Journal of Data Science **5**(2007), 471-490)

Distortion Diagnostics for Covariate-adjusted Regression: Graphical Techniques Based on Local Linear Modeling

Danh V. Nguyen[1] and and Damla Şentürk[2]
[1] *University of California, Davis and* [2] *Pennsylvania State University*

Abstract: Linear regression models are often useful tools for exploring the relationship between a response and a set of explanatory (predictor) variables. When both the observed response and the predictor variables are contaminated/distorted by unknown functions of an observable confounder, inferring the underlying relationship between the latent (unobserved) variables is more challenging. Recently, Şentürk and Müller (2005) proposed the method of covariate-adjusted regression (CAR) analysis for this distorted data setting. In this paper, we describe graphical techniques for assessing departures from or violations of specific assumptions regarding the type and form of the data distortion. The type of data distortion consists of multiplicative, additive or no-distortion. The form of the distortion encompasses a class of general smooth distorting functions. However, common confounding adjustment methods in regression analysis implicitly make distortion assumptions, such as assuming additive or multiplicative linear distortions. We illustrate graphical detection of departures from such assumptions on the distortion. The graphical diagnostic techniques are illustrated with numerical and real data examples. The proposed graphical assessment of distortion assumptions is feasible due to the CAR estimation method, which utilizes a local regression technique to estimate a set of transformed distorting functions (Şentürk and Nguyen, 2006).

Key words: Covariate-adjusted regression, distortion, graphical diagnostics, local regression modeling, multiplicative effect, varying-coefficient models.

1. Introduction

1.1 Examples of covariate adjustments in the health sciences

Regression modeling is a useful tool for exploring possible relationships between the primary response and explanatory variables of interest, especially for observational studies. For situations where both the predictors and the response in a regression model are not directly observed, but instead are observed after being contaminated by unknown functions of a common confounder, straightforward applications of regression models may result in misleading conclusions. Adjustment for the effect of the observed confounder is needed. Observable confounders, such as body mass index (BMI) and/or other measures of body configuration, are common in medical or health related studies because they are known confounding variables that affect the primary variables of interest.

The method of covariate-adjusted regression (CAR), proposed by Şentürk and Müller (2005), was designed to infer the underlying relationship between the (latent) primary variables of interest under a general multiplicative data distortion setting. Their method was originally motivated by data on inflammation protein markers in haemodialysis patients. More specifically, a primary outcome variable is elevated plasma fibrinogen level (Kaysen et al., 2003; Şentürk and Müller, 2005). Fibrinogen is a protein found in blood plasma and it is a risk factor for cardiovascular disease in haemodialysis patients. It is of interest to examine the relationship between fibrinogen concentration and other predictors, such as serum transferrin protein level. However, both primary variables of interest, fibrinogen and transferrin protein levels, are known to depend on body mass index (BMI), which exerts a confounding effect on the protein measurements. A common approach to adjust for the confounders, like BMI, is to normalize the primary variables of interest by simply dividing (by the confounder BMI). Şentürk and Nguyen (2006) provide another example of adjustment for BMI in exploring the underlying regression relationship between hypertensive variables and glycosolated hemoglobin (a diagnostic measurement for diabetes).

Adjustment for confounding/distorting covariates is also common in the assessment of environmental contaminants on human health risks from observational or epidemiological studies. For example, the relationship between exposure to lipophilic agents, such as polychlorinated biphenyls (PCBs), and health outcomes is often analyzed after adjustment for the distorting effect of serum lipid (SL; Schisterman et al., 2005). The covariate adjustment here involves the ratio PCB/SL^{ρ}, where the power ρ allows for a more general relationship between PCB and SL.

To provide a more formal description of the above examples, in terms of the multiplicative distortion framework considered here, some notations are needed.

Denote the observed response, the p predictors and the confounder by $\widetilde{Y}$, $\{\widetilde{X}_1, \ldots, \widetilde{X}_p\}$, and U, respectively. The confounder U in the two examples above is BMI and SL. The above examples further suggest that the distortion is believed to be specific cases of multiplicative distortion of the following type:

$$\widetilde{Y} = \psi(U)Y, \quad \text{and} \quad \widetilde{X}_r = \phi_r(U)X_r, \qquad r = 1, \ldots, p, \tag{1.1}$$

where Y and $\{X\}_{r=1}^p$ are the underlying (latent) variables of interest. The functions $\psi(\cdot)$, $\phi_1(\cdot), \ldots, \phi_p(\cdot)$ are unknown (smooth) distorting functions. The above distortions can induce an artificial relationship between the observed variables, $\widetilde{Y}$ and $\{\widetilde{X}_r\}$, that may not be reflective of the true underlying regression relationship between Y and $\{X_r\}$ of interest. The underlying/latent regression relationship is given by $E(Y) = \gamma_0 + \sum_{r=1}^p \gamma_r X_r$, where $\{\gamma_r\}_{r=0}^p$ are the parameters of interest. It is of interest to estimate this latent regression relationship based on the available distorted data, namely $\widetilde{Y}$, $\widetilde{X}_r$, and the confounder U.

Note that the distortion framework (1.1) accommodates various forms of covariate adjustments. For instance, it allows for linear and/or possibly nonlinear distortion on both Y and $X_1, \ldots, X_r$. In the PCB example, the predictor distortion is assumed to be nonlinear: $\phi(U) = U^\rho$. In the inflammation protein marker example, the distortion on the response and predictors are assumed to be both linear: $\psi(U) = \phi(U) = U = BMI$. This assumption of a common linear distortion is used in practice for its simplicity.

We emphasize that the distortion framework (1.1) allows for the *unknown* contaminating functions. This is an appealing aspect, from a practical point of view. This is because, in practice, the precise nature of the multiplicative relationships between the confounder and the primary variables of interest is unknown. Lacking this precise knowledge, the practice of dividing by the confounder U, or equivalently assuming the specific linear distortion form, $\psi(U) = U$ and $\phi_r(U) = U$ in (1.1), imposes unnecessarily rigid constraints on the form of the data distortion. Also, the assumption of a specific linear form under multiplicative distortion may be incorrect. We suggest simple graphical techniques that can be used to check if this specific assumption does not hold, as well as other assumptions regarding the data distortion.

We point out here that CAR, an adjustment method under distortion framework (1.1), does not restrict the form of the distorting functions, assuming only that they are smooth functions. Using CAR, the regression relationship between the unobserved variables, Y and $\{X_r\}_{r=1}^p$, can be consistently estimated based on the distorted data. In addition to allowing the forms of the distorting functions to be more general, CAR also accommodates different types of distortion models, namely: (a) multiplicative distortion (i.e. $\widetilde{Y} = \psi(U)Y$, $\widetilde{X}_r = \phi_r(U)X$), (b) additive distortion (i.e. $\widetilde{Y} = \psi(U) + Y$, $\widetilde{X}_r = \phi_r(U) + X_r$), (c) and no-

distortion (i.e. $\widetilde{Y} = Y$, $\widetilde{X}_r = X_r$). Under suitable identifiability conditions, given in Section 2, the consistency of the CAR estimators holds under these three types of distortion (Şentürk and Nguyen 2006; Şentürk and Müller 2005). Covariate-adjusted regression was originally proposed in Şentürk and Müller (2005) using a rough binning approach for estimation. A more refined estimation method to reduce the variance, based on local regression modeling, was proposed in Şentürk and Nguyen (2006). The asymptotic distributions of the CAR estimators were established in Şentürk and Müller (2006).

In this work, we examine graphical approaches for assessing specific assumptions regarding the types and forms of the data distortion. For example, violations of the assumption of a specific linear distortion, $\psi(U) = \phi_r(U) = U$, under multiplicative distortion can be checked graphically. Another example is the assumption that the above distortion only affects the predictors (i.e. $\psi(U) = 1$). Also, in some cases, it is possible to fully characterize the types of distortion (i.e. no-distortion, additive, or multiplicative) graphically. We describe graphical techniques to assess these and other related assumptions regarding the data distortion in the context of covariate-adjusted regression.

Finally, we note here that the multiplicative distortion framework (1.1) has similarities with measurement error modeling if the distortion by U is thought of as an error affecting both the response and the predictors. However, a distinct difference with the measurement error literature is that the "measurement" error is a function of an *observable* confounder U. Although there is a vast literature on additive measurement error modeling, the work on multiplicative measurement error modeling is limited. Estimation procedures targeting the regression coefficients under multiplicative measurement error in the predictor variables were considered by, for example, Hwang (1986) and Iturria, Carroll and Firth (1999). The case of multiplicative measurement errors in both the response and predictors has not been considered previously to our knowledge.

1.2 An example of the distortion effects

To further introduce and illustrate the potential distortion effects on the underlying regression relationship between Y and $\{X_r\}_{r=1}^p$, we consider the following numerical example. Suppose that the underlying (unobserved) regression model of interest is

$$Y = 2 - 1.5X_1 + 0.8X_2 + e, \tag{1.2}$$

where the predictors $\{X_1, X_2\}$ are bivariate normal with means $(2, 4)$, variances $(2^2, 1.8^2)$ and with correlation $r(X_1, X_2) = 0.2$. Also, assume that the error term e is normally distributed with mean 0 and variance $\sigma^2 = 0.5^2$. Suppose that we have $n = 500$ observations from model (1.2). Then the simple ordinary least

squares (OLS) estimators will target the underlying regression parameters of interest: $\boldsymbol{\gamma}^{\mathrm{T}} = (\gamma_0, \gamma_1, \gamma_2) = (2, -1.5, 0.8)$. However, estimation of the relationship between Y and $\{X_1, X_2\}$ is more difficult when the available data has been contaminated. More precisely, suppose that the observed response and predictor values for n observations are $\{\widetilde{Y}_i, (\widetilde{X}_{i1}, \widetilde{X}_{i2})\}_{i=1}^n$. The observed (available) data is the result of multiplicative distortions on the response and the predictors:

$$\widetilde{Y}_i = \psi(U_i)Y_i, \quad \widetilde{X}_{i1} = \phi_1(U_i)X_{i1}, \quad \text{and} \quad \widetilde{X}_{i2} = \phi_2(U_i)X_{i2},$$

where the unknown smooth distorting functions are $\psi(U_i) \propto U_i^3$, $\phi_1(U_i) \propto \exp(U_i - 4)$ and $\phi_2(U_i) \propto (U_i + 4)^2$. For illustration, we take the confounder U_i to be uniformly distributed on the interval $[1, 6]$.

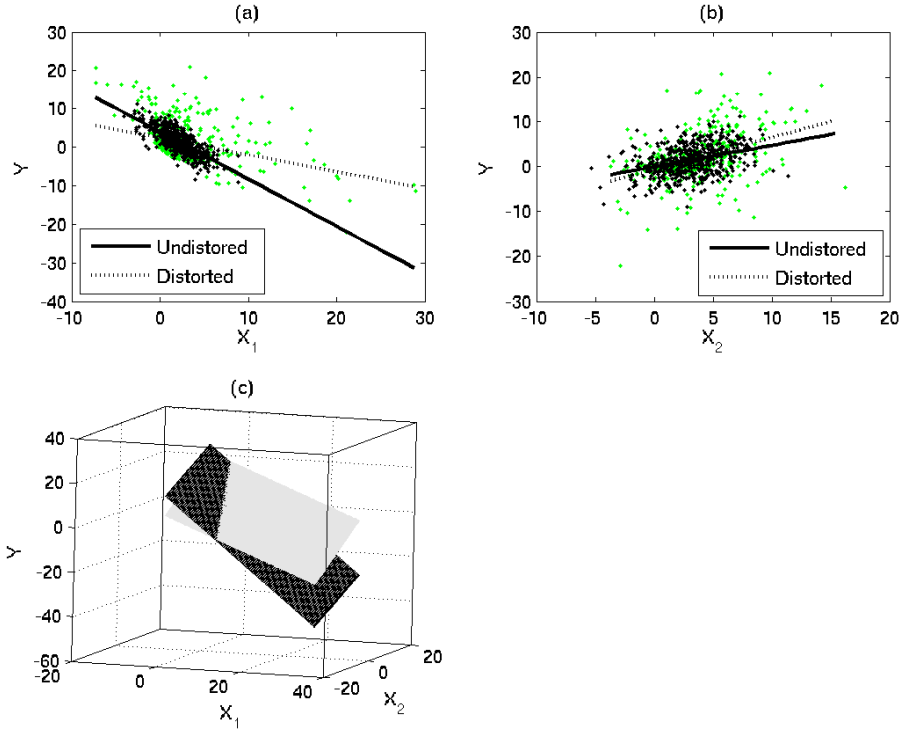

Figure 1: **Example of distortion effects**. Effects of the distorting functions $\psi(U) \propto U^3$, $\phi_1(U) \propto \exp(U - 4)$ and $\phi_2(U) \propto (U + 4)^2$ on Y, X_1 and X_2, respectively. The solid lines are the (estimated) marginal relationship of (a) Y on X_1 and (b) Y on X_2 obtained using ordinary least squares with unobserved/undistorted data (black dots). The dotted lines are the OLS fits based on the distorted data $\widetilde{Y}$, $\widetilde{X}_1$ and $\widetilde{X}_2$ (gray dots/green color in online version). (c) The true underlying regression relationship between Y and $\{X_1, X_2\}$ (black) and the corresponding incorrect relationship estimated using OLS based on distorted data (gray).

Figures 1(a) and (b) show the distortion effects on the marginal relation of the response Y to X_1 and Y to X_2, respectively. Displayed are the undistorted data (black) and the available distorted data (gray/green in online version) along with the OLS regression fits. Although there is a strong negative marginal relationship between Y and X_1, with $\widehat{r}(X_1, Y) = -0.76$, the strength of this relationship is substantially diminished after the distortion by $\psi(\cdot)$ and $\phi_1(\cdot)$. This is also reflected in the reduced estimated correlation based on the distorted observations: $\widehat{r}(\widetilde{X}_1, \widetilde{Y}) = -0.38$. On the other hand, the distortion can also artificially strengthen (or weaken) the observed relationship between the response and the predictor(s), when in fact the strength of association is weak (or completely lacking). For instance, in this example, the estimated sample correlation between Y and X_2 is $\widehat{r}(X_2, Y) = 0.37$. However, based on the distorted data, the estimated correlation is higher ($\widehat{r}(\widetilde{X}_2, \widetilde{Y}) = 0.44$; see 1(b)). The estimated relationship between the response and the predictors using OLS, if we were to have the original data $\{Y_i, (X_{i1}, X_{i2})\}_{i=1}^n$, is $y = 1.955 - 1.510x_1 + 0.815x_2$. This is close to the true relationship given by (1.2), as expected. However, the estimated relationship based on the distorted data $\{\widetilde{Y}_i, (\widetilde{X}_{i1}, \widetilde{X}_{i2})\}_{i=1}^n$ is $y = -0.435 - 0.740x_1 + 1.104x_2$. The overall distortion effect on the relationship between Y and $\{X_1, X_2\}$ is illustrated in Figure **??**(c). CAR provides consistent estimation of the underlying relationship based on the distorted data, as detailed in the next Section.

2. Estimation in covariate-adjusted regression

2.1 The basic CAR model

We formally describe the basic CAR model and review the estimation method based on local (linear) regression (Şentürk and Nguyen, 2006). The regression parameters of interest are $\{\gamma_r\}_{r=0}^p$ in the underlying (unobserved) regression model,

$$Y_i = \gamma_0 + \sum_{r=1}^{p} \gamma_r X_{ir} + e_i, \tag{2.1}$$

where Y_i and $\{X_{ir}\}_{r=0}^p$ are the response and predictor values corresponding to the ith subject, respectively. The error variable e_i is assumed to have $E(e_i) = 0$ and $\text{var}(e_i) = \sigma^2$. Parameter estimation is based on n distorted predictor and response observations, $\{\widetilde{Y}_i, \widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip}\}_{i=1}^n$, along with the confounding covariate U, where

$$\widetilde{Y}_i = \psi(U_i)Y_i, \quad \text{and} \quad \widetilde{X}_{ir} = \phi_r(U_i)X_{ir}, \quad r = 1, \ldots, p. \tag{2.2}$$

Also, let $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)^{\mathrm{T}}$, $\widetilde{\mathbf{X}}_i = (\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip})^{\mathrm{T}}$, and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$. It is assumed that $\{(\mathbf{X}_i, U_i, e_i)\}_{i=1}^n$ are independent and identically distributed, where $\mathbf{X}$, e and U are mutually independent for the underlying model (2.1) only.

The problem of estimating the parameters, $\{\gamma_r\}_{r=0}^p$, is identifiable under some constraints on the unknown smooth distorting functions. A set of reasonable constraints for $\psi(\cdot)$ and $\{\phi_r(\cdot)\}_{r=1}^p$ is implied by the natural assumption that the mean distorting effect should correspond to no distortion (Şentürk and Müller, 2005), i.e.

$$E\{\psi(U)\} = 1 \qquad \text{and} \qquad E\{\phi_r(U)\} = 1. \tag{2.3}$$

The multiplicative distortion model described collectively by (2.1)-(2.3) is referred to as the covariate-adjusted regression (CAR) model. From the CAR model (2.1)-(2.3), it appears that targeting the underlying regression parameters will first require the difficult task of estimating the distorting functions directly. However, a connection between the CAR models and varying coefficient models still allows for consistent estimation of the underlying parameters without directly estimating $\psi(\cdot)$ and $\phi_r(\cdot)$. This relationship results from the following regression of $\widetilde{Y}$ on $\{\widetilde{X}_r\}_{r=1}^p$ (Şentürk and Müller, 2005; 2006),

$$\begin{aligned} E(\widetilde{Y}_i|\widetilde{\mathbf{X}}_i^{\mathrm{T}}, U_i) &= \psi(U_i)\gamma_0 + \psi(U_i)\sum_{r=1}^{p}\gamma_r\frac{\phi_r(U_i)X_{ir}}{\phi_r(U_i)} \\ &= \beta_0(U_i) + \sum_{r=1}^{p}\beta_r(U_i)\widetilde{X}_{ir}, \end{aligned} \tag{2.4}$$

where

$$\beta_0(U_i) = \gamma_0\psi(U_i), \qquad \text{and} \qquad \beta_r(U_i) = \gamma_r\frac{\psi(U_i)}{\phi_r(U_i)}. \tag{2.5}$$

Therefore, a direct regression of the observed response on the set of observed predictors leads to the following multiple varying coefficient model,

$$\widetilde{Y} = \beta_0(U_i) + \beta_1(U_i)\widetilde{X}_{i1} + \cdots + \beta_p(U_i)\widetilde{X}_{ip} + \epsilon_i, \tag{2.6}$$

with $\epsilon_i \equiv \psi(U_i)e_i$. Cleveland et al. (1991) and Hastie and Tibshirani (1993) proposed varying coefficient models to allow for more flexible regression modeling where the variable U changes the coefficient of $\widetilde{X}_r$ through the unspecified function $\beta_r(U)$. Consequently, because the varying coefficient model (2.6) is completely observable, estimation techniques for varying coefficient models can be utilized in the CAR model estimation. One efficient approach is based on local regression modeling (Fan and Gijbels, 1996; Fan and Zhang, 1999; Cai, Fan and Li, 2000), as proposed in Şentürk and Nguyen (2006) and Şentürk (2006). We note that there is a vast literature on the theory and application of varying coefficient models. The literature includes Chen and Tsay (1993) for nonlinear time series, Chiang, Rice, and Wu (2001) for repeatedly measured response, and Hoover et al. (1998), Wu and Chiang (2000), Wu and Yu (2002), and Şentürk (2006) for longitudinal data.

Based on the relationships in (2.5) between the varying coefficient functions, $\{\beta_r(\cdot)\}_{r=0}^p$, and the distorting functions, $\{\psi(\cdot), \phi_r(\cdot)\}_{r=1}^p$, the CAR method provides consistent estimation of the underlying (unobserved) regression relationship between Y and $\{X_r\}_{r=1}^p$. Note that we can consider the $\{\beta_r(\cdot)\}$ as a set of transformed distorting functions. If we denote the estimators of the varying coefficient functions as $\{\widehat{\beta}_r(\cdot)\}_{r=0}^p$, then the CAR estimators of the underlying regression parameters are

$$\widehat{\gamma}_r = \frac{1}{\overline{\widetilde{X}}_r} \sum_{i=1}^{n} \frac{1}{n} \widehat{\beta}_r(U_i) \widetilde{X}_{ir}, \qquad r = 0, \ldots, p \tag{2.7}$$

where $\overline{\widetilde{X}}_r = n^{-1} \sum_{i=1}^n \widetilde{X}_{ir}$ and $X_{i0} \equiv 1$. (More details are given in Section 2.2 below.) The consistency of the estimators has been shown (Şentürk and Nguyen, 2006).

Furthermore, because of the relationships given by (2.5), we can directly use the estimated varying coefficient functions, $\{\widehat{\beta}_r(\cdot)\}_{r=0}^p$, for diagnosing various types and forms of the distortion. We provide details of these graphical techniques in Section 2.2 below. However, we first provide a brief summary of the local linear regression estimator of $\beta_r(U)$, as they are the main quantities used for the graphical assessment of violations of distortion assumptions.

2.2 CAR estimators based on local linear regression

Graphical assessment of specific assumptions regarding the forms and types of data distortion can be implemented based on the relationships described by (2.5). This involves the estimated varying coefficient functions $\{\widehat{\beta}_r(\cdot)\}_{r=0}^p$. We use a simple local linear regression estimator (Fan and Gijbels, 1996) for estimating the varying coefficient functions as follows. For a given point u, the function $\beta_r(\cdot)$ can be approximated locally as

$$\beta_r(U) \approx b_r + c_r(U - u), \qquad r = 0, 1, \ldots, p,$$

for U in a neighborhood of u. For simplicity, we consider local linear fits, although higher order polynomial approximation for $\beta_r(U)$ can also be used. However, our previous experience suggests that local linear fits are sufficiently accurate for CAR estimation purposes.

Explicit expressions for the local linear estimators of $\{\beta_r(\cdot)\}$ are obtained by minimizing the sum

$$\sum_{i=1}^{n} \left[\widetilde{Y}_i - \sum_{r=0}^{p} \{b_r + c_r(U_i - u)\widetilde{X}_{ir}\} \right]^2 K_h(U_i - u),$$

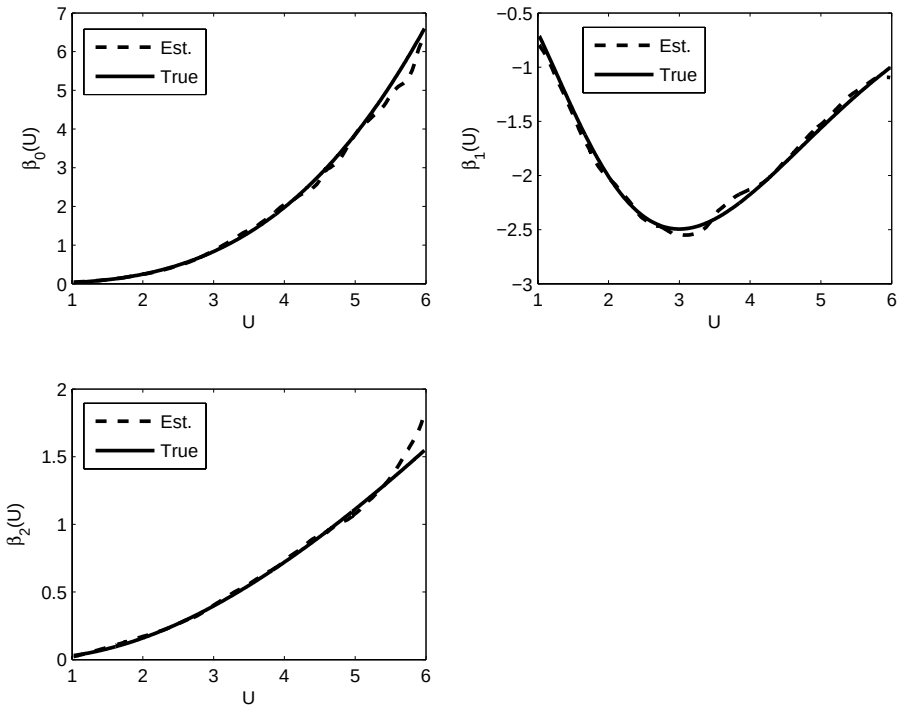

Figure 2: **Local linear estimation.** Given are the estimated (dashed lines)and true (solid lines) varying coefficient functions, $\{\widehat{\beta}_r(U)\}_{r=0}^2$, for model (2.6). The estimates are obtained using local linear regression ($h = 0.5$) based on the distorted data. The specific varying coefficient functions are $\beta_0(U) \propto U^3$, $\beta_1(U) \propto U^2/\exp(U-4)$, and $\beta_2(U) \propto U^3/(U+4)^2$, which are functions/transformations of the smoothed distorting functions $\{\psi(U), \phi_r(U), r = 1, 2\}$.

with respect to the coefficients $\{b_r, c_r\}$ and for a specified kernel function $K(\cdot)$ with bandwidth h where $K_h(\cdot) = K(\cdot/h)/h$. The minimization problem is simply a weighted least squares problem; hence, the estimators can be obtained explicitly. More precisely, let $\widehat{\boldsymbol{\alpha}} \equiv (\widehat{b}_0, \ldots, \widehat{b}_p, \widehat{c}_0, \ldots, \widehat{c}_p)^{\mathrm{T}}$. The local linear regression estimates $\widehat{\boldsymbol{\alpha}}$ is given by

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{\Sigma}(u)\mathcal{X}(u)^{\mathrm{T}}\mathbf{W}(u)\widetilde{\mathbf{Y}},$$

where $\mathbf{W}(u) = \mathrm{diag}\{K_h(U_1-u), \ldots, K_h(U_n-u)\}$, $\boldsymbol{\Sigma}(u) = \{\mathcal{X}(u)^{\mathrm{T}}\mathbf{W}(u)\mathcal{X}(u)\}^{-1}$, and $\mathcal{X}(u)$ is the following $n \times 2(p+1)$ data matrix

$$\mathcal{X}(u) = \begin{bmatrix} 1 & \widetilde{X}_{11} & \cdots & \widetilde{X}_{1p} & (U_1-u) & (U_1-u)\widetilde{X}_{11} & \cdots & (U_1-u)\widetilde{X}_{1p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \widetilde{X}_{n1} & \cdots & \widetilde{X}_{np} & (U_n-u) & (U_n-u)\widetilde{X}_{n1} & \cdots & (U_n-u)\widetilde{X}_{np} \end{bmatrix}.$$

Then an estimate of $\widehat{\beta}_r(\cdot)$ is $\widehat{\beta}_r(u) = \widehat{b}_r$ and an estimate of the derivative of $\widehat{\beta}_r(\cdot)$ is $\widehat{\beta}'_r(u) = \widehat{c}_r$. The bandwidth h can be chosen, for instance, by generalized cross-validation (Wahba, 1977; Craven and Wahba, 1979). We will elaborate on the bandwidth choice subsequently.

To illustrate the distorting functions, $\{\psi(U), \phi_r(U)\}_{r=1}^p$, and the estimation of the set of transformed distorting functions, $\{\beta_r(U)\}_{r=0}^p$, consider the example introduced in Section 1.2. The distorting functions are $\psi(U) = U^3/\omega_1$, $\phi_1(U) = \exp(U-4)/\omega_2$, and $\phi_2(U) = (U_1)^2/\omega_3$. The constants $(\omega_1, \omega_2, \omega_3) \approx (64.81, 1.47, 58.32)$ are chosen so that the distorting functions satisfy the identifiability constraints in (2.3). The local linear regression estimators of the corresponding varying coefficient functions, namely $\beta_0(U) = \gamma_0\psi(U)$, $\beta_1(U) = \gamma_1\{\psi(U)/\phi_1(U)\}$ and $\beta_2(U) = \gamma_2\{\psi(U)/\phi_2(U)\}$, are displayed in Figure 2.

In the next section we describe how the estimated transformed distorting functions, namely $\{\widehat{\beta}_r(U)\}$, can be used to graphically check for violation of specific assumptions about the form and type of data distortion. More precisely, we identify the structures of $\beta_r(\cdot)$ under various distortion assumptions. These structures can then be used to check for violations of model assumptions related to the distortion.

3. Graphical Assessment of Distortion Assumptions

3.1 Assessing violation of assumptions on the type of distortion

We first consider the assessment of specific assumptions regarding the type of distortion, namely (1) multiplicative, (2) additive, or (3) no-distortion. Although the CAR model, described in Section 2.1, can account for all three types of distortion, it may be of interest in practice to examine the violation of specific assumptions on the types of distortion. This may lead to the use of simpler estimation procedures than CAR, such as ordinary least squares regression in some special cases. Graphical assessment of distortion assumptions in the context of CAR essentially makes use of the relationship between the unknown distorting functions and their transformed versions, namely the varying coefficient functions in (2.5).

For simplicity of exposition, let us first consider the case of a single predictor. The multiple predictors case is similar and we will address it at the end of this section. For a single predictor, the unobserved model is $Y_i = \gamma_0 + \gamma_1 X_i + e_i$. Under the assumption of additive distortion, the available observations are $\{\widetilde{Y}_i, \widetilde{X}_i\}_{i=1}^n$, where $\widetilde{Y}_i = \psi(U_i) + Y_i$ and $\widetilde{X}_i = \phi(U_i) + X_i$. Thus, from the underlying model,

it follows that

$$\begin{aligned} \widetilde{Y}_i &= \gamma_0 + \gamma_1 \widetilde{X}_i + \{\psi(U_i) - \gamma_1 \phi(U_i)\} + e_i \\ &= \gamma_0 + \gamma_1 \widetilde{X}_i + \nu(U_i) + e_i, \end{aligned} \tag{3.1}$$

where $\nu(U_i) \equiv \{\psi(U_i) - \gamma_1 \phi(U_i)\}$. One may recognize the above model to be a partly linear model (PLM; Heckman, 1986). Note that the PLM is a special case of the varying coefficient model , $\widetilde{Y}_i = \beta_0(U_i) + \beta_1(U_i)\widetilde{X}_i + e_i$ (Section 2.1, equation (2.6)), where $\beta_0(U_i) = \nu(U_i) + \gamma_0$ and $\beta_1(U_i) = \gamma_1$. Because the resulting slope varying coefficient function is constant under additive distortion, i.e. $\beta_1(U_i) = \gamma_1$, departures from the assumption of additive distortion can be detected graphically by examining $\widehat{\beta}_1(U_i)$ for constancy.

In fact, more information can be obtained from the graphical examination of $\widehat{\beta}_1(U_i)$ regarding the type of distortion. More precisely, if $\beta_1(U_i)$ is not constant then the distortion type is consistent with a multiplicative form where $\psi(U_i) \neq \phi(U_i)$. This follows directly from the relationships in (2.5) and resulting regression equation under additive distortion given in (3.1) above.

In the special case where the distortion processes on the response and predictor are the same, i.e. when $\psi(U_i) = \phi(U_i)$, then the constancy of $\beta_1(U_i)$ implies that the distortion can be additive or multiplicative with $\psi(U_i) = \phi(U_i)$. More precisely, under multiplicative distortion with $\psi(U_i) = \phi(U_i) \equiv \varphi(U_i)$, we have the varying coefficient model $\widetilde{Y}_i = \beta_0(U_i) + \beta_1(U_i)\widetilde{X}_i + \epsilon_i$ with $\beta_0(U_i) = \gamma_0\varphi(U_i)$ and $\beta_1(U_i) = \gamma_1$.

Next, consider the case of no-distortion under the additive model. This is, $\psi(U_i) = \phi(U_i) = 0$. Consequently, we have $\nu(U_i) = 0$, so both the intercept and slope varying coefficient functions are constants: $\beta_0(U_i) = \gamma_0$ and $\beta_1(U_i) = \gamma_1$. These varying coefficient functions are constants under multiplicative distortion model (2.2) as well (i.e. $\psi(U_i) = \phi(U_i) = 1$). Thus, we can graphically check the estimated intercept and slope varying coefficient functions for no-distortion. Again, this is feasible under additive or multiplicative distortion models. Clearly, under the no-disortion case, measurements on U can be ignored.

For the case of multiple predictors in the context of additive distortion, it follows similarly as in (3.1) that $\widetilde{Y}_i = \beta_0(U_i) + \sum_{r=1}^{p} \gamma_r \widetilde{X}_{ir} + \nu(U_i) + e_i$, where $\nu(U_i) = \psi(U_i) - \sum_{r=1}^{p} \gamma_r \phi_r(U_i)$. Graphical examination of the estimated coefficient functions $\{\widehat{\beta}_r(U)\}_{r=1}^{p}$ for constancy, as in the single predictor case, is sufficient for diagnosing departures from the additive distortion assumption.

The following key points summarize our discussion of the graphical assessment of assumptions on the types of data distortion based on $\{\widehat{\beta}_r(U_i)\}$.

- To detect departures from the additive distortion assumption, it is sufficient to examine whether the estimated slope varying coefficient functions $\widehat{\beta}_r(U_i)$ are constants.

- If $\beta_r(U_i) \neq$ constant, then the distortion is consistent with a multiplicative form where $\psi(U_i) \neq \phi_r(U_i)$.
- If $\beta_r(U_i)$ are constants, then the distortion can be (a) additive or (b) multiplicative with $\psi(U_i) = \phi_r(U_i)$.
- If $\beta_0(U_i)$ *and* $\beta_r(U_i)$ are all constants, then there is no distortion.

3.2 Assessing violations of assumptions on the form of distortion

We next consider the graphical assessment of some specific and common assumptions regarding the functional form of the data distortion using the estimated varying coefficient functions. Of specific interest is the assessment of whether the distortion form is linear under additive distortion. Under linear additive distortion, the distortion functions are $\psi(U_i) = a + bU_i$ and $\phi(U_i) = c + dU_i$. Thus, we have that

$$\widetilde{Y}_i = \alpha_0 + \alpha_1 U_i + \gamma_1 \widetilde{X}_i + e_i \tag{3.2}$$

where the parameters are $\alpha_0 = \gamma_0 + (a - c\gamma_1)$ and $\alpha_1 = b - d\gamma_1$. As in (3.1), the resulting regression in (3.2) can be viewed as a varying coefficient model: $\widetilde{Y}_i = \beta_0(U_i) + \beta_1(U_i)\widetilde{X}_i + e_i$. As before, the slope varying coefficient $\beta_1(U_i)$ is constant. Additionally, the corresponding intercept varying coefficient function is linear: $\beta_0(U_i) = \alpha_0 + \alpha_1 U_i$. Thus, departures from or violations of the assumption of a linear additive distortion can be graphically examined by checking for constancy of $\widehat{\beta}_1(U_i)$ and linearity of $\widehat{\beta}_0(U_i)$. Note that the resulting model (3.2) implies that inclusion of the observable confounder U into a direct regression model based on the distorted data (response and predictor data) will provide a consistent estimate of the underlying slope coefficient γ_1. Therefore, the commonly used adjustment method of including U as an additional predictor is justified only under an additive linear distortion assumption.

Next, we consider two common assumptions on the functional form of the distortion under multiplicative distortion. The first case is linear regression models based on the adjusted response and predictor variables obtained via division by the confounder U. Such models implicitly assume that the distortion type is multiplicative and that the form is a special case of the linear distortion: $\psi(U) = \phi(U) \propto U$. One example, provided in the Introduction Section, involves dividing the observed response $\widetilde{PFL}$ (plasma fibrinogen level) and the predictor $\widetilde{STP}$ (serum transferrin protein) by the confounder $U = BMI$. The adjusted variables assumed to be free of the effect of BMI are $\widetilde{PFL}/BMI$ and $\widetilde{STP}/BMI$. This assumption would hold if, in fact, the distortion on the protein markers are of the form $\psi(U) = \phi(U) \propto U$. Other examples that make

this assumption are neurological studies comparing volumetric structures, such as amygdala and hippocampal volumes, obtained from magnetic resonance imaging (MRI). (See, for example, Pinter et al. (2001).) Typically, to compare across patients, the volumetric structures are normalized via division by total cranial volume ($TCV = U$) or total brain volume ($TBV = U$). This practice of division by the confounder implicitly assumes the special linear-multiplicative distortion of the forms $\psi(U) \propto U$ and $\phi(U) \propto U$. Thus, it follows directly from the relationships given by (2.5) that $\beta_r(U) \propto \gamma_r$ for $r \geq 1$ and $\beta_0(U) \propto \gamma_0 U$. Thus, violations of the assumption of this specific multiplicative linear distortion can be detected by checking for departures from linearity of $\beta_0(U)$ and the constancy of $\beta_r(U)$.

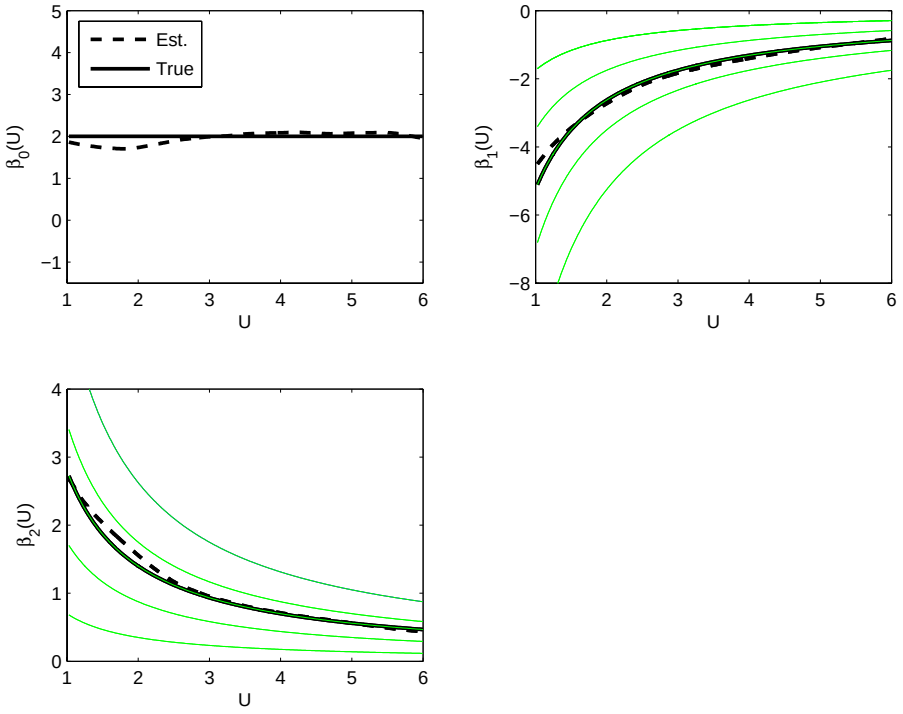

Figure 3: **Linear distortion on the predictor.** Displayed are the estimated varying coefficient functions, $\{\widehat{\beta}_r(U)\}_{r=0}^2$ (with $h = 2.0$), for the case of special linear distortion on the predictor variable only: $\psi(U) = 1$, $\phi_r(U) \propto U$. These distorting functions correspond to $\beta_0(U) = \gamma_0$ and $\beta_r(U) \propto U^{-1}$ under the covariate-adjusted regression model. The light thin (green for online version) cures are reference curves cU^{-1} for various constants of proportionality c. Departure from the distortion assumption occurs when the the estimated curve (dashed) deviates from a reference curve.

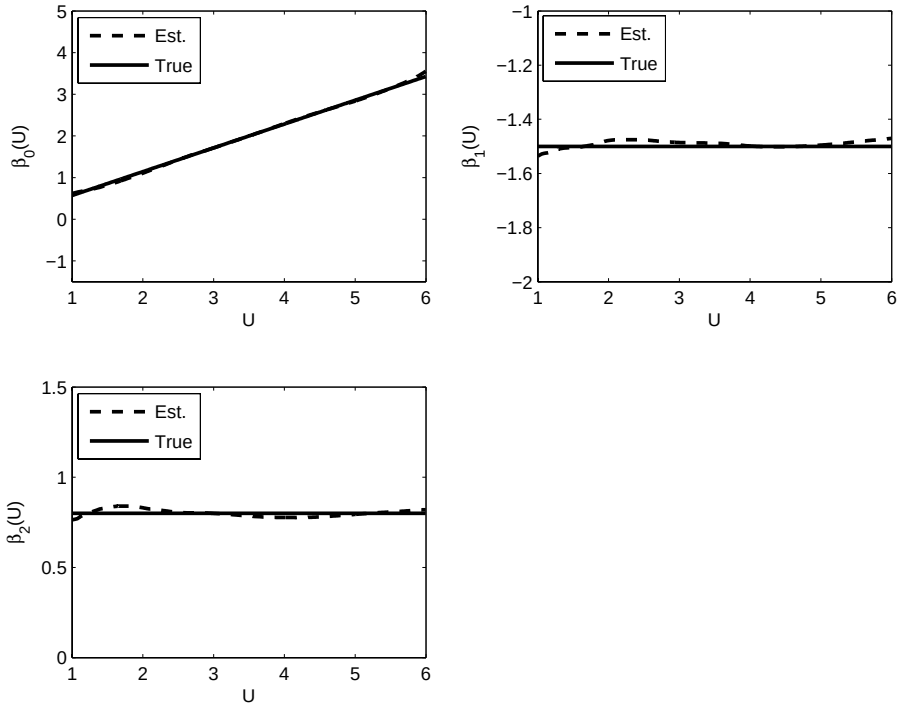

Figure 4: **Linear distortion on both response and predictor.** Special linear distortion on both the response and predictors, namely $\psi(U) = \phi_r(U) \propto U$, leads to constant functions for $\beta_r(U)$, $r \geq 1$. The dashed lines are the corresponding local linear estimates with $h = 2.0$. Note that when the *same nonlinear* distortion affects both the response and the predictors (i.e., $\psi(U) = \phi_r(U) \equiv \varphi(U)$ and nonlinear), only the plot of $\beta_0(U)$ (top left) will change, reflecting this nonlinearity of $\varphi(U)$. The remaining plots of $\beta_r(U)$, $r > 0$, are still constants.

The second common assumption used in practice is the assumption that $\phi(U) \propto U$ and $\psi(U) = 1$. That is, the special linear distortion is believed to only affect the predictor variable and the response variable is unaffected by the confounder. In this case violation of this assumption can be determined by checking for departures from $\beta_0(U) =$ constant and $\beta_1(U) \propto U^{-1}$. Figure 3 illustrates the estimated varying coefficient functions, $\widehat{\beta}_0(U)$ and $\widehat{\beta}_r(U)$, $r = 1, 2$, for this case and Figure 4 illustrates the above case where the distorting functions are proportional to the confounder U: $\psi(U) = \phi_r(U) \propto U$. In both cases, the data were generated using the same parameters as the motivating example introduced earlier in Section 1.2 (and also summarized in Figure 1). The examples displayed in Figures 3 and 4 use the local linear regression estimation procedure described in Section 2.2. As discussed earlier, the local linear regression modeling require selection of the bandwidth h. We used

generalized cross-validation (Wahba, 1977; Craven and Wahba, 1979) as previously described in Şentürk and Nguyen (2006) with the the Epanechnikov kernel, $K(t) = 0.75(1-t)_+^2$. That is, the bandwidth h is chosen to minimize the generalized cross-validation criterion: $n^{-1}||\widetilde{\mathbf{Y}} - \widehat{\widetilde{\mathbf{Y}}}||^2/[1 - n^{-1}tr(\mathbf{V})]^2$, where $\mathbf{V}$ is the "hat" matrix in $\widehat{\widetilde{\mathbf{Y}}} = (\widehat{\widetilde{Y}}_1, \ldots, \widehat{\widetilde{Y}}_n)^{\mathrm{T}} = \mathbf{V}\widetilde{\mathbf{Y}}$ and $\widehat{\widetilde{Y}}_i = \sum_{r=0}^{p} \widehat{\beta}_r(U_i)\widetilde{X}_{ir}$ (with $\widetilde{X}_{i0} \equiv 1$).

Finally, we note that under the multiplicative distortion, if the distortion processes on the response and predictors are the same, whether they are linear or nonlinear, then $\beta_r(U)$ $(r \geq 1)$ are constants. Consequently, plotting the estimated intercept function $\widehat{\beta}_0(U)$ provides the functional form of the common distortion, since $\beta_0(U) = \gamma_0\varphi(U)$, where $\varphi(U) \equiv \psi(U) = \phi_r(U)$.

3.3 A data example: Graphical assessment of the distorting effect of BMI on cholesterol and blood pressure measurements

In this data example, we examine the distortion effect of body mass index (BMI) on the regression relationship between serum cholesterol (SC; mg/100ml) and blood pressure (BP) measurements, namely systolic BP (SBP; mm Hg) and diastolic BP (DBP; mm Hg). The underlying relationship under exploration is $SC = \gamma_0 + \gamma_1 SBP + \gamma_2 DBP + e$. It is postulated that both response and predictor measurements may be affected by each individual's body mass index, resulting in the observed data $\{\widetilde{SC}_i, \widetilde{SBP}_i, \widetilde{DBP}_i\}_{i=1}^n$ and $\{BMI_i\}_{i=1}^n$ are the measurements on the confounder for n individuals. As we discussed above, the distortion effect may be null (i.e., $\psi(BMI) = \phi_r(BMI) = 1$ for multiplicative distortion or $\psi(BMI) = \phi_r(BMI) = 0$ for additive distortion), additive, or multiplicative. We explore some of these possibilities as well as the functional form of the distortion.

The data that will be examined here was obtained from the National Health and Nutrition Survey (NHANES) and is available from Hosmer and Lemeshow (2000). For illustration, we analyzed a random subset of $n = 1000$ observations (from 7,344 complete observations available for male subjects). Based on the observed data, we fit the varying coefficient model $\widetilde{SC}_i = \beta_0(BMI_i) + \beta_1(BMI_i)\widetilde{SBP}_{i1} + \beta_2(BMI_i)\widetilde{DBP}_{i2} + \epsilon(BMI_i)$, $i = 1, \ldots, 1000$. Using covariate-adjusted regression (Section **??**), the estimated relationship between serum cholesterol and blood pressure, adjusted for the effect of BMI, is given by $\widehat{SC} = 131.21 + 0.3199SBP + 0.3904DBP$, i.e. with $\widehat{\boldsymbol{\gamma}} = (131.21, 0.3199, 0.3904)^{\mathrm{T}}$. The standard error estimates for the $\widehat{\gamma}_r$ can be obtained using the bootstrap, as described in Şentürk and Nguyen (2006). Based on 300 bootstrap samples, the standard error estimates corresponding to $\widehat{\boldsymbol{\gamma}}$ are (13.196, 0.0996, 0.1486). Not

surprisingly, predicted SC is positively related to BP, after adjusting for BMI.

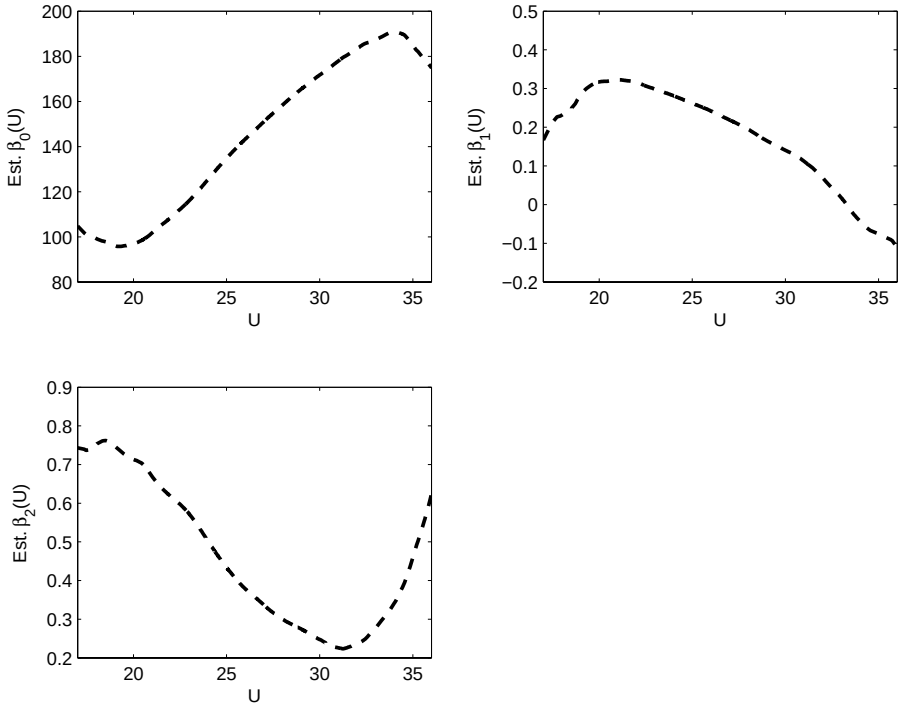

Figure 5: **Distortion in cholesterol-blood pressure data example.** Displayed are the estimate of the distorting function $\beta_0(U) \propto \psi(U)$ on cholesterol (top left) and the transformed distortion functions (varying coefficient functions) corresponding to SBP (right) and DBP (below), with bandwidth $h = 8.0$.

To explore the type and form of the distortion effects of BMI on SC, SBP, and DBP, we examine the corresponding estimated varying coefficient functions $\widehat{\beta}_r(BMI)$, $r = 0, 1, 2$. Figure 5 displays these estimated functions, obtained using a bandwidth of $h = 8$ from generalized cross-validation. Because the estimated varying coefficient functions corresponding to $\widetilde{SBP}$ and $\widetilde{DBP}$ (i.e. $\widehat{\beta}_1(BMI)$ and $\widehat{\beta}_2(BMI)$) are not both constants, the hypothesis of no-distortion effects of BMI on both the response and predictors is not supported. In fact, the estimated varying coefficient functions vary significantly with BMI, so the hypothesis/assumption that the distortion is additive is also not tenable. The estimated functions suggest a multiplicative distortion where $\psi(BMI) \neq \phi_r(BMI)$. Under multiplicative distortion, the assumption of no-distortion on the response variable only (i.e. $\psi(BMI) = 1$) and the assumption of a special linear distortion (i.e. $\psi(BMI) = \phi_r(BMI) \propto BMI$) are not compatible with the observed data. Furthermore, because $\beta_0(BMI) \propto \psi(BMI)$, the form of the distortion on

the response variable, serum cholesterol, can be inferred directly from the plot of $\widehat{\beta}_0(BMI)$. As can be seen from Figure **??** the distortion on the response is approximately linear and increasing in BMI in a wide range of observed body mass index (mean $BMI \pm 1.5$ standard deviation: 19.6-33.4). Thus, increasing BMI has an overall monotonic increasing and linear-multiplicative effect on serum cholesterol in this range. The estimated varying coefficient functions corresponding to $\widetilde{SBP}$ and $\widetilde{DBP}$ suggest that the distortion structure on SBP and DBP are more complex and may not be strictly linear throughout the range of BMI.

Finally, we note that the assumption of a *common distortion form* that affects both the response and predictors, *whether linear or nonlinear*, is not compatible with the observed data. That is, the distortion effect of BMI on cholesterol appears to be different than the distortion on blood pressure measurements (SBP and DBP).

4. Discussion

The covariate-adjusted regression model framework (2.1)-(2.3) provides a consistent estimation procedure that is automatically adaptive to the case of no-distortion as well as linear and nonlinear additive or multiplicative distortion. Using this consistent estimation procedure as a basis, we have proposed simple graphical techniques to further assess violations of specific assumptions on the forms and types of distortion under the CAR model framework. In real data applications, various simpler adjustment methods are commonly used under specific assumptions on the distortion form and type. Diagnostic techniques presented here can be used to better understand the distortion structures and facilitate interpretation, as well as checking for departures from specific model assumptions. As illustrated with various examples, the approach is feasible due the simple local linear regression estimation of the varying coefficient functions.

When estimating the varying coefficient functions, $\beta_r(U)$, selection of the bandwidth h can be chosen using the generalized cross-validation (GCV) criterion, for example. Generally, the choice of h is a trade-off between bias and variance. For estimation of the underlying parameters γ_r, GCV works well to balance the bias and variance (Şentürk and Nguyen, 2006). However, even with the use of GCV, the estimates $\widehat{\beta}_r(U)$ may not be sufficiently smooth for the graphical uses described here. There are various reasons for this, one of which is the different degrees of smoothness of the functions $\beta_r(U)$, $r = 0, \ldots, p$. For the graphical diagnostic purposes, one can reduce the variability by oversmoothing, using the chosen GCV choice of h as an initial guideline for the amount of oversmoothing. For the data example above, oversmoothing gave similar results as the GCV choice of $h = 8.0$. However, our experience with other data sets suggests that this

"oversmoothing" after GCV selection may work better for graphical assessment of distortion assumptions. Alternatively, one can also use a two-step local linear approach to estimate the varying coefficient functions (Fan and Zhang, 1999), where the initial (first-step) estimate of $\beta_r(U)$ is obtained by undersmoothing so that the bias is small. A re-estimation (re-smoothing) is done in the second step. Such an approach can be incorporated into the CAR estimation method and graphical diagnosis of assumptions on the data distortion.

Acknowledgement

DVN was partially supported by NIEHS grant 2 P01 ES011269-6.

References

Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying coefficient models. *Journal of the American Statistical Association* **95**, 888-902.

Chen, R., and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**, 298-308.

Chiang, C., Rice, J. A., and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association* **96**, 605-17.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1991). Local regression models. In *Statistical Models in S* (Chambers, J. M., and Hastie, T. J., eds), 309-376. Wadsworth & Brooks.

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics* **31**, 377-403.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall.

Fan, J., and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics* **27**, 1491-1518.

Hastie, T., and Tibshirani, R. (1993). Varying coefficient models. *Journal of the Royal Statistical Society, Series B* **55**, 757-96.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Statistical Society, Series B* **48**, 244-248.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edition. John Wiley and Sons.

Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. *Journal of the American Statistical Association* **8**, 680-688.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809-822.

Iturria, S., Carroll, R. J. and Firth, D. (1999). Polynomial regression and estimating functions in the presence of multiplicative measurement error. *Journal of the Royal Statistical Society, Series B* **61**, 547-561.

Kaysen, G. A., Dubin, J. A., Müller, H. G., Mitch, W. E., Rosales, L. M., Levin, N. W., and the Hemo Study Group (2003). Relationship among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. *Kidney International* **61**, 2240-2249.

Pinter, J. D., Brown, W. E., Eliez, S., Schmitt, J. E., Capone, G. T., and Reiss, A. L. (2001). Amygdala and hippocampal volumes in children with Down syndrome: A high-resolution MRI study. *Neurology* **56**, 972-974.

Şentürk, D. (2006). Covariate-adjusted varying coefficient models. *Biostatistics* **7**, 235-251.

Şentürk, D., and Müller, H. G. (2005). Covariate-adjusted regression. *Biometrika* **92**, 75-89.

Şentürk, D., and Nguyen, D. V. (2006). Estimation in covariate-adjusted regression. *Computational statistics and Data Analysis*, in-press.

Şentürk, D., and Müller, H. G. (2006). Inference for covariate-adjusted regression via varying coefficient models. *Annals of Statistics*, in-press.

Schisterman, E. F., Whitcomb, B. W., Louis, G. M. B., Louis, T. A. (2005) Lipid adjustment in the analysis of environmental contaminants and human health risks. *Environmental Health Perspectives* **113**, 853-857.

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (Edited by P. R. Krisnaiah), 507-523. Amsterdam: North Holland.

Wu, C. O., and Chiang, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica* **10**, 433-456.

Wu, C. O., and Yu, K. F. (2002). Nonparametric varying coefficient models for the analysis of longitudinal data. *International Statistical Review* **70**, 373-393.

Received May 1, 2005; accepted September 10, 2006.

Danh V. Nguyen
Division of Biostatistics
University of California School of Medicine
Davis, CA 95616, U.S.A.
ucdnguyen@ucdavis.edu

Damla Şentürk
Department of Statistics
Pennsylvania State University
University Park, PA 16802, U.S.A.
dsenturk@stat.psu.edu