REGRESSION MODELING OF A PHYSIOLOGICAL PARAMETER AFFECTED BY BOTH INTRINSIC AND EXTRINSIC FACTORS

Peihua Qiu
School of Statistics
University of Minnesota

Abstract

We discuss regression modeling of a physiological parameter affected by both intrinsic and extrinsic factors. Such parameter commonly exists in our daily life (e.g., the sleeping quality of human beings or other animals). An additive regression model is suggested which consists of two parts. The first part is for explaining the effect of intrinsic factors and the second part is for describing the effect of extrinsic factors. The fitted model is proved to be statistically consistent. Hypothesis tests about the model coefficients are also discussed. Some simulation results are presented and the modeling procedure is applied to a rat sleep data set.

Key Words: Additive regression model; Circadian rhythm; Hypothesis tests; Kernel estimation; Least squares estimation; Piecewise polynomials; Strong consistency.

1 Introduction

Many physiological parameters are affected by both intrinsic and extrinsic factors. For example, our sleeping process is affected by the so-called circadian rhythm (an intrinsic factor, see e.g., Strogatz 1986). It is also affected by the lighting condition and other environmental conditions (extrinsic factors). This paper discusses statistical modeling of such parameters.

Suppose that the response variable is Y and the independent variable is x. Both of them are univariate. In the rat sleep example presented in Section 5, Y denotes the percentage of time in each 5-minute interval of a day that a rat is in sleep and x denotes the middle time point of each 5-minute interval (see Benca *et al.* (1993) for more background introduction about this example). When the extrinsic factors are in "normal" condition, Y is believed to be affected mainly by intrinsic factors. In such a case, the regression model for the physiological process is assumed to be

$$Y_i^* = f_1(x_i^*) + \varepsilon_i^*, \quad i = 1, 2, \ldots, n_1, \tag{1.1}$$

where $\{x_i^*\}$ are the design points which are the middle time points of 288 5-minute intervals of the day in the rat sleep example, $\{Y_i^*\}$ are observations of the physiological parameter at $\{x_i^*\}$, $\{\varepsilon_i^*\}$ are i.i.d. random errors with mean 0 and variance σ_*^2, and f_1 is an unknown regression function. For simplicity, we assume that the design points $\{x_i^*\}$ are from design space $[0,1]$.

When the extrinsic factors are at some specific "test" level, the regression model for the physiological process is assumed to be

$$Y_i = cf_1(x_i) + f_2(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n_2, \tag{1.2}$$

where design points $\{x_i\}$ could be different from $\{x_i^*\}$ used in (1.1) but they both come from the same design space $[0,1]$, c is a coefficient, f_1 is the same function as that used in (1.1), f_2 is a function that is different from f_1 and $\{\varepsilon_i\}$ are random errors with mean 0 and variance σ^2. $\{\varepsilon_i\}$ are assumed to be independent of $\{\varepsilon_i^*\}$.

In model (1.2), the first term cf_1 explains the effect of the intrinsic factors under the test condition. The coefficient c is related to the magnitude of this effect. For example, "$c = 0$" implies that the intrinsic factors have no effect on the related physiological parameter under the test condition. The second term f_2 denotes the effect of the extrinsic factors which is assumed to be additive to the effect of the intrinsic factors. Conventionally there is a constant term in a model like (1.2). For simplicity of presentation, this term has been included in the function f_2.

By the idea of local linear kernel estimation, it can be checked that the sample from model (1.2) does not help much in estimating f_1. Fortunately, it is often convenient to obtain observations under the normal condition of the extrinsic factors. So the sample size n_1 of the sample from the model (1.1) is often moderately large and consequently f_1 can be estimated well based on the model (1.1) alone. After replacing f_1 by its estimator $\widehat{f_1}$, the model (1.2) looks similar to a semiparametric model if f_2 is assumed to be nonparametric (see e.g., Chen 1988; Speckman 1988). But it is different from the conventional semiparametric model in the sense that the estimator $\widehat{f_1}$ depends on a sample from another model (i.e., from the model (1.1)). At this moment such model is still too complicated for us to study its statistical properties. As a first step, we assume that the function f_2 has the following parametric expression:

$$f_2(x) = a_0 + a_1 g_1(x) + a_2 g_2(x) + \cdots + a_p g_p(x), \tag{1.3}$$

where $\{a_i, i = 0, 1, \ldots, p\}$ are unknown coefficients, p is a known positive integer, and $\{g_i, i =$

$1, 2, \ldots, p\}$ are known base functions. In applications, the base functions can be simply chosen to be the power functions $\{g_i(x) = x^i, i = 1, 2, \ldots, p\}$. Based on the Weierstrass Theorem in mathematics (e.g., Stoll 2001, Chapter 8), any continuous function defined in [0,1] can be approximated uniformly by a series of polynomials. In the numerical examples presented in Sections 4 and 5, the order p is determined by the backward model selection procedure which will be further explained there.

The coefficients in model (1.2) can then be estimated by the ordinary least squares (LS) procedure (see e.g., Draper and Smith 1980; Seber 1977). Let us first introduce some matrix notation. Define

$$\widehat{X} = (\underline{\widehat{f}_1}, \underline{1}, \underline{g_1}, \ldots, \underline{g_p}), \quad \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_{n_2} \end{pmatrix}, \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{n_2} \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} c \\ a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix},$$

where $\underline{\widehat{f}_1}, \underline{g_1}, \ldots, \underline{g_p}$ are vectors obtained by evaluating the functions $\widehat{f}_1, g_1, \ldots, g_p$ at the design points $\{x_i\}$ and $\underline{1}$ denotes a vector of length n_2 with all elements equal to 1. Then the LS estimator of $\underline{\beta}$ is

$$\underline{\widehat{\beta}}(n_1, n_2) = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{Y} \tag{1.4}$$

Under some regularity conditions on the model (1.1) (see some related discussions in Section 2), it is not difficult to find an estimation procedure such that $\widehat{f}_1$ converges to f_1 almost surely. Let $\underline{\widehat{\beta}}(n_2) := \lim_{n_1 \to \infty} \underline{\widehat{\beta}}(n_1, n_2) = (X'X)^{-1}X'\underline{Y}$ where $X = (\underline{f_1}, \underline{1}, \underline{g_1}, \ldots, \underline{g_p})$. Then $\underline{\widehat{\beta}}(n_2)$ is an ordinary LS estimator of $\underline{\beta}$ when f_1 is known. Therefore under some regularity conditions on the models (1.1) and (1.2), we have

$$\lim_{n_2 \to \infty} \lim_{n_1 \to \infty} \underline{\widehat{\beta}}(n_1, n_2) = \underline{\beta}, \quad a.s. \tag{1.5}$$

However both (1.5) and $\underline{\widehat{\beta}}(n_2)$ are not much helpful in applications because it is not realistic to let $n_1 \to \infty$ first and then let $n_2 \to \infty$ second. A more useful result would be

$$\lim_{n_2 \to \infty, n_1 \to \infty} \underline{\widehat{\beta}}(n_1, n_2) = \underline{\beta}, \quad a.s. \tag{1.6}$$

In Section 2, we will sketch the estimation of f_1. Conditions on the models are given in Section 3 under which the equation (1.6) is true. The asymptotic normality of $\underline{\widehat{\beta}}(n_1, n_2)$ is also discussed

there which can be used for testing $H_0 : c = 0$ vs $H_a : c \neq 0$. In Section 4, some simulation results are presented. The modeling procedure suggested by this paper is applied to the rat sleep data set in Section 5. Some remarks conclude the article in Section 6.

Sine/cosine waves are traditionally used for describing circadian rhythms in sleeping (Pavlidis 1973; Strogatz 1986). But such models could not separate circadian rhythms from extrinsic factors such as the lighting condition. They can not incorporate abrupt shifts either which may occur at transitions between light and dark. Wang and Brown (1996) studied human circadian rhythms by suggesting a periodic spline model. But they did not consider separating the effect of the intrinsic factors from that of the extrinsic factors either in their model.

2 Estimation of f_1

If f_1 is linear, then it can be estimated by the LS procedure. The almost sure convergence rate of its LS estimator is $o(n^{-1/2} \log n)$ under some regularity conditions (see e.g., Lai and Robbins 1977; Wu 1980). In this paper we assume that f_1 is nonparametric. There are several nonparametric smoothing methods available in the literature for estimating f_1. These methods include the smoothing spline method (Wahba 1991), the locally weighted scatter plot smoothing method (Cleveland 1979), the kernel smoothing method (Härdle 1991)), the local polynomial kernel smoothing method (Fan and Gijbels 1996), and several others. For simplicity, the kernel smoothing method is used in this paper for estimating f_1.

Let $K(\cdot)$ be a kernel function with support $[-1/2, 1/2]$. The Nadaraya-Watson kernel estimator of $f_1(x)$ is defined by:

$$\widehat{f_1}(x) = \frac{1}{n_1 h_{n_1}} \sum_{i=1}^{n_1} K(\frac{x_i^* - x}{h_{n_1}}) Y_i^* \tag{2.1}$$

where h_{n_1} is a bandwidth parameter. From (2.1), $\widehat{f_1}(x)$ is a weighted average of the observations in a neighborhood of a given point x where the weights are determined by the kernel function.

In the literature, there are several data-driven bandwidth selection procedures including the plug-in procedures, the cross-validation procedure, the Mellow's C_p and the Akaike's information criterion (cf. Loader 1999). In some of the numerical examples presented in Section 4, the cross-

validation procedure is used for determining the bandwidth h_{n_1}. Let

$$CV(h_{n_1}) = \frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i^* - \widehat{f}_{1,-i}(x_i^*))^2$$

where $\widehat{f}_{1,-i}(\cdot)$ is the "leave-1-out" estimator of $f_1(\cdot)$. Namely, the observation (x_i^*, Y_i^*) is left out in constructing $\widehat{f}_{1,-i}(\cdot)$. Then the optimal value of h_{n_1} is chosen by minimizing $CV(h_{n_1})$.

3 Properties of the Fitted Model

Lemma 3.1 Let ν be a positive number and γ_{n_1} be a series of numbers which satisfy $\lim_{n_1\to\infty}\gamma_{n_1} = \infty$. In model (1.1), suppose that $f_1 \in Lip(\alpha)$, $\alpha > 0$; $E|\varepsilon_1^*|^\rho < \infty$, $\rho \geq 2$; and $\max_{1\leq i\leq n_1+1}|\frac{1}{n_1} - (x_i^* - x_{i-1}^*)| = O(n_1^{-1-\lambda})$, $\lambda > 0$, where $x_0^* = 0$ and $x_{n_1+1}^* = 1$. The kernel function $K(\cdot)$ is a non-negative bounded function which satisfies $\int_{-1/2}^{1/2} K(x)\,dx = 1$ and $K(\cdot) \in Lip(\beta)$, for some $\beta > 0$. The bandwidth h_{n_1} satisfies $\lim_{n_1\to\infty} h_{n_1} = 0$, $\lim_{n_1\to\infty} n_1 h_{n_1} = \infty$ and the following additional conditions (when n_1 is large enough) (1) $\frac{n_1^\nu}{\gamma_{n_1}\log n_1}[h_{n_1}^\alpha + \frac{1}{n_1^\lambda h_{n_1}} + \frac{1}{n_1^\beta h_{n_1}^{\beta+1}}] = o(1)$; (2) $\frac{n_1^{2\nu}}{n_1\gamma_{n_1}h_{n_1}} = O(1)$; and (3) $\frac{n_1^{\nu+1/\rho-1}}{h_{n_1}\gamma_{n_1}\log n_1} = o(1)$. Then

$$\lim_{n_1\to\infty}\frac{n_1^\nu}{\gamma_{n_1}\log n_1}\|\widehat{f}_1 - f_1\|_{[h_{n_1}/2,1-h_{n_1}/2]} = 0,\ a.s.,$$

where $\|f_1\|_\Omega$ denotes $\max_{x\in\Omega}|f_1(x)|$.

The above lemma establishes the strong consistency of the Nadaraya-Watson kernel estimator $\widehat{f}_1$, which was originally discussed by Cheng and Lin (1981) and Qiu (1994). The following lemma is its direct conclusion.

Lemma 3.2 In Lemma 3.1, if $\alpha = \beta = 1$, $\rho \geq 3, h_{n_1} = O(n_1^{-1/3}), \lambda \geq 2/3, \nu = 1/3$, then the conditions (1)-(3) are all satisfied and consequently under the other conditions we have $\|\widehat{f}_1 - f_1\|_{[h_{n_1}/2,1-h_{n_1}/2]} = O(n_1^{-1/3}\log(n_1)), a.s.$

Theorem 3.1 Besides the conditions stated in Lemma 3.1, suppose that $g_1, g_2, \cdots, g_p$ in (1.3) are known continuous functions and that there exists an integer $n_0 > 0$ such that the matrix X is of full rank when $n_2 > n_0$ where $X = (\underline{f_1}, \underline{1}, \underline{g_1}, \cdots, \underline{g_p})$. If $\lim_{n_1\to\infty, n_2\to\infty}\frac{n_2^2}{n_1^{1/2-\zeta}} = 0$ for some $0 < \zeta < 1/2$, then for $i = 1, 2, \cdots, p+2$,

$$\widehat{\beta}_i(n_1, n_2) - \beta_i = o(\sqrt{v_{ii}^{(n_2)}}\log(v_{ii}^{(n_2)})),\ a.s.,$$

where $v_{ii}^{(n_2)}$ is the i-th diagonal element of $(X'X)^{-1}$. Consequently, the equation (1.6) is true if $\lim_{n_2\to\infty} v_{ii}^{(n_2)} = 0$ for all $i = 1, 2, \cdots, p+2$.

Proof It is not difficult to check that

$$\begin{aligned}
\widehat{\underline{\beta}}(n_1, n_2) &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{Y} \\
&= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'(X\underline{\beta} + \underline{\varepsilon}) \\
&= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'(\widehat{X}\underline{\beta} + (X - \widehat{X})\underline{\beta} + \underline{\varepsilon}) \\
&= \underline{\beta} + (\widehat{X}'\widehat{X})^{-1}\widehat{X}'(\underline{\delta}, \underline{0}, \cdots, \underline{0})\underline{\beta} + (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{\varepsilon} \\
&= \underline{\beta} + (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{\delta}c + (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{\varepsilon} \qquad (3.1)
\end{aligned}$$

where $\underline{\delta} = (f_1(x_1) - \widehat{f}_1(x_1), \cdots, f_1(x_{n_2}) - \widehat{f}_1(x_{n_2}))'$ and $\underline{0}$ is a vector of length n_2 with all elements equal to 0. By Lemma 3.1, $\max_{1\le i\le n_2} |\delta_i| \le M n_1^{-1/2+\zeta}$ for some constants $M > 0$ and $0 < \zeta < 1/2$.

Now

$$\begin{aligned}
\widehat{X}'\widehat{X} &= \begin{pmatrix}
\underline{x_1}'\underline{x_1} + 2\underline{x_1}'\underline{\delta} + \underline{\delta}'\underline{\delta} & \underline{x_1}'\underline{x_2} + \underline{\delta}'\underline{x_2} & \cdots & \underline{x_1}'\underline{x_{p+2}} + \underline{\delta}'\underline{x_{p+2}} \\
\underline{x_1}'\underline{x_2} + \underline{\delta}'\underline{x_2} & \underline{x_2}'\underline{x_2} & \cdots & \underline{x_2}'\underline{x_{p+2}} \\
\vdots & \vdots & & \vdots \\
\underline{x_1}'\underline{x_{p+2}} + \underline{\delta}'\underline{x_{p+2}} & \underline{x_2}'\underline{x_{p+2}} & \cdots & \underline{x_{p+2}}'\underline{x_{p+2}}
\end{pmatrix} \\
&\triangleq \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}
\end{aligned}$$

where $\underline{x_i}$ denotes the i-th column of the matrix X and the last equation defines a matrix partition with A_{11} the submatrix of dimension 1×1. Then

$$(\widehat{X}'\widehat{X})^{-1} = \begin{pmatrix}
A_{11.2}^{-1} & -A_{11.2}^{-1}A_{12}A_{22}^{-1} \\
-A_{11.2}^{-1}A_{22}^{-1}A_{21} & A_{22}^{-1} + A_{22}^{-1}A_{21}A_{11.2}^{-1}A_{12}A_{22}^{-1}
\end{pmatrix}$$

where $A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$.

It is quite obvious that

$$\begin{aligned}
A_{11.2}^{-1} &= \frac{1}{A_{11} - A_{12}A_{22}^{-1}A_{21}} \\
&= \frac{1}{B_{11} - B_{12}B_{22}^{-1}B_{21} + 2\underline{\delta}'\underline{x_1} + \underline{\delta}'\underline{\delta} - 2C_{12}A_{22}^{-1}B_{21} - C_{12}A_{22}^{-1}C_{12}'} \qquad (3.2)
\end{aligned}$$

where the matrices B's are submatrices obtained by partitioning $X'X$ in the same way as we partitioned $\widehat{X}'\widehat{X}$ and $C_{12} = (\underline{\delta}'\underline{x_2}, \cdots, \underline{\delta}'\underline{x_{p+2}})$. Obviously, $A_{12} = B_{12} + C_{12}$.

By using the Chebyshev's inequality,

$$C_{12}A_{22}^{-1}B_{21} \leq \sqrt{(C_{12}A_{22}^{-1}C_{12}') \cdot (B_{12}A_{22}^{-1}B_{21})} \tag{3.3}$$

Suppose that the singular value decomposition of A_{22}^{-1} is

$$A_{22}^{-1} = Q\ diag(\lambda_2, \cdots, \lambda_{p+2})\ Q'$$

where λ's are eigenvalues and the columns of Q are the corresponding eigenvectors of A_{22}^{-1}. Then

$$C_{12}A_{22}^{-1}C_{12}' \leq (\max_{2\leq j\leq p+2}\lambda_j) \cdot (p+1)M_1^2 \cdot \frac{n_2^2}{n_1^{1-2\zeta}} \tag{3.4}$$

In the above equation, we use the fact that the absolute value of each element of C_{12} is less than or equal to $M_1\frac{n_2}{n_1^{1/2-\zeta}}$ for some positive constant M_1. Similarly,

$$B_{12}A_{22}^{-1}B_{21} \leq (\max_{2\leq j\leq p+2}\lambda_j) \cdot (p+1)M_2^2 n_2^2 \tag{3.5}$$

for some positive constant M_2. After combining (3.3)-(3.5), we have $C_{12}A_{22}^{-1}B_{21} = o(1),\ a.s.$, and $C_{12}A_{22}^{-1}C_{12}' = o(1),\ a.s.$ So from (3.2),

$$A_{11.2}^{-1} = B_{11.2}^{-1} + o(\Delta),\quad a.s.,$$

where $B_{11.2} = B_{11} - B_{12}B_{22}^{-1}B_{21}$ and $\Delta = \frac{n_2^2}{B_{11.2}^2 n_1^{1/2-\zeta}}$. Similarly, we can prove that

$$(\widehat{X}'\widehat{X})^{-1} = (X'X)^{-1} +$$
$$\begin{pmatrix} o(\Delta), & o(\Delta)B_{12}B_{22}^{-1} + B_{11.2}^{-1}C_{12}B_{22}^{-1} + o(\Delta)C_{12}B_{22}^{-1} \\ o(\Delta)B_{22}^{-1}B_{21} + B_{11.2}^{-1}B_{22}^{-1}C_{21} + o(\Delta)B_{22}^{-1}C_{21}, & B_{22}^{-1}C_{21}B_{11.2}^{-1}B_{12}B_{22}^{-1} + \cdots + B_{22}^{-1}C_{21}o(\Delta)C_{12}B_{22}^{-1} \end{pmatrix}.$$

It is not hard to check that the $(1,1)$ element of $[(\widehat{X}'\widehat{X})^{-1} - (X'X)^{-1}](X'X)$ is

$$o(\Delta)B_{11} + o(\Delta)B_{12}B_{22}^{-1}B_{21} + B_{11.2}^{-1}C_{12}B_{22}^{-1}B_{21} + o(\Delta)C_{12}B_{22}^{-1}B_{21}.$$

It can be seen that each term of the above expression tends to zero by the fact that the functions $f_1, g_1, g_2, \cdots, g_p$ are all bounded in $[0,1]$ and each element of C_{12} is of order $O(n_2/n_1^{1/2-\zeta})$. Similarly, we can prove that

$$(\widehat{X}'\widehat{X})^{-1} = (X'X)^{-1} + o((X'X)^{-1}),\quad a.s., \tag{3.6}$$

where $o((X'X)^{-1})$ means that each element of $o((X'X)^{-1})(X'X)$ tends to 0 when $n_1, n_2 \to \infty$.

It is not hard to check that

$$\lim_{n_1\to\infty, n_2\to\infty} \underline{\delta}'\widehat{X}(X'X)^{-2}\widehat{X}'\underline{\delta} = 0,\quad a.s.$$

By this equation and the equation (3.6), we have

$$\lim_{n_1\to\infty,n_2\to\infty} \underline{\delta}'\widehat{X}(\widehat{X}'\widehat{X})^{-2}\widehat{X}'\underline{\delta} = 0, \quad a.s.$$

Therefore

$$\lim_{n_1\to\infty,n_2\to\infty} (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{\delta}c = 0, \quad a.s. \tag{3.7}$$

Now

$$(\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{\varepsilon} = (\widehat{X}'\widehat{X})^{-1}X'\underline{\varepsilon} - (\widehat{X}'\widehat{X})^{-1}(\underline{\delta}'\underline{\varepsilon}, 0, \cdots, 0)', \quad a.s. \tag{3.8}$$

It is not difficult to check that

$$\lim_{n_1\to\infty,n_2\to\infty} (\widehat{X}'\widehat{X})^{-1}(\underline{\delta}'\underline{\varepsilon}, 0, \cdots, 0)' = 0, \quad a.s. \tag{3.9}$$

and

$$(\widehat{X}'\widehat{X})^{-1}X'\underline{\varepsilon} = (X'X)^{-1}X'\underline{\varepsilon} + [(\widehat{X}'\widehat{X})^{-1} - (X'X)^{-1}]X'\underline{\varepsilon} \tag{3.10}$$

By the results in Lai, Robbins and Wei (1979), the i-th element of the first term in the right hand side of (3.10) is of order $o(\sqrt{v_{ii}^{(n_2)}}\log(v_{ii}^{(n_2)}))$, $a.s.$ By (3.6), the second term is of higher order. Hence the i-th element of $(\widehat{X}'\widehat{X})^{-1}X'\underline{\varepsilon}$ is of order $o(\sqrt{v_{ii}^{(n_2)}}\log(v_{ii}^{(n_2)}))$, $a.s.$ It is not difficult to check that the convergence rates of (3.7) and (3.9) are both of higher orders. These facts along with (3.1) give us the results of the theorem.

Remark 3.1 If the functions $g_1, g_2, \cdots, g_p$ are all continuous and the functions $\{1, f_1, g_1, \cdots, g_p\}$ are linearly independent in the sense that there does not exist constants $k_1, k_2, \cdots, k_{p+2}$ in R^1 such that $k_1 + k_2 f_1(x) + \sum_{i=1}^{p} k_{i+2}g_i(x) = 0$ for all $x \in R^1$, then it could be checked that the conditions stated in the theorem on X are all satisfied.

Theorem 3.2 Under the conditions stated in Theorem 3.1, $\widehat{\underline{\beta}}(n_1, n_2)$ is also a L^2 consistent estimator of $\underline{\beta}$ and $\sqrt{n_2}(\widehat{\underline{\beta}}(n_1, n_2) - \underline{\beta})$ is asymptotically normally distributed with mean $\underline{0}$ and covariance matrix $\sigma^2 I_{p+2}$.

Proof It is not hard to see that

$$E(\widehat{\underline{\beta}}(n_1, n_2)|Y_i^*, i = 1, 2, \cdots, n_1) = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'X\underline{\beta}$$

$$Var(\widehat{\underline{\beta}}(n_1, n_2)|Y_i^*, i = 1, 2, \cdots, n_1) = \sigma^2(\widehat{X}'\widehat{X})^{-1}$$

From the proof of Theorem 3.1,

$$\lim_{n_1\to\infty,n_2\to\infty} E(\widehat{\underline{\beta}}(n_1,n_2)|Y_i^*, i=1,2,\cdots,n_1) = \underline{\beta}, \quad a.s.$$

and

$$\lim_{n_1\to\infty,n_2\to\infty} Var(\widehat{\underline{\beta}}(n_1,n_2)|Y_i^*, i=1,2,\cdots,n_1) = 0, \quad a.s.$$

On the other hand, we can check that there exists a positive integer N_1 such that when $n_1 > N_1$

$$|E(\widehat{\underline{\beta}}(n_1,n_2)|Y_i^*, i=1,2,\cdots,n_1)| \le 2|\underline{\beta}|, \quad a.s.$$

and

$$|Var(\widehat{\underline{\beta}}(n_1,n_2)|Y_i^*, i=1,2,\cdots,n_1)| \le \sigma^2 \underline{1}\underline{1}', \quad a.s.$$

where $|A|$ denotes $(|a_{ij}|)$ for a matrix $A=(a_{ij})$ and $A \le B = (b_{ij})$ implies $a_{ij} \le b_{ij}$ for all i and j. By using the dominated convergence theorem (see e.g., Chapter 2, Loève, 1977), we have

$$\lim_{n_1\to\infty,n_2\to\infty} E(\widehat{\underline{\beta}}(n_1,n_2)) = \lim_{n_1\to\infty,n_2\to\infty} E(E(\widehat{\underline{\beta}}(n_1,n_2)|Y_i^*, i=1,2,\cdots,n_1)) = \underline{\beta}$$

and

$$\begin{aligned} & \lim_{n_1\to\infty,n_2\to\infty} Var(\widehat{\underline{\beta}}(n_1,n_2)) \\ = & \lim_{n_1\to\infty,n_2\to\infty} E(Var(\widehat{\underline{\beta}}(n_1,n_2)|Y_i^*, i=1,2,\cdots,n_1)) + \\ & \lim_{n_1\to\infty,n_2\to\infty} Var(E(\widehat{\underline{\beta}}(n_1,n_2)|Y_i^*, i=1,2,\cdots,n_1)) \\ = & 0 \end{aligned}$$

Thus, $\widehat{\underline{\beta}}(n_1,n_2)$ is L^2 consistent.

From (3.1), (3.8) and (3.10), we know that

$$\begin{aligned} & \sqrt{n_2}(\widehat{\underline{\beta}}(n_1,n_2) - \underline{\beta}) \\ = & \sqrt{n_2}\left((X'X)^{-1}X'\underline{\varepsilon} + (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{\delta}c - (\widehat{X}'\widehat{X})^{-1}(\underline{\delta}'\underline{\varepsilon}, 0, \cdots, 0)' + [(\widehat{X}'\widehat{X})^{-1} - (X'X)^{-1}]X'\underline{\varepsilon}\right). \end{aligned}$$

Similar to (3.7) and (3.9), it is not difficult to check that

$$\lim_{n_1\to\infty,n_2\to\infty} \sqrt{n_2}(\widehat{X}'\widehat{X})^{-1}\widehat{X}'\underline{\delta}c = \lim_{n_1\to\infty,n_2\to\infty} \sqrt{n_2}(\widehat{X}'\widehat{X})^{-1}(\underline{\delta}'\underline{\varepsilon}, 0, \cdots, 0)' = 0, \quad a.s.$$

By (3.6) and the fact that $\sqrt{n_2}(X'X)^{-1}X'\underline{\varepsilon}$ is asymptotically normally distributed with mean $\underline{0}$ and covariance matrix $\sigma^2 I_{p+2}$,

$$\begin{aligned} & \lim_{n_1\to\infty,n_2\to\infty} \sqrt{n_2}[(\widehat{X}'\widehat{X})^{-1} - (X'X)^{-1}]X'\underline{\varepsilon} \\ = & \lim_{n_1\to\infty,n_2\to\infty} [(\widehat{X}'\widehat{X})^{-1} - (X'X)^{-1}](X'X)[\sqrt{n_2}(X'X)^{-1}X'\underline{\varepsilon}] = 0, \quad \text{in probability.} \end{aligned}$$

Therefore, $\sqrt{n_2}(\widehat{\underline{\beta}}(n_1, n_2) - \underline{\beta})$ is asymptotically normally distributed with mean $\underline{0}$ and covariance matrix $\sigma^2 I_{p+2}$.

4 Simulation Study

In this section, we present some simulation results. In models (1.1) and (1.2), it is assumed that $f_1(x) = 3\sin(2\pi x)$, $x \in [0,1]$; $f_2(x) = -3 - 2x + 8x^2, x \in [0,1]$; $c = 1$; $\varepsilon_1^* \sim N(0, \sigma_*^2)$ and $\varepsilon_1 \sim N(0, \sigma^2)$. All design points are equally spaced in the design space $[0,1]$.

Figure 4.1(a) shows the function f_1 and a set of observations from the model (1.1) with $\sigma_* = 1$ and $n_1 = 1000$. The regression function $f_1 + f_2$ and a set of observations from the model (1.2) are shown in Figure 4.1(b) with $\sigma = 1$ and $n_2 = 500$. The function $f_1(x)$ is then estimated by the procedure (2.1) with the Epanechnikov kernel function $K(x) = \frac{12}{11}(1 - x^2)I_{\{x \in [-.5,.5]\}}$ (see e.g., Härdle 1991, page 45) and a bandwidth h_{n_1} determined by the cross-validation procedure. The estimator of $f_1(x)$ is shown in plot (c) with a dotted curve. In (1.3), the base functions $g_i(x) = x^i$, $i = 1, 2, \ldots, p$, are used for approximating $f_2(x)$, where p is determined by the backward model selection procedure with its initial value 10 (which is often good enough in applications based on our experience), a significance level 0.15 (which is default in most statistical softwares like SAS), and the hierarchy principal (namely, if x^k is included in the model, then all lowers of x should be included in the model). The estimated $f_2(x)$ is $\widehat{f}_2(x) = -3.224 - 1.769x + 8.163x^2$ which is shown in plot (d), and $\widehat{c} = 1.088$.

Next we study the effect of the sample sizes n_1 and n_2 on the behavior of the estimators. In the above example, let $\sigma = \sigma_* = 1$, n_1 vary among $10^2, 20^2, \cdots, 200^2$, and n_2 take the values of $10^2, 40^2$ or 70^2. To simplify the computation, the bandwidth h_{n_1} is chosen to be $0.2n_1^{-1/5}$ based on the following considerations. According to Härdle (1991, page 135), $\widehat{f}_1(x)$ could reach the optimal mean squared error if $h_{n_1} = O(n_1^{-1/5})$. The number 0.2 before $n_1^{-1/5}$ is an adjustment factor which makes the bandwidth reasonably large when n_1 is small to moderate. For each combination of n_1 and n_2, the simulation is repeated 100 times. After each simulation, the difference between $\widehat{\underline{\beta}}(n_1, n_2)$ and $\underline{\beta}$, which is measured by the Euclidean distance between the two vectors, is recorded. The averaged difference based on the 100 replications is displayed in Figure 4.2(a). From the plot, several conclusions can be made. First, for a fixed value of n_2, the difference between $\widehat{\underline{\beta}}(n_1, n_2)$ and

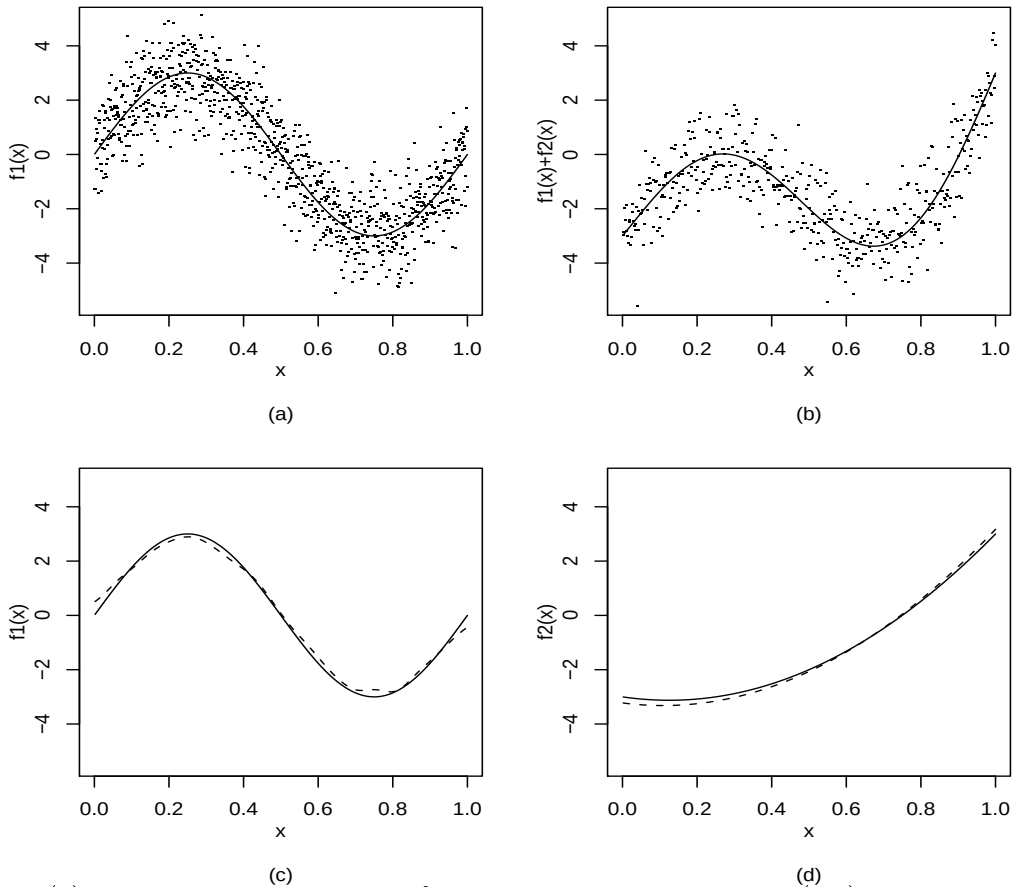

Figure 4.1: (a) The regression function f_1 and a sample from model (1.1) with $n_1 = 1000$ and $\sigma_* = 1$. (b) The regression function $f_1 + f_2$ and a sample from model (1.2) with $n_2 = 500$ and $\sigma = 1$. (c) The dotted curve represents the kernel estimator of f_1 and the solid curve is f_1. (d) The dotted curve represents the LS estimator of f_2 and the solid curve denotes f_2.

$\underline{\beta}$ first decreases quite dramatically and then becomes stable when n_1 increases, which implies that for a fixed sample size in model (1.2), moderately large sample from model (1.1) is helpful for model fitting but it is not necessary to acquire very large sample from model (1.1). Second, the curve with larger n_2 is below the one with smaller n_2, which demonstrates the statistical consistency of the fitted models as discussed in Section 3.

Then the values of n_1 and n_2 are switched and other quantities are kept unchanged. The corresponding results are shown in Figure 4.2(b). It can be seen that this plot has a similar pattern to that of Figure 4.2(a), which implies that the sample sizes of the models (1.1) and (1.2) are equally important for estimating $\underline{\beta}$. In this example, the value of c and the coefficients of f_2 are comparable. So it is reasonable to put the same weights on the first and the remaining components of $\underline{\widehat{\beta}}(n_1, n_2)$ in computing the Euclidean distance between $\underline{\widehat{\beta}}(n_1, n_2)$ and $\underline{\beta}$. If the values of c and the

coefficients of f_2 are not comparable in some cases, then it might be more reasonable to consider the distance between the first components of $\underline{\widehat{\beta}}(n_1, n_2)$ and $\underline{\beta}$ and the distance between the remaining components of the two vectors separately.

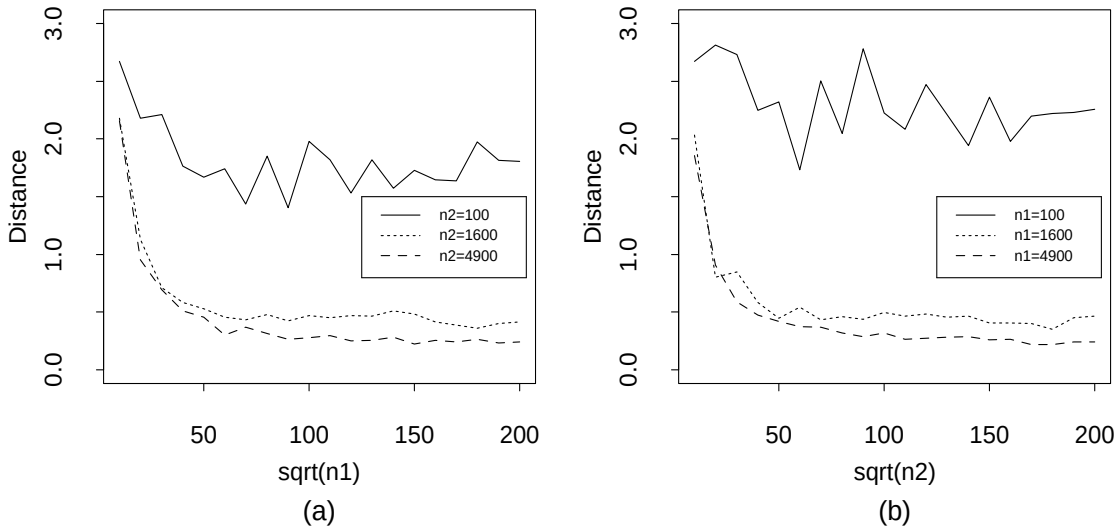

Figure 4.2: The averaged Euclidean distance between $\underline{\widehat{\beta}}(n_1, n_2)$ and $\underline{\beta}$ based on 100 replications. (a) The sample size n_2 varies among $10^2, 40^2$ and 70^2 and n_1 takes the values of $10^2, 20^2, \ldots, 200^2$; (b) the values of n_1 and n_2 in plot (a) are switched and other quantities are unchanged.

Next we study the sensitivity of the estimated parameters to the noise levels in models (1.1) and (1.2). In the above example, the sample sizes are fixed at $(n_1, n_2) = (1000, 500)$. We let the standard deviations σ_* and σ both vary among .2, .4, .6, .8 and 1. The averaged Euclidean distance between $\underline{\widehat{\beta}}(n_1, n_2)$ and $\underline{\beta}$ based on 100 replications is shown in Figure 4.3. The plot shows that the accuracy of the estimated parameters is as sensitive to σ_* as to σ.

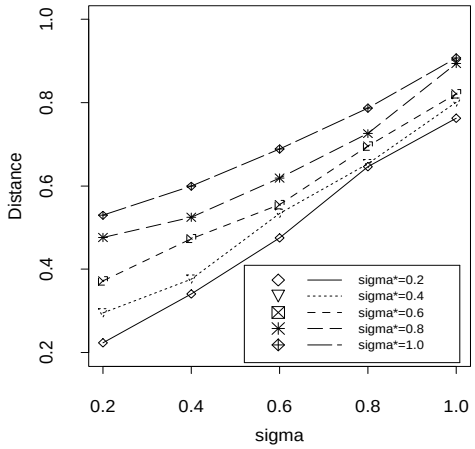

Figure 4.3: The Euclidean distance between $\underline{\widehat{\beta}}(n_1, n_2)$ and $\underline{\beta}$ when σ_* and σ both vary among .2, .4, .6, .8 and 1.

In Section 3, we established the asymptotic normality of $\widehat{\underline{\beta}}(n_1, n_2)$. In finite sample case, it might be natural to expect that the distribution of $\widehat{\underline{\beta}}(n_1, n_2)$ is close to $N(\underline{\beta}, \sigma^2(X'X)^{-1})$. Next we investigate the distribution of $\widehat{c}$ (which is the first component of $\widehat{\underline{\beta}}(n_1, n_2)$) by repeating the algorithm 1000 times with $\sigma_* = \sigma = 1$, $n_1 = 1000$ and $n_2 = 500$. The histogram of the 1000 $\widehat{c}$ values is displayed in Figure 4.4, which looks bell-shaped. The sample mean and sample variance of these $\widehat{c}$ values are 1.0103 and .0017, respectively, which are quite close to their true values $c = 1.0$ and $\sigma^2 v_{11}^{(n_2)} = .0012$.

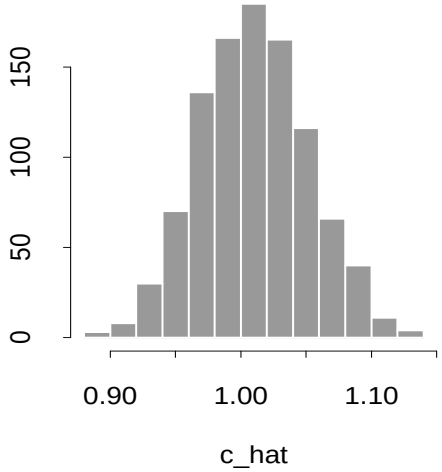

Figure 4.4: The histogram of $\widehat{c}$ values based on 1000 replications with $\sigma_* = \sigma = 1$, $n_1 = 1000$ and $n_2 = 500$.

5 An Application

In this section we apply the method discussed in the previous sections to a rat sleep data set. In a brainwave experiment, one group of eight Brown Norway rats were exposed to the normal lighting condition which was simply 12 hours on beginning at midnight followed by 12 hours off beginning at noon. Another group of eight rats of the same strain were exposed to the test lighting condition which changed conditions every 3 hours beginning with light at midnight, then darkness at 3:00 am, and so forth. For each rat, the percentage of time in sleep state in each 5-minute interval of a day was recorded by a computer after its brainwaves were analysed. By following the convention of the psychiatric literature, the investigators then averaged the data over the group of rats exposed to the same lighting condition.

The normal lighting condition with 24-hour cycle was used to approximate the daily natural

lighting under which the circadian rhythm is believed to play a central role in animal sleep. The test condition with 6-hour cycle was used to add a stimula to rats' sleeping environment. An important question here is whether or not the effect of the circadian rhythm is still statistically significant under the test lighting condition.

Rats are nocturnal animals. They tend to go to sleep when exposed to light and become active in dark environment. This can be seen from Figure 5.1(a) in which the data under the normal lighting condition are shown. From the plot and our experience, there is an obvious discontinuity in the data at the time when the light was switched (i.e., time=12 in the plot). Therefore in model (1.1), f_1 is assumed to be piecewisely continuous with a possible jump at time=12. It is then fitted in two separated regions [0,12) and [12,24] by using the Splus function ksmooth() with a default bandwidth. The estimated f_1 is shown in Figure 5.1(a) by the solid curves.

The data under the test lighting condition are shown in Figure 5.1(b). In model (1.2), we assume that f_2 is a periodic fuction with period equal to 6 hours. In each period, it is assumed to be piecewise polynomial with a possible discontinuity in the middle of the interval. The orders of the polynomials are determined by the backward model selection procedure with an initial order 10, a significance level 0.15 and the hierarchy principal, as mentioned in Section 4. The estimated orders of the polynomials in the first and second halfs of each period are 5 and 2, respectively.

We present $\widehat{c}\widehat{f}_1$ in plot (c) and $\widehat{f}_2$ in plot (d). Then the fitted model of (1.2) (namely, $\widehat{c}\widehat{f}_1 + \widehat{f}_2$) is shown in plot (b) by the solid curves. The standardized residuals of the fitted model of (1.2) are presented in plot (e). It can be seen that overall the model (1.2) fits the data reasonably well. Several residuals at the beginning of the first and second halfs of the first 6-hour period are quite large though, the reason of which is unknown to the scientists yet.

We then check for the autocorrelation of the residuals presented in plot (e). The Akaike's Information Criterion (AIC) values of the autocorrelation models of various orders are displayed in plot (f) (please notice that the x-axis denotes the order of the autocorrelation model + 1, which is a convention in Splus). Based on this plot, the first-order autocorrelation model seems appropriate for modeling the autocorrelation of the residuals. The estimate of its autocorrelation coefficient is 0.6349. If we want to take the existing autocorrelation into account in the model fitting, then the LS estimate (1.4) needs to be replaced by a generalized least squares (GLS) estimator which depends on the estimated autocorrelation coefficient (see e.g., Judge *et al.* 1980, Chapter 5). Generally

speaking, the LS estimator is still an unbiased estimator when autocorrelation in the error exists. But it will lose some efficiency compared to the GLS estimator. We computed the GLS estimator for this data and found that the difference between the fitted models of (1.2) with and without the autocorrelation consideration was too small to be visually noticeable. Therefore this part of the results are not presented.

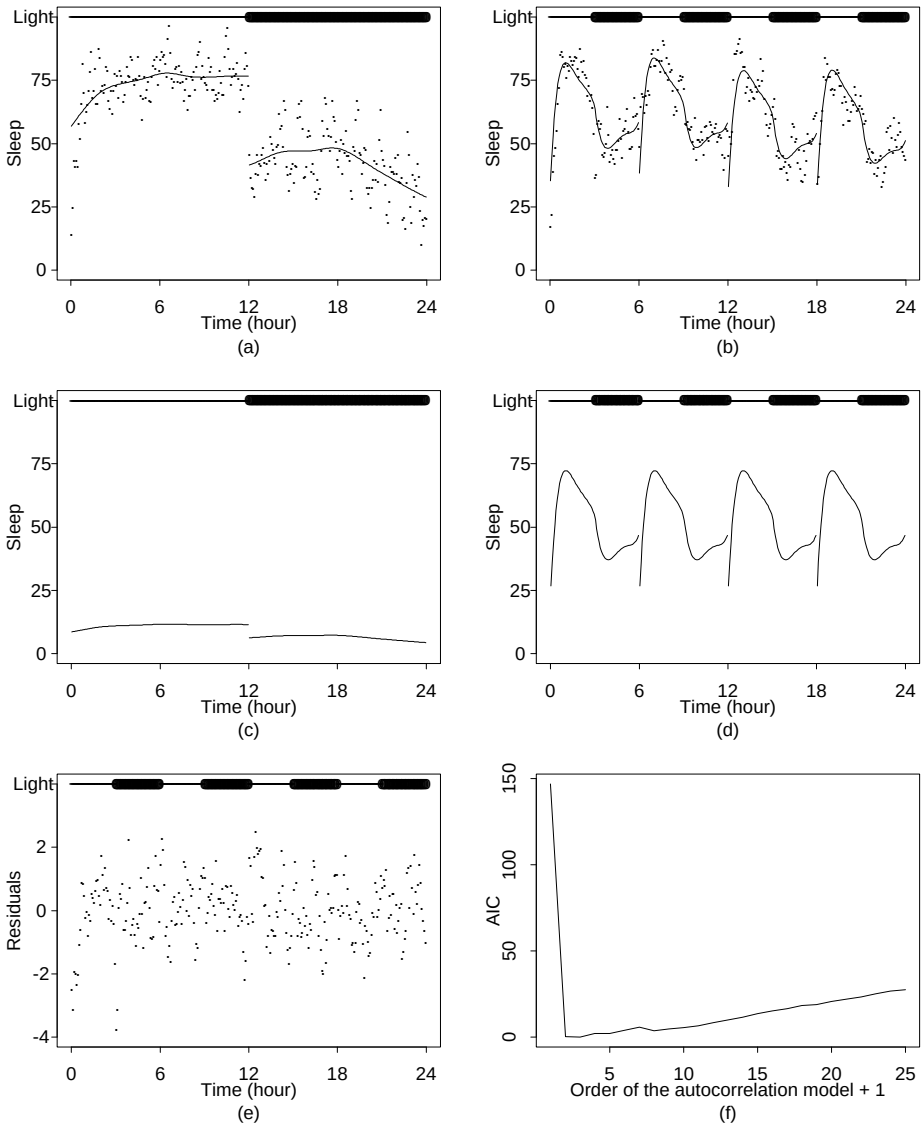

Figure 5.1: (a) The rat sleep data under the "normal" lighting condition. The solid curves represent the fitted model of (1.1). Thin line at top indicates period of light. Thick line represents darkness. (b) The rat sleep data under the "test" lighting condition. The curves represent the fitted model of (1.2). (c) The curves denote $\widehat{c}\widehat{f}_1$. (d) The curves denote $\widehat{f}_2$. (e) Standardized residuals of the fitted model of (1.2). (f) AIC criterion for modeling the autocorrelation of the residuals.

The estimated value of c is $\widehat{c} = .1491$. By a simple T-test, the p-value is $< .0001$ for testing

$H_0 : c = 0$ *vs* $H_a : c \neq 0$, with or without the autocorrelation consideration. Therefore we reject H_0 and conclude that the 24-hour cycle effect (i.e., the effect of circadian rhythm) is statistically significant under the 6-hour cycle lighting condition for the Brown Norway rats.

6 Some Concluding Remarks

We have presented a procedure for modeling physiological parameters affected by both intrinsic and extrinsic factors. It is proved that the fitted model is statistically consistent. Simulation results show that it works reasonably well in practice.

In the current procedure, f_1 is estimated by the sample from the model (1.1) alone. The estimator $\widehat{f}_1$ is then used to replace f_1 in the model (1.2). Because $\widehat{f}_1$ is random, the model (1.2) looks like an "error-in-covariates" model (see e.g., Fuller 1987). But they are different in the sense that the error components of the related covariates in a conventional "error-in-covariates" model are often assumed to have a fixed but unknown distribution while the variability of $\widehat{f}_1$ in the current problem will diminish when n_1 increases. The current modeling procedure has made use of this property of $\widehat{f}_1$ by substituting it for f_1 in (1.2).

In the rat sleep example, we assume in our model that discontinuities may exist at the times when the lighting conditions are switched. However, the change of the lighting condition may have a time lag effect on rat sleep, which is not explained by the current model. Besides the 24-hour cycle effect and the 6-hour cycle effect on rat sleep under the test lighting condition, scientists doubt that some acyclic effects may exist. It is still an open problem how to modify our modeling procedure to accomodate this kind of acyclic effects. The rat sleep data set presented in this paper is an average data set averaged by the investigators over a group of eight rats under the same lighting condition. If the data were not averaged, then a random effect term might be needed in (1.2) to explain the random effect of different rats. The statistical properties of the estimated coefficients of this "mixed effects" model are not available yet.

Acknowledgements The author would like to thank two referees for many helpful comments and suggestions. This research is supported in part by a Grant-in-Aid of Research, Artistry and Scholarship from the University of Minnesota.

REFERENCES

Benca, R., Obermeyer, W., Bergmann, B., Lendvai, N., and Gilliland, M. (1993). Failure to induce rapid eye movement sleep by dark pulses in pigmented inbred rat strains. *Physiology and Behavior* **54**, 1211-1214.

Chen, H. (1988), "Convergence rates for parametric components in a partly linear model," *The Annals of Statistics* **16**, 136-146.

Cheng, K.F., and Lin, P.E. (1981), "Nonparametric estimation of a regression function," *Z. Wahrscheinlichkeitstheorie view. Gebiete* **57**, 223-233.

Cleveland, W.S. (1979), "Robust locally weighted regression and smoothing of scatterplots," *Journal of the American Statistical Association* **74**, 829-836.

Draper, N.R., and Smith, H. (1980), *Applied Regression Analysis*, 2nd edition, Wiley: New York.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman & Hall: London.

Fuller, W.A. (1987), *Measurement Error Models*, John Wiley & Sons: New York.

Härdle, W. (1991), *Smoothing Techniques: with implementation in S,* Springer-Verlag: New York.

Judge, G., Griffiths, W., Hill, R., and Lee, T. (1980), *The Theory and Practice of Econometrics*, John Wiley & Sons: New York.

Lai, T.L., and Robbins, H. (1977), "Strong consistency of least squares estimators in regression models," *Proceedings of the National Academic Sciences of USA* **74**, 2667-2669.

Lai, T.L., Robbins, H., and Wei, C.Z. (1979), "Strong consistency of least squares estimates in multiple regression II," *Journal of Multivariate Analysis* **9**, 343-361.

Loader, C.R. (1999), "Bandwidth selection: Classical or plug-in?", *The Annals of Statistics* **27**, 415-438.

Loève, M. (1977), *Probability Theory I,* 4th Edition, Springer-Verlag: New York.

Pavlidis, T. (1973), *Biological oscillators: their mathematical analysis*, Academic Press: New York.

Qiu, P. (1994), "Estimation of the number of jumps of the jump regression functions," *Communications in Statistics-Theory and Methods* **23**, 2141-2155.

Seber, G.A.F. (1977), *Linear Regression Analysis,* Wiley: New York.

Speckman, P. (1988), "Kernel smoothing in partial linear models," *Journal of the Royal Statistical Society (Series B)* **50**, 413-436.

Stoll, M. (2001), *Real Analysis*, 2nd edition, Addison Wesley Longman, Inc.: Boston.

Strogatz, S.H. (1986), *The Mathematical Structure of the Human Sleep-Wake Cycle*, Lecture Notes in Biomathematics 69, Springer-Verlag: Berlin.

Wahba, G. (1991), *Spline Models For Observational Data,* SIAM: Philadelphia, PA.

Wang, Y., and Brown, M. (1996), "A flexible model for human circadian rhythms," *Biometrics* **52**, 588-596.

Wu, C.F. (1980), "Characterizing the consistent directions of least squares estimates", *The Annals of Statistics* **8**, 789-801.

Peihua Qiu
School of Statistics
University of Minnesota
313 Ford Hall
224 Church St. S.E.
Minneapolis, MN 55455, USA
qiu@stat.umn.edu