The Pennsylvania State University

The Graduate School

Department of Statistics

ESTIMATION IN COVARIATE-ADJUSTED NONLINEAR REGRESSION

A Thesis in

Statistics

by

Esra Kurum

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2009

The thesis of Esra Kurum was reviewed and approved* by the following:

Damla Senturk
Assistant Professor in Statistics
Thesis Advisor

Runze Li
Professor of Statistics

Bruce Lindsay
Willaman Professor of Statistics,
Head, Department of Statistics

*Signatures are on file in the Graduate School

ABSTRACT

We propose a new estimation procedure for covariate adjusted nonlinear regression models for situations where both the predictors and response in a nonlinear regression model are not directly observed, however distorted versions of the predictors and response are observed. The distorted versions are assumed to be contaminated with a multiplicative factor that is determined by the value of an unknown function of an observable covariate. We demonstrate how the regression coefficients can be estimated by establishing a connection to nonlinear varying coefficient models. Simulation studies are used to illustrate the efficacy of the proposed estimation algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Literature Review: Covariate Adjusted Regression

Senturk and Muller (2005, 2006) have proposed covariate-adjusted regression (CAR) for cases where the predictors and response are not directly observed in a regression model. Instead, their distorted versions along with a univariate covariate is observed. This distortion is multiplicative with a factor that is a smooth unknown function of an observed covariate. This methodology is illustrated by a study of haemodialysis patients (Kaysen et al., 2003) where the main interest was to uncover the relationship between plasma fibrinogen concentration and serum transferrin level. Body mass index (BMI) was identified as the covariate with multiplicative effects on both of the variables of interest. The interest is the relationship between plasma fibrinogen concentration and serum transferrin level adjusted for the effects of BMI.

Consider the multiple regression model,

$$Y = \beta_0 + \sum_{r=1}^{p} \beta_r X_r + \varepsilon \tag{1.1}$$

where X_r's are the predictors, Y is the response and ε is the error. The departure from this model is that X and Y are not directly observed, instead their distorted versions $\tilde{X}$ and $\tilde{Y}$, along with a univariate covariate U are observed,

$$\tilde{X}_r = \phi_r(U)\, X_r \,, \tag{1.2}$$

$$\tilde{Y} = \psi(U)\, Y \,.$$

Here, $\psi(U)$ and $\phi_r(U)$ are unknown smooth functions of U.

In the estimation procedure, Senturk and Muller (2005) utilized the connection to varying-coefficient models that holds between the observed variables. Let $\tilde{X} = \left(\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p\right)^T$, then we can write,

$$E\left(\tilde{Y}\middle|\tilde{X}, U\right) = \psi(U)\beta_0 + \psi(U)\sum_{r=1}^{p}\beta_r\frac{\phi_r(U)X_r}{\phi_r(U)} = \alpha_0(U) + \sum_{r=1}^{p}\alpha_r(U)\tilde{X}_r\,,$$

where $\alpha_0(U) = \psi(U)\beta_0$ and $\alpha_r(U) = \beta_r\frac{\psi(U)}{\phi_r(U)}$.

As a result,

$$\tilde{Y} = \alpha_0(U) + \sum_{r=1}^{p}\alpha_r(U)\tilde{X}_r + \psi(U)\varepsilon\ . \tag{1.3}$$

Model (1.3) is a multiple varying coefficient model, where the regression coefficients are allowed to vary smoothly with the value of the covariate U (Hastie and Tibshirani, 1993). Common applications of the model to longitudinal data take covariate U to be time (Hoover et al., 1998; Brumback and Rice, 1998; Fan and Zhang, 2000).

Kernel-local polynomial smoothing (Hoover et al., 1998; Fan and Zhang, 2008), polynomial spline (Huang et al., 2002; 2004; Huang and Shen, 2004) and smoothing spline (Hastie and Tibshirani, 1993; Hoover et al., 1998 and Chiang et al., 2001) are the major estimation procedures proposed for varying coefficient models. Fan and Zhang (2008) indicate that kernel smoothing method is the most natural, since varying coefficient models are inherently local models. All three approaches can be used to find the estimators of regression coefficients in covariate adjusted regression. Senturk and Muller (2005) target the underlying regression coefficients in (1.1) using estimated varying-coefficient functions $\left(\alpha_1, \alpha_2, \dots, \alpha_p\right)$ in model (1.3).

Cui et al. (2008) have extended CAR to nonlinear regression models. The methodology has been motivated by the Modification of Diet in Renal Disease (MDRD) study. Glomerular filtration rate (GFR) is used as an indicator of kidney health and the stage of kidney disease. Because of the difficulty of the other techniques to measure GFR, serum creatinine (SCr) has become a traditional way to estimate GFR. It is known that there is a nonlinear relationship between GFR and SCr. In addition, GFR and SCr are both affected by Body Surface Area (BSA). Therefore, Cui et al. (2008) has suggested covariate-adjusted nonlinear models to adjust for the nonparametric effects of BSA on GFR and SCr in the regression relationship of GFR and SCr.

Consider the following nonlinear regression model,

$$Y = f(X, \beta) + \varepsilon \ , \tag{1.4}$$

where $E(\varepsilon)=0$ and $Var(\varepsilon)=\sigma^2$. The distorted versions of the predictor and response are defined as in (1.2). The estimation procedure suggested by Cui et al. (2008) starts with regressing the observed response and predictors on the covariate U respectively, to obtain estimators of the smooth functions, $\widehat{\psi}(u)$ and $\widehat{\phi}(u)$,

$$\widehat{\psi}(u) \cong \frac{\hat{g}_Y(u)}{\hat{p}(u)} \frac{1}{\overline{\overline{Y}}}$$

$$\widehat{\phi}_r(u) \cong \frac{\hat{g}_r(u)}{\hat{p}(u)} \frac{1}{\overline{\overline{X}}_r},$$

where $\overline{\overline{Y}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{Y}_i$, $\overline{\overline{X}}_r = \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_{ri}$, $g_Y(u) = E(\tilde{Y}|U)p(U),\ g_r(u) = E(\tilde{X}_r|U)p(U)$ and p(U) is the density function of U. These estimators are used to estimate observable predictors and response,

$$\hat{Y} = \frac{\tilde{Y}}{\widehat{\psi}(u)} \ \text{ and } \ \hat{X}_r = \frac{\tilde{X}_r}{\widehat{\phi}_r(u)} \ , \tag{1.5}$$

where r =1, …, p. Then, the estimators in (1.5) are used to minimize a nonlinear least squares criterion to estimate regression coefficients in (1.4). See Cui et al. (2008) for details.

In our research, we are proposing a new estimation procedure similar to the connection proposed by Senturk and Muller (2005). We will base the new estimation procedure on the connection between varying-coefficient models and covariate adjusted nonlinear regression to obtain estimators of the regression coefficients in nonlinear regression models. We introduce nonlinear varying-coefficient models and review their estimation procedures in Chapter 2. The proposed estimation algorithm is outlined in Chapter 3. In Chapter 4, we present a simulation study to demonstrate the efficacy of the proposed method.

Chapter 2

Nonlinear Varying Coefficient Models

Varying coefficient models are used to explore the dynamic feature of the data set and they are extensions of linear regression models. Varying coefficient models allow the regression coefficients to vary over certain covariates, such as time and temperature, to increase the flexibility of linear regression models and to reduce modeling bias. The nonlinear varying-coefficient model can be given as,

$$Y_i = f\{X_i, \alpha(U_i)\} + \varepsilon_i \ , \tag{2.1}$$

where Y is the response variable, $X_i = \left(X_{1i}, \ldots, X_{pi}\right)^T$ are the predictors, U is the covariate, $\alpha(U_i) = \left\{\alpha_1(U_i), \ldots, \alpha_p(U_i)\right\}^T$ are the unknown regression coefficient functions, $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2$, $f\{X_i, \alpha(U_i)\}$ is a known nonlinear function of α and i=1, …, n. Nonlinear varying coefficient models establish a unique interaction relationship between the covariate U and predictors. They become nonlinear regression models with constant coefficients when we condition on the covariate U.

Wang (2007) proposed to use local linear regression to estimate the coefficients in a nonlinear varying coefficient model. The estimation procedure proposed by Wang (2007) starts with approximating the regression coefficients for any u in the neighborhood of u_0 by Taylor's expansion,

$$\alpha_r(u) \approx \alpha_r(u_0) + \alpha_r(u_0)(u - u_0) \equiv a_r + b_r(u - u_0) \ , \qquad r = 1, \ldots, p \ .$$

Define the vectors, $a = \left(a_1, a_2, \ldots, a_p\right)^T$, $b = \left(b_1, b_2, \ldots, b_p\right)^T$, $Y = (Y_1, \ldots, Y_n)^T$ and $\alpha(U_i) = \left\{\alpha_1(U_i), \ \ldots, \ \alpha_p(U_i)\right\}^T$. Local linear regression estimator of $\left(a^T, b^T\right)^T$ is obtained by minimizing,

$$l(a, b) = \frac{1}{2}\sum_{i=1}^{n}[Y_i - f\{X_i, a + b(U_i - u_o)\}]^2 K_h(U_i - u_0), \qquad (2.2)$$

where $K_h(\,.\,) = h^{-1}K(./h)$, $K(\,.\,)$ is the kernel function and h is the bandwidth.

An iterated least squares algorithm is used to compute the value that minimizes (2.2). First, $f\{X_i, a + b(U_i - u_o)\}$ is approximated for a given initial value $\left(a_0^T, b_0^T\right)^T$ of $\left(a^T, b^T\right)^T$ by using Taylor's expansion,

$$f\{X_i, a + b(U_i - u_o)\} \approx f\{X_i, a_0 + b_0(U_i - u_o)\} + \{(a - a_0) + (b - b_0)(U_i - u_0)\}^T f^{'}\{X_i, a_0 + b_0(U_i - u_0)\},$$

where $f^{'}(X, \alpha) = \partial f(X, \alpha)/\partial\alpha$. Then, they defined the matrices necessary for updating a and b vectors. At the j^{th} step of the iteration, define the following matrices,

$$F_j^T = \begin{pmatrix} f^{'}\left\{X_1, a^j + b^j(U_1 - u_o)\right\} & \ldots & f^{'}\left\{X_n, a^j + b^j(U_n - u_o)\right\} \\ (U_1 - u_0)f^{'}\left\{X_1, a^j + b^j(U_1 - u_o)\right\} & \ldots & (U_n - u_0)f^{'}\left\{X_n, a^j + b^j(U_n - u_o)\right\} \end{pmatrix},$$

$$Y_{i,j} = Y_i - f\{X_i, a_0 + b_0(U_i - u_o)\} - \left\{a^j + b^j(U_i - u_0)\right\}f^{'}\left\{X_i, a^j + b^j(U_i - u_0)\right\},$$

$$Y_j = \left(Y_{1,j}, \ldots., Y_{n,j}\right)^T,$$

$$W_i = \mathrm{diag}\left(K_h(U_1 - u_0), \ldots K_h(U_n - u_0)\right),$$

where $a^{(j)}$ and $b^{(j)}$ are the current values of a and b. Using the matrices defined above, we update a and b by,

$$\begin{pmatrix} a^{j+1} \\ b^{j+1} \end{pmatrix} = \left(F_j^T W_i F_j\right)^{-1} F_j W_i Y_j\ .$$

When the algorithm converges, it should satisfy, $l^{'}(a, b) = 0$. Hence, we obtain, $\widehat{\alpha}(u_o) = \hat{a}$ and $\widehat{\alpha}^{'}(u_0) = \hat{b}$. In the next chapter, we will describe the connection between nonlinear varying coefficient model and covariate adjusted nonlinear regression.

Chapter 3

Estimation in Covariate Adjusted Nonlinear Regression via Nonlinear Varying-Coefficient Models

Consider the nonlinear regression model,

$$Y_i = f(X_i, \beta) + \varepsilon_i \ , \tag{3.1}$$

where $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2$ and $f(X_i, \beta)$ is a nonlinear function of β. Distorted versions of the predictors and response are denoted as follows,

$$\tilde{X}_{ri} = \phi_r(U_i)\, X_{ri} \ , \tag{3.2}$$

$$\tilde{Y}_i = \psi(U_i)\, Y_i \ .$$

Here, $\psi(U)$ and $\phi_r(U)$ are unknown smooth functions of U, r =1, …, p and i=1, …, n.

Senturk and Muller (2005) has imposed identifiability constraints on the smoothing functions,

$$E\big(\phi_r(U_i)\big) = 1 \quad \text{and} \quad E\big(\psi(U_i)\big) = 1 \ ,$$

i.e the mean of the distorted variables are same with the adjusted variables.

According to (3.1) and (3.2),

$$E\big(\tilde{Y}_i \big| \tilde{X}_i, U_i\big) = f\big\{\tilde{X}_i, \alpha(U_i)\big\} \ ,$$

where $\tilde{X} = \big(\tilde{X}_1, \dots \tilde{X}_p\big)^T$, $\alpha(U_i) = \big\{\alpha_1(U_i), \dots, \alpha_p(U_i)\big\}^T$, $\alpha_1(U_i) = \psi(U_i)\beta_1$ and $\alpha_r(u) = \frac{\beta_r}{\phi_r(U)}$, $r = 2, \dots, p$. Therefore,

$$\tilde{Y}_i = f\big\{\tilde{X}_i, \alpha(U_i)\big\} + \varepsilon_i \ , \tag{3.3}$$

which is a nonlinear varying-coefficient model.

To obtain the estimator of β in (3.1), we first need to estimate $\alpha(U_i)$ in (3.3) using the procedures in Chapter 2. First, using Taylor's expansion in the neighborhood of a given u_i, i=1, … , n; we redefine $\alpha_r(u)$ as follows,

$$\alpha_r(u) \approx \alpha_r(U_i) + \alpha_r'(U_i)(u - U_i) = a_r + b_r(u - U_i),$$

where $a = \left(a_1, \ldots, a_p\right)^T$, $b = \left(b_1, \ldots, b_p\right)^T$, and r =1, ... , p. Define the vectors,

$a = \left(a_1, a_2, \ldots, a_p\right)^T$, $b = \left(b_1, b_2, \ldots, b_p\right)^T$, $\tilde{Y} = \left(\tilde{Y}_1, \ldots, \tilde{Y}_n\right)^T$ and $\alpha(U_i) = \left\{\alpha_1(U_i), \ldots \alpha_p(U_i)\right\}^T$.

We obtain local linear regression estimator of $\left(a^T, b^T\right)^T$ by minimizing,

$$l(a,b) = \frac{1}{2}\sum_{i=1}^{n}\left[\tilde{Y}_i - f\left\{\tilde{X}_i, a + b(U_m - U_i)\right\}\right]^2 K_h(U_m - U_i),$$

where $K_h(\,.\,) = h^{-1}K(./h)$, $K(\,.\,)$ is the kernel function, h is the bandwidth and m=1, …, n.

As in Chapter 2, we minimize l(a,b) by solving an iterated least squares algorithm. For a given initial value $\left(a_0^T, b_0^T\right)^T$ of $\left(a^T, b^T\right)^T$, we approximate $f\left\{\tilde{X}_i, a + b(U - U_i)\right\}$ by Taylor's expansion,

$$f\left\{\tilde{X}_i, a + b(U - U_i)\right\} \approx f\left\{\tilde{X}_i, a_0 + b_0(U - U_i)\right\} + \\ \{(a - a_0) + (b - b_0)(U - U_i)\}^T f'\left\{\tilde{X}_i, a_0 + b_0(U - U_i)\right\},$$

where $f'(X, \alpha) = \partial f(X, \alpha)/\partial \alpha$. Then, we define the matrices necessary for updating a and b vectors. At the jth step of the iteration, define the following matrices,

$$F_j^T = \begin{pmatrix} f'\left\{\tilde{X}_1, a^j + b^j(U_1 - U_i)\right\} & \ldots & f'\left\{\tilde{X}_n, a^j + b^j(U_n - U_i)\right\} \\ (U_1 - U_i)f'\left\{\tilde{X}_1, a^j + b^j(U_1 - U_i)\right\} & \ldots & (U_n - U_i)f'\left\{\tilde{X}_n, a^j + b^j(U_n - U_i)\right\} \end{pmatrix},$$

$$\tilde{Y}_{m,j} = \tilde{Y}_m - f\left\{\tilde{X}_m, a^j + b^j(U_m - U_i)\right\} - \left\{a^j + b^j(U_m - U_i)\right\}f'\left\{\tilde{X}_m, a^j + b^j(U_m - U_i)\right\},$$

$$\tilde{Y}_j = \left(\tilde{Y}_{1,j}, \ldots, \tilde{Y}_{n,j}\right)^T,$$

$W_i = \text{diag}\big(K_h(U_1 - U_i), \dots K_h(U_n - U_i)\big)$.

where $a^{(j)}$ and $b^{(j)}$ are the current values of a and b. As in Chapter 2, we update a and b by,

$$\begin{pmatrix} a^{j+1} \\ b^{j+1} \end{pmatrix} = \left(F_j^T W_i F_j\right)^{-1} F_j W_i \tilde{Y}_j. \tag{3.4}$$

We continue updating a and b vectors until they satisfy, $l^{'}(a, b) = 0$. Therefore, we obtain the following estimators, $\hat{\alpha}(u_i) = \hat{a}$ and $\hat{\alpha}^{'}(u_i) = \hat{b}$. This algorithm is repeated for each U_i . Finally, the estimators obtained for β_1 and β_r are given as,

$$\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}_1(U_i) \quad \text{and} \quad \hat{\beta}_r = \frac{1}{n\hat{\mu}_{\tilde{X}_r}}\sum_{i=1}^{n} \hat{\alpha}_r(U_i)\tilde{X}_r \qquad r = 2, \dots, p \ . \tag{3.5}$$

These estimators are motivated by the equalities $E\{\alpha_1(U)\} = \beta_1$ and $E\{\alpha_r(U)\tilde{X}_r\} = \beta_r E(\tilde{X}_r)$. It can be shown that the estimators in (3.5) satisfy these equations using identifiability conditions and the definitions of α_1 and α_r.

Chapter 4

Simulation Study

Consider the following nonlinear regression model,

$$Y_i = \beta_1\{1 - \exp(-\beta_2 X_i)\} + \varepsilon_i \,, \tag{4.1}$$

where $\beta_1 = 4.31$, $\beta_2 = 0.21$, $\varepsilon \sim N(0,0.02)$ and X is from a normal distribution with mean 5 and variance 3. We also assume that the covariate, $U \sim \text{Uniform}(3,5)$. The distortion functions are $\psi(U_i) = (U_i + 1)^2/25.3$ and $\phi(U_i) = (U_i + 1)/4.9$, which satisfy the identifiability conditions, $E(\phi(U)) = 1$ and $E(\psi(U)) = 1$. Therefore, the distorted versions of the predictor and response are as follows,

$$\tilde{X}_i = \phi(U_i)X_i \ \text{, where } \phi(U_i) = (U_i + 1)/4.9 \,,$$

$$\tilde{Y}_i = \psi(U_i)X_i \ \text{, where } \psi(U_i) = (U_i + 1)^2/25.3 \ .$$

We can write the regression model of observed response on the observed predictors,

$$\tilde{Y}_i = \alpha_1(U_i)\left[1 - \exp\left\{-\alpha_2(U_i)\tilde{X}_i\right\}\right] + \varepsilon \,, \tag{4.2}$$

where $\alpha_1(U_i) = \psi(U_i)\beta_1$ and $\alpha_2(U_i) = \beta_2/\phi(U_i)$. Model (4.2) is a multiple nonlinear varying coefficient model (Wang, 2007), which has the following form,

$$\tilde{Y}_i = f\left\{\tilde{X}_i, \alpha(U_i)\right\} + \varepsilon_i \,,$$

in our case, $f\left\{\tilde{X}_i, \alpha(U_i)\right\} = \alpha_1(U_i)\left[1 - \exp\left\{-\alpha_2(U_i)\tilde{X}_i\right\}\right]$ and $\alpha(U_i) = \begin{pmatrix}\alpha_1(U_i)\\ \alpha_2(U_i)\end{pmatrix}$. We first estimate $\alpha(U_i)$ using the techniques given in Wang (2007). Then, following Senturk and Muller (2005), we use them to estimate β_1 and β_2.

As described in Chapter 3, at the kth step of our iteration algorithm, we approximate $\alpha_1(u_i)$ and $\alpha_2(u_i)$ by using Taylor's expansion in the neighborhood of a given U_i, i=1, …, n;

$$\alpha_1(u) \approx \alpha_1(U_i) + \alpha_1'(U_i)(u - U_i) = a_{11}^j + b_{11}^j(u - U_i)$$

$$\alpha_2(u) \approx \alpha_2(U_i) + \alpha_2'(U_i)(u - U_i) = a_{21}^j + b_{21}^j(u - U_i)\ ,$$

where $a^j = \begin{pmatrix} a_{1j} \\ a_{2j} \end{pmatrix}$ and $b^j = \begin{pmatrix} b_{1j} \\ b_{2j} \end{pmatrix}$. We will update a and b vectors as follows,

$$\begin{pmatrix} a^{j+1} \\ b^{j+1} \end{pmatrix} = \left(F_j^T W_i F_j\right)^{-1} F_j W_i \widetilde{Y}_j\ ,$$

where F_j, W_i and $\widetilde{Y}_j$ are defined as in Chapter 3. However, we need to update these matrices according to the regression model given in (4.2). Therefore,

$$F_j^T = \begin{pmatrix} f'\{\widetilde{X}_1, a^j + b^j(U_1 - U_i)\} & \ldots & f'\{\widetilde{X}_n, a^j + b^j(U_n - U_i)\} \\ (U_1 - U_i)f'\{\widetilde{X}_1, a^j + b^j(U_1 - U_i)\} & \ldots & (U_n - U_i)f'\{\widetilde{X}_n, a^j + b^j(U_n - U_i)\} \end{pmatrix},$$

$$\widetilde{Y}_{m,j} = \widetilde{Y}_m - f\{\widetilde{X}_m, a^j + b^j(U_m - U_i)\} - \{a^j + b^j(U_m - U_i)\}f'\{\widetilde{X}_m, a^j + b^j(U_m - U_i)\},$$

where

$$f'\{\widetilde{X}_i, a^j + b^j(U_m - U_i)\} = \begin{pmatrix} 1 - \exp\{-\widehat{\alpha}_2(U_m - U_i)\widetilde{X}_i\} \\ \widehat{\alpha}_1(U_m - U_i)\widetilde{X}_i \exp\{-\widehat{\alpha}_2(U_m - U_i)\widetilde{X}_i\} \end{pmatrix},$$

for the mth column, $\widehat{\alpha}_1(U_m - U_i) = a_{11}^j + b_{11}^j(U_m - U_i)$ and

$\widehat{\alpha}_2(U_m - U_i) = a_{21}^j + b_{21}^j(U_m - U_i)$. For W_i matrix, defined in Chapter 3, we used the Epanechnikov Kernel Function,

$$K_h(U_m - U_i) = \frac{1}{h}.\frac{3}{4}\left\{1 - \left(\frac{U_m - U_i}{h}\right)^2\right\},\quad -1 < \frac{U_m - U_i}{h} < 1\ ,$$

where h is the bandwidth. To use this kernel function, we need to select a bandwidth. We will use multiple fold cross validation to select a bandwidth. We exclude the data 10 at a time and then fit the model. We computed the cross validation score as follows,

$$CV(h) = \sum_i \left\{\widetilde{Y}_i - \widehat{\widetilde{Y}}_{-}\right\}^2\ , \tag{4.4}$$

where $\widehat{\widetilde{Y}}_{-}$ denotes the fitted value for the excluded observations. We choose the bandwidth that gives us the minimum value in (4.4). For our analysis, bandwidths chosen are 0.4, 0.36, 0.32 for sample sizes 50, 100 and 200, respectively.

When the algorithm converges, we assign the values of the estimators of $\alpha(U_i)$ using the a vector,

$$\widehat{\alpha}(U_i) = \begin{pmatrix} \widehat{\alpha}_1(U_i) \\ \widehat{\alpha}_2(U_i) \end{pmatrix} = \begin{pmatrix} \hat{a}_{1(j+1)} \\ \hat{a}_{2(j+1)} \end{pmatrix}, \qquad (4.5)$$

assuming the iteration ends at (j+1). Then, using the estimators in (4.5) for each u_i, we estimate β_1 and β_2, as follows,

$$\widehat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}_1(U_i)$$

$$\widehat{\beta}_2 = \frac{1}{n\widehat{\mu}_{\widetilde{X}}}\sum_{i=1}^{n} \widehat{\alpha}_2(U_i)\widetilde{X}_i \,.$$

For our simulation study, we simulated 100 samples of $\widetilde{Y}$ and $\widetilde{X}$ from the distributions we specified before, for sample sizes 50, 100 and 200. Table 4.1 summarizes the simulation results, where the bias, variance and mean squared error are all decreasing with sample size.

Table 4.1 Estimated bias, variance and mean squared error of the proposed estimators

Sample size	$\hat{\beta}_1$			$\widehat{\beta}_2$		
	Bias	Variance	MSE	Bias	Variance	MSE
50	0.01460	0.06453	0.06474	0.00453	0.00033	0.00035
100	0.00540	0.03254	0.03257	0.00390	0.00013	0.00015
200	0.00333	0.01230	0.01232	0.00300	0.00005	0.00006

Figure 4.1 shows the cross-sectional mean (dash-dotted), and 5th (dotted) and 95th (dotted) percentiles of the 100 estimators overlapping the true intercept coefficient functions (solid).

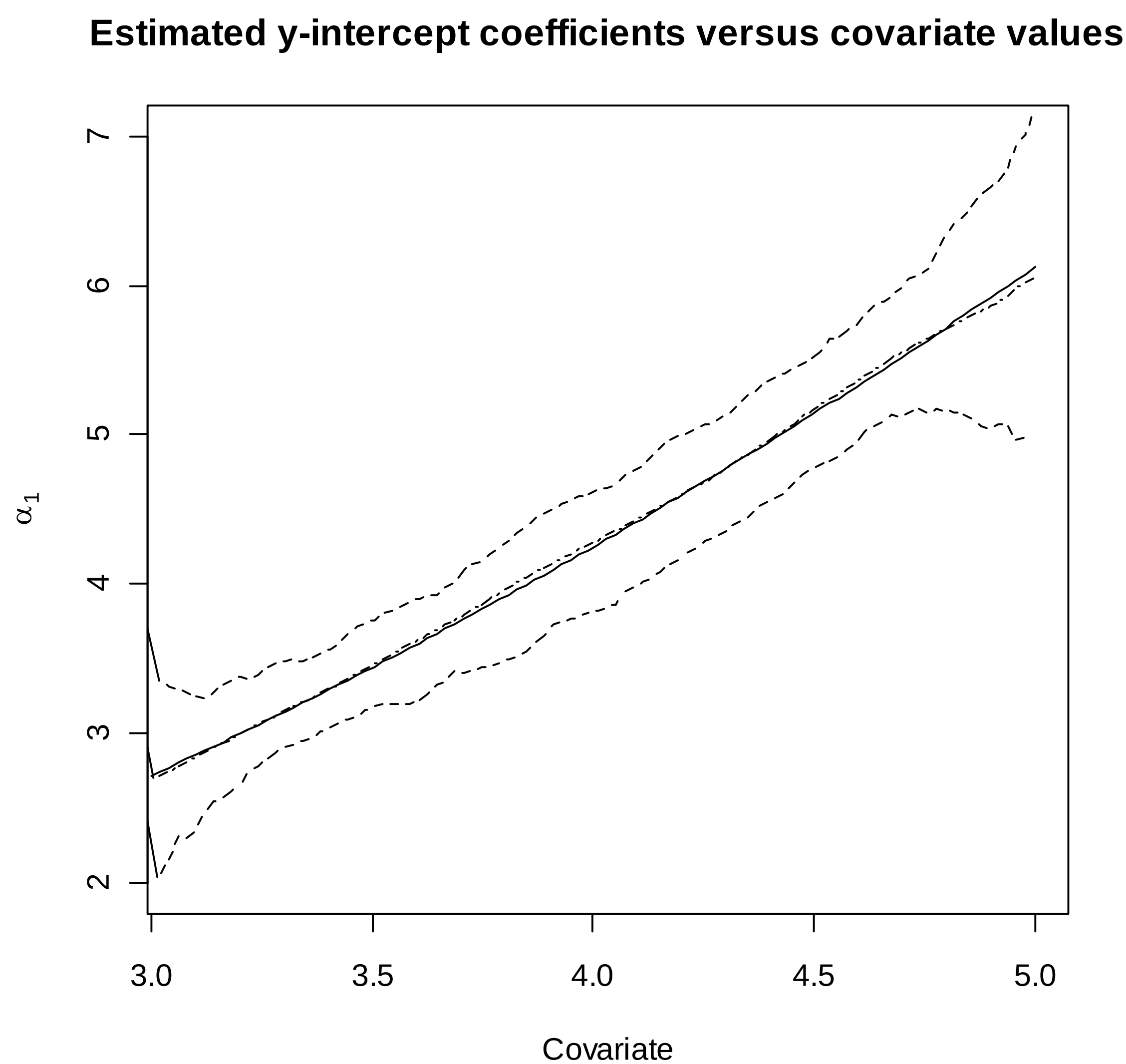

Figure 4.1. Estimated y-intercept coefficients versus covariate values

Similarly, Figure 4.2 corresponding quantities for α_2.

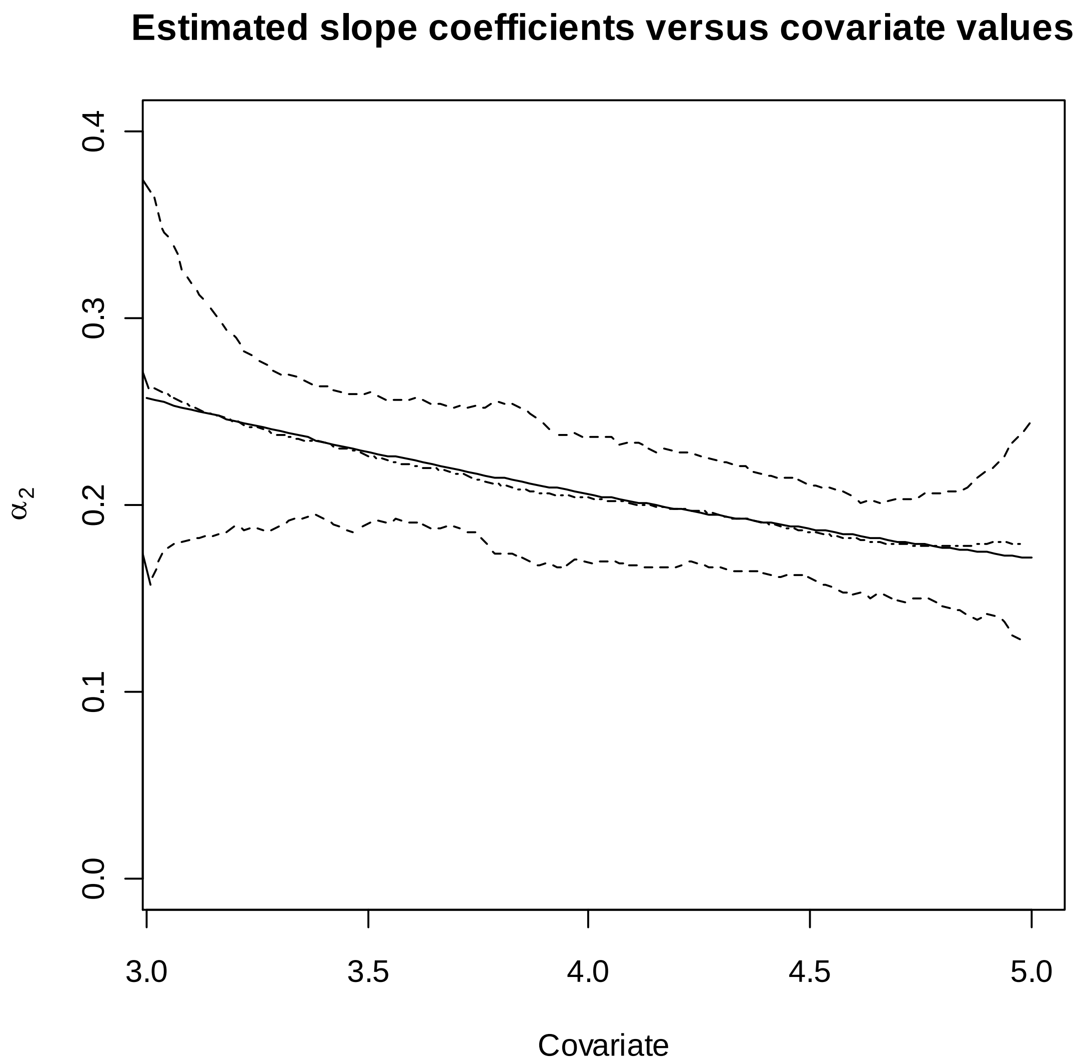

Figure 4.2. Estimated slope coefficients versus covariate values

Chapter 5

Conclusions

We developed a statistical procedure to estimate the regression coefficients in a covariate adjusted nonlinear regression via varying coefficient models. We used the connection between covariate adjusted regression and varying coefficient models established by Senturk and Muller (2005) and extend it to nonlinear regression models. The estimated method is based on an iterated least squares algorithm to target the regression coefficients of the varying coefficient model. The estimated coefficients are then used to target the regression coefficients of the underlying nonlinear regression. We next hope to compare the estimation algortihm proposed by Cui et al. (2008) to the currently proposed estimation procedure. One apparent disadvantage of the earlier estimation algorithm proposed by Cui et al. (2008) is that it contains a considerable number of divisions as given in equation (1.5). These divisions may cause problems in the algorithm if the estimated distorted functions $\widehat{\psi}(U)$ and $\widehat{\phi}_r(U)$ approach zero. The currently proposed algorithm does not get back to the unobserved predictors and response values and hence avoid these divisions.

References

Brumback, B. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.* **93** 961–994

Chiang, C.-T., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* **96** 605–619.

Cui, X**.,** Guo, W. Lin, L. and Zhu, L. (2008) Covariate-adjusted nonlinear regression. *Annals of Statistics (to appear).*

Fan, J. and Zhang, J. T. (2000). Functional linear models for longitudinal data. *J. Roy. Statist. Soc. B* **62** 303–322.

Hastie,T.J. and Tibshirani,R.J. (1993), Varying-coefficient models (with discussion), *J.Royal Statist. Soc., Ser. B,* **55** , 757-796.

Hoover, D. R., Rice, J. A., Ww, C. O. And Yang, L.-P. (1998). Nonparametric smoothing estimates of timevarying coefficient models with longitudinal data. Biometrika 85, 809–822.

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128.

Huang, J. Z. and Shen, H. (2004). Functional coefficient regression models for nonlinear time series: A polynomial spline approach. *Scandinavian Journal of Statistics* **31** 515–534.

Huang, J. Z.,Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica* **14** 763–788.

Senturk, D. and Muller, H. G. (2005a). Covariate adjusted regression. *Biometrika* 92, 59–74.

Senturk, D. and Muller, H. G. (2006). Inference for covariate adjusted regression via varying coefficient models. *Annals of Statistics*, 34, 654-679.

Wang,Y. (2007) Varying-coefficient Models:New models, inference procedures and applications, unpublished thesis (PhD), Penn State University.