Diagnostics for functional regression via residual processes

Jeng-Min Chiou

Academia Sinica, Taiwan

E-mail: *jmchiou@stat.sinica.edu.tw*

Hans-Georg Müller*

University of California, Davis, USA

E-mail: *mueller@wald.ucdavis.edu*

Version December 2007, with typos corrected that appeared in eq. (15) and the definition of the D_{ikj} on p. 4857 of the published article [Chiou, J.M., Müller, H.G. (2007). Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis* **51**, 4849-4863].

Abstract

We develop regression diagnostics for functional regression models which relate a functional response to predictor variables that can be multivariate vectors or random functions. For this purpose, we define a residual process by subtracting the predicted from the observed response functions. This residual process is expanded into functional principal components, and the corresponding functional principal component scores are used as natural proxies for the residuals in functional regression models. For the case of a univariate covariate, we propose a randomization test based on these scores to examine if the residual process depends on the covariate. If this is the case, it indicates lack of fit of the model. Graphical methods based on the functional principal component scores of observed and fitted functions can be used to complement more formal tests. The methods are illustrated with data from a recent study of *Drosophila* fruit flies regarding life-cycle gene expression trajectories as well as functional data from a dose-response experiment for Mediterranean fruit flies (*Ceratitis capitata*).

Keywords: Cook's distance; Eigenvalue weighting, Functional data analysis; Gene expression profile; Goodness-of-fit; Hat matrix; Principal component; Randomization test; Residuals.

*Corresponding author

1 Introduction

Methodology of regression analysis for functional data, where either predictors or responses can be viewed as random functions, is receiving increasing attention. Functional data are generated when the value of a variable is repeatedly recorded on a dense grid of time points for a sample of subjects. Each string of repeated measurements may then be represented as a function, where one usually assumes that the underlying trajectories are smooth. Functional data are becoming more common due to rapid advancements in modern technology and increasingly complex research questions which involve the dynamics of time-dependent processes. The earliest statistical study of linear functional regression appears to be by Ramsay and Dalzell (1991), while the idea of regression for stochastic processes dates back at least to Grenander (1950). Reviews of various regression models for functional data concerning different combinations of functions and vectors as response and predictor variables are provided in the monograph by Ramsay and Silverman (2005), as well as in several articles (e.g., Chiou, Müller and Wang 2004; Müller 2005). Recent relevant works on functional regression include Faraway (1997), Cuevas, Febrero and Fraiman (2002), Cardot et al. (2003) and Ferraty and Vieu (2004), and articles on the related generalized functional linear model include James (2002), Escabias, Aguilera and Valderrama (2004), Cardot and Sarda (2005) and Müller and Stadtmüller (2005), among others.

Classical regression diagnostics are based on residuals (Anscombe and Tukey 1963) and have an important place in applied statistics for the task of checking model assumptions that underlie statistical analysis. Such techniques have been largely limited to classical linear and nonlinear regression models, where response and predictor variables are scalars or vectors. A primary application of diagnostics has traditionally been bias detection, besides checking for homoscedasticity and distributional assumptions. It is certainly of interest to extend and develop basic regression diagnostics for functional regression analysis, in particular regression models with functional responses. This paper provides some ideas for such an extension. Models that specifically address functional responses include the above-mentioned functional linear regression model of Ramsay and Dalzell (1991), where both response and predictor variables are random functions, the functional response model (Chiou, Müller and Wang 2003a,b), and the functional dose-response model (Chiou, Müller and Wang 2004), where in the latter models the predictors are assumed to be scalars or vectors.

An ubiquitous approach in regression diagnostics is to ascertain whether the residuals depend in any way on the fitted responses, i.e., the values predicted by the assumed model for given predictor levels. The most commonly used residual plot, the scatterplot of residuals versus fitted values, serves this purpose. Any deviations from a random scattering of the residuals, which are supposed to lie in a "band" around the abscissa, are taken to indicate deviations caused by lack of fit. The most serious deviation is typically model bias. Such bias usually invalidates statistical inference drawn from fitting the model and often reveals itself in trends visible in the residual plot. When such bias is detected, a next step is to find an improved model that provides a satisfactory fit to the data.

We address here the question of how to extend this diagnostic principle to the case of a regression model with functional responses. The main idea is to replace the customary residuals that are used for one-dimensional responses with residual processes. Residual processes are obtained analogously to classical residuals by subtracting the fitted response function from the observed response function for each subject or experimental unit. Of interest is whether these residual processes depend in any way on predicted values. If such dependence is detected, this points to bias and accordingly lack of fit of the underlying functional regression model. The lack of fit could be due to misspecified parametric or structural components. While fully parametric models are not often used in typical functional regression applications, such models are common in longitudinal data analysis, which can also be brought under the umbrella of functional regression (compare Müller 2005 for a review of recent developments in this direction).

A typical example for structural constraints built into functional regression models is the linear functional regression model (Ramsay and Dalzell 1991), which is inherently nonparametric (i.e., it has no parametric components) but is based on structural linear assumptions about the relationship of the response function with the predictor function. These assumptions reduce the dimensionality of the regression problem and allow the representation of this model with basis function expansions that for example can be conveniently implemented with functional principal components. Functions are represented with an infinite number of functional principal components that are usually ordered according to the size of their corresponding eigenvalues, and in statistical applications will be truncated at a finite number of components. It is then of interest to ascertain whether such structural model assumptions and the associated "smaller" model are satisfied for a functional regression model, or whether a less structured and therefore "larger" model must be used.

The general form of the functional regression model with functional response that we consider can be described as

$$Y(t) = f(t, \eta(Z)) + Q(t), \tag{1}$$

where $Y(t)$ is the random response function defined on domain $t \in \mathcal{T}$ where $\mathcal{T}$ is a closed interval, $f(t, \eta(Z))$ is a function of t and $\eta(Z)$ where Z is a covariate (possibly multivariate or functional), and $\eta(\cdot)$ is a function mapping Z to $\mathbb{R}$, for example a single index for the case of a covariate vector Z. Furthermore, Q is an error process. We consider model (1) to be normalized in such a way that the overall mean function is on target, i.e.,

$$E[f(t, \eta(Z))] = EY(t) = \mu_Y(t), \quad EQ(t) = 0, \quad t \in \mathcal{T}.$$

We include the case where the predictor variable is a function $Z(s)$, $s \in \mathcal{S}$, for a closed interval $\mathcal{S}$. For example, in the functional linear model the conditional mean function is of the form

$$E(Y(t)|Z) = f(t, \eta(Z)) = \mu_Y(t) + \int_{\mathcal{S}} \beta(s,t)(Z(s) - \mu_Z(s))\, ds, \tag{2}$$

where $\mu_Z(s) = EZ(s)$ and the bivariate regression parameter function $\beta(s,t)$ is either parametrically specified or (more commonly) assumed to be just smooth and square integrable.

Functional regression models (1) incorporate the influence of the predictors Z through the function $f(t, \eta(Z))$, which can assume various forms. An example is the additive model $f(t, \eta(Z)) = g(\mu_Y(t) + \eta(Z))$, where $g(\cdot)$ is a known link function and $\mu_Y(t)$ stands for an overall mean response function. Another useful choice is a multiplicative effects model, as implemented in the functional response models described in Chiou et al. (2003b, 2004). In these models, one assumes a multiplicative form $f(t, \eta(Z)) = \mu_Y(t)\eta(Z)$ with the constraint $E\eta(Z) = 1$, i.e., an overall mean time trend $\mu_Y(t)$ is multiplied with a covariate-dependent smooth effect. This approach was shown to provide parsimonious descriptions for functional dose-response data.

The remainder of the paper is organized as follows. The role played by residual processes in diagnostics is described in Section 2. Section 3 presents functional residual plots in the context of an application to life-cycle gene expression trajectories for *Drosophila* fruit flies. A randomization test for goodness-of-fit diagnostics is described in Section 4. Section 5 presents additional details of the proposed diagnostic procedures in connection with an application to functional response models for data obtained from a dose-response experiment that was carried out with Mediterranean fruit flies (*Ceratitis capitita*). Discussion and concluding remarks are in Section 6.

2 Diagnostics via functional principal component analysis of residual processes

When fitting model (1) to data, the fitted model $\hat{Y}(t)$ and residual process $R(t)$ are given by

$$\hat{Y}(t) = f(t, \hat{\eta}(Z)), \qquad R(t) = Y(t) - \hat{Y}(t), \tag{3}$$

where $\hat{\eta}$ is an estimate of a single index, i.e., the effect that Z has in the assumed model. This estimate is specific to the particular model considered. If the model fits reasonably well, we expect

$$\mu_R(t) = ER(t) \approx 0, \quad t \in \mathcal{T},$$

and more importantly, that $R(t)$ does not depend on the predictor Z, i.e.,

$$E(R(t)|Z) \approx 0, \quad t \in \mathcal{T}.$$

Goodness-of-fit diagnostics and tests for residuals naturally focus on checks for these desirable properties of residual processes.

2.1 *Functional principal component analysis*

Generally, under mild assumptions, observed square integrable processes X defined on support $[0, T]$ with mean $EX(t) = \mu_X(t)$ allow the covariance expansion $\Gamma(s,t) = E[\{X(s) - \mu_X(s)\}\{X(t) - \mu_X(t)\}] = \sum_k \lambda_k \rho_k(s)\rho_k(t)$, for $s, t \in [0, T]$, where λ_k and ρ_k, $k = 1, 2, \ldots$, are eigenvalue-eigenfunction pairs. This leads to the Karhunen-Loève expansion of X in $L^2([0, T])$,

$$X(t) = \mu_X(t) + \sum_{k=1}^{\infty} A_k \rho_k(t). \tag{4}$$

Here, the random variables A_k are the functional principal component (FPC) scores, defined as $A_k = \int (X(t) - \mu_X(t))\rho_k(t)\, dt$. They satisfy $E(A_k) = 0$ and $\mathrm{cov}(A_k, A_l) = \delta_{kl}\lambda_k$, for all k and l, with the Kronecker symbol $\delta_{kl} = 1$ if $k = l$, and $\delta_{kl} = 0$ otherwise. The associated eigenvalues $\lambda_k = \mathrm{var}(A_k)$, assuming that $\lambda_1 \geq \lambda_2 \geq \ldots$, with $\sum \lambda_k < \infty$. The eigenfunctions $\rho_k(t)$ are orthonormal in $L^2([0, T])$ and are assumed to be smooth (twice continuously differentiable).

Given a sample of observed random trajectories X_i, $i = 1, \ldots, n$, of processes X, estimates $\hat{\mu}_X$ of the mean function of X can be obtained by cross-sectional averaging over all observed trajectories, and estimates $\hat{\Gamma}$ of the covariance function Γ by smoothing the cross-products $(X_i(s) - \hat{\mu}_X(s))(X_i(t) - \hat{\mu}_X(t))$, $i = 1, \ldots, n$, followed by spectral decomposition of the resulting discrete covariance matrix. These steps lead to eigenfunction estimates $\hat{\rho}_k$ and eigenvalue estimates $\hat{\lambda}_k$, implementing for example the method described in Chiou, Müller and Wang (2003); compare also Rice and Silverman (1991). Once mean and eigenfunction estimates have been determined, FPC score estimates are obtained as

$$\tilde{A}_{ik} = \int_0^T (X_i(t) - \hat{\mu}_X(t))\; \hat{\rho}_k(t)\, dt, \tag{5}$$

via numerical integration, or alternatively, by a conditioning step (these two approaches are compared in Yao, Müller and Wang 2005a and Müller 2005).

Alternative eigenvalue estimates may also be obtained from the empirical variances of the estimated FPC scores $\tilde{A}_{ik}$, $i = 1, \ldots, n$. In situations where the observed functions are contaminated with measurement errors or where one wishes to ensure non-negative definiteness of the covariance matrices, modifications using shrinkage estimates of functional principal component scores and projections on positive definite covariance surfaces (by omitting components with negative estimated eigenvalues) have been described in Yao et al. (2003, 2005a). In the applications of this study, the shrinkage estimators of functional principal component scores are used, where the optimal shrinkage factor is obtained by minimizing cross-validated squared prediction errors.

In the Karhunen-Loève expansion of X (4), the orthonormal set of eigenfunctions ρ_k plays the role of basis functions in $L^2([0, T])$. Other basis functions in $L^2([0, T])$ can also be used instead to span the process. Among all expansions of X with K basis functions, however, the truncated Karhunen-Loève expansion maximizes the percentage of total variance explained by the K components, which is $\sum_{k=1}^{K} \lambda_k / \sum_{k=1}^{\infty} \lambda_k$, and therefore the eigenbasis is often preferable. We note that the eigenfunctions form just one convenient basis among many possible bases, and although they are estimated, are fixed functions that are solely determined by the covariance structure of processes X. A consequence is that we do not need to worry about dependency of the eigenfunctions of the processes that we consider on the predictor variable Z. To make the present approach feasible, it is on the contrary important that eigenfunctions are chosen irrespective of the covariates Z, so that any dependency on Z is entirely carried by the FPC scores A_k which are the random components; see Chiou et al. (2003a). In this setting, the sequence of FPC scores $A_1, A_2, \ldots$ and processes X are equivalent, so that any regression relation of X with a variable Z can then be expressed in terms of regression relations of the A_j with Z.

We focus here on practical aspects, and it is a common practical experience that for functional data the first K eigenfunctions ρ_k, $k = 1, \ldots, K$, effectively span the processes for values of K that are small to moderate. As more and more curves are sampled, the value of included components K typically will increase. In asymptotic theory, one assumes typically that $K = K_n$ increases with sample size n (Yao et al. 2005a, Hall and Hosseini-Nasab 2006), while in practical applications this is reflected by the fact that K is chosen from the data, in dependence on sample size and sample characteristics. However, it will always be a finite choice that is made for a given data set. The situation is similar to the bandwidth choice in nonparametric curve estimation where a bandwidth has to converge to zero asymptotically but in practical situations a fixed value will be chosen.

Choices $K = K_n$ then lead to the fitted model of a truncated Karhunen-Loève expansion for individual i, using the first K random components,

$$\hat{X}_i^K(t) = \hat{\mu}_X(t) + \sum_{k=1}^{K} \tilde{A}_{ik}\hat{\rho}_k(t) \ , \tag{6}$$

where $\hat{\mu}_X(t)$ and $\tilde{A}_{ik}$ are as in (5). Since the functional linear regression we consider (at least in theory) can be decomposed into uncorrelated simple linear regressions, as explained at the end of section 3.1, the exact choice of K will not impact the diagnostics for each component. Therefore, we do not need to worry about the total number of components when interpreting the goodness-of-fit for the components that are included, one at a time.

To determine the number of unknown components K from the data one has several options. These include use of the scree graph or the cumulative percentage of total variance (CPV), as in conventional multivariate principal components analysis, the pseudo-AIC criterion (Yao et al. 2005a), and minimizing cross-validation prediction errors. The cumulative percentage of total variance explained by the first M functional principal components is

$$\text{CPV}(M) = \sum_{k=1}^{M} \lambda_k / \sum_{k=1}^{\infty} \lambda_k, \quad M = 1, 2, \ldots \tag{7}$$

With a predetermined threshold value, 90% say, the number of components is chosen such that K is chosen as the minimal value M satisfying $\text{CPV}(M) \geq 90\%$. Let $\tilde{X}_i = (X_i(t_{i1}), \ldots, X_i(t_{im_i}))^T$ be the observed trajectories for process X_i, where m_i is the number of observations for individual i. The number of components determined by the pseudo-AIC criterion is chosen by minimizing $\text{AIC}(M)$ with respect to M,

$$\text{AIC}(M) = \sum_{i=1}^{n} \left\{ \frac{1}{2\hat{\sigma}_X^2} (\tilde{X}_i - \tilde{X}_i^M)^T (\tilde{X}_i - \tilde{X}_i^M) + \frac{m_i}{2}\log(2\pi) + \frac{m_i}{2}\log(2\hat{\sigma}_X^2) \right\} + K \ , \tag{8}$$

where $\tilde{X}_i^M = (\hat{X}_i^M(t_{i1}), \ldots, \hat{X}_i^M(t_{im_i}))$, and $\hat{\sigma}_X^2$ is the estimated measurement error variance (see Yao et al. 2005a for details). Another automatic selection criterion is obtained by minimizing the cross-validation prediction errors $\text{CV}(M)$ with respect to M,

$$\text{CV}(M) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\tilde{X}_i(t_{ij}) - \hat{X}_{(i)}^M(t_{ij}))^2 dt, \tag{9}$$

where $\hat{X}^M_{(i)}(t_{ij})$ is the fitted function for X_i by leaving out the ith observed trajectory, using M components for model fits.

Functional principal component analysis (FPCA) can be applied to various processes which play a role in functional regression. These include the response functions $Y(t)$, the residual processes $R(t)$, the fitted processes $\hat{Y}(t) = f(t, \hat{\eta}(Z))$ and predictor processes $Z(t)$ (in the case of functional predictors). Accordingly, we denote the estimated FPC scores obtained from FPCA for each of these processes by A_Y for processes Y, A_R for processes R, A_F for processes $\hat{Y}$ and A_Z for processes Z. The number of included components K to be determined for each process separately will be denoted by K_Y, K_R, K_F and K_Z, respectively. In this study, these numbers of components are chosen according to the CPV criterion (7).

2.2 *Diagnostics via residual processes*

The residual process $R(t) = Y(t) - \hat{Y}(t)$ plays a role analogous to that of the ordinary residual in classical regression models. Diagnostics based on nonparametric smoothing using residuals for testing the fit of various parametric and nonparametric models have been investigated by many authors. A review is provided in the monograph of Hart (1997). A basic goodness-of-fit check consists of determining whether the residual process $R(t)$ depends on the covariate Z. Writing $R(t, Z)$ to denote dependence on the covariate Z, we look for graphical and formal evidence regarding the null hypothesis

$$H_0: \quad E(R(t) \mid Z = z) = 0 \quad \text{for all } z, \tag{10}$$

indicating that R does not depend on Z. As residual processes are infinite-dimensional, dimension reduction is necessary to derive suitable graphical diagnostics and test statistics, and is most easily implemented by considering only the first K_R components of the residual process as in (6).

We propose to use the functional principal component (FPC) scores A_R (5) as proxies for residual processes. The mean of a residual process is expected to be zero approximately and the eigenfunctions serve as a set of orthonormal basis functions that span the residual process. Notably, any residual process in L^2 with continuous covariance can be spanned by all of the eigenfunctions and dependency on covariate is then entirely expressed in terms of its FPC scores. Therefore, it is assumed that the eigenfunctions ρ_{Rk} of residual processes are independent of covariates Z, and any dependencies in the residual processes on covariates will lead to dependencies in at least some of the A_{Rk} on predictors Z, or alternatively, on fitted processes $\hat{Y}(t)$, which can also be represented by their FPC scores. This means there will be some indices k for which $E(A_k \mid Z)$ depends on Z, resp., $E(A_k \mid \hat{Y}(t))$ depends on $\hat{Y}(t)$. This is explored in more detail in the next section.

3 Residual plots for functional linear regression models, with application to gene expression profile data

The case where the covariate Z is a function as in the functional linear model (2) is considered here. If this model is assumed to be the true underlying model, one would estimate the regression parameter function $\beta(s,t)$ by $\hat{\beta}(s,t)$ to obtain the fitted response function

$$\hat{Y}(t) = \hat{\mu}_Y(t) + \int_{\mathcal{S}} \hat{\beta}(s,t)(Z(s) - \hat{\mu}_Z(s))\, ds \tag{11}$$

for predictor function Z. With eigenfunctions $\rho_{Y\,k}$ and $\rho_{Z\,j}$ for processes Y and Z, and corresponding (true) FPC scores B_{Yk} for processes Y and B_{Zj} for processes Z, the regression parameter function can be represented as

$$\beta(s,t) = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{E(B_{Zj}B_{Yk})}{E(B_{Zj}^2)} \rho_{Z\,j}(s)\rho_{Y\,k}(t), \tag{12}$$

(see He, Müller and Wang 2000) with corresponding estimates

$$\hat{\beta}(s,t) = \sum_{k=1}^{K_Y}\sum_{j=1}^{K_Z} \frac{\hat{E}(B_{Zj}B_{Yk})}{\hat{E}(B_{Zj}^2)} \hat{\rho}_{Zj}(s)\hat{\rho}_{Yk}(t). \tag{13}$$

Details regarding the calculation of $\hat{\beta}(s,t)$ and the fitted response function can be found in Yao, Müller and Wang (2005b).

3.1 *Functional residual plots and functional predictor-response plots of FPC scores*

A popular plot for multivariate predictors is to plot residuals versus fitted values. To check goodness of fit in the functional case, we plot the estimated residual FPC scores A_{Rik} obtained for the residual processes against the estimated FPC scores A_{Fij} of the fitted processes. The proposed *functional residual plots* consist of all pairwise plots of residual FPC scores versus fitted FPC scores, i.e., all pairwise plots A_{Rik} versus A_{Fij}, $1 \leq k \leq K_R$, $1 \leq j \leq K_F$, $1 \leq i \leq n$, with a total of $K_R K_F$ plots. Each of these scatterplots contains n points. These functional residual plots are a natural extension of the classical residual plots and play a similar role; that is, highlighting whether there is any pattern or dependency of the residual FPC scores A_{Rik} on the fitted FPC scores A_{Fij}, corresponding to goodness-of-fit checking. Each of the plots is examined in the same way as a classical residual plot. None of the point clouds among these plots should show a trend and ideally all point clouds should lie in a band around the abscissa, similar to classical plots of residuals versus fitted values.

Additional plots of interest can be obtained by plotting FPC scores A_{Yik} of the response trajectories against FPC scores A_{Zil} of the predictor trajectories. If the linear model assumption is correct, a simple calculation shows that (11) and (12) imply that

$$E(A_{Yk} \mid A_{Zj}) = \sum_{j=1}^{K_Z} \frac{E(B_{Zj} B_{Yk})}{E(B_{Zj}^2)} A_{Zj}.$$

Since the FPC scores corresponding to different components are uncorrelated theoretically, the above conditional expectation implies simple linear regressions without intercept (see also Yao et al. 2005b). This means that at a minimum, these functional predictor-response plots of FPC scores (A_{Zj}, A_{Yk}) should show linear trends if the functional linear regression model is correct.

3.2 *Application to gene expression profile data*

We illustrate functional residual plots through an application to the analysis of *Drosophila* life-cycle gene expression profile data. Time course microarray gene expression experiments are becoming increasingly common, as the assessment of changes of gene expression over time is essential for analyzing temporal gene regulation and the molecular organization of time-dynamic processes such as growth or aging of organisms. We use a subset of the data generated in the *Drosophila* life cycle gene expression experiment described by Arbeitman et al. (2002) to illustrate the proposed diagnostic methods for functional regression. This subset consists of genes that were identified to be related to eye development in Drosophila. For each of the 30 eye-related genes, the experiment provided (among other data) a pair of trajectories of adult and embryonic gene expression. The experimental units correspond here to individual genes. The eye-specific genes were identified by hierarchical clustering by Arbeitman et al. (2002), and the gene expression trajectory measurements consist of 31 normalized measurements of gene expression level at the embryonal stage and of eight measurements at the adult stage, measured at a grid of non-equidistant times. We treat the measurement times as equidistant, which corresponds to a simple time warping transformation. For issues regarding the dynamics of regulatory processes such as time-warping for these data, we refer to Liu and Müller (2003).

We conducted a functional regression analysis for these data. The gene expression profiles at the adult stage are taken as the response functions Y and those at the embryo stage are treated as the predictor functions Z in this regression analysis. We adopt the functional linear regression model (2) as the null model, where the fitted versions of (11) and (13) were obtained with the method described in Yao et al. (2005b). The number of components chosen for predictor and response functions based on the CPV criterion was $K_Y = 2$ (CPV(1) = 86.2%, CPV(2) = 99.1%), $K_Z = 3$ (CPV(2) = 92.5%, CPV(3) = 97.3%), respectively, and $K_F = 2$ (CPV(1) = 96.6%, CPV(2) = 98.7%) $K_R = 2$ (CPV(1) = 81.0%, CPV(2) = 98.4%) for fitted functions and residual processes.

Figure 1 illustrates the relationship between the response and predictor FPC scores in the resulting six scatterplots of A_{Yk} vs A_{Zj} for $k = 1, 2$ and $j = 1, 2, 3$. These pairwise scatterplots indicate linear relationships with the exception of the scatterplot plotting the first FPC scores against each other (lower left panel), where a quadratic component is discernible. The functional residual plots using the first two leading FPC scores from both residual processes and fitted functions, i.e., plotting A_{Rk} vs A_{Fj} for $j, k = 1, 2$ are shown in Figure 2. In these plots, again an indication of dependency of the residual FPC scores on the fitted FPC scores is found for the first FPC scores A_{Ri1} vs A_{Fi1} (lower left panel), with a roughly quadratic structure, and to a lesser degree also for the plot of the second FPC scores A_{Ri2} vs A_{Fi2}, while none of the other plots show a clear trend. Overall, this residual analysis points to some areas of concern regarding the goodness-of-fit of the functional linear model for these data.

3.3 *Functional leverage and functional Cook's distance*

Under the functional linear regression model, all relationships between FPC scores for response and predictor processes must correspond to simple linear regressions through the origin. Therefore, all diagnostic tools that are known to be useful for simple linear regression can be applied to the functional case as well. Besides the basic diagnostic plots, this includes the concepts of hat matrix and leverage points, among other diagnostics, such as Cook's distance. The question then arises how to combine quantitative diagnostic devices such as a hat matrix over the various pairwise simple regressions that correspond to the pairwise regression plots as described above, in order to arrive at one overall summary diagnostic. For each FPC of the response process, we propose to sum the numerical values obtained for each simple linear regression over the various predictor FPCs. Then an overall diagnostic can be formed by taking a weighted average over the aggregated values that are obtained in this first step.

In order to define a hat matrix for functional regression, with the goal to determine predictor functions that exercise high leverage analogous to leverage points in ordinary regression, we proceed as follows. First, compute the $n \times n$ hat matrices $\mathbf{H}_{kj}$ for the simple linear regressions of A_{Yik} versus A_{Zij}, without intercept, which are given by $\mathbf{H}_{kj} = \mathbf{X}_j(\mathbf{X}_j^T\mathbf{X}_j)^{-1}\mathbf{X}_j^T$, where the design matrices $\mathbf{X}_j$ in this case do not depend on k and are $n \times 1$ vectors $\mathbf{X}_j = (A_{Z1j}, \dots, A_{Znj})^T$. Defining $\mathbf{H}_k = \sum_{j=1}^{K_Z} \mathbf{H}_{kj}$, the vector of fitted curves $\tilde{\mathbf{Y}}(t) = (\hat{Y}_1(t) - \hat{\mu}_Y(t), \dots, \hat{Y}_n(t) - \hat{\mu}_Y(t))^T$ and the vectors of underlying and of fitted FPC scores for processes Y, $\mathbf{A}_{Yk} = (A_{Y1k}, \dots, A_{Ynk})^T$, $\quad \hat{\mathbf{A}}_{Yl} = (\hat{A}_{Y1l}, \dots, \hat{A}_{Ynl})^T$, a simple calculation shows

$$\tilde{\mathbf{Y}}(t) = \sum_{k=1}^{K_Y} \mathbf{H}_k \mathbf{A}_{Yk} \rho_{Y k}(t), \tag{14}$$

which provides the linear mapping from Y to $\hat{Y}$, as befits a hat matrix.

At this point, one has various options. One option is to plot the elements of the matrix on the r.h.s. of (14) as a function of t, since this hat matrix is a function of t. This may be of interest to detect leverage for some parts of the domain of the response curves, as would be evidenced by large diagonal elements of the hat matrix curves for the corresponding times t. Another option is to seek a combined hat matrix. Such a matrix is not uniquely defined, and there are several ways to reduce the above hat matrix functions to a single hat matrix. By multiplying (14) on both sides with ρ_{Yk} and integrating over t, we find for the relation of fitted FPC scores for response processes to the underlying FPC scores,

$$\hat{\mathbf{A}}_{Yk} = \mathbf{H}_k \mathbf{A}_{Yk}.$$

An overall hat matrix $\mathbf{H}$ may be obtained by targeting the weighted average $\sum_k \omega_k \hat{\mathbf{A}}_{Yk}$, where $\omega_k = \hat{\mathbf{A}}_{Yk} \mathrm{var}(A_{Yk}) / \sum_l \mathrm{var}(A_{Yl})$, suggesting

$$\mathbf{H} = \frac{\sum_{k=1}^{K_Y} \lambda_{Yk} \mathbf{H}_k}{\sum_{k=1}^{K_Y} \lambda_{Yk}} = \mathbf{H}_k, \tag{15}$$

as here $\mathbf{H}_k$ does not depend on k. In general, for eigenvalue weighting, unknown quantitites such as λ_{Yk} will be replaced by their estimates in practical implementation. Eigenvalue weighted statistics have been suggested before in the context of defining a global value for the coefficient of determination R^2 for functional regression in Yao et al. (2005b).

An example where eigenvalue weighting leads to a simple non-time dependent diagnostic is a functional version of Cook's distance (Cook 1977). The functional Cook's distance for the i-th predictor can be defined similarly to (15) as

$$D_i = \frac{\sum_{k=1}^{K_Y} \lambda_{Yk} \sum_{j=1}^{K_Z} D_{ikj}}{\sum_{k=1}^{K_Y} \lambda_{Yk}}, \tag{16}$$

where

$$D_{ikj} = (\hat{\beta}_{kj} - \hat{\beta}_{kj(i)})^T (\mathbf{X}_j^T \mathbf{X}_j)(\hat{\beta}_{kj} - \hat{\beta}_{kj(i)}) \Big/ \frac{1}{n-1} \sum_{i=1}^{n} (A_{Yik} - A_{Zij}\hat{\beta}_{kj})^2,$$

where $\hat{\beta}_{kj}$ is the fitted regression slope coefficient of the simple linear regression without intercept fitted to the scatterplot A_{Yik} versus A_{Zij}. The coefficients $\hat{\beta}_{kj(i)}$ are obtained analogously when leaving out the data for the i-th subject.

We show the diagonal elements of the functional hat matrix (15) that indicate leverage of individual predictor genes on the left panel and the functional Cook's distances for the predictor genes on the right panel of Figure 3. Notably, a few predictor gene expressions have high leverage, while there is one gene expression profile with a particular large Cook's distance. This indicates that the corresponding gene has large influence on the fitted functional regression, and may warrant further analysis.

4 A simple test statistic for functional goodness-of-fit

Functional residual plots provide a graphical tool for visual inspection of the residual process. In addition, it is desirable to have quantitative statistics for the assessment of functional goodness-of-fit, specifically, to assess whether the residual process depends on vector or functional covariates as in (10). We discuss here some heuristic extensions of statistics that are related to classical lack-of-fit testing in linear regression models and may prove useful for the functional case. Rather than deriving asymptotic distributions, we consider a randomization procedure to obtain inference.

Using the functional principal component (FPC) scores as proxies for residual processes, and assuming as before that only the first K_R components matter, the null hypothesis (10) of goodness-of-fit (6) is that for each residual FPC score A_{Rk} we must have

$$H_0^{(1)}:\ E(A_{Rk} \mid Z) = 0, \quad k = 1, \ldots, K_R, \tag{17}$$

and this must hold for each component of a multivariate predictor Z and for each of the first K_Z relevant FPC scores A_{Zj} of the predictor processes for the case of a functional predictor. Hypothesis $H_0^{(1)}$ can then be rewritten as

$$H_0^{(1)'}:\ E(A_{Rk} \mid A_{Zj}) = E(A_{Rk}) = 0,\ \text{for}\ k = 1, \ldots, K_R,\ j = 1, \ldots, K_Z.$$

We provide details for the case of a univariate predictor Z. The case of a functional predictor can be dealt with by adding up the test statistics obtained for the various uncorrelated univariate predictor FPC scores as described at the end of the previous section for hat matrix and Cook's distance.

The null hypothesis $H_0^{(1)}$ (17) implies that the variances of the conditional expectation vanish,

$$H_0^{(2)}: \ \mathrm{var}(E(A_{Rk} \mid Z)) = 0, \quad \text{for } k = 1, \dots, K. \tag{18}$$

On the other hand, $\mathrm{var}(E(A_{Rk} \mid Z)) = 0$, $k = 1, \dots, K_R$ implies that $E(A_{Rk} \mid Z) \equiv \mathrm{const}$ and therefore $E(A_{Rk} \mid Z) = E(A_{Rk}) = 0$. Therefore, null hypotheses $H_0^{(1)}$ and $H_0^{(2)}$ are equivalent.

We propose to construct a heuristic statistic to test $H_0^{(2)}$ as follows. The starting point is sample data $\{(Z_i, A_{Rik}),\ i = 1, \dots, n\}$ for a fixed k. For a suitable $L > 1$, we define L bins based on the values of the Z_i, by assembling neighboring Z_i's into the same bin, such that each bin contains approximately the same number of observations. The sizes of the bins are assumed small enough so that within each bin the conditional mean and variance of the A_{Rik} are approximately constant. Let $\mathcal{I}_\ell$ be the set of indices $\{i\}$ for which Z_i fall into the ℓth bin, and denote the A_{Rik} that fall into the ℓth bin by $A_{R_{ik}}^{(\ell)}$, i.e., $A_{R_{ik}}^{(\ell)} \in \{A_{Rik}, i \in \mathcal{I}_\ell\}$. Let $m_\ell = \sum\limits_{i=1}^{n} I(i \in \mathcal{I}_\ell)$, the number of data falling into the ℓth bin.

To test the hypothesis $H_0^{(2)}$ in (18), we define statistics T_k for $k = 1, \dots, K_R$, such that

$$T_k = \frac{1}{L-1} \sum_{\ell=1}^{L} (b_{k\ell} - \bar{b}_k)^2, \tag{19}$$

where $b_{k\ell} = m_\ell{}^{-1} \sum_{i \in \mathcal{I}_\ell} A_{R_{ik}}^{(\ell)}$ and $\bar{b}_k = \sum_{\ell=1}^{L} b_{k\ell}/L$. Here, the bin means $b_{k\ell}$ target $E(A_{Rk} \mid Z^{(\ell)})$ for Z in the ℓth bin and $\bar{b}_k$ target the overall mean $E(A_{Rk})$. Therefore T_k may be interpreted as an empirical estimate of $\mathrm{var}(E(A_k \mid Z))$. This type of test statistic is closely related to the so-called "lack-of-fit sum of squares" (Draper and Smith 1998, compare Green 1971). We note that under the null hypothesis, one expects $E(A_{Rk} \mid Z^{(\ell)}) = 0$; the centering of the test statistic in (19) may help to improve power (Hall and Wilson 1991). It remains to define an overall test statistic combining the above statistics defined for each of the components $k = 1, \dots, K_R$, and this is done through another application of eigenvalue-weighting, obtaining the overall test statistic

$$T_0 = \sum_{k=1}^{K_R} \lambda_{Rk} T_k \Big/ \sum_{k=1}^{K_R} \lambda_{Rk}. \tag{20}$$

The empirical version $\hat{T}_0$ is obtained by replacing λ_{Rk} with $\hat{\lambda}_{Rk}$.

The null distribution of the test statistic $\hat{T}_0$ can be approximated by randomizing the binning scheme; in each randomized version, bins are composed of a matching number of randomly selected A_{Rik} for which the test statistics is computed, ignoring the neighborhood relation of the associated Z_i. This is done many times so as to obtain a null distribution of the test statistic under the situation of no association. The empirical p-value for the observed test statistics can then be easily calculated from this null distribution. More specifically, we obtain an approximation to the null distribution by repeatedly dividing $\{A_{Rik}\}$ into L bins via random selection for each of $m = 1, \ldots, M$ random samples, leading to the value $T_0^{(m)*}$ of the test statistic at each iteration. The empirical p-value for the resulting randomization test is

$$\tilde{p} = \frac{1}{M} \sum_{m=1}^{M} I\{T_0^{(m)*} \geq \hat{T}_0\} \ ,$$

where I stands for the indicator function.

5 Diagnostics for functional dose-response models

In a second illustration, we apply both the proposed functional residual plots and the randomization tests to data from a functional dose-response experiment. In this experiment the responses are random trajectories of daily egg-laying recorded for Mediterranean fruit flies (medflies) and the predictors are scalars (Carey et al. 2002). Daily egg-laying was recorded for a sample of 874 female medflies in response to one of ten dietary doses. Each dose was provided to about 100 medflies each and the daily egg counts were recorded throughout the life of each fly. Thus the predictor Z is dietary intake and the functional response is the individual daily egg-laying profile, treated as a random trajectory. Since the predictor is univariate, we have $\eta(Z) = Z$ in the basic model (1).

We consider three functional models for these data. In the first model the mean egg-laying function $\mu_Y(t)$ is assumed to be always the same, irrespective of dose, i.e., within the framework of (1) we assume

(M1) *No covariate effects model:* $f(t, \eta(Z)) = \mu_Y(t)$, where $\mu_Y(\cdot)$ is a smooth function. In this model, the covariate Z has no effect.

(M2) *Functional smooth surface model:* $f(t, Z) = \mu(t, Z)$, where $\mu(\cdot, \cdot)$ is a 2-dimensional function of t and Z.

(M3) *Functional multiplicative effects model:* $f(t, \eta(Z)) = \mu_Y(t)\ \theta(Z)$, where $\mu_Y(\cdot)$ is a smooth function of time as in (M1) and $\theta(\cdot)$ is a nonparametric smooth function of the covariate function $\eta(Z)$, which has a multiplicative modulating effect such that $E\theta(Z) = 1$.

Model (M1) is the simplest and also most restrictive model, while (M2) is the most flexible. Model (M3), like model (M2), allows for a nonparametric impact of the covariate on the response function, but in a more restricted manner. For further details regarding these models and issues of estimation specific to these models, we refer to Chiou et al. (2003b, 2004). We use the above models to illustrate both the proposed graphical method and the test for functional goodness-of-fit that was proposed in the previous section.

Notably for model (M1), the FPC scores A_{Yk} of the response processes are the same as the residual FPC scores A_{Rk}, since the conditional expectation of the process is assumed not to depend on the covariate. As a goodness-of-fit check, we therefore directly assess the dependency of the response function on the covariate, the dose level Z, by plotting the first three FPC scores $A_{Yi1}, A_{Yi2}, A_{Yi3}$, $i = 1, \ldots, n$, against the dose levels Z_i, as shown in Figure 4, where $K_Y = 3$ for the number of components chosen by CPV criterion (CPV(2) = 83.4%, CPV(3) = 92.2%). One finds that the first FPC scores A_{Yi1} exhibit an obvious increasing trend with increasing dose level, indicating dependency of the response function on the covariate. This reveals model (M1) as overly simplistic. Functional residual plots for Models (M2) and (M3) are displayed in Figures 5 ($K_R = 2$ with CPV(1) = 95.7% and CPV(2) = 99.7%) and 6 ($K_R = 1$ with CPV(1) = 99.9%), respectively, where A_{Rik} are plotted against A_{Fik}. No obvious pattern for the dependency of the residual FPC scores on the fitted FPC scores can be discerned for either of these two models.

The proposed test can be used to further investigate these findings. We choose $L = 10$, i.e., each dietary dose level forms one bin, and $K_Y = 3$ for the number of random components. The approximate null distributions for the test statistics T_0 (20) are displayed in Figure 7, with the corresponding values of the test statistic and the p-values. The p-value of the test for Model (M1) is very small, indicating lack of fit for this simple model, and corroborating the graphical evidence. In contrast, the p-value of the test for Model (M2) is quite large, indicating no problems with lack of fit for this very flexible model. Similarly, the test does not provide evidence for lack of fit for Model (M3) with its multiplicative structure for the covariate effect on the overall mean function. Among the models considered, this diagnostic analysis therefore points to model (M3) as the overall best model, since it is simpler than model (M2), due to the structural constraint that is imposed by the multiplicative structure.

6 Discussion and concluding remarks

Residual processes are a useful tool for functional regression diagnostics. We reduce the high dimensionality of residual processes to a manageable level by reducing them to their first few functional principal component scores. These FPC scores can then be entered into functional residual plots that are similarly interpreted as classical residual plots. For example, one may check for dependency of the functional principal component scores of residual processes on fitted values or the scores of fitted processes. By forming eigenvalue-weighted averages one can combine various one-dimensional diagnostics such as hat matrix or Cook's distance to obtain functional versions.

A randomization test of model goodness-of-fit based on residual processes compares the variance of the means of residual functional principal scores in small bins that are constructed by ordering the data according to the predictor variable with the variance that would be obtained by a random ordering of the data. The resulting p-values provide an indication of the dependency of the functional residuals on the predictor. While we developed this test for the case of a one-dimensional predictor, it can be extended to functional and higher-order predictors by invoking a single index assumption so that the relevant predictor would be $\eta(Z)$, where η is a suitable single index. Alternatively, predictor processes can be represented by their functional principal component scores and the test statistics for each of those can be combined to an overall test statistics by simply adding the component statistics to an overall statistic.

If lack-of-fit is discovered, one may consider various extensions of a model. For example, in the case of model (M1), one could consider extensions that include a certain number S of additional functional components,

$$Y(t) - \hat{Y}(t) = \sum_{k=1}^{S} \eta_k(Z)\rho_{Y\,k}(t),$$

where the residual processes on the l.h.s. would be regressed on the predictor variable Z or the components of a functional predictor, similarly to the modeling of response processes in Chiou et al. (2003a). This is likely to lead to improved fits, at the expense of more complex models.

Table 1 summarizes the results of such a procedure for the medfly dose-response data. It is not surprising that the p-values for Models (M2) and (M3), when considering additional random components, move from large to very large, as already without additional components these models do not suffer from a lack of fit. In contrast, for Model (M1), which only contains an overall mean component, the p-value increases markedly from very small to large by adding just one random component to the model. We conclude that the proposed diagnostics are a useful addition to the toolkit of functional regression.

References

Anscombe, F.J. and Tukey, J.W. (1963). The examination and analysis of residuals. Technometrics **5**, 141-160.

Arbeitman, M.N., Furlong, E.E.M., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W. and White, K.P. (2002). Gene expression during the life cycle of *Drosophila melanogaster.* Science **297**, 72-83.

Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003). Testing hypotheses in the functional linear model. Scand. J. Statist. **30**, 241-255.

Cardot H. and Sarda P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. J. Multiv. Anal. **92**, 2441.

Carey, J.R., Liedo, P., Harshman, L., Zhang, Y., Müller, H.G., Partridge, L. and Wang, J.L. (2002). Life history response of Mediterranean fruit flies to dietary restriction. Aging Cell **1**, 140-148.

Chiou, J.M., Müller, H.G. and Wang, J.L. (2003a). Functional quasi-likelihood regression models with smooth random effects. J. Royal Statist. Soc. B **65**, 405-423.

Chiou, J.M., Müller, H.G., Wang, J.L. and Carey, J.R. (2003b). A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies. Statist. Sinica **13**, 1119-1133.

Chiou, J.M., Müller, H.G. and Wang, J.L. (2004). Functional response models. Statist. Sinica **14**, 675-693.

Cook, R.D. (1977). Detection of influential observations in linear regression. Technometrics **19**, 15-18.

Cook, R.D. and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, London.

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. Canadian J. Statist. **30**, 285-300.

Draper, N.R. and Smith, H. (1998). Applied Regression Analysis (Third Edition). Wiley, New York.

Escabias M., Aguilera, A.M. and Valderrama, M.J. (2004). Principal component estimation of functional logistic regression: Discussion of two different approaches. J. Nonpara. Statist. **16**, 365-384.

Faraway, J.J. (1997). Regression analysis for a functional response. Technometrics **39**, 254-261.

Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. J. Nonpara. Statist. **16**, 111-125.

Green, J.R. (1971). Testing departures from a regression without using replication. Technometrics **13**, 609-615.

Grenander, U. (1950). Stochastic processes and statistical inference. Arkiv för Matematik, 195-

276.

Hall, P. and Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. Biometrics **47**, 757-762.

Hart, J.D. (1997). Nonparametric Smoothing and Lack-of-Fit Tests. Springer, New York.

He, G., Müller, H.G., and Wang, J.L. (2000). Extending correlation and regression from multivariate to functional data. *Asymptotics in Statistics and Probability* (Edited by M. Puri), 197-210.

James, G.M. (2002). Generalized linear models with functional predictors. J. Royal Statist. Soc. B **64**, 411-432.

Liu, X. and Müller, H.G. (2003). Modes and clustering for time-warped gene expression profile data. Bioinformatics **19**, 1937-1944.

Müller, H.G. and Stadtmüller, U. (2005). Generalized functional linear models. Ann. of Statist. **33**, 774-805.

Müller, H.G. (2005). Functional modelling and classification of longitudinal data. Scand. J. Statist. **32**, 223-240.

Ramsay, J.O. and Dalzell, C.J. (1991). Some tools for functional data analysis. J. Royal Statist. Soc. B **53**, 539-572.

Ramsay, J.O. and Silverman, B.W. (2005). Functional Data Analysis. Springer, New York.

Rice, J.A. and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. J. Royal Statist. Soc. B **53**, 233-243.

Yao, F., Müller, H.G., Clifford, A.J., Dueker, S.R., Follett, J., Lin, Y., Buchholz, B. and Vogel, J.S. (2003). Shrinkage estimation for functional principal component scores, with application to the population kinetics of plasma folate. Biometrics **59**, 676-685.

Yao, F., Müller, H.G. and Wang, J.L. (2005a). Functional data analysis for sparse longitudinal data. J. Amer. Statist. Assoc. **100**, 577-590.

Yao, F., Müller, H.G. and Wang, J.L. (2005b). Functional linear regression analysis for longitudinal data. Ann. Statist., to appear.

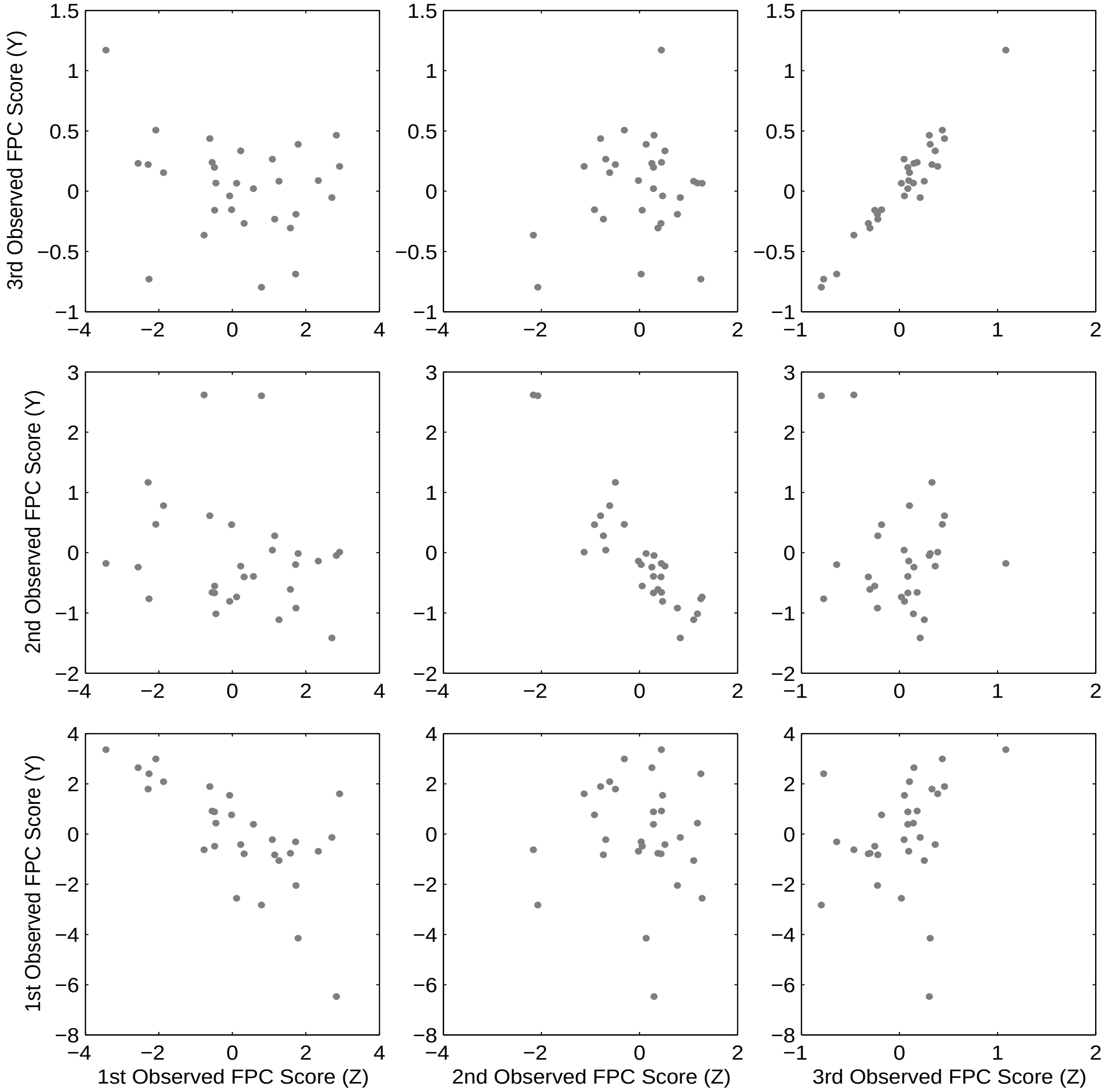

Figure 1: Plotting functional principal component scores of response processes versus those of predictor processes, choosing $K_Y = K_Z = 3$ components, for the *Drosophila* gene expression profile data.

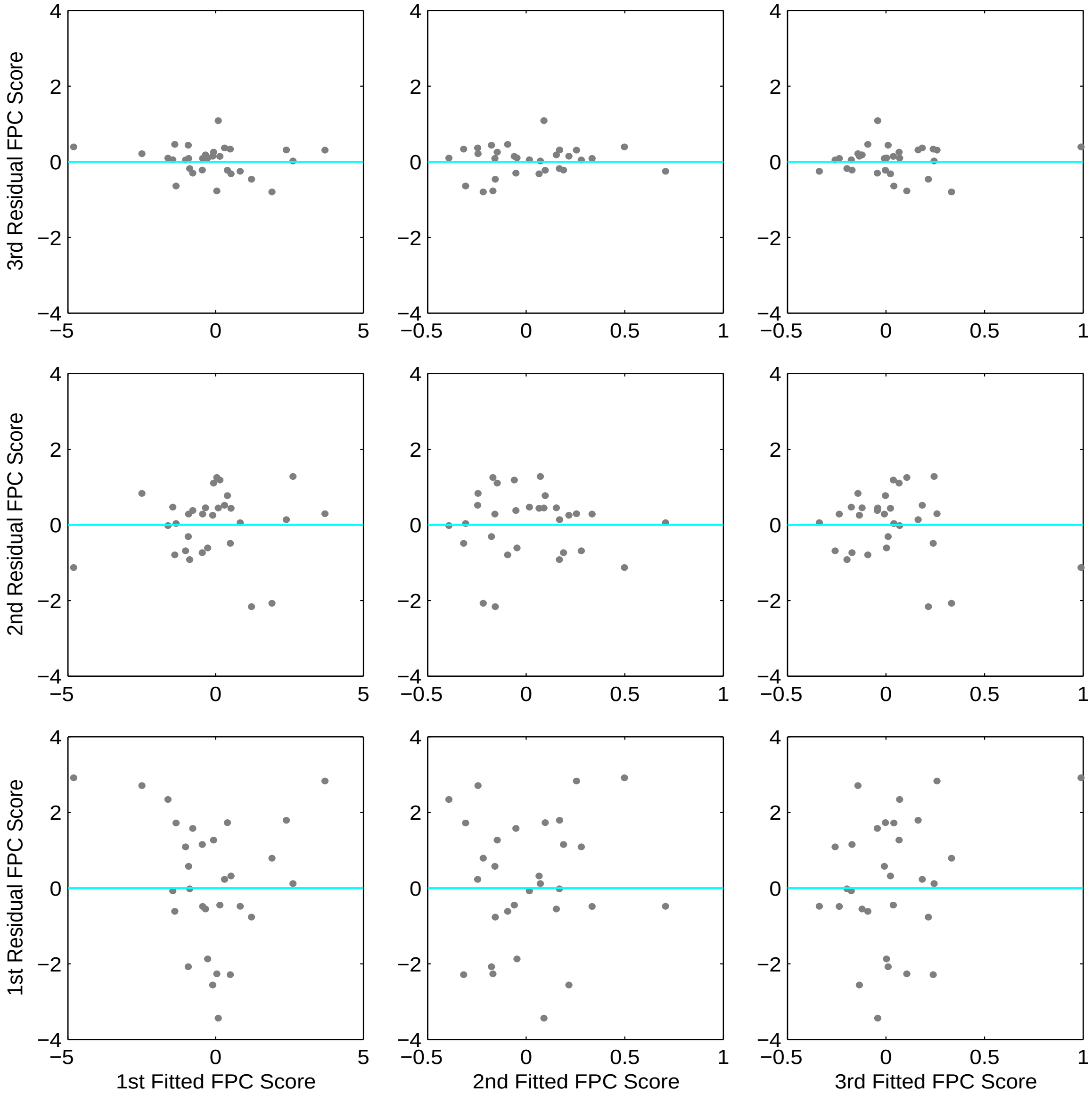

Figure 2: Functional residual plots of functional principal component scores of residual versus fitted processes for the *Drosophila* gene expression profile data.

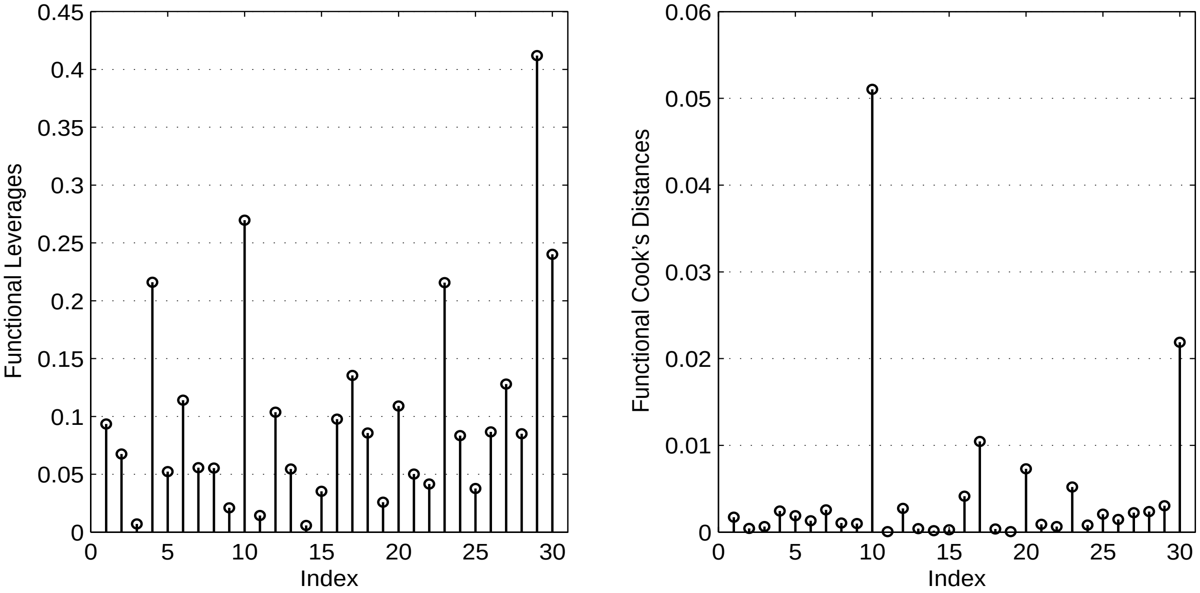

Figure 3: Functional leverages obtained as diagonal elements of functional hat matrix **H** (15) (left panel) and functional Cook's distances (16) of the predictor trajectories for the *Drosophila* gene expression profile data.

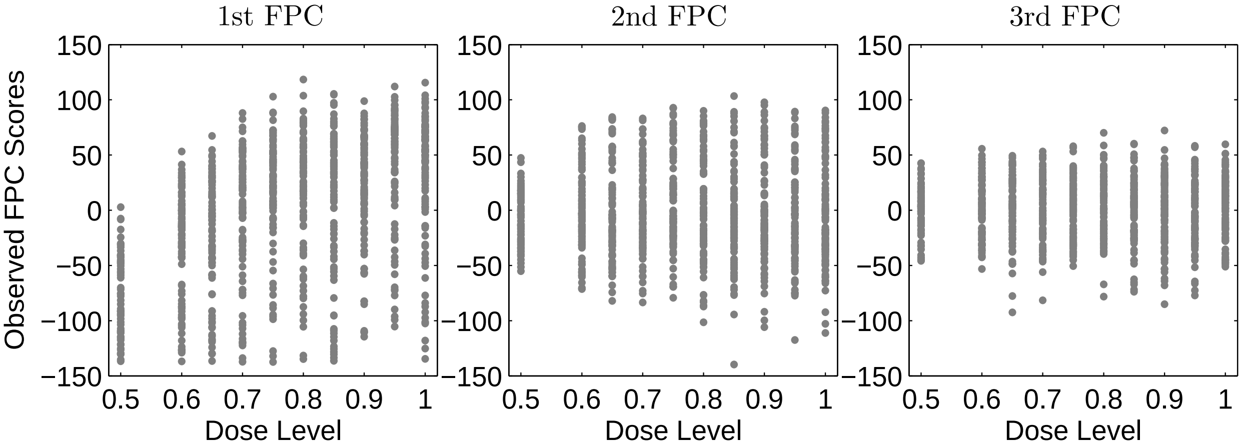

Figure 4: The first three observed FPC scores plotted against dietary dose level for Model (M1) (medfly dose-response data).

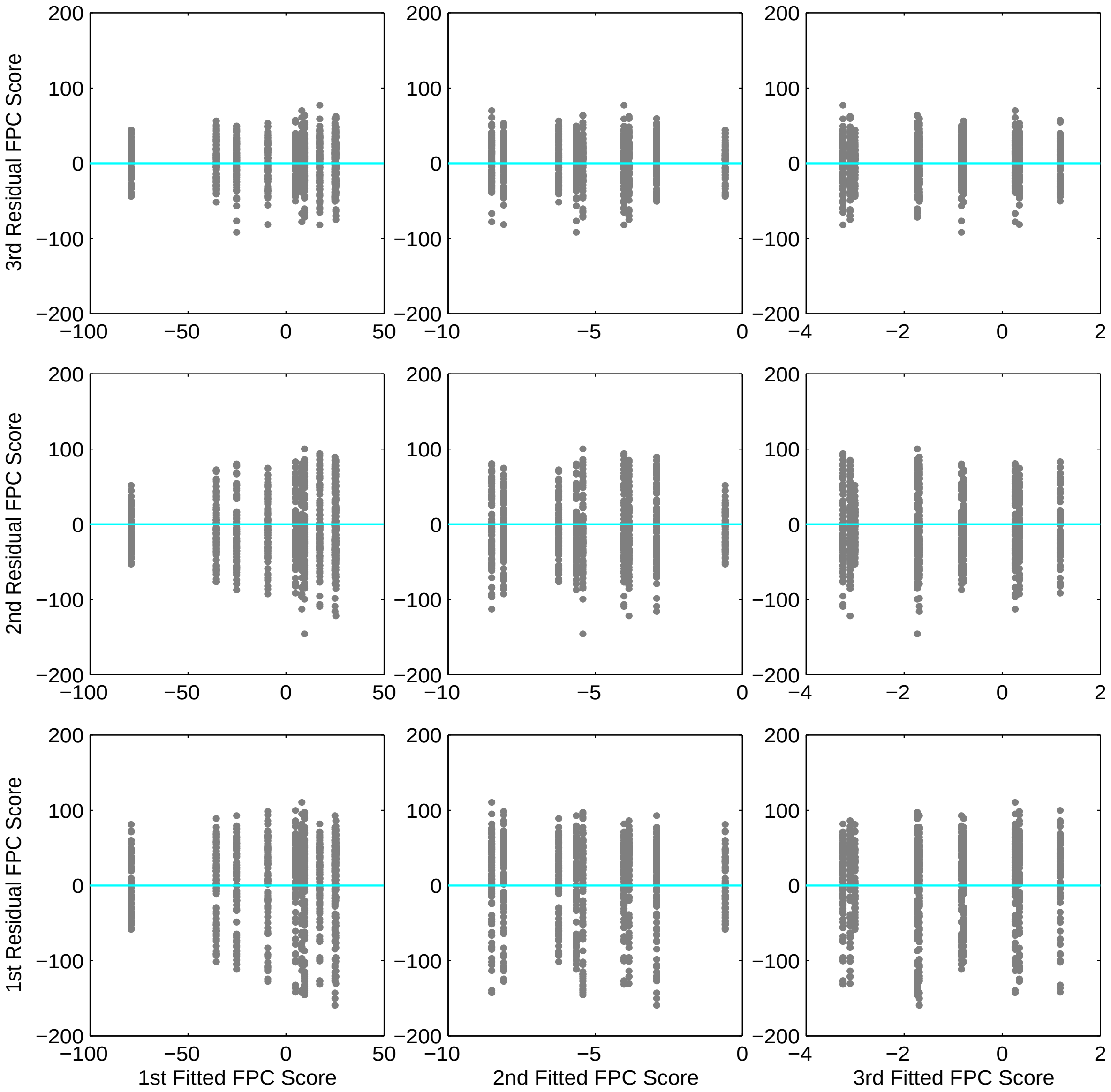

Figure 5: Functional residual plots for Model (M2) (medfly dose-response data).

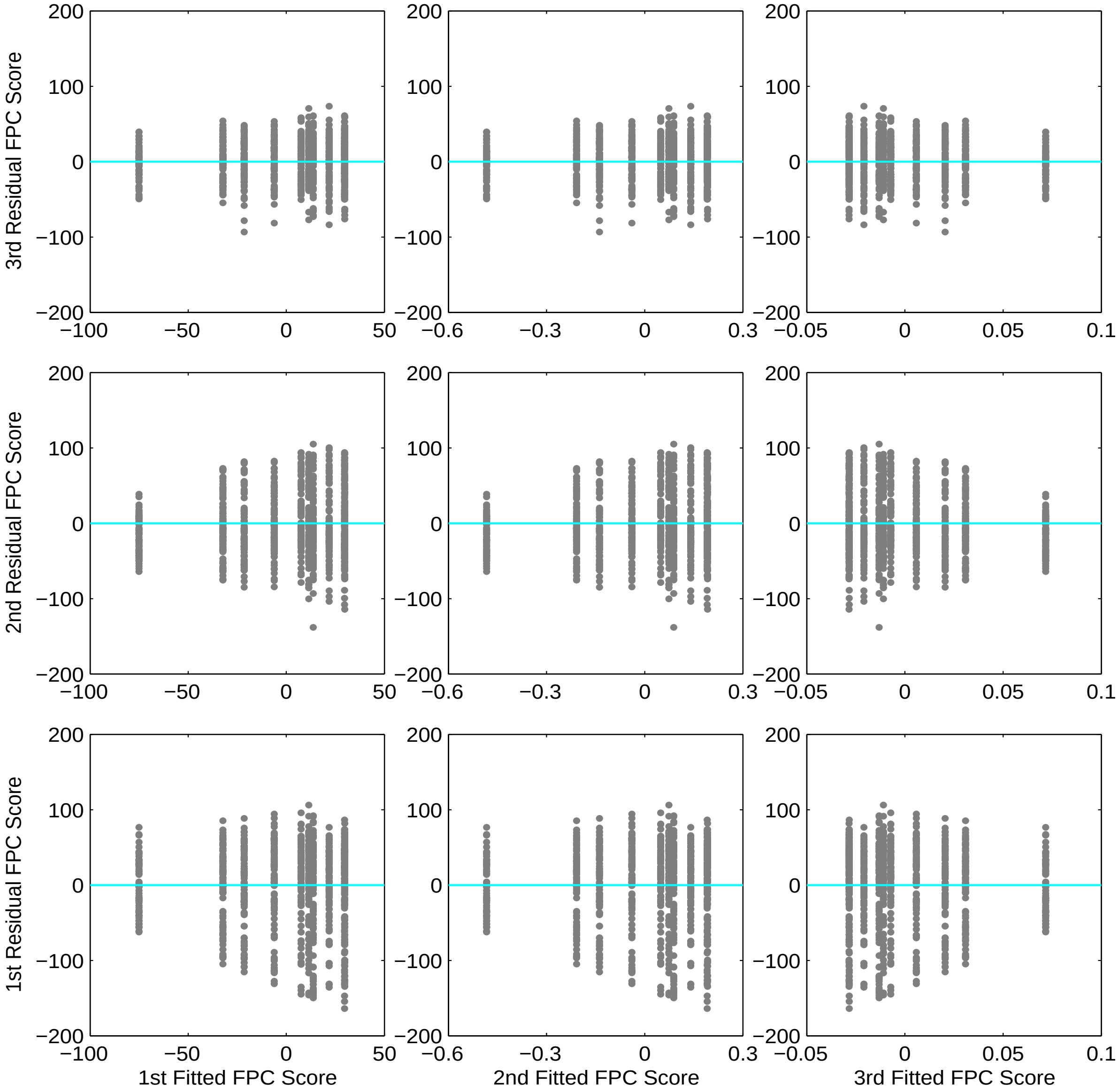

Figure 6: Functional residual plots for Model (M3) (medfly dose-response data).

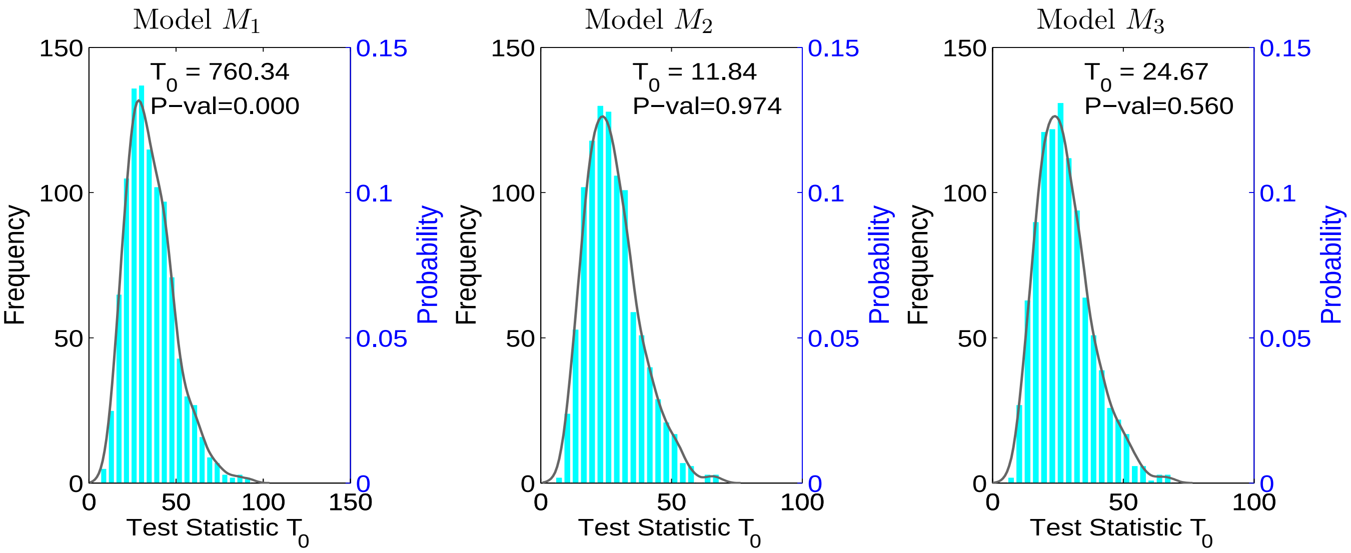

Figure 7: Approximate null distributions of the test statistics T_0 for models (M1)-(M3) (medfly dose-response data).

Table 1: Summary of p-values obtained for models (M1)-(M3) as S additional functional components are added to the respective model, for $S = 0, 1, 2, 3$, for the medfly dose-response data.

Model	$S = 0$	$S = 1$	$S = 2$	$S = 3$
(M1)	0.000	0.874	0.980	0.991
(M2)	0.979	1.000	1.000	1.000
(M3)	0.561	0.816	0.881	0.900