij}) = \sigma^2 = 0.3$, a value similar to the empirical variance seen in the original data. The weights are varied between 0 (corresponding to the reduced model) and 1 (corresponding to the full model) with an increment of 0.1. For each weight, 1000 sets of data were generated.

Shen and Faraway [11] compared functional F test with multivariate likelihood ratio test (Wilks' lambda) and the B-spline multivariate test. Here we investigated the performance of functional F test in comparison with linear mixed-effects models and Wilks' lambda test within the Fourier

DOI: 10.1002/sim

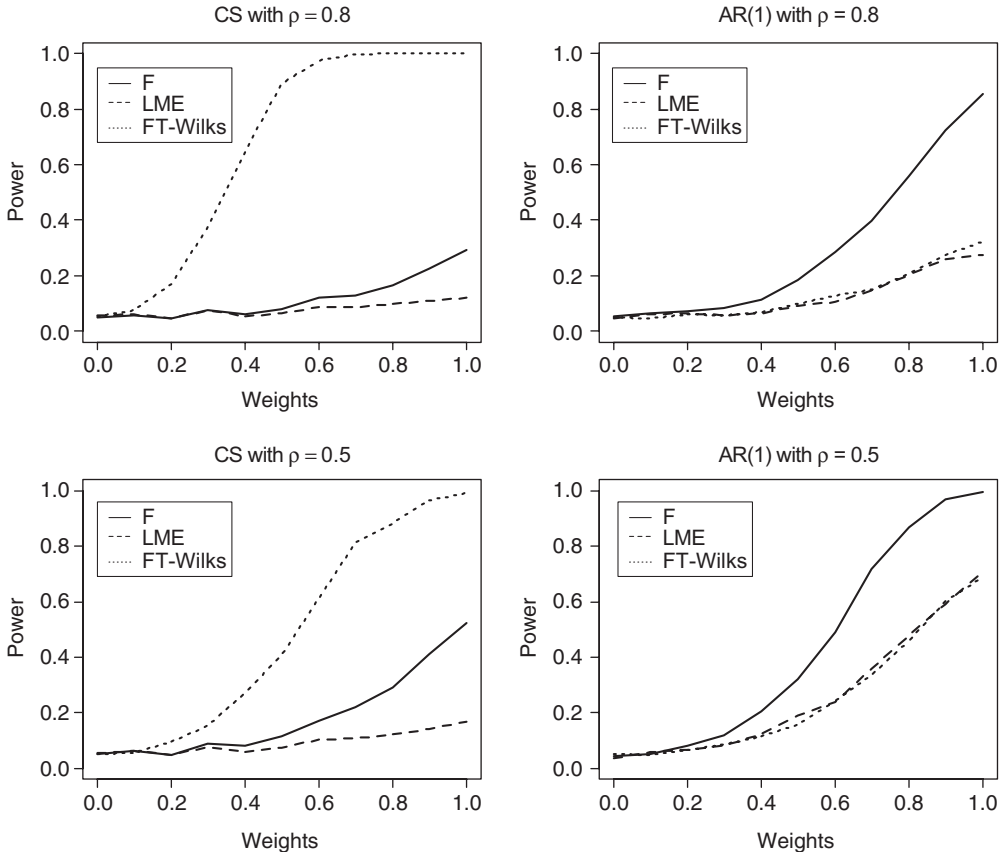

Figure 5. Statistical power of the F-test, Wilks' lambda with Fourier transform, and mixed-effects model with AIC. The four plots present the statistical power curves of the three methods for two covariance structures (CS and AR(1)) with two correlation levels ($\rho = 0.5$ and 0.8). On each plot, the x-axis indicates the weights (0–1 with increment of 0.1) used for simulating 1000 data sets at each weight and the y-axis corresponds to the power, i.e. the probability of correctly rejecting the null model (i.e. the reduced model), for each method on the 1000 data sets. F, F test; LME, linear mixed-effects model; FT-Wilks, Wilks' lambda test after Fourier transform.

frequency domain. Linear mixed-effects models were estimated by maximum likelihood estimation and compared by AIC. The analysis with Fourier transformation is of interest here because it is a useful approach for multivariate data analysis with many appealing features [22]. The fast Fourier transformation (FFT) for discrete data [29] was first applied to each simulated data set, and then the first five Fourier coefficients corresponding to the low frequencies were kept to fit a multivariate regression model. These five coefficients capture around 97% of the 'energy' (i.e. $\sum_{i=1}^{n} \|y_i\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}^2$) defined in the original time space. In addition to dimension reduction, the temporal correlations among repeated measures were also reduced by the orthogonalization of the Fourier transformation.

The plots in Figure 5 show the powers of the three methods for the two covariance structures with correlation set at $\rho = 0.8$ and 0.5. When weight is 0, the reduced model is the true model and

DOI: 10.1002/sim

the power is the size of the test. The simulated sizes are well around the specified significance level 0.05 for all three tests, indicating the functional F test, as well as other two tests, has an accurate size. For the CS covariance structure, it is seen that the Fourier Wilks' lambda test is the most powerful, while the mixed-effects model with AIC is the least. As ρ elevates up, the powers of the functional F test and the mixed-effects model with AIC degrade, whereas the power of Fourier Wilks' test is strengthened. For the AR(1) case, the power of functional F test is higher than those of the other two methods when weights are larger than 0.3. It is also observed that the linear mixed-effects model and Fourier Wilks' lambda test are comparable to each other with similar patterns in terms of power. It is of our special interest to notice that the functional regression model with F test has overall significantly higher power than the linear mixed-effects model for these simulated smoking data.

Our simulation partially confirms the finding reported by Shen and Faraway [11], that is, the covariance structure of the error process is influential to the power of the tests. When $\rho = 0.5$, the ordered eigenvalues for the CS and AR(1) structures are (5.55, 0.15, 0.15, 0.15, ...) and (0.888, 0.856, 0.806, 0.745, ...), respectively. It is clear that the decreasing rate of the eigenvalues is slower for the AR(1) structure, thus making the F test more powerful because the degrees-of-freedom-adjustment-factor is determined not by the actual size of the eigenvalues but their decreasing rate.

We observed that the functional F test is very efficient in computation. For the 1000 simulated data sets with AR(1) structure and $\rho = 0.5$, it took 13 h to fit and compare functional regression models with F test, whereas it took 225 h, about 17 times longer, to fit and compare linear mixed-effects models. The simulation was done on a 1 GHz CPU Mac Xserve.

We also used simulation to investigate the possible effect of smoothing on the performance of the diagnostics introduced in Section 2.2. In the simulation, we multiplied the CS or AR(1) covariance structure by some factor (e.g. 100, 10, 1, 0.1, 0.01) and counted the chance that a specific curve (e.g. the first curve) was detected as an outlier by using jackknife residual and F test when it is not. We found that the chance was around the specified significance level, regardless the covariance structures and the multiplication factors used, indicating that smoothing has little effect on detecting outliers and influential cases. This is consistent with theoretical results by Shen and Xu [25], because jackknife residuals and Cook's distances are functions of studentized residuals that are scale free in the sense that they do not depend on the overall variance.

5. DISCUSSION

As a companion to the work of Shen and Faraway [11], this paper demonstrates the functional regression analysis with an F test to analyse a longitudinal data with a fairly large number of repeated observations measured per subject. The method, as a complement to mixed-effects models, helped us gain better understandings of the efficacy of the behavioural therapies in a smoking cessation study. The simulation study indicates that the F test for the functional regression model has acceptable statistical power when the first few eigenvalues are not predominantly larger than the rest (e.g. as seen in a CS covariance matrix). By estimating the time-varying regression coefficients and making overall significance test, the F test approach provides a medium to strengthen the power of traditional functional data analysis, which was basically exploratory in nature, primarily aiming to represent and display data to highlight interesting characteristics. In clinical trials with repeated measure design, when causal inference is of most concern, the F method could be applied.

DOI: 10.1002/sim

When applying functional regression analysis, the empirical covariance structures are targeted, requiring no specific structures. This does not imply that its performance would be independent of the actual correlation structure. In fact, our simulation study shows that the method is much more powerful when the correlation structure is auto-regressive rather than CS. For the simulated data under conditions of the smoking cessation data, it turns out that the F-test method is more powerful than the linear mixed-effects model. When fitting a mixed-effects model, it is important to correctly specify the intra-subject correlation and correlations among the random effects. An extreme case would be comparing models with and without random effects, which may end up with different estimates of the fixed parameters [30]. Missing values or unbalanced longitudinal data can be handled naturally by applying smoothing techniques that do not require a common fixed time-grid. Functional data analysis for sparse and unbalanced longitudinal data was specially considered in Reference [31]. Additionally, functional regression coefficients provide both intuitive and time-dependent estimators thereby yielding insights for studying time-varying relationships.

Similar ideas on the functional F test could be traced to Box [32], where the property of the F-test statistic in the two-way ANOVA for correlated data was studied in detail. Other ways of functional data analysis were provided by Fan and Lin [22], who used adaptive Neyman or thresholding tests on the Fourier or wavelet expansion coefficients of the estimated parameter function in order to compare groups of curves. As suggested by Eubank [23], these transform-based methods are complicated and may not ultimately boost power. Since the functional regression model, restricted to the finite time grid, becomes a standard multivariate problem, it is natural to try multivariate-based tests. Shen and Faraway [11] carefully compared the performance of the functional F test with a traditional multivariate likelihood ratio test and its variation, such as a B-spline coefficient test, and found that the functional F test had at least the following advantages: (i) it works when the grid size becomes large; (ii) it is stable and not easily influenced by unimportant variation directions; (iii) it is fairly powerful; and (iv) it is computationally cheap. These reasons provide strong rationale for applying functional regression analysis with functional F test in practice.

There are several limitations with the current F-test method for functional regression analysis. First, it does not handle functional or time-varying predictors, which restricts its application in many practical settings when a battery of covariates are measured repeatedly along the outcome variable. Second, the method models repeated measures that are assumed of Gaussian distribution. Although the large sample theory ensures the use of functional regression analysis in wider applications, more specific or generalized forms of functional F-test statistics need to be developed for other types of longitudinal data, e.g. generalized functional linear models [33]. Another limitation comes from missing data problems, which is a common problem also for standard longitudinal modelling in practice. In the smoking cessation data, missing data were assumed 'ignorable' [27], so that multiple imputations could be created using an MCMC algorithm. When assuming such a process, analyses based only on observed data, while ignoring missing values, would provide unbiased estimates. Unfortunately, this assumption of ignorability could not be verified in this smoking cessation study without follow-up investigations [34]. It is urgent that functional regression analysis be developed to analyse longitudinal data sets with informative missing values [4].

ACKNOWLEDGEMENTS

This work was partially supported by the National Institute of Drug Abuse through an SBIR contract N44 DA35513 and two research grants: R03 DA016721 and P50 DA 18185. We especially thank Hamutahl Cohen for her editorial assistance. We also thank two referees for their helpful comments.

DOI: 10.1002/sim

REFERENCES

1. Hand DJ, Crowder MJ. *Practical Longitudinal Data Analysis*. Chapman & Hall: London, 1996.
2. Shoptaw S, Rotheram-Fuller E, Yang X, Frosch D, Nahom D, Jarvik ME, Rawson RA, Ling W. Smoking cessation in methadone maintenance. *Addiction* 2002; **97**:1317–1328.
3. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: Oxford, 1994.
4. Davidian M, Lin X, Wang J-L. Introduction (Emerging Issues in Longitudinal and Functional Data Analysis). *Statistica Sinica* 2004; **14**:613–614.
5. Rice JA. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* 2004; **14**:631–647.
6. Singer SCJ, Willett JB. *Applied Longitudinal Data Analysis*: *Modeling Change and Event Occurrence*. Oxford University Press: Oxford, 2003.
7. Bryk AS, Raudenbush SW. *Hierarchical Linear Models*: *Applications and Data Analysis Methods* (2nd edn). Sage Publications: Beverley Hills, CA, 2001.
8. Kaplan D. *Structural Equation Modeling*: *Foundations and Extensions*. Sage Publication: Beverley Hills, CA, 2000.
9. Ramsay JO, Silverman BW. *Functional Data Analysis* (2nd edn). Springer: New York, 2005.
10. Faraway JJ. Regression analysis for a functional response. *Technometrics* 1997; **39**:254–261.
11. Shen Q, Faraway J. An F test for linear models with functional responses. *Statistica Sinica* 2004; **14**:1239–1257.
12. Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves (with Discussion). *Journal of the American Statistical Association* 1998; **93**:961–994.
13. Zhao X, Marron JS, Wells MT. The functional data analysis view of longitudinal data. *Statistica Sinica* 2004; **14**:789–808.
14. Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* 1991; **53**:233–243.
15. Yao F, Müller H-G, Wang J-L. Functional linear regression analysis for longitudinal data. *Annals of Statistics* 2005; **33**:2873–2903.
16. Hastie TJ, Tibshirani RJ. Varying-coefficient models (with Discussion). *Journal of the Royal Statistical Society of London*, *Series B* 1993; **55**:757–796.
17. Wahba G. *Spline Models for Observational Data*, *Society for Industrial and Applied Mathematics*. SIAM: Philadelphia, PA, 1990.
18. Cleveland W. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 1979; **74**:829–836.
19. Seber GAF. *Multivariate Observations*. Wiley: New York, 1984.
20. Johnson R, Wichern D. *Applied Multivariate Statistical Analysis* (5th edn). Prentice Hall: New Jersey, 2002.
21. Rencher AC. *Methods of Multivariate Analysis* (2nd edn). Wiley: New York, 2002.
22. Fan J, Lin S-K. Tests of significance when data are curves. *Journal of the American Statistical Association* 1998; **93**:1007–1021.
23. Eubank RL. Testing for no effect by cosine series methods. *Scandinavian Journal of Statistics* 2000; **27**:747–763.
24. Abramovich F, Antoniadis A, Sapatinas T, Vidakovic B. Optimal testing in a fixed-effects functional analysis of variance model. *International Journal of Wavelets*, *Multiresolution and Information Processing* 2004; **2**:323–349.
25. Shen Q, Xu H. Diagnostics for linear models with functional responses. UCLA Statistics Preprint 439. http://preprints.stat.ucla.edu/
26. Hsu J. *Multiple Comparison*: *Theory and Methods*. Chapman & Hall/CRC: London, 1996.
27. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
28. Schafer JL. *Analysis of Incomplete Multivariate Data.* Chapman & Hall: London, 1997.
29. Bracewell R. *The Fourier Transform and Its Application* (3rd edn). New York: McGraw-Hill, 1999.
30. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-plus*. Springer: New York, 2000.
31. Yao F, Müller H-G, Wang J-L. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005; **100**:577–590.
32. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems: II. The effect of inequality of variance and correlation between errors in the two-way classification. *Annals of Mathematical Statistics* 1954; **25**:484–498.
33. Müller H-G, Stadtmüller U. Generalized functional linear models. *Annals of Statistics* 2005; **33**:774–805.
34. Yang X, Shoptaw S. Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial. *Drug and Alcohol Dependence* 2005; **77**:213–225.

DOI: 10.1002/sim