Partial Covariate Adjusted Regression

By DAMLA ŞENTÜRK

Department of Statistics, Pennsylvania State University, University Park

dsenturk@stat.psu.edu

And DANH V. NGUYEN

Division of Biostatistics, University of California School of Medicine, Davis

ucdnguyen@ucdavis.edu

Summary

Covariate adjusted regression (CAR) is a recently proposed adjustment method for regression analysis where both the response and predictors are not directly observed (Şentürk and Müller, 2005). The available data has been distorted by unknown functions of an observable confounding covariate. CAR provides consistent estimators for the coefficients of the regression between the variables of interest, adjusted for the confounder. We develop a broader class of partial covariate adjusted regression (PCAR) models to accommodate undistorted as well as distorted predictors. The PCAR model allows for the modeling of undistorted predictors, such as age, gender and demographic variables which are commonly included in medical and biological data applications. The estimation and inference procedures developed earlier for CAR are shown not to be valid for the proposed PCAR model. We propose new estimators and develop new inference tools for the more general PCAR setting. In particular, we establish the asymptotic normality of the proposed estimators and propose consistent estimators of their asymptotic variances. Finite sample properties of the proposed estimators are investigated using simulation studies and the method is also illustrated with a Pima Indians diabetes data set.

Some key words: Asymptotic normality; Binning; Confidence intervals; Multiple regression; Varying-coefficient models.

1. Introduction and background

Covariate adjusted regression has been recently proposed to adjust for the distorting effects of a confounder in a regression setting. It was motivated by a common adjustment method in medical and health related studies. The adjustment entails normalization by anthropometric measurements, such as body mass index (BMI) and/or other measures of body configuration, as confounding variables that affect the primary variables of interest. For example, in a study involving haemodialysis patients, it is of interest to examine the relationship between elevated plasma fibrinogen level (a risk factor for cardiovascular disease in haemodialysis patients) and other predictors, such as serum transferrin protein level (Kaysen et al., 2003; Şentürk and Müller, 2005). However, both primary variables, fibrinogen and transferrin protein levels, are known to depend on body mass index, 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 BMI). Note that this adjustment by division implies that the assumed contamination is of a multiplicative form. Let $\widetilde{Y}$, $\widetilde{X}$, and U denote the observed fibrinogen concentration, serum transferrin level, and confounder BMI, respectively. Using these notations, the adjusted primary variables that are thought to be free from the confounding effect of BMI are,

$$Y = \frac{\widetilde{Y}}{U} \qquad \text{and} \qquad X = \frac{\widetilde{X}}{U}.$$

The basic motivation for the above adjustment is to obtain normalized versions of the observed primary variables by removing the confounder effects, so that the measurements are comparable across patients. Other examples include normalizations by BMI in studies on diabetes, and division of brain volumetric structures by total brain volume in neurological studies (Pinter et al., 2001).

Şentürk and Müller (2005, 2006) proposed a more flexible adjustment, by modeling the confounding through *unknown functions* of the confounder instead of the confounder itself. This reflects the uncertainty encountered in many applications about the precise nature of the commonly assumed multiplicative relation between the confounder and the variables. For the case of p predictors, Şentürk and Müller model the underlying variables as

$$Y = \frac{\widetilde{Y}}{\psi(U)}, \quad X_1 = \frac{\widetilde{X}_1}{\phi_1(U)}, \quad \ldots, \quad X_p = \frac{\widetilde{X}_p}{\phi_p(U)},$$

where they are defined to be the parts of the observed variables, $\widetilde{Y}$, $\widetilde{X}_1, \ldots, \widetilde{X}_p$, that are

independent of the observable confounder U. In the haemodialysis data example, the latent variables would be defined to be serum protein levels adjusted for body mass index. Here, $\phi_1(\cdot), \ldots, \phi_p(\cdot)$ and $\psi(\cdot)$ denote unknown smooth contaminating functions of U. CAR gives consistent estimators of the coefficients in the unobserved regression model which can be expressed as

$$Y = \gamma_0 + \sum_{r=1}^{p} \gamma_r X_r + e,$$

where e is the error term, assumed to be independent of $\{X_r\}_{r=1}^p$ and U. The estimation procedure is based on the observed data: the distorted response, $\widetilde{Y}$, distorted predictors, $\{\widetilde{X}_r\}_{r=1}^p$, and confounder U.

Note that a main attraction of CAR is that under the identifiability conditions introduced in Section 2, it yields consistent estimates whether the distortion is multiplicative or additive (i.e. $Y = \widetilde{Y} - \psi(U)$, $X_r = \widetilde{X}_r - \phi_r(U)$). Additive distortions can be handled by the method of nonparametric partial regression; however, there existed no consistent estimation procedure targeting the γ's under multiplicative distortion of both the response and the predictors. In addition, not having to decide on the form of distortion (additive or multiplicative) a priori adds to the flexibility of the CAR methodology.

The main goal of this paper is to construct estimation and inference procedures needed to model some of the variables as undistorted predictors denoted by $Z_1, \ldots, Z_s$. The proposed underlying regression model is of the form

$$Y = \gamma_0 + \sum_{r=1}^{p} \gamma_r X_r + \sum_{s=1}^{q} \delta_s Z_s + e,$$

where $\widetilde{X}_r = \phi(U)X_r$ and $\widetilde{Y} = \psi(U)Y$ denote the distorted predictors and response, respectively, and Z_s denotes the undistorted predictors. The observed data is $\widetilde{Y}$, $\widetilde{X}_r$, Z_s and U. Furthermore, the confounding covariate, U, is allowed to depend on the undistorted predictors $\{Z_s\}$.

The flexibility of modeling both distorted and undistorted predictors is needed especially in the regression analysis of biomedical data, since commonly encountered predictors in medical data, including age, gender, obesity measures, and ethnicity among others, are directly observable. Modeling these variables as distorted yields latent variables that are not easily interpretable. For instance, consider the additional predictor age in the haemodialysis example discussed earlier. If age is modeled as distorted by BMI, then the corresponding

latent variable, defined as the part of age that is independent of the confounder BMI, does not have a clear interpretation. Thus, an adjustment method is needed that is flexible enough to model undistorted predictors as well as distorted ones, to enhance the applicability of CAR.

Under the more general partial covariate adjusted regression (PCAR) setting, formally presented in Section 2, the original CAR estimators (Şentürk and Müller, 2005) for $\{\delta_s\}_{s=1}^q$ are inconsistent. More precisely, they target $\delta_s C_s$, where the factor C_s can get arbitrarily large as shown in Section 3. We propose alternative estimators that are consistent under this extended CAR setting where the issues of estimation are discussed in Section 3. The inference procedures developed for the CAR modeling are not valid for the PCAR setting, mainly due to the different dependence structure needed for PCAR. This new structure is explained in detail in Sections 2 and 3. Thus, new theoretical tools has to be developed for inference in the PCAR model. We derive at the asymptotic distributions of the proposed estimators, and present them in Section 4. Consistent estimators of the asymptotic variance are also derived in Section 4. Simulation studies to characterize the finite sample properties of the proposed estimators are summarized in Section 5.2. The method is further illustrated with a Pima Indians diabetes data set, where the response, plasma glucose concentration and the predictor, diastolic blood pressure are distorted by body mass index (Section 5.1). The undistorted predictors in the regression analysis are age and triceps skin fold thickness, which are directly observable and are allowed to depend on body mass index. The proofs of the main results are assembled in Section 6, where some technical conditions and auxiliary results are deferred to the Appendix.

Note that the proposed distortion setting has similarities with the well developed area of measurement error modeling; however a main difference is that in the proposed distortion setting the error is a function of an observable covariate U. Even though additive measurement error modeling has been considered extensively in literature, work on multiplicative measurement errors is limited. Hwang (1986) and Iturria, Carroll, and Firth (1999) propose estimation procedures targeting the regression coefficients under multiplicative measurement error in the predictors. However, the case of multiplicative measurement errors that affect both the predictors and the response has not been considered previously to our knowledge.

2. Partial covariate adjusted regression models

We consider the underlying (unobserved) regression model

$$Y_{ni} = \gamma_0 + \sum_{r=1}^{p} \gamma_r X_{nri} + \sum_{s=1}^{q} \delta_s Z_{nsi} + e_{ni} = \boldsymbol{\chi}_{ni}^{\mathrm{T}} \boldsymbol{\alpha} + e_{ni}, \tag{1}$$

where Y_{ni}, e_{ni}, $\boldsymbol{\chi}_{ni} = (1, X_{n1i}, \ldots, X_{npi}, Z_{n1i}, \ldots, Z_{nqi})^{\mathrm{T}}$ and $\boldsymbol{\alpha} = (\gamma_0, \ldots, \gamma_p, \delta_1, \ldots, \delta_q)^{\mathrm{T}}$ are the response, error, $p+q$ predictors and unknown regression coefficients, respectively. The error variable e has mean zero and variance σ^2. The goal is estimation and inference for the parameter vector $\boldsymbol{\alpha}$ of the unobserved regression model (1). Estimation is based on available distorted predictor and response data, namely $\{\widetilde{Y}_{ni}, \widetilde{\boldsymbol{\chi}}_{ni}, U_{ni}\}_{i=1}^{n}$, where

$$\begin{aligned} \widetilde{Y}_{ni} = \psi(U_{ni}) Y_{ni} \quad \text{and} \quad \widetilde{\boldsymbol{\chi}}_{ni} &= (1, \widetilde{X}_{n1i}, \ldots, \widetilde{X}_{npi}, Z_{n1i}, \ldots, Z_{nqi})^{\mathrm{T}} \\ &= \{1, \phi_1(U_{ni}) X_{n1i}, \ldots, \phi_p(U_{ni}) X_{npi}, Z_{n1i}, \ldots, Z_{nqi}\}^{\mathrm{T}}. \end{aligned} \tag{2}$$

The unknown distorting functions $\{\psi(\cdot), \phi_r(\cdot)\}_{r=1}^{p}$ are assumed to be smooth functions of the confounder, U.

Some constraints on the unknown smooth distortion functions are needed for the identifiability of the estimation problem. A set of reasonable constraints for $\psi(\cdot)$ and $\{\phi_r(\cdot)\}$ is implied by the natural assumption that the mean distorting effect should correspond to no distortion (Şentürk and Müller, 2005), i.e.

$$E\{\psi(U)\} = 1 \qquad \text{and} \qquad E\{\phi_r(U)\} = 1. \tag{3}$$

These conditions directly imply that the means of adjusted variables are the same as the means of the observed variables, i.e. $E(\widetilde{X}_r) = E(X_r)$ and $E(\widetilde{Y}) = E(Y)$.

We consider the following dependence structure. The underlying predictors X_r and the undistorted predictors Z_s are allowed to be dependent. The error, e, is assumed to be mutually independent of X_r, Z_s, and U. We depart from the original CAR model (Şentürk and Müller, 2005), where U is independent of all the latent predictors, by allowing U to depend on Z_s, while still being independent of X_r. We will elaborate further on this important difference of the proposed setting from CAR at the end of Section 3. This is an important flexibility of the proposed method, since the common confounder correlates with all the observed variables in these distortion settings. This is consistent with the assumption that the observed predictors $\widetilde{X}_r$ and Z_s are dependent on the confounder U, and that the latent variable X_r is defined to be the part of $\widetilde{X}_r$ that is independent of U.

The assumption that the underlying predictors, $\{X_r\}_{r=1}^p$, and response, Y, are independent of the contaminating variable U is a fundamental assumption for the estimation procedure. It defines the proposed contamination setting through defining the unobserved, underlying variables. This independence assumption cannot be checked in practice since X_r and Y are unobservable. Instead, the question of more relevance in practice is whether the independence conditions help define interpretable latent variables of interest from their observable counterparts. In the haemodialysis data example, the latent variables are defined to be serum protein levels adjusted for body mass index, which are commonly used in medical studies.

We refer to the model described by (1)-(3) as the partial covariate adjusted regression (PCAR) model, since only a partial set of the predictors are adjusted for the confounder. Note also that the CAR model is a special case of the PCAR model.

For the estimation, we show that a regression of $\widetilde{Y}$ on $\widetilde{\boldsymbol{\chi}} = (1, \widetilde{X}_1, \ldots, \widetilde{X}_p, Z_1, \ldots, Z_q)^{\mathrm{T}}$ leads to a fully observable varying coefficient model. Our estimation (and inference) procedure for the unobserved model relies on this observable varying coefficient model. More precisely, the regression of $\widetilde{Y}$ on $\widetilde{\boldsymbol{\chi}}$ leads to the following relation,

$$\begin{aligned} E(\widetilde{Y}|\widetilde{\boldsymbol{\chi}}, U) &= E\left\{Y\psi(U)|\phi_1(U)X_1, \ldots, \phi_p(U)X_p, Z_1, \ldots, Z_q, U\right\} \\ &= \psi(U)E\left\{\gamma_0 + \sum_{r=1}^{p}\gamma_r X_r + \sum_{s=1}^{q}\delta_s Z_s + e\Big|\phi_1(U)X_1, \ldots, \phi_p(U)X_p, Z_1, \ldots, Z_q, U\right\}. \end{aligned}$$

Straightforward simplifications of the above expression by making use of (2) and the mutual independence of $\{e \text{ and } U\}$, $\{e \text{ and } (X_r, Z_s)\}$, and $\{U \text{ and } X_r\}$, for $r = 1, \ldots, p$, $s = 1, \ldots, q$, give

$$\begin{aligned} E(\widetilde{Y}|\widetilde{\boldsymbol{\chi}}, U) &= \psi(U)\gamma_0 + \psi(U)\sum_{r=1}^{p}\gamma_r\frac{\phi_r(U)X_r}{\phi_r(U)} + \psi(U)\sum_{s=1}^{q}\delta_s Z_s \\ &= \beta_0(U) + \sum_{r=1}^{p}\beta_r(U)\widetilde{X}_r + \sum_{s=1}^{q}\eta_s(U)Z_s, \end{aligned}$$

where

$$\beta_0(u) = \gamma_0\psi(u), \qquad \beta_r(u) = \gamma_r\frac{\psi(u)}{\phi_r(u)} \qquad \text{and} \qquad \eta_s(u) = \delta_s\psi(u). \tag{4}$$

Therefore, the regression of $\widetilde{Y}$ on $\widetilde{\boldsymbol{\chi}}$ leads to the following varying coefficient model (Cleveland, Grosse and Shyu, 1991; Hastie and Tibshirani, 1993),

$$\widetilde{Y}_{ni} = \beta_0(U_{ni}) + \sum_{r=1}^{p} \beta_r(U_{ni})\widetilde{X}_{nri} + \sum_{s=1}^{q} \eta_s(U_{ni})Z_{nsi} + \epsilon_{ni} \tag{5}$$

with $\epsilon_{ni} \equiv \psi(U_{ni})e_{ni}$. Note that in (5), the varying coefficient functions, $\{\beta_r(\cdot)\}_{r=1}^{p}$, are proportional to the quotient of the original distorting functions, $\{\psi(\cdot)/\phi_r(\cdot)\}$; both the intercept function, $\beta_0(\cdot)$, and the functions $\{\eta_s(\cdot)\}_{s=1}^{q}$ are proportional to $\psi(\cdot)$. The constants of proportionality are precisely the underlying regression parameters, $\{\gamma_r, \delta_s\}$, of interest. These connections allow estimation of the underlying model through the varying coefficient functions. In Section 3 below, we describe an estimation procedure which targets $\{\gamma_r, \delta_s\}$ and mitigates the effects of the distorting functions $\{\psi(\cdot), \phi_r(\cdot)\}$.

Varying coefficient models are popular in many applications. They are appealing extensions of the regression models where the coefficients are allowed to vary as smooth functions of a covariate possibly different than the predictors. The literature includes, among others, Ramsay and Silverman (1997) on functional data analysis, Nicholls and Quinn (1982) and Chen and Tsay (1993) on nonlinear time series, and Wu and Yu (2002) on applications to longitudinal data. Some approaches to estimation in varying coefficient models for independent and identically distributed data are described in Hoover, Rice, Wu and Yang (1998), Wu and Chiang (2000), Chiang, Rice and Wu (2001), and Cai, Fan and Li (2000).

3. Estimation procedure

The estimation of the regression coefficients, γ_0, $\{\gamma_r\}_{r=1}^{p}$ and $\{\delta_s\}_{s=1}^{q}$, in the underlying regression model $E(Y) = \gamma_0 + \sum_{r=1}^{p} \gamma_r X_r + \sum_{r=1}^{q} \delta_s Z_s$ is a two-step procedure. The first step involves estimation of the varying coefficient functions in model (5), namely $\beta_0(\cdot)$, $\{\beta_r(\cdot)\}_{r=1}^{p}$ and $\{\eta_s(\cdot)\}_{s=1}^{q}$ using a binning approach. These varying functions are estimable because $\widetilde{Y}$, $\widetilde{X}_r$, Z_s, and U are all observable. The underlying regression coefficients are targeted in the second step, with weighted averages of the estimated $\beta_0(\cdot)$, $\beta_r(\cdot)$ and $\eta_s(\cdot)$ for γ_0, γ_r and δ_s, respectively. The estimation makes use of the relations between the varying coefficient functions and the regression coefficients given by (4) and the identifiability conditions (3), as will be described next.

The binning approach for the estimation of the varying coefficient functions involves dividing the support of U into m equidistant bins and then fitting linear regressions of $\widetilde{Y}$ on $\widetilde{\boldsymbol{\chi}}$ using the data falling within each bin. The observed data is the collection of n samples: $\{\widetilde{Y}_{ni}, \widetilde{\boldsymbol{\chi}}_{ni}, U_{ni}\}_{i=1}^{n}$. It is assumed that the confounding covariate, U, is bounded below and

above, $a \leq U \leq b$, where $a < b$ are real numbers. The estimation procedure initially divides the interval $[a, b]$ into m equidistant intervals, denoted $B_{n1}, \ldots, B_{nm}$ and referred to as bins. Let L_{nj} be the number of U_{ni}'s falling into bin j. Furthermore, denote the data for which $U_{ni} \in B_{nj}$ by the collection $\{(U'_{njk}, \widetilde{X}'_{nrjk}, Z'_{nsjk}, \widetilde{Y}'_{njk}, X'_{nrjk}, Y'_{njk}),\ r = 1, \ldots, p,\ s = 1, \ldots, q,\ k = 1, \ldots, L_{nj}\} = \{(U_{ni}, \widetilde{X}_{nri}, Z_{nsi}, \widetilde{Y}_{ni}, X_{nri}, Y_{ni}), i = 1, \ldots, n,\ r = 1, \ldots, p,\ s = 1, \ldots, q : U_{ni} \in B_{nj}\}$, where $(U'_{njk}, \widetilde{X}'_{nrjk}, Z'_{nsjk}, \widetilde{Y}'_{njk}, X'_{nrjk}, Y'_{njk})$ is the kth data element in the jth bin, B_{nj}. Data elements in any given bin are marked by a prime.

After the initial binning of the data, a linear regression is fitted to the data observed within each bin B_{nj}, $j = 1, \ldots, m$. The least squares estimator of the multiple regression of the data in the jth bin is

$$(\widehat{\beta}_{n0j}, \ldots, \widehat{\beta}_{npj}, \widehat{\eta}_{n1j}, \ldots, \widehat{\eta}_{nqj})^{\mathrm{T}} = (\widetilde{\boldsymbol{\chi}}'^{\mathrm{T}}_{nj} \widetilde{\boldsymbol{\chi}}'_{nj})^{-1} \widetilde{\boldsymbol{\chi}}'^{\mathrm{T}}_{nj} \widetilde{\mathbf{Y}}'_{nj}, \tag{6}$$

where the response vector is $\widetilde{\mathbf{Y}}'_{nj} = (\widetilde{Y}'_{nj1}, \ldots, \widetilde{Y}'_{njL_{nj}})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\chi}}'_{nj} = (\widetilde{\boldsymbol{\chi}}'_{nj1}, \ldots, \widetilde{\boldsymbol{\chi}}'_{njL_{nj}})^{\mathrm{T}}$ is the $L_{nj} \times (p+q+1)$ data matrix in bin j, with the kth observation $\widetilde{\boldsymbol{\chi}}'_{njk} = (1, \widetilde{X}'_{n1jk}, \ldots, \widetilde{X}'_{npjk}, Z'_{n1jk}, \ldots, Z'_{nqjk})^{\mathrm{T}}$. The estimated regression coefficients in each bin (6) are the raw estimators of the varying coefficient functions.

In the second step of the estimation procedure, the estimators of the targeted regression parameters, γ_0, $\{\gamma_r\}_{r=1}^p$ and $\{\delta_s\}_{s=1}^q$, are obtained as weighted averages of the raw estimators $\{\widehat{\beta}_{n0j}, \ldots, \widehat{\beta}_{npj}, \widehat{\eta}_{n1j}, \ldots, \widehat{\eta}_{nqj}\}_{j=1}^m$ from the m bins. The proposed PCAR estimators for γ_0, $\{\gamma_r\}_{r=1}^p$ and $\{\delta_s\}_{s=1}^q$ are

$$\widehat{\gamma}_{n0} = \sum_{j=1}^{m} \frac{L_{nj}}{n} \widehat{\beta}_{n0j}, \qquad \widehat{\gamma}_{nr} = \frac{1}{\overline{\widetilde{X}}_{nr}} \sum_{j=1}^{m} \frac{L_{nj}}{n} \widehat{\beta}_{nrj} \overline{\widetilde{X}'}_{nrj} \qquad \text{and} \qquad \widehat{\delta}_{ns} = \sum_{j=1}^{m} \frac{L_{nj}}{n} \widehat{\eta}_{nsj}, \tag{7}$$

where $\overline{\widetilde{X}}_{nr} = n^{-1} \sum_{i=1}^{n} \widetilde{X}_{nri}$ and $\overline{\widetilde{X}'}_{nrj} = L_{nj}^{-1} \sum_{k=1}^{L_{nj}} \widetilde{X}'_{nrjk}$. The above weighted averages are motivated by the relations

$$\begin{aligned} E\{\beta_0(U)\} &= \gamma_0 E\{\psi(U)\} = \gamma_0, \\ E\{\beta_r(U)\widetilde{X}_r\} &= \gamma_r E\{\psi(U) X_r\} = \gamma_r E(X_r) = \gamma_r E(\widetilde{X}_r) \qquad \text{and} \\ E\{\eta_s(U)\} &= \delta_s E\{\psi(U)\} = \delta_s. \end{aligned}$$

Note that the weights in equation (7) depend on the number of data points in each bin, namely L_{nj} for $j = 1, \ldots, m$. The above relations follow directly from identifiability condition (3) and the derived connection between the underlying regression parameters and the varying coefficient functions (4).

We note that the estimators $\widehat{\gamma}_{n0}$ and $\widehat{\gamma}_{nr}$ have the same form as the CAR estimators (Şentürk and Müller, 2005), whereas $\widehat{\delta}_{ns}$ are different. Furthermore, a straightforward application of the CAR estimators give inconsistent estimators for δ_s under the more general PCAR model. To see this, denote the original CAR estimators for δ_s by $\{\widehat{\delta}^*_{ns}\}_{s=1}^q$. It follows from Şentürk and Müller (2005), that

$$\widehat{\delta}^*_{ns} = \frac{1}{\overline{Z}_{ns}} \sum_{j=1}^{m} \frac{L_{nj}}{n} \widehat{\eta}_{nsj} \overline{Z'}_{nsj},$$

where $\overline{Z}_{ns} = n^{-1} \sum_{i=1}^{n} Z_{nsi}$ and $\overline{Z'}_{nsj} = L_{nj}^{-1} \sum_{k=1}^{L_{nj}} Z'_{nsjk}$. The estimators $\widehat{\delta}^*_{ns}$ do not target δ_s, instead they target $E\{\eta_s(U)Z_s\}/E(Z_s) = \delta_s E\{\psi(U)Z_s\}/E(Z_s) = \delta_s C_s$, where $C_s \equiv E\{\psi(U)Z_s\}/E(Z_s) = [cov\{\psi(U), Z_s\}/E(Z_s)]+1$ can get arbitrary large as $E(Z_s)$ approaches zero.

The bias of the CAR estimator is mainly due to the important assumption needed for the undistorted predictors, namely that they are dependent on U. (Note that if Z_s is assumed to be independent of U, $C_s = 1$ because of the identifiability conditions.) CAR is not designed to handle this dependence structure, because it was proposed for the modeling of only distorted predictors and under the assumption that their latent counterparts are independent of U. The independence assumption does not pose a problem if the observed predictor is modeled as distorted. Note that this is because the assumption of independence is between the unobserved predictor and the confounder not the observed predictor and U. However if the predictor is modeled as undistorted, then the assumption requires that the observed predictor itself be independent of the confounder. This is not realistic, since U is the common confounder that correlates with all the observed variables in these distortion settings. In the data examples given in the Introduction, the undistorted predictor age is not independent of the confounders like body mass index, body weight or height. On the other hand, the proposed estimators, namely $\widehat{\delta}_{ns}$, are designed to handle this new dependence structure and yield consistent estimators for δ_s. Next we develop new inference tools for the proposed estimators.

4. Asymptotic properties

We present the asymptotic distribution of the estimators $\widehat{\gamma}_{n0}$, $\widehat{\gamma}_{nr}$ and $\widehat{\delta}_{ns}$ in (7) for γ_0, γ_r and δ_s in model (1), when the number of subjects n tends to infinity. Even though the proposed point estimators for γ_0 and γ_r have the same form as the previously proposed CAR

estimators, they will have different asymptotic distributions due to the new dependence structure. As in typical smoothing applications, the number of bins $m = m(n)$ is required to satisfy $m \to \infty$, $n/(m \log n) \to \infty$ and $m/\sqrt{n} \to \infty$ as $n \to \infty$. We denote convergence in distribution by $\xrightarrow{\mathcal{D}}$ and convergence in probability by $\xrightarrow{p}$.

Recall that the raw coefficient estimators coming from the least squares regression of the data in the jth bin were denoted by $(\widehat{\beta}_{n0j}, \ldots, \widehat{\beta}_{npj}, \widehat{\eta}_{n1j}, \ldots, \widehat{\eta}_{nqj})^{\mathrm{T}} = (\widetilde{\boldsymbol{\chi}}_{nj}'^{\mathrm{T}}\widetilde{\boldsymbol{\chi}}_{nj}')^{-1}\widetilde{\boldsymbol{\chi}}_{nj}'^{\mathrm{T}}\widetilde{\mathbf{Y}}_{nj}'$, given in (6). Similarly, let

$$(\widetilde{\gamma}_{n0j}, \ldots, \widetilde{\gamma}_{npj}, \widetilde{\delta}_{n1j}, \ldots, \widetilde{\delta}_{nqj})^{\mathrm{T}} = (\boldsymbol{\chi}_{nj}'^{\mathrm{T}}\boldsymbol{\chi}_{nj}')^{-1}\boldsymbol{\chi}_{nj}'^{\mathrm{T}}Y_{nj}'^{\mathrm{T}} \tag{8}$$

denote the least squares estimators of the multiple regression of the unobserved data falling into B_{nj}, where the vectors $\boldsymbol{\chi}_{nj}'$ and Y_{nj}' are defined the same way as $\widetilde{\boldsymbol{\chi}}_{nj}'$ and $\widetilde{Y}_{nj}'$, with X_{nrjk}' and Y_{njk}' replacing $\widetilde{X}_{nrjk}'$ and $\widetilde{Y}_{njk}'$, respectively. This quantity is not estimable, but will be used in the proof of the main results.

For the PCAR estimators given in (7) to be well defined, the least squares estimators given in (6) must exist for each bin B_{nj}. This requires that the inverse of $\widetilde{\boldsymbol{\chi}}_{nj}'^{\mathrm{T}}\widetilde{\boldsymbol{\chi}}_{nj}'$ is well defined, i.e. $\det(\widetilde{\boldsymbol{\chi}}_{nj}'^{\mathrm{T}}\widetilde{\boldsymbol{\chi}}_{nj}') \neq 0$. Correspondingly, the estimators in (8) will exist under the condition that $\det(\boldsymbol{\chi}_{nj}'^{\mathrm{T}}\boldsymbol{\chi}_{nj}') \neq 0$. Therefore, we define the events

$$\begin{aligned} \widetilde{A}_n &= \{\omega \in \Omega : \inf_j |\det(L_{nj}^{-1}\widetilde{\boldsymbol{\chi}}_{nj}'^{\mathrm{T}}\widetilde{\boldsymbol{\chi}}_{nj}')| > \zeta \text{ and } \min_j L_{nj} > p+q\}, \\ A_n &= \{\omega \in \Omega : \inf_j |\det(L_{nj}^{-1}\boldsymbol{\chi}_{nj}'^{\mathrm{T}}\boldsymbol{\chi}_{nj}')| > \zeta \text{ and } \min_j L_{nj} > p+q\}, \end{aligned} \tag{9}$$

where $\zeta = \min\{\rho/2, [\inf_j\{\phi_1^2(U_{nj}'^*), \ldots, \phi_p^2(U_{nj}'^*)\}]^p\rho/2\}$, ρ is as defined in (C5), $U_{nj}'^* = L_{nj}^{-1}\sum_{k=1}^{L_{nj}} U_{njk}'$ is the average of the U's in B_{nj}, and $(\Omega, \mathcal{F}, P)$ is the underlying probability space. The estimators in (7) and (8) are well defined on events $\widetilde{A}_n$ and A_n, respectively. The event E_n in Theorems 1 and 2 is defined to be the intersection of A_n and $\widetilde{A}_n$, i.e., $E_n = A \cap \widetilde{A}_n$. It is shown in the Appendix that $\mathrm{pr}(E_n) \to 1$ as $n \to \infty$.

For the following theorems and the proofs of the main results, we define the following notations:

1. $\lambda_\psi = E\{\psi^2(U)\}$, $\lambda_\phi = E\{\phi^2(U)\}$, $\lambda_{\psi\phi_r} = E\{\psi(U)\psi_r(U)\}$, $\sigma_\psi^2 = \mathrm{var}\{\psi(U)\}$,
2. $m_{r,k} = E(X_r^k)$,
3. $\boldsymbol{\chi}^{\mathrm{T}} = (1, X_1, \ldots, X_p, Z_1, \ldots, Z_q)$,

4. $\mathbf{\Gamma} = E(\boldsymbol{\chi}\boldsymbol{\chi}^{\mathrm{T}}|U)$,

5. $\mathcal{M} = \mathbf{\Gamma}^{-1}\boldsymbol{\chi}\boldsymbol{\chi}^{\mathrm{T}}\mathbf{\Gamma}^{-1}$,

6. $\widetilde{\mathbf{\Theta}}_{nj} = L_{nj}^{-1}\widetilde{\boldsymbol{\chi}}_{nj}^{\prime\mathrm{T}}\widetilde{\boldsymbol{\chi}}_{nj}^{\prime}$,

7. $\omega_{n,\ell,k} = (\widetilde{\mathbf{\Theta}}_{nj})_{\ell 1}^{-1} + (\widetilde{\mathbf{\Theta}}_{nj})_{\ell 2}^{-1}\widetilde{X}_{n1jk}^{\prime} + \ldots + (\widetilde{\mathbf{\Theta}}_{nj})_{\ell,p+q+1}^{-1}Z_{nqjk}^{\prime}$, for $\ell = 1, \ldots, p+s+1$ and $k = 1, \ldots, L_{nj}$.

Theorem 1. *Under the technical conditions (C1)-(C7) in Section 6, on event E_n with $pr(E_n) \to 1$ as $n \to \infty$,*

$$\begin{aligned} \sqrt{n}(\widehat{\gamma}_{nr} - \gamma_r) &\xrightarrow{\mathcal{D}} \mathbb{N}(0, \sigma_r^2), \qquad 0 \le r \le p, \\ \sqrt{n}(\widehat{\delta}_{ns} - \delta_s) &\xrightarrow{\mathcal{D}} \mathbb{N}(0, \sigma_s^2), \qquad 1 \le s \le q, \end{aligned}$$

where

$$\sigma_0^2 = \gamma_0^2\sigma_\psi^2 + \sigma^2 E\{\psi^2(U)\mathcal{M}_{11}\},$$

$$\sigma_r^2 = \frac{\gamma_r^2(\lambda_\psi m_{r,2} - m_{r,1}^2) + \sigma^2 m_{r,1}^2 E\{\psi^2(U)\mathcal{M}_{r+1,r+1}\} - 2\gamma_r^2(\lambda_{\psi\phi_r} m_{r,2} - m_{r,1}^2) - \gamma_r^2 var(\widetilde{X}_r)}{m_{r,1}^2},$$

$$\sigma_s^2 = \delta_s^2\sigma_\psi^2 + \sigma^2 E\{\psi^2(U)M_{p+s+1,p+s+1}\} \quad \textit{for } 1 \le r \le p \textit{ and } 1 \le s \le q.$$

Theorem 1 establishes the asymptotic normality of the proposed PCAR estimators. The following theorem provides consistent estimators of the asymptotic variances given in Theorem 1.

Theorem 2. *Under the technical conditions (C1)-(C7) in Section 6, on event E_n with $pr(E_n) \to 1$ as $n \to \infty$,*

$$\begin{aligned} \widehat{\sigma}_{nr}^2 &\xrightarrow{p} \sigma_r^2, \qquad 0 \le r \le p, \\ \widehat{\sigma}_{ns}^2 &\xrightarrow{p} \sigma_s^2, \qquad 1 \le s \le q, \end{aligned}$$

where

$$\widehat{\sigma}_{n0}^2 = \sum_{j=1}^{m} \frac{L_{nj}}{n}\widehat{\beta}_{n0j}^2 - \widehat{\gamma}_{n0}^2 + \widehat{\sigma}^2 \frac{n^{-1}\sum_{j=1}^{m}\widehat{\beta}_{n0j}^2 \sum_{k=1}^{L_{nj}}\omega_{n,1,k}^2}{\sum_{j=1}^{m} n^{-1}L_{nj}\widehat{\beta}_{n0j}^2},$$

$$
\begin{aligned}
\widehat{\sigma}^2_{nr} &= \frac{n^{-1}\sum_{j=1}^{m}\widehat{\beta}^2_{nrj}\sum_{k=1}^{L_{nj}}\widetilde{X}'^2_{nrjk}+\widehat{\gamma}^2_{nr}\overline{\widetilde{X}}^2_{nr}-2\widehat{\gamma}_{nr}n^{-1}\sum_{j=1}^{m}\widehat{\beta}_{nrj}\sum_{k=1}^{L_{nj}}\widetilde{X}'^2_{nrjk}+\widehat{\gamma}^2_{nr}s^2_{\widetilde{X}_r}}{\overline{\widetilde{X}}^2_{nr}} \\
&+ \widehat{\sigma}^2\frac{n^{-1}\sum_{j=1}^{m}\widehat{\beta}^2_{n0j}\overline{\widetilde{X}'}^2_{nrj}\sum_{k=1}^{L_{nj}}\omega^2_{n,r+1,k}}{\overline{\widetilde{X}}^2_{nr}\sum_{j=1}^{m}n^{-1}L_{nj}\widehat{\beta}^2_{n0j}} \\
\widehat{\sigma}^2_{ns} &= \sum_{j=1}^{m}\frac{L_{nj}}{n}\widehat{\eta}^2_{nsj}-\widehat{\delta}^2_{ns}+\widehat{\sigma}^2\frac{n^{-1}\sum_{j=1}^{m}\widehat{\beta}^2_{n0j}\sum_{k=1}^{L_{nj}}\omega^2_{n,p+s+1,k}}{\sum_{j=1}^{m}n^{-1}L_{nj}\widehat{\beta}^2_{n0j}}, \\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}(\widetilde{Y}'_{njk}-\widehat{\beta}_{n0j}-\sum_{r=1}^{p}\widehat{\beta}_{nrj}\widetilde{X}'_{nrjk}-\sum_{s=1}^{q}\widehat{\eta}_{nsj}Z'_{nsjk})^2 \quad \textit{and} \\
s^2_{\widetilde{X}_r} &= \frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{X}_{nri}-\overline{\widetilde{X}}_{nr})^2.
\end{aligned}
$$

Remark. These proposed variance estimators are motivated by the identifiability conditions, the definition of the smooth varying coefficients functions given in (4), Lemma 3 and Lemma 4 (a.). Using the consistency of $\widehat{\beta}_{nrj}$ and $\widehat{\eta}_{nsj}$ for the values of the functions β_r and η_s at the midpoint of the jth bin and the definitions of $\widetilde{Y}'_{njk}$ and $\widetilde{X}'_{nrjk}$, we target the quantities $\sigma^2\lambda_\psi$, $\gamma_0^2\lambda_\psi$, $\delta_s^2\lambda_\psi$, $\gamma_r^2\lambda_\psi m_{r,2}$ and $\gamma_r^2\lambda_{\psi\phi_r}m_{r,2}$ with the estimators $\widehat{\sigma}^2$, $\sum_{j=1}^{m}n^{-1}L_{nj}\widehat{\beta}^2_{n0j}$, $\sum_{j=1}^{m}n^{-1}L_{nj}\widehat{\eta}^2_{nsj}$, $n^{-1}\sum_{j=1}^{m}\widehat{\beta}^2_{nrj}\sum_{k=1}^{L_{nj}}\widetilde{X}'^2_{nrjk}$ and $n^{-1}\widehat{\gamma}_{nr}\sum_{j=1}^{m}\widehat{\beta}_{nrj}\sum_{k=1}^{L_{nj}}\widetilde{X}'^2_{nrjk}$, respectively. Furthermore, relying mainly on Lemma 3 and Lemma 4 (a.), we target $\gamma_0^2E\{\psi^2(U)\mathcal{M}_{11}\}$, $\gamma_0^2m_{r,1}^2E\{\psi^2(U)\mathcal{M}_{r+1,r+1}\}$ and $\gamma_0^2E\{\psi^2(U)\mathcal{M}_{p+s+1,p+s+1}\}$ with $n^{-1}\sum_{j=1}^{m}\widehat{\beta}^2_{n0j}\sum_{k=1}^{L_{nj}}\omega^2_{n,1,k}$, $n^{-1}\sum_{j=1}^{m}\widehat{\beta}^2_{n0j}\overline{\widetilde{X}'}^2_{nrj}\sum_{k=1}^{L_{nj}}\omega^2_{n,r+1,k}$ and $n^{-1}\sum_{j=1}^{m}\widehat{\beta}^2_{n0j}\sum_{k=1}^{L_{nj}}\omega^2_{n,p+s+1,k}$, respectively.

5. Application and numerical studies

Theorem 1 and 2 provide the foundation for inference. From Theorem 1, we have that

$$
\frac{\sqrt{n}}{\sigma_r}(\widehat{\gamma}_{nr}-\gamma_r)\xrightarrow{\mathcal{D}}\mathbb{N}(0,1),\quad 0\le r\le p,\quad \text{and}\quad \frac{\sqrt{n}}{\sigma_s}(\widehat{\delta}_{ns}-\delta_s)\xrightarrow{\mathcal{D}}\mathbb{N}(0,1),\quad 1\le s\le q, \tag{10}
$$

as $n\to\infty$. Using the consistent estimators, $\widehat{\sigma}^2_{nr}$ and $\widehat{\sigma}^2_{ns}$, proposed in Theorem 2, it follows from (10) and Slutsky's theorem that

$$
\frac{\sqrt{n}}{\widehat{\sigma}_{nr}}(\widehat{\gamma}_{nr}-\gamma_r)\xrightarrow{\mathcal{D}}\mathbb{N}(0,1),\quad 0\le r\le p,\quad \text{and}\quad \frac{\sqrt{n}}{\widehat{\sigma}_{ns}}(\widehat{\delta}_{ns}-\delta_s)\xrightarrow{\mathcal{D}}\mathbb{N}(0,1),\quad 1\le s\le q.
$$

Therefore, the approximate $(1-\alpha)100\%$ asymptotic confidence intervals for γ_r and δ_s have the endpoints

$$
\widehat{\gamma}_{nr}\pm z_{\alpha/2}\frac{\widehat{\sigma}_{nr}}{\sqrt{n}},\qquad \text{and}\qquad \widehat{\delta}_{ns}\pm z_{\alpha/2}\frac{\widehat{\sigma}_{ns}}{\sqrt{n}}, \tag{11}
$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard Gaussian distribution.

5.1 Application to the Pima Indians Diabetes Data

We illustrate the proposed partial covariate adjusted regression methodology with an application to the Pima Indians diabetes data set, available at http://www.ics.uci.edu/∼mlearn. Obesity is an important contributing factor to diabetes and has been widely studied in the Pima Indians population (Smith et al., 1988; Knowler et al., 1991; Hansen et al., 1998). One-half of adult Pima Indians have diabetes and 95% of those with diabetes are overweight (National Institute of Diabetes and Digestive and Kidney Diseases, http://diabetes.niddk.nih.gov). The available data comes from a larger database, where the subgroup used consists of females at least 21 years old and of Pima Indian heritage. (The population lives near Phoenix, Arizona, U.S.A.) An oral glucose tolerance test is one of the diagnostic tests for type II diabetes. We consider a linear regression model with response variable, plasma glucose concentration (PGC; from a oral glucose tolerance test), and predictors diastolic blood pressure (DBP), triceps skin fold thickness ($TSFT$) and age. More precisely, the underlying regression model is $PGC = \gamma_0+\gamma_1 DBP+\delta_1 Age+\delta_2 TSFT+e$. We incorporate the confounding effect of body mass index (BMI) on both the plasma glucose concentration (the response) and diastolic blood pressure (the predictor) using the proposed PCAR method. It would be plausible to model the remaining two predictors, namely Age and $TSFT$, as undistorted by $U \equiv BMI$, as they are observed directly. The observed data is $\{\widetilde{PGC}_i, \widetilde{DBP}_i, Age_i, TSFT_i, BMI_i\}_{i=1}^{n=524}$.

Table 1 gives the regression coefficient estimates for $(\gamma_0, \gamma_1, \delta_1, \delta_2)$ using the proposed PCAR method, CAR method, the ordinary least squares (OLS) estimates from regressing the observed $\widetilde{PGC}$ on $(\widetilde{DBP}, Age, TSFT)$ without adjusting for the confounder BMI, and adjustment via division, i.e. regressing $\widetilde{PGC}/BMI$ on $(\widetilde{DBP}/BMI, Age, TSFT)$. The approximate 95% asymptotic confidence intervals for the regression parameters obtained through all three methods are also displayed. The approximate confidence intervals for PCAR estimates were obtained as proposed in (11).

The implementation of the binning algorithm allows for merging of sparsely populated bins. Bin widths were chosen such that there are at least $(p+q+1)$ points, enough to fit the linear regression with $(p+q)$ predictors in each bin. If there were bins with less than $(p+q+1)$ elements, such bins were randomly merged with neighboring bins. For

this example $n = 524$ (after the removal of outliers) and the average number of points per bin was 15, yielding a total of 34 bins after merging. Note that CAR estimates have been shown to be sufficiently robust regarding different choices of m, under the rate conditions given in Section 4 (Şentürk and Müller, 2006). We have found this property to hold for the proposed PCAR estimates as well.

Note that coefficients obtained by adjustment via division are quite different from the other three methods applied. In this adjustment the coefficient of DBP becomes quite pronounced compared to the other two predictors. This is most likely due to the pseudo dependence created between $\widetilde{PGC}/BMI$ and $\widetilde{DBP}/BMI$ via division by the common variable BMI. This is an example of the misleading conclusions that adjustment by division may suggest. In other words, if the original contamination is not exactly multiplication by the confounder (BMI in this example), then normalization by division may create further confounding, or "coupling" (as defined in Archie, 1981), creating a pseudo dependence that does not exist in the original data.

Even though OLS estimates for blood pressure and age are different from the PCAR and CAR estimates, all are found statistically significant at the usual 5% level. Thus, diastolic blood pressure and age are still important predictors of PGC even after adjusted for body mass index. However, using OLS, $TSFT$ is a significant predictor of PGC, but it is not significant using PCAR and CAR at the 5% significance level. This result is not too surprising, since both $TSFT$ and body mass index are indicators of obesity. They are positively correlated (Pearson correlation 0.67). Thus, adjusting for one, the other becomes an insignificant factor for predicting plasma glucose concentration. We note that even though estimation via CAR leads to the same conclusion as PCAR on the significance of the predictors for this analysis, the estimates from these two methods are different for $TSFT$. This is again to be expected, since CAR estimates are shown to be biased for undistorted modeling of predictors. The bias is directly related to the correlation between BMI and $TSFT$, which is quite high for this example.

5.2 Numerical studies

To examine the numerical properties of the estimators, we implemented the following simulation studies. The underlying multiple regression model is

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \delta Z + e, \tag{12}$$

where the parameters of interest are $(\gamma_0, \gamma_1, \gamma_2, \delta)^{\mathrm{T}} = (4, -1, 0.3, 3)$. The error variable is $e \sim N(0, .5)$, and the confounder variable U is generated from a uniform distribution on $[2, 6]$. We considered the joint distribution of the predictors to be multivariate normal: $(X_1, X_2, Z)^{\mathrm{T}} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with a general covariance structure

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.490 & 0.168 & 0.280 \\ 0.168 & 1.440 & -0.360 \\ 0.280 & -0.360 & 1.000 \end{bmatrix}.$$

The mean vector is $\boldsymbol{\mu} = (0.7, 1.2, |U| - 3.5)^{\mathrm{T}}$, so that the undistorted predictor Z is dependent on U. To simulate the distorted (observed) data, we consider the following distorting functions, $\psi(U) = (U + 3)/7$, $\phi_1(U) = (U + 1)^2/26.3333$, and $\phi_2(U) = (U + 10)/14$, satisfying the identifiability constraints that $E\{\psi(U)\} = 1$ and $E\{\phi_r(U)\} = 1$. The distorted response and predictors are $\widetilde{Y} = \psi(U)Y$, $\widetilde{X}_1 = \phi_1(U)X_1$, and $\widetilde{X}_2 = \phi_2(U)X_2$.

We conducted 1000 Monte Carlo simulation runs for sample sizes $n = 100$, 150, 350, 800, and 1400 to study the approximate asymptotic confidence intervals given in (11). Table 2 summarizes the coverage and interval lengths, averaged over the 1000 simulation runs, for the approximate 95% asymptotic confidence intervals for the parameter vector $(\gamma_0, \gamma_1, \gamma_2, \delta)^{\mathrm{T}} = (4, -1, .3, 3)$. The numerical study indicates that the estimated non-coverage percentages are close to the target value of 0.05, as the sample size n increases. The estimated interval lengths are decreasing as n increases, as expected.

We also examined the bias, variance and mean squared error (MSE) of the proposed estimators. For example, the estimated (absolute bias, variance, MSE) values at the smallest sample size $n = 100$ are (0.0112, 0.2223, 0.2224), (0.0120, 0.1392, 0.1393), (0.0028, 0.0421, 0.0421) and (0.0167, 0.0651, 0.0654) for $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\delta}$, respectively. These values are averages over 1000 Monte Carlo runs. The results are similar for other sample sizes, where the variance seems to be the dominating factor contributing to the MSE. The estimated MSE is decreasing as sample size increases, as expected.

In addition, we compared partial covariate adjusted estimates with CAR estimates for δ (even though their asymptotic distributions are different, the three other point estimates for γ_0, γ_1 and γ_2 are the same for the two methods). The multiplicative bias factor of the CAR estimate for δ, shown to be $C_s = E\{\psi(U)Z_s\}/E(Z_s)$ in Section 3, is equal to 1.416 for this simulation set-up. As expected, the CAR estimate $\widehat{\delta}^*$ is off target for $\delta = 3$ (with a

mean of 4.294 at $n = 100$, and 4.146 at $n = 1400$). The estimated (absolute bias, variance, MSE) values for $\widehat{\delta}^*$ are (1.294, 1.694, 3.368) for $n = 100$. Thus, the absolute bias of $\widehat{\delta}^*$ is about 77 times that of $\hat{\delta}$ at $n = 100$. In addition to being biased, note that the CAR estimator $\widehat{\delta}^*$ has substantially larger variance relative to the PCAR estimator.

6. Proofs of the main results

We provide the major steps of the proofs of the main results (Theorem 1 and 2) here and defer the auxiliary results for these proofs to the Appendix, where they are listed as lemmas 1 to 4. We introduce the following technical conditions:

(C1) The covariate U is bounded below and above, $-\infty < a \leq U \leq b < \infty$ for real numbers $a < b$. The density $f(u)$ of U satisfies $\inf_{a\leq u\leq b} f(u) > c_1 > 0$, $\sup_{a\leq u\leq b} f(u) < c_2 < \infty$ for real c_1, c_2, and is uniformly Lipschitz continuous, i.e., there exists a real number M such that $\sup_{a\leq u\leq b} |f(u + c) - f(u)| \leq M|c|$ for any real number c.

(C2) The variables (e, U, X_r) are mutually independent for $r = 1, \ldots, p$. In addition, (e, Z_s) are assumed to be independent.

(C3) For the predictors, $\sup_{1\leq i\leq n, 1\leq r\leq p, 1\leq s\leq q}\{|X_{nri}|, |Z_{nsi}|\} \leq B$ for some bound $B \in \mathbb{R}$. In addition, the predictors X_r satisfy the condition that $E(X_r) \neq 0$.

(C4) Contamination functions $\psi(\cdot)$ and $\phi_r(\cdot)$, $1 \leq r \leq p$, are twice continuously differentiable, satisfying

$$E\psi(U) = 1, \quad E\phi_r(U) = 1, \quad \phi_r(\cdot) > 0 \quad 1 \leq r \leq p.$$

(C5) The matrices $\mathbf{\Gamma}_{nj}$, $j = 1, \ldots, m$ are nonsingular, i.e. $\rho = |\inf_j \det(\mathbf{\Gamma}_{nj})| > 0$, where $\mathbf{\Gamma}_{nj} = E(L_{nj}^{-1}\boldsymbol{\chi}_{nj}'^{\mathrm{T}}\boldsymbol{\chi}_{nj}'|U_{nj}'^*)$, $\boldsymbol{\chi}_{nj}' = (\boldsymbol{\chi}_{nj1}', \ldots, \boldsymbol{\chi}_{njL_{nj}}')^{\mathrm{T}}$ is a $L_{nj} \times (p+q+1)$ undistorted data matrix in bin j, and $\boldsymbol{\chi}_{njk}' = (1, X_{n1jk}', \ldots, X_{npjk}', Z_{n1jk}', \ldots, Z_{nqjk}')^{\mathrm{T}}$ denotes the kth observation.

The technical conditions above are similar to those introduced in Şentürk and Müller (2006), except for the new independence structure outlined in (C2), the boundedness of the undistorted predictors Z_s in (C3), and the bin dependent limiting matrices $\mathbf{\Gamma}_{nj}$ in (C5), resulting from the dependence structure between Z_s and U. Bounded covariates are standard in asymptotic theory for least squares regression, as are conditions (C2) and

(C5) (see Lai, Robbins and Wei, 1979). The identifiability conditions stated in (C4) are equivalent to $E(\widetilde{Y}|X) = E(Y|X)$ and $E(\widetilde{X}_r|X_r) = X_r$. Note that this means that the confounding of Y by U does not change the mean regression function.

In the proofs of the main results, the following notations will be utilized.

1. $A \boxdot B$: The Hadamard product of two matrices, A and B, of the same dimension. The matrix $A \boxdot B$ is also of the same dimension with (i,j)th element equal to the product of the (i,j)th elements of matrices A and B.

2. $\mathbf{1}_{a\times b}$: A matrix of size $a \times b$ with all entries equal to one.

3. $\widehat{\boldsymbol{\theta}}_{nj} = (\widehat{\beta}_{n0j}, \widehat{\beta}_{n1j}, \ldots, \widehat{\beta}_{npj}, \widehat{\eta}_{n1j}, \ldots, \widehat{\eta}_{nqj})^{\mathrm{T}}$.

4. $\widetilde{\boldsymbol{\theta}}_{nj} = (\widetilde{\gamma}_{n0j}\psi(U'^{*}_{nj}), \widetilde{\gamma}_{n1j}\psi(U'^{*}_{nj})/\phi_1(U'^{*}_{nj}), \ldots, \widetilde{\gamma}_{npj}\psi(U'^{*}_{nj})/\phi_p(U'^{*}_{nj}), \widetilde{\delta}_{n1j}\psi(U'^{*}_{nj}), \ldots, \widetilde{\delta}_{nqj}\psi(U'^{*}_{nj}))^{\mathrm{T}}$.

5. $\boldsymbol{\theta}_{nj} = (\gamma_0\psi(U'^{*}_{nj}), \gamma_1\psi(U'^{*}_{nj})/\phi_1(U'^{*}_{nj}), \ldots, \gamma_p\psi(U'^{*}_{nj})/\phi_p(U'^{*}_{nj}), \delta_1\psi(U'^{*}_{nj}), \ldots, \delta_q\psi(U'^{*}_{nj}))^{\mathrm{T}}$.

6. We use $\boldsymbol{\chi}'_{nj(i)}$ to denote the matrix $\boldsymbol{\chi}'_{nj}$ and $L_{nj(i)}$ to denote the number of points in the jth bin such that $U_{ni} \in B_{nj}$, and $\kappa_{rk(i)} \equiv \{(L^{-1}_{nj(i)}\boldsymbol{\chi}'^{\mathrm{T}}_{nj(i)}\boldsymbol{\chi}'_{nj(i)})^{-1}\boldsymbol{\chi}'^{\mathrm{T}}_{nj(i)}\}_{rk(i)}$ is the (r,k)th element of the matrix $\{(L^{-1}_{nj}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}\boldsymbol{\chi}'_{nj})^{-1}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}\}$ for $1 \le r \le p+q+1$, where $U_{ni} = U'_{njk}$ is the kth element in the ordered sample $(U'_{nj1}, \ldots, U'_{njL_{nj}}) \in B_{nj}$.

7. $\boldsymbol{\Theta}_{nj} = L^{-1}_{nj}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}\boldsymbol{\chi}'_{nj}$.

Proof of Theorem 1. From Lemma 4 (*b.*), we have that

$$\sup_j |(L^{-1}_{nj}\widetilde{\boldsymbol{\chi}}'^{\mathrm{T}}_{nj}\widetilde{Y}'_{nj}) - \{\Delta \boxdot (L^{-1}_{nj}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}Y'_{nj})\}| = O_p(m^{-1})\mathbf{1}_{(p+q+1)\times 1}, \tag{13}$$

where $\Delta = \{\psi(U'^{*}_{nj}), \psi(U'^{*}_{nj})\phi_1(U'^{*}_{nj}), \ldots, \psi(U'^{*}_{nj})\phi_p(U'^{*}_{nj}), \psi(U'^{*}_{nj}), \ldots, \psi(U'^{*}_{nj})\}^{\mathrm{T}}$. Lemma 3 together with (13) implies that, on event E_n,

$$\sup_j \left|\widehat{\boldsymbol{\theta}}_{nj} - \widetilde{\boldsymbol{\theta}}_{nj}\right| = O_p(m^{-1})\mathbf{1}_{(p+q+1)\times 1}. \tag{14}$$

We first consider the case of $r = 0$ and show that $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$ is asymptotically normal. Using Lemma 4, (14), and some algebra, $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0) = \sqrt{n}(\sum_{j=1}^{m} L_{nj}n^{-1}\widehat{\beta}_{n0j} - \gamma_0)$ can be expressed as

$$\sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}\left[\frac{\gamma_0\psi(U'_{njk})}{\sqrt{n}} + \frac{\psi(U'_{njk})e'_{njk}}{\sqrt{n}}\{(L^{-1}_{nj}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}\boldsymbol{\chi}'_{nj})^{-1}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}\}_{1k}\right] - \sqrt{n}\gamma_0 + O_p(\sqrt{n}/m).$$

Since the above sum is over all bins indexed by j, and over all points within the bins indexed by k, it is equal to the sum over all data points indexed by i, summed up in a random order. Thus, the above expression for $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$ can be further simplified to

$$\sum_{\substack{i=1\\j,k}}^{t}\left[\frac{\gamma_0\psi(U_{ni})}{\sqrt{n}} + \frac{\psi(U_{ni})e_{ni}\kappa_{1k(i)}}{\sqrt{n}} - \frac{\gamma_0}{\sqrt{n}}\right] + O_p(\sqrt{n}/m) \equiv \sum_{i=1}^{t} W_{n0i} + O_p(\sqrt{n}/m). \tag{15}$$

Therefore, $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$ is asymptotically equivalent to $S_{n0t} \equiv \sum_{i=1}^{t} W_{n0i}$ because the second term $O_p(\sqrt{n}/m)$ is negligible when $m/\sqrt{n} \to \infty$ as $n \to \infty$. Next, let F_{n0t} be the σ-field generated by $\{e_{n1}, \ldots, e_{nt}, U_{n1}, \ldots, U_{nt}, L_{nj(1)}, \ldots, L_{nj(t)}, \boldsymbol{\chi}'_{nj(1)}, \ldots, \boldsymbol{\chi}'_{nj(t)}\}$. Then $\{S_{n0t}, F_{n0t}, 1 \le t \le n\}$ is a mean zero martingale for $n \ge 1$, since $E(S_{n0t}) = 0$, $E(S_{n0,t+1}|F_{n0t}) = S_{n0t}$, and S_{n0t} is adapted to F_{n0t}. Furthermore, note that the σ-fields are nested, that is, $F_{n0t} \subseteq F_{n0,t+1}$ for all $t \le n$. Hence, it follows from Lemma 1 that $S_{n0n} \to \mathbb{N}(0, \sigma_0^2)$ in distribution (McLeish, 1974, Theorem 2.3 and subsequent discussion). This establishes the asymptotic normality of $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$.

We proceed next to establish the asymptotic normality of $\sqrt{n}(\widehat{\gamma}_{nr} - \gamma_r)$ for $r = 1, \ldots, p$. Let $\widehat{\nu}_{nr} = \sum_{j=1}^{m} L_{nj} n^{-1} \widehat{\beta}_{nrj} \overline{\widetilde{X}'}_{nrj}$ and $\overline{\nu}_{nr} = \sum_{j=1}^{m} L_{nj} n^{-1} \overline{\widetilde{X}'}_{nrj}$ and note that $\widehat{\gamma}_{nr} = \widehat{\nu}_{nr}/\overline{\nu}_{nr}$. We first show that

$$\sqrt{n}\begin{pmatrix} \widehat{\nu}_{nr} - \gamma_r E(X_r) \\ \overline{\nu}_{nr} - E(X_r) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathbb{N}_2(\underline{0}, \Sigma_r). \tag{16}$$

For (16) to hold, by the Cramer-Wald device, it is enough to show the asymptotic normality of

$$\sqrt{n}\left[a\left\{\widehat{\nu}_{nr} - \gamma_r E(X_r)\right\} + b\left\{\overline{\nu}_{nr} - E(X_r)\right\}\right] \tag{17}$$

for real a, b. The asymptotic normality of $\sqrt{n}(\widehat{\gamma}_{nr} - \gamma_r)$ $(1 \le r \le p)$ will follow from (16) by applying the δ-method with $\widehat{\gamma}_{nr} = \widehat{\nu}_{nr}/\overline{\nu}_{nr}$. Again applying Lemma 4 together with (14) and some simple algebra, we can express $\widehat{\nu}_{nr}$ and $\overline{\nu}_{nr}$ as

$$\begin{aligned}
\widehat{\nu}_{nr} &= \sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}\left[\frac{\gamma_r}{n}\psi(U'_{njk})X'_{nrjk} + \frac{\overline{X'}_{nrj}}{n}\psi(U'_{njk})e'_{njk}\{(L_{nj}^{-1}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}\boldsymbol{\chi}'_{nj})^{-1}\boldsymbol{\chi}'^{\mathrm{T}}_{nj}\}_{rk}\right] + O_p(m^{-1}) \text{ and} \\
\overline{\nu}_{nr} &= \sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}\frac{1}{n}\phi_r(U'_{njk})X'_{nrjk} + O_p(m^{-1}).
\end{aligned}$$

Thus, using similar simplifications as was done for the case of $r = 0$ in (15), the linear

combination (17), namely $\sqrt{n}[a\{\widehat{\nu}_{nr} - \gamma_r E(X_r)\} + b\{\overline{\nu}_{nr} - E(X_r)\}]$, can be expressed as

$$\sum_{\substack{i=1\\j,k}}^{n}\Bigg[a\frac{\gamma_r}{\sqrt{n}}\psi(U_{ni})X_{nri} + a\frac{\overline{X'}_{nrj(i)}}{\sqrt{n}}\psi(U_{ni})e_{ni}\kappa_{rk(i)} - a\frac{\gamma_r}{\sqrt{n}}E(X_r) + \frac{b}{\sqrt{n}}\phi_r(U_{ni})X_{nri}$$
$$-b\frac{E(X_r)}{\sqrt{n}}\Bigg] + O_p(\sqrt{n}/m) \equiv \sum_{i=1}^{t} W_{nri} + O_p(\sqrt{n}/m).$$

The second term $O_p(\sqrt{n}/m)$ is asymptotically negligible and it is straightforward to verify that $\{S_{nrt} \equiv \sum_{i=1}^{t} W_{nri}, F_{n0t}, 1 \le t \le n\}$ is a mean zero martingale for $n \ge 1$. Analogous to the case of $r = 0$, described in more details earlier, it follows from Lemma 2 that $S_{nrn} \xrightarrow{\mathcal{D}} \mathbb{N}(0, (a,b)\Sigma_r(a,b)^{\mathrm{T}})$. Finally, a direct application of the δ-method gives $\sqrt{n}(\widehat{\gamma}_{nr} - \gamma_r) \xrightarrow{p} \mathbb{N}(0, \sigma_r^2)$ for $1 \le r \le p$, where σ_r^2 is explicitly given in Theorem 1.

The asymptotic normality of $\sqrt{n}(\widehat{\delta}_{ns} - \delta_s) \to \mathbb{N}(0, \sigma_s^2)$ follows similarly to the case of $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$, since they have similar forms in (14). (See also definition/notation 4.) The asymptotic variance σ_s^2 which has a similar form as σ_0^2, is given explicitly in Theorem 1. This completes the proof of Theorem 1.

Proof of Theorem 2. The following relation holds on event A_n

$$\sup_j |(\widetilde{\gamma}_{n0j} - \gamma_0, \ldots, \widetilde{\gamma}_{npj} - \gamma_p, \widetilde{\delta}_{n1j} - \delta_1, \ldots, \widetilde{\delta}_{nqj} - \delta_q)^{\mathrm{T}}| = o_p(1)\mathbf{1}_{(p+q+1)\times 1}. \tag{18}$$

It follows from Lemma 4 (a.) and (b.). Utilizing (18) together with (14) gives

$$\sup_j \left|\widehat{\boldsymbol{\theta}}_{nj} - \boldsymbol{\theta}_{nj}\right| = o_p(1)\mathbf{1}_{(p+q+1)\times 1}. \tag{19}$$

By the Law of Large Numbers, (19), and boundedness considerations

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}\left(\widetilde{Y}'_{njk} - \widehat{\beta}_{n0j} - \sum_{r=1}^{p}\widehat{\beta}_{nrj}\widetilde{X}'_{nrjk} - \sum_{s=1}^{q}\widehat{\eta}_{nsj}Z'_{nsjk}\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}\{\psi(U'^{*}_{nj})e'_{njk} + o_p(1)\}^2 = \frac{1}{n}\sum_{i=1}^{n}\psi(U^2_{ni})e^2_{ni} + o_p(1) \\
&= \sigma^2\lambda_\psi + o_p(1).
\end{aligned}$$

It has been shown in Şentürk and Müller (2006) that

$$A_1 \equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{nrj}^2\sum_{k=1}^{L_{nj}}\widetilde{X}_{nrjk}'^2 = \gamma_r^2\lambda_\psi m_{r,2} + o_p(1),$$

$$A_2 \equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{nrj}\sum_{k=1}^{L_{nj}}\widetilde{X}_{nrjk}'^2 = \gamma_r\lambda_{\psi\phi_r} m_{r,2} + o_p(1),$$

$$A_3 \equiv \sum_{j=1}^{m}\frac{L_{nj}}{n}\widehat{\beta}_{n0j}^2 = \gamma_0^2\lambda_\psi + o_p(1),$$

and it can be shown similarly that $A_4 \equiv \sum_{j=1}^{m} n^{-1}L_{nj}\widehat{\eta}_{nsj}^2 = \delta_s^2\lambda_\psi + o_p(1)$. Also, using Lemma 3, Lemma 4 $(a.)$, $(b.)$ and the Law of Large Numbers, we have

$$A_5 \equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{n0j}^2\sum_{k=1}^{L_{nj}}\omega_{n,1,k}^2 \xrightarrow{p} \gamma_0^2 E\{\psi^2(U)\mathcal{M}_{11}\},$$

$$A_6 \equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{n0j}^2\overline{\widetilde{X}'}_{nrj}^2\sum_{k=1}^{L_{nj}}\omega_{n,r+1,k}^2 \xrightarrow{p} \gamma_0^2\{E(X_r)\}^2 E\{\psi^2(U)\mathcal{M}_{r+1,r+1}\}, \text{ and}$$

$$A_7 \equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{n0j}^2\sum_{k=1}^{L_{nj}}\omega_{n,p+s+1,k}^2 \xrightarrow{p} \gamma_0^2 E\{\psi^2(U)\mathcal{M}_{p+s+1,p+s+1}\}.$$

The estimators of the asymptotic variances given in Theorem 2, in terms of the above quantities, are: (1) $\widehat{\sigma}_{n0}^2 = A_3 - \widehat{\gamma}_{n0}^2 + \widehat{\sigma}^2 A_5/A_3$, (2) $\widehat{\sigma}_{nr}^2 = (A_1 + \widehat{\gamma}_{nr}^2\overline{\widetilde{X}}_{nr}^2 - 2\widehat{\gamma}_{nr}A_2 + \widehat{\gamma}_{nr}^2 s_{\widetilde{X}_r}^2)/\overline{\widetilde{X}}_{nr}^2 + (\widehat{\sigma}^2 A_6)/(\overline{\widetilde{X}}_{nr}^2 A_3)$, and (3) $\widehat{\sigma}_{ns}^2 = A_4 - \widehat{\delta}_{ns}^2 + (\widehat{\sigma}^2 A_7)/A_3$. Thus, the results of Theorem 2 follow by noting that $\widehat{\gamma}_{n0} \xrightarrow{p} \gamma_0$, $\widehat{\gamma}_{nr} \xrightarrow{p} \gamma_r$, $\widehat{\delta}_{ns} \xrightarrow{p} \delta_s$, $s_{\widetilde{X}_r}^2 \xrightarrow{p} \text{var}(\widetilde{X}_r)$ and $\overline{\widetilde{X}}_{nr} \xrightarrow{p} E(X_r)$.

Appendix: Additional lemmas and proofs

We introduce some additional technical conditions that are needed for the proof of Lemma 4 given below:

(C6) The functions $h_1(u) = \int x g_1(x,u)dx$ and $h_2(u) = \int x g_2(x,u)dx$ are uniformly Lipschitz, where $g_1(\cdot,\cdot)$ and $g_2(\cdot,\cdot)$ are the joint density functions of $(\boldsymbol{\chi}, U)$ and $(\boldsymbol{\chi}e, U)$, respectively.

(C7) The error term satisfies $E|e^\tau| < \infty$ for $\tau > 4$.

Lemma 1. *Under the technical conditions (C1)-(C7), on event A_n (9), the martingale differences W_{n0t} satisfy the conditions*

$$(a.) \qquad \sum_{t=1}^{n} E\{W_{n0t}^2 I(|W_{n0t}| > \epsilon)\} \to 0 \qquad \text{for all} \quad \epsilon > 0,$$

$$(b.) \qquad \Delta_{n0}^2 = \sum_{t=1}^{n} W_{n0t}^2 \xrightarrow{p} \sigma_0^2 \qquad \text{for} \;\; \sigma_0^2 > 0.$$

Proof. Let $W_{n0t} = w_{n0t} v_{n0t}$, where $w_{n0t} = 1/\sqrt{n}$, $v_{n0t} = \gamma_0 \psi(U_{nt}) + \psi(U_{nt}) e_{nt} \kappa_{1k(t)} - \gamma_0 \equiv \alpha_{1nt} + \alpha_{2nt} e_{nt}$, $\alpha_{1nt} = \gamma_0 \psi(U_{nt}) - \gamma_0$, $\alpha_{2nt} = \psi(U_{nt}) \kappa_{1k(t)}$, and $E(v_{n0t}) = 0$. Using (C1), (C3) and (C4), it holds on event A_n that $\sup_{1 \le t \le n} |\alpha_{1nt}| < c_1$ and $\sup_{1 \le t \le n} |\alpha_{2nt}| < c_2$ for some $c_1, c_2 > 0$. Thus, it holds for $\epsilon > 0$ that

$$\begin{aligned}
&\sum_{t=1}^{n} E\{W_{n0t}^2 I(|W_{n0t}| > \epsilon)\} = \sum_{t=1}^{n} \int x^2 I(|x| > \epsilon) dF_{w_{n0t} v_{n0t}}(x) \\
&= \sum_{t=1}^{n} \int x^2 I(|x| > \epsilon / |w_{n0t}|) w_{n0t}^2 dF_{v_{n0t}}(x) = n^{-1} \sum_{t=1}^{n} \int x^2 I(|x| > \sqrt{n}\epsilon) dF_{v_{n0t}}(x) \\
&\le n^{-1} \sum_{t=1}^{n} \{E(v_{n0t}^4)\}^{1/2} \{\mathrm{pr}(v_{n0t}^2 > n\epsilon^2)\}^{1/2}.
\end{aligned}$$

Now, $E(v_{n0t}^4)$ is bounded uniformly in n and t, since e_{nt} has finite fourth moment by (C7). Also note that $\mathrm{pr}(v_{n0t}^2 > n\epsilon^2) = \mathrm{pr}((\alpha_{1nt} + \alpha_{2nt} e_{nt})^2 > n\epsilon^2) \le \mathrm{pr}(\alpha_{1nt}^2 + \alpha_{2nt}^2 e_{nt}^2 + 2|\alpha_{1nt} \alpha_{2nt} e_{nt}| > n\epsilon^2) \le \mathrm{pr}(c_1^2 + c_2^2 e_{nt}^2 + 2c_1 c_2 |e_{nt}| > n\epsilon^2)$. Lemma 1 (a.) follows, since $\mathrm{pr}(c_1^2 + c_2^2 e_{nt}^2 + 2c_1 c_2 |e_{nt}| > n\epsilon^2) \to 0$ uniformly in n and t, e_{nt}^2 and $|e_{nt}|$ being i.i.d. with finite fourth moments.

Next, consider the term Δ_{n0}^2 given in Lemma 1 (b.). It is equal to

$$\Delta_{n0}^2 = \gamma_0^2 \left\{ n^{-1} \sum_t \psi^2(U_{nt}) \right\} + \gamma_0^2 - 2\gamma_0^2 \left\{ n^{-1} \sum_t \psi(U_{nt}) \right\} + 2\gamma_0 n^{-1} \sum_t \psi^2(U_{nt}) e_{nt} \kappa_{1k(t)}$$

$$\begin{aligned}
&- \; 2\gamma_0 n^{-1} \sum_t \psi(U_{nt}) e_{nt} \kappa_{1k(t)} + n^{-1} \sum_{j=1}^{m} \sum_{k=1}^{L_{nj}} \Big\{ (\boldsymbol{\Theta}_{nj})_{11}^{-1} e'_{njk} \psi(U'_{njk}) \\
&+ \; (\boldsymbol{\Theta}_{nj})_{12}^{-1} e'_{njk} \psi(U'_{njk}) X'_{n1jk} + \ldots + (\boldsymbol{\Theta}_{nj})_{1,p+q+1}^{-1} e'_{njk} \psi(U'_{njk}) Z'_{nqjk} \Big\}^2 \equiv T_1 + \ldots + T_6.
\end{aligned}$$

Using Law of Large Numbers, it holds that $T_1 + T_2 + T_3 \xrightarrow{p} \gamma_0^2 \mathrm{var}\{\psi(U)\}$. Since T_4 and T_5 have expected values zero and variances $O(n^{-1})$, they are both $O_p(n^{-1/2})$. By Lemma 4

$(a.)$ and the Law of Large Numbers, term T_6 is equal to

$$\begin{aligned}
& \sigma^2 E\Big\{\psi^2(U)[(\mathbf{\Gamma}^{-1})^2_{11} + (\mathbf{\Gamma}^{-1})^2_{12}X_1^2 + \ldots + (\mathbf{\Gamma}^{-1})^2_{1,p+q+1}Z_q^2 \\
+ \; & \{2(\mathbf{\Gamma}^{-1})_{11}(\mathbf{\Gamma}^{-1})_{12}X_1 + \ldots + 2(\mathbf{\Gamma}^{-1})_{11}(\mathbf{\Gamma}^{-1})_{1,p+q+1}Z_q\} \\
+ \; & \{2(\mathbf{\Gamma}^{-1})_{12}(\mathbf{\Gamma}^{-1})_{13}X_1X_2 + \ldots + 2(\mathbf{\Gamma}^{-1})_{12}(\mathbf{\Gamma}^{-1})_{1,p+q+1}X_1Z_q\} + \ldots \\
+ \; & \{2(\mathbf{\Gamma}^{-1})_{1,p+q}(\mathbf{\Gamma}^{-1})_{1,p+q+1}Z_{q-1}Z_q\}]\Big\} + o_p(1) \\
= \; & \sigma^2 E\{\psi^2(U)\mathcal{M}_{11}\} + o_p(1),
\end{aligned}$$

where $\mathbf{\Gamma}$ and $\mathcal{M}$ are as defined prior to Theorem 1. Thus, $\Delta^2_{n0} \xrightarrow{p} \gamma_0^2\sigma_\psi^2 + \sigma^2 E\{\psi^2(U)\mathcal{M}_{11}\} \equiv \sigma_0^2$ and Lemma 1 (b.) follows.

Lemma 2. *Under the technical conditions (C1)-(C7), on event A_n (9), the martingale differences W_{nrt} satisfy the conditions*

$$(a.) \qquad \sum_{t=1}^{n} E\{W^2_{nrt} I(|W_{nrt}| > \epsilon)\} \to 0 \qquad \text{for all} \quad \epsilon > 0,$$

$$(b.) \qquad \Delta^2_{nr} = \sum_{t=1}^{n} W^2_{nrt} \xrightarrow{p} (a,b)\Sigma_r(a,b)^T \qquad \text{for} \;\; (a,b)\Sigma_r(a,b)^T > 0.$$

Proof. Part (a.) of Lemma 2 follows in a similar fashion as part (a.) of Lemma 1. Therefore, we focus on the proof of part (b.). The term Δ^2_{nr} in Lemma 2 (b.) is equal to

$$\begin{aligned}
\Delta^2_{nr} \; = \; & a^2\gamma_r^2\Big\{n^{-1}\sum_t \psi^2(U_{nt})X^2_{nrt}\Big\} + a^2\gamma_r^2\{E(X_r)\}^2 + b^2\Big\{n^{-1}\sum_t \phi_r^2(U_{nt})X^2_{nrt}\Big\} \\
+ \; & b^2 m^2_{r,1} - 2a^2\gamma_r^2 m_{r,1}\Big\{n^{-1}\sum_t \psi(U_{nt})X_{nrt}\Big\} + 2ab\gamma_r m^2_{r,1} \\
+ \; & 2ab\gamma_r\Big\{n^{-1}\sum_t \psi(U_{nt})\phi_r(U_{nt})X^2_{nrt}\Big\} - 2b^2 m_{r,1}\Big\{n^{-1}\sum_t \phi_r(U_{nt})X_{nrt}\Big\} \\
- \; & 2ab\gamma_r m_{r,1}\Big\{n^{-1}\sum_t \psi(U_{nt})X_{nrt}\Big\} - 2ab\gamma_r m_{r,1}\Big\{n^{-1}\sum_t \phi_r(U_{nt})X_{nrt}\Big\} \\
+ \; & 2a^2\gamma_r n^{-1}\sum_t \psi^2(U_{nt})e_{nt}\bar{X}'_{nrj(t)}X_{nrt}\kappa_{rk(t)} - 2a^2\gamma_r E(X_r)n^{-1}\sum_t \psi(U_{nt})e_{nt}\bar{X}'_{nrj(t)}\kappa_{rk(t)} \\
+ \; & 2abn^{-1}\sum_t \psi(U_{nt})\phi_r(U_{nt})e_{nt}\bar{X}'_{nrj(t)}X_{nrt}\kappa_{rk(t)} - 2abE(X_r)n^{-1}\sum_t \psi(U_{nt})e_{nt}\bar{X}'_{nrj(t)}\kappa_{rk(t)} \\
+ \; & a^2 n^{-1}\sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}\Big\{(\mathbf{\Theta}_{nj})^{-1}_{r1}\overline{X}'_{nrj}e'_{njk}\psi(U'_{njk}) + (\mathbf{\Theta}_{nj})^{-1}_{r2}\overline{X}'_{nrj}e'_{njk}\psi(U'_{njk})X'_{n1jk} + \ldots \\
+ \; & (\mathbf{\Theta}_{nj})^{-1}_{r,p+q+1}\overline{X}'_{nrj}e'_{njk}\psi(U'_{njk})Z'_{nqjk}\Big\}^2 \equiv T_1 + \ldots + T_{15}.
\end{aligned}$$

Using the Law of Large Numbers, it holds that $T_1+\ldots+T_{10} \xrightarrow{p} a^2\gamma_r^2[m_{r,1}^2\sigma_\psi^2+\mathrm{var}(X_r)\lambda_\psi]+2ab\gamma_r[\lambda_{\psi\phi_r}m_{r,2}-m_{r,1}^2]+b^2\mathrm{var}(\widetilde{X}_r)$. Since T_{11}, T_{12}, T_{13} and T_{14} have expected values zero and variances $O(n^{-1})$, they are all $O_p(n^{-1/2})$. By Lemma 4 $(a.)$ and the Law of Large Numbers, term T_{15} is equal to

$$\begin{aligned}
& a^2\sigma^2m_{r,1}^2E\Big\{\psi^2(U)[(\mathbf{\Gamma}^{-1})_{r+1,1}^2+(\mathbf{\Gamma}^{-1})_{r+1,2}^2X_1^2+\ldots+(\mathbf{\Gamma}^{-1})_{r+1,p+q+1}^2Z_q^2 \\
+ \;& \{2(\mathbf{\Gamma}^{-1})_{r+1,1}(\mathbf{\Gamma}^{-1})_{r+1,2}X_1+\ldots+2(\mathbf{\Gamma}^{-1})_{r+1,1}(\mathbf{\Gamma}^{-1})_{r+1,p+q+1}Z_q\} \\
+ \;& \{2(\mathbf{\Gamma}^{-1})_{r+1,2}(\mathbf{\Gamma}^{-1})_{r+1,3}X_1X_2+\ldots+2(\mathbf{\Gamma}^{-1})_{r+1,2}(\mathbf{\Gamma}^{-1})_{r+1,p+q+1}X_1Z_q\}+\ldots \\
+ \;& \{2(\mathbf{\Gamma}^{-1})_{r+1,p+q}(\mathbf{\Gamma}^{-1})_{r+1,p+q+1}Z_{q-1}Z_q\}]\Big\}+o_p(1) \\
= \;& a^2\sigma^2m_{r,1}^2E\{\psi^2(U)\mathcal{M}_{r+1,r+1}\}+o_p(1).
\end{aligned}$$

Thus

$$\Delta_{nr}^2 \xrightarrow{p} (a,b)\Sigma_r(a,b)^{\mathrm{T}}=(a,b)\begin{bmatrix}\Sigma_{r11} & \Sigma_{r12}\\ \Sigma_{r12} & \Sigma_{r22}\end{bmatrix}(a,b)^{\mathrm{T}},$$

where $\Sigma_{r11}=\gamma_r^2[m_{r,1}^2\sigma_\psi^2+\mathrm{var}(X_r)\lambda_\psi]+\sigma^2m_{r,1}^2E\{\psi^2(U)\mathcal{M}_{r+1,r+1}\}$, $\Sigma_{r12}=\gamma_r[\lambda_{\psi\phi_r}m_{r,2}-m_{r,1}^2]$, and $\Sigma_{r22}=\mathrm{var}(\widetilde{X}_r)$. Hence Lemma 2 (b.) follows.

Lemma 3. *Under the technical conditions (C1)-(C6), it holds on event E_n that,*

$$\sup_j|(\widetilde{\mathbf{\Theta}}_{nj})^{-1}-\{\Phi_{nj}\boxdot(\mathbf{\Theta}_{nj})^{-1}\}|=O(m^{-1})\mathbf{1}_{(p+q+1)\times(p+q+1)},$$

where

$$\Phi_{nj}=\begin{bmatrix}\Phi_{nj}^{11} & \Phi_{nj}^{21^T}\\ \Phi_{nj}^{21} & \mathbf{1}_{q\times q}\end{bmatrix},\quad \Phi_{nj}^{21}=\begin{bmatrix}1 & 1/\phi_1(U_{nj}'^*) & \ldots & 1/\phi_p(U_{nj}'^*)\\ \vdots & \vdots & & \vdots\\ 1 & 1/\phi_1(U_{nj}'^*) & \ldots & 1/\phi_p(U_{nj}'^*)\end{bmatrix}_{p+1\times q},\quad \textit{and}$$

$$\Phi_{nj}^{11}=\begin{bmatrix}1 & 1/\phi_1(U_{nj}'^*) & \ldots & 1/\phi_p(U_{nj}'^*)\\ 1/\phi_1(U_{nj}'^*) & 1/\phi_1^2(U_{nj}'^*) & \ldots & 1/(\phi_p(U_{nj}'^*)\phi_1(U_{nj}'^*))\\ \vdots & & \ddots & \\ 1/\phi_p(U_{nj}'^*) & 1/(\phi_p(U_{nj}'^*)\phi_1(U_{nj}'^*)) & \ldots & 1/\phi_p^2(U_{nj}'^*)\end{bmatrix}_{p+1\times p+1}.$$

The proof follows from Lemma 3 of Şentürk and Müller (2006) by substituting 1 in place of $\phi_{p+1},\ldots,\phi_{p+q}$.

Lemma 4. *Under the technical conditions (C1)-(C7), for a sequence r_n such that $r_n = O_p\{\sqrt{(m \log n)/n}\}$, on event A_n*

$$(a.) \qquad \sup_j \left| (\mathbf{\Theta}_{nj})^{-1} - \mathbf{\Gamma}_{nj}^{-1} \right| = O_p(r_n)\mathbf{1}_{(p+q+1)\times(p+q+1)},$$

$$(b.) \qquad \sup_j \left| L_{nj}^{-1} \boldsymbol{\chi}_{nj}'^T e'_{nj} \right| = O_p(r_n)\mathbf{1}_{(p+q+1)\times 1},$$

where $\mathbf{\Gamma}_{nj}$ is assumed to be nonsingular by (C5), and $e'_{nj} = (e'_{nj1}, \ldots, e'_{njL_{nj}})^T$.

The proof is similar to the proof of Lemma 4 given in Şentürk and Müller (2006). However a key difference is that the limiting term in part (a.), $\mathbf{\Gamma}_{nj}^{-1}$, contains expectations taken conditional on U. The conditioning on U does not disappear because of the dependence between Z_s and U in the case of PCAR.

Proof that pr$(E_n) \to 1$. The formula given in (32) in Şentürk and Müller (2006) can be extended to $\sup_j |\det(\mathbf{\Theta}_{nj}) - \det(\mathbf{\Gamma}_{nj})| = O_p(r_n)$, where r_n is as defined in Lemma 4. This implies, on event A_n, that $\text{pr}(\inf_j \det(\mathbf{\Theta}_{nj}) > \zeta) \to 1$ as $n \to \infty$, where $\zeta = \min\{\rho/2, [\inf_j(\phi_1^2(U'^*_{nj}), \ldots, \phi_p^2(U'^*_{nj}))]^p \rho/2\}$ and ρ is as defined in (C5). Similarly, it can be shown that $\text{pr}(\min_j L_{nj} \leq p+q) \to 0$ as $n \to \infty$, where $p+q$ denotes the number of predictors. Thus, $\text{pr}(A) \to 1$ as $n \to \infty$. Furthermore, Lemma 3 implies that

$$\sup_j |\det(\widetilde{\mathbf{\Theta}}_{nj}) - \phi_1^2(U'^*_{nj} \ldots \phi_p^2(U'^*_{nj}) \det(\mathbf{\Theta}_{nj})| = O_p(m^{-1}).$$

This shows that $\text{pr}(\inf_j \det(\widetilde{\mathbf{\Theta}}_{nj}) > \zeta) \to 1$ as $n \to \infty$, which implies $\text{pr}(\widetilde{A}_n) \to 1$ as $n \to \infty$. Thus $\text{pr}(E_n) \to 1$ as $n \to \infty$.

References

Archie, J. P. (1981). Mathematical coupling of data: a common source of error. *Annals of Surgery* 193, 296-303.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying coefficient models. *J. Am. Statist. Assoc.* 95, 888–902.

Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *J. Am. Statist. Assoc.* 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. *J. Am. Statist. Assoc.* 96, 605–17.

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

Hanson, R. L., Ehm, M. G., Pettitt, D. J., Prochazka, M., Thompson, D. B., Timberlake, D., Foroud, T., Kobes, S., Baier, L., Burns, D. L., Almasy, L., Blangero, J., Garvey, W. T., Bennett, P. H. and Knowler, W. C. (1998). An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *Am. J. Hum. Genet.* 63, 1130–1138.

Hastie, T. and Tibshirani, R. (1993). Varying coefficient models. *J. R. Statist. Soc.* B 55, 757–96.

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.

Hwang, J.T. (1986). Multiplicative Errors-in-Variables Models With Applications to Recent Data Released by the U.S. Department of Energy. *J. Am. Statist. Assoc.* 81, 680–688.

Iturria, S., Carroll, R. J. and Firth, D. (1999). Polynomial Regression and Estimating Functions in the Presence of Multiplicative Measurement Error. *J. R. Statist. Soc.* 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 Int.* 61, 2240–2249.

Knowler, W. C., Pettitt, D. J., Saad, M. F., Charles, M. A., Nelson, R. G., Howard, B. V., Bogardus, C. and Bennett P. H. (1991). Obesity in the Pima Indians: its magnitude and relationship with diabetes. *Am. J. Clin. Nutr.* 53, 1543S–1551S.

Lai, T. L., Robbins, H. and Wei, C. Z. (1979). Strong consistency of least-squares estimates in multiple regression 2. *J. Mult. Anal.* 9, 343-361.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Statist.* 2, 620-628.

Nicholls, D. F. and Quinn, B. G. (1982). Random coefficient autoregressive models: an introduction. *Lect. Notes Stat.* 11.

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.

Ramsay, J. O. and Silverman, B. W. (1997). *The Analysis of Functional Data.* New York: Springer.

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

Şentürk, D. and Müller, H. G. (2006). Inference for covariate adjusted regression via varying coefficient models. *Ann. Statist.* In Press.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care* 261–265.

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

Wu, C. O. and Yu, K. F. (2002). Nonparametric varying coefficient models for the analysis of longitudinal data. *Int. Statist. Rev.* 70, 373–393.

Table 1: Parameter estimates for the regression model $PGC = \gamma_0 + \gamma_1 DBP + \delta_1 Age + \delta_2 TSFT + e$, obtained by least squares regression of $\widetilde{Y} = \widetilde{PGC}$ (plasma glucose concentration) on $\widetilde{X}_1 = \widetilde{DBP}$ (diastolic blood pressure), $Z_1 = Age$ and $Z_2 = TSFT$ (triceps skin fold thickness), and by adjustment through division, i.e. regressing $\widetilde{PGC}/BMI$ on $\widetilde{DBP}/BMI$, Age and $TSFT$, and alternatively by PCAR and CAR, for $n = 524$ subjects. Confidence intervals at the 95% level were obtained by the standard t-statistic for least squares regression and adjustment through division, by the proposed asymptotic intervals (11) for PCAR and by asymptotic intervals given in Şentürk and Müller (2006) for CAR.

	Least sq. reg.		Adj. by division		PCAR		CAR	
Coeff.	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
γ_0	70.6	(54.9, 86.2)	2.2	(1.7,2.8)	75.7	(51.2, 100.1)	75.7	(53.4, 97.9)
DBP	0.25	(0.02, 0.48)	0.72	(0.54,0.90)	0.42	(0.13, 0.71)	0.42	(0.14,0.70)
Age	0.64	(0.38, 0.89)	0.01	(0.00,0.02)	0.47	(0.17, 0.78)	0.48	(0.18, 0.78)
$TSFT$	0.42	(0.16, 0.68)	-0.02	(-0.03,-0.01)	0.22	(-0.22, 0.66)	0.08	(-0.36,0.51)

Table 2: Estimated coverage (in percent) and average interval lengths for the approximate 95% confidence intervals formed for the parameters of the regression model (12), corresponding to sample sizes n = 100, 150, 350, 800, and 1400. Results are based on 1000 Monte Carlo simulation runs for each sample size.

	γ_0		γ_1		γ_2		δ	
n	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
100	87.5	1.15	90.4	0.99	90.8	0.52	92.2	0.68
150	88.6	0.92	92.3	0.75	91.3	0.41	92.4	0.54
350	90.0	0.46	93.9	0.36	91.2	0.19	93.0	0.26
800	93.7	0.26	94.0	0.19	92.1	0.10	93.2	0.15
1400	94.3	0.19	94.3	0.14	94.7	0.08	94.8	0.11