NIH Public Access
Author Manuscript
*J Stat Plan Inference*. Author manuscript; available in PMC 2010 February 1.

Published in final edited form as:
*J Stat Plan Inference*. 2009 February 1; 139(2): 454. doi:10.1016/j.jspi.2008.04.030.

# Partial covariate adjusted regression

**Damla Şentürk**[a] and **Danh V. Nguyen**[b,*]

[a]Department of Statistics, Pennsylvania State University, University Park Pennsylvania 16802, U.S.A.

[b]Division of Biostatistics, University of California, Davis, California 95616, U.S.A.

## Abstract

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 both distorted and undistorted (adjusted/unadjusted) predictors. The PCAR model allows for unadjusted predictors, such as age, gender and demographic variables, which are common in the analysis of biomedical and epidemiological data. The available estimation and inference procedures for CAR are shown to be invalid 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.

### Keywords

Asymptotic normality; Binning; Confidence intervals; Multiple regression; Varying-coefficient models

## 1 Introduction

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

© 2008 Elsevier B.V. All rights reserved.

*Corresponding author. Email addresses: ucdnguyen@ucdavis.edu (D. Nguyen), dsenturk@stat.psu.edu (D. Şentürk).

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 $\tilde{Y}$, $\tilde{X}$, 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,

$$Y=\frac{\tilde{Y}}{U} \quad \text{and} \quad X=\frac{\tilde{X}}{U}.$$

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

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

where they are defined to be the parts of the observed variables, $\tilde{Y}$, $\tilde{X}_1,\ldots,\tilde{X}_p$, 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, $\tilde{Y}$, distorted predictors, $\left\{\tilde{X}_r\right\}_{r=1}^{p}$, and confounder $U$.

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

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

where $\tilde{X}_r = \phi(U)X_r$ and $\tilde{Y} = \psi(U)Y$ denote the adjusted/distorted predictors and response, respectively, and $Z_s$ denotes the unadjusted predictors. The observed data is $\tilde{Y}$, $\tilde{X}_r$, $Z_s$ and $U$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Furthermore, the confounding covariate, $U$, is allowed to depend on the unadjusted predictors $\{Z_s\}$. The flexibility of allowing both distorted and undistorted predictors is needed, particularly in the regression analysis of biomedical data. In many of these applications, which motivated our current work, the researcher is directly interested in the effects of $Z_s$ = age, gender, obesity measures and/or ethnicity on $Y$. Therefore, the predictors $Z_s$ are unadjusted/undistorted. A specific example is the study of Kaysen et al. (2002) where albumin turnover and protein catabolic rate were adjusted for body surface area ($U$) via division, while age and gender were among the unadjusted variables. This is an example of model (1) above, where an adjustment method is needed that allows for unadjusted predictors. In this example, model (1) is used to reflect (a) the known dependence of albumin turnover and protein catabolic rate on $U$ and (b) the specific interest in the direct effects of age and gender in the regression relationship. This issue is further discussed in Section 3 in the context of estimation as well as in Section 8.

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 and may have an arbitrarily large asymptotic bias 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 proposed PCAR methodology, like CAR, provides consistent estimators not only under multiplicative but also under additive distortion as discussed also 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 Section 2 and Section 3. Thus, we develop new theoretical tools for valid inference in the PCAR model. We derive 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 6. The method is further illustrated with a Pima Indians diabetes data set given in Section 5. The proofs of the main results are assembled in Section 7, where some technical conditions and auxiliary results are deferred to the Appendix. We conclude with a brief discussion in Section 8.

We note here that an advantage of CAR and PCAR is that, under the identifiability conditions introduced in Section 2, it yields consistent estimates whether the distortion is multiplicative or additive, i.e. $Y = \tilde{Y} - \psi(U)$ and $X_r = \tilde{X}_r - \phi_r(U)$. (A more detailed discussion of the additive distortion case is given in Section 3.) Additive distortions can be handled by the method of nonparametric partial regression; however, there existed no consistent estimation procedure targeting the $\gamma$'s under multiplicative distortion of both the response and the predictors. Also, the proposed distortion setting has similarities with measurement error modeling if one views $\psi(U)$ and $\phi_r(U)$ as unobserved errors affecting the response and predictors. There is an extensive literature on additive measurement error modeling, which dates back to Berkson (1950). See Carroll et al. (2006) and references therein for a comprehensive overview. A key difference between the proposed framework and traditional measurement error is that, in the proposed distortion setting, the error is a function of an observable covariate $U$. Also, since $U$ is observable this information is incorporated into the estimation. While much work is devoted to additive measurement error, work on multiplicative measurement errors is limited. Relevant work includes Hwang (1986) and Iturria, Carroll and Firth (1999) who proposed 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. Therefore, the method and theory presented here may potentially be of interest to the area of measurement error modeling, despite the differences in the methodological motivation.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

## 2 Partial covariate adjusted regression models

We consider the underlying (unobserved) regression model

$$Y_{ni}=\gamma_0+\sum_{r=1}^{p}\gamma_r X_{nri}+\sum_{s=1}^{q}\delta_s Z_{nsi}+e_{ni}=\boldsymbol{\chi}_{ni}^{\mathrm{T}}\boldsymbol{\alpha}+e_{ni}, \tag{2}$$

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, variance $\sigma^2$ and a finite moment that is higher than the 4th moment. The goal is estimation and inference for the parameter vector $\boldsymbol{\alpha}$ of the unobserved regression model (2). Estimation is based on available distorted predictor and response data, namely $\left\{\tilde{Y}_{ni},\tilde{\boldsymbol{\chi}}_{ni}, U_{ni}\right\}_{i=1}^{n}$, where

$$\begin{aligned}\tilde{Y}_{ni}=\psi(U_{ni})Y_{ni} \text{ and } \tilde{\boldsymbol{\chi}}_{ni} &=(1,\tilde{X}_{n1i},\ldots,\tilde{X}_{npi},Z_{n1i},\ldots,Z_{nqi})^{\mathrm{T}}\\ &=\left\{1,\phi_1(U_{ni})\,X_{n1i},\ldots,\phi_p(U_{ni})\,X_{npi},Z_{n1i},\ldots,Z_{nqi}\right\}^{\mathrm{T}}.\end{aligned} \tag{3}$$

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

$$E\{\psi(U)\}=1 \quad \text{and} \quad E\{\phi_r(U)\}=1. \tag{4}$$

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 $\tilde{X}_r$ and $Z_s$ are dependent on the confounder $U$, and that the latent variable $X_r$ is defined to be the part of $\tilde{X}_r$ 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.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We refer to the model described by (2)–(4) 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, note that it follows from (2) and the mutual independence of $\{e \text{ and } U\}$, $\{e \text{ and } (X_r, Z_s)\}$, and $\{U \text{ and } X_r\}$, for $r = 1,\ldots, p$, $s = 1,\ldots, q$, that the regression of $\tilde{Y}$ on $\tilde{\boldsymbol{\chi}} = (1, \tilde{X}_1,\ldots, \tilde{X}_p, Z_1,\ldots, Z_q)^{\mathrm{T}}$ leads to a fully observable varying coefficient model (Cleveland, Grosse and Shyu, 1991;Hastie and Tibshirani, 1993),

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

where

$$\beta_0(U_{ni})=\gamma_0\psi(U_{ni}),\quad \beta_r(U_{ni})=\gamma_r\frac{\psi(U_{ni})}{\phi_r(U_{ni})},\quad \eta_s(U_{ni})=\delta_s\psi(U_{ni}), \tag{6}$$

and $\epsilon(U_{ni}) = \psi(U_{ni})e_{ni}$. For an extensive overview of the estimation procedures proposed for varying coefficient models, see Wu and Yu (2002). Note that in (6), 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 $\{\psi(\cdot), \phi_r(\cdot)\}$.

## 3 Estimation procedure

The estimation of the regression coefficients, $\gamma_0$, $\{\gamma_r\}_{r=1}^{p}$ 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 $\tilde{Y}$, $\tilde{X}_r$, $Z_s$, 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 $\gamma_0$, $\gamma_r$ and $\delta_s$, respectively. The estimation makes use of the relations between the varying coefficient functions and the regression coefficients given by (6) and the identifiability conditions (4), 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 $\tilde{Y}$ on $\tilde{\boldsymbol{\chi}}$ using the data falling within each bin. The observed data is the collection of $n$ samples: $\left\{\tilde{Y}_{ni}, \tilde{\boldsymbol{\chi}}_{ni}, U_{ni}\right\}_{i=1}^{n}$. It is assumed that the confounding covariate, $U$, is bounded below and above, $a \le U \le b$, where $a < b$ are real numbers. In practice $a$ and $b$ would be taken to be $\min_i U_{ni}$ and $\max_i U_{ni}$, respectively. 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, let $\left(U'_{njk}, \tilde{X}'_{nrjk}, Z'_{nsjk}, \tilde{Y}'_{njk}, X'_{nrjk}, Y'_{njk}\right)$ be the $k$th data element in the $j$th bin, $B_{nj}$, where $k = 1,\ldots, L_{nj}$. Data elements in any given bin are marked by a prime.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 $j$th bin is

$$\left(\widehat{\beta}_{n0j},\ldots,\widehat{\beta}_{npj},\widehat{\eta}_{n1j},\ldots,\widehat{\eta}_{nqj}\right)^{\mathrm{T}}=\left(\tilde{\mathbf{X}}'^{\mathrm{T}}_{nj}\tilde{\mathbf{X}}'_{nj}\right)^{-1}\tilde{\mathbf{X}}'^{\mathrm{T}}_{nj}\tilde{\mathbf{Y}}'_{nj}, \tag{7}$$

where the response vector is $\tilde{\mathbf{Y}}'_{nj}=\left(\tilde{Y}'_{nj1},\ldots,\tilde{Y}'_{njL_{nj}}\right)^{\mathrm{T}}$ and $\tilde{\mathbf{X}}'_{nj}=\left(\tilde{\mathbf{X}}'_{nj1},\ldots,\tilde{\mathbf{X}}'_{njL_{nj}}\right)^{\mathrm{T}}$ is the $L_{nj}\times(p+q+1)$ data matrix in bin $j$, with the $k$th observation $\tilde{\mathbf{X}}'_{njk}=\left(1,\tilde{X}'_{n1jk},\ldots,\tilde{X}'_{npjk},Z'_{n1jk},\ldots,Z'_{nqjk}\right)^{\mathrm{T}}$. The estimated regression coefficients in each bin (7) are the estimators of the varying coefficient functions.

In the second step of the estimation procedure, the estimators of the targeted regression parameters, $\gamma 0$, $\{\gamma_r\}_{r=1}^{p}$ and $\{\delta_s\}_{s=1}^{q}$, are obtained as weighted averages of the estimators $\left\{\widehat{\beta}_{n0j},\ldots,\widehat{\beta}_{npj},\widehat{\eta}_{n1j},\ldots,\widehat{\eta}_{nqj}\right\}_{j=1}^{m}$ from the $m$ bins. The proposed PCAR estimators for $\gamma_0$, $\{\gamma_r\}_{r=1}^{p}$ and $\{\delta_s\}_{s=1}^{q}$ are

$$\widehat{\gamma}_{n0}=\sum_{j=1}^{m}\frac{L_{nj}}{n}\widehat{\beta}_{n0j},\quad \widehat{\gamma}_{nr}=\frac{1}{\overline{\tilde{X}}_{nr}}\sum_{j=1}^{m}\frac{L_{nj}}{n}\widehat{\beta}_{nrj}\overline{\tilde{X}'}_{nrj}\quad\text{and}\quad \widehat{\delta}_{ns}=\sum_{j=1}^{m}\frac{L_{nj}}{n}\widehat{\eta}_{nsj}, \tag{8}$$

where $\overline{\tilde{X}}_{nr}=n^{-1}\sum_{i=1}^{n}\tilde{X}_{nri}$ and $\overline{\tilde{X}'}_{nrj}=L_{nj}^{-1}\sum_{k=1}^{L_{nj}}\tilde{X}'_{nrjk}$. The weights in (8) depend on the number of data points in each bin, namely $L_{nj}$ for $j = 1,\ldots, m$. Note that the estimators proposed in (8) are method of moments estimators targeting $E\{\beta_0(U)\}$, $E\{\beta_r(U)\tilde{X}_r\}/E(\tilde{X}_r)$ and $E\{\eta_s(U)\}$, respectively. It follows then that they are consistent for the underlying parameters, formally stated in Section 4, since $E\{\beta_0(U)\} = \gamma_0$, $E\{\beta_r(U)\tilde{X}_r\}/E(\tilde{X}_r) = \gamma_r$ and $E\{\eta_s(U)\} = \delta_s$, by the relations in (6) and the identifiability conditions.

We note that the estimators $\hat{\gamma}_{n0}$ and $\hat{\gamma}_{nr}$ have the same form as the CAR estimators (Şentürk and Müller, 2005), whereas $\hat{\delta}_{ns}$ are different. Furthermore, a straightforward application of the CAR algorithm yield inconsistent estimators for $\delta_s$ under the more general PCAR model. To see this, denote the original CAR estimators for $\delta_s$ 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 $\delta_s$, 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. Note here that since $Z_s$ is correlated with $U$, one can argue that a latent variable can also be created for $Z_s$ that would be independent of $U$, similar to the construction of $X_r$ and $Y$. An argument can be made that the dependence structure suggests $Z_s=\alpha(U,Z^*_s)$, for some latent variable $Z^*_s$ that is independent of $U$, which can be ed as $Z_{nsi}=\phi^*_s(U_{ni})\,Z^*_{nsi}$. In this case, $C_s$ further simplifies to equal $cov\ \{\psi(U),\phi^*_s(U)\}+1$ (This insight was provided by a reviewer.) From this last form of $C_s$, it follows immediately that if (1) the response is not distorted by $U$, i.e. $\psi(U) = 1$, or if (2) $Z_s$ is

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

independent of $U$, i.e. $\phi_s^*(U)=1$, $C_s=1$, in which case the CAR estimator for $\delta_s$ would be consistent. For the latter mentioned case (2) of assuming $Z_s$ is independent of $U$, this is not realistic in most data applications, 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.

For the former mentioned case (1) when the response is undistorted, simpler adjustments given by Hwang (1986) and Iturria et al. (1999), for multiplicative measurement error only in the predictors, would also be applicable. Hwang (1986) proposes a consistent estimator for the regression coefficients by estimating and adjusting for the bias of the regular least squares estimator. The estimation assumes that consistent estimates of the moments of the measurement error are available. Iturria (1999) proposes two estimation methods, where the first considers specific distributional forms for the measurement error and the second also models the distribution of the unobserved predictors. The two approaches of Hwang and Iturria are similar to PCAR in assuming that the error is independent of the unobserved predictors and that it has a mean of one. The difference of PCAR from the previous two approaches is that no knowledge of the distributional forms or the specific moments are assumed. Instead, information from the observed covariate $U$, of which the measurement error $\phi_r(U)$ is a function of, is utilized in the proposed estimation procedure.

The CAR estimators are biased for only the case of undistorted predictors. In other words, if all variables are considered as distorted then CAR provides consistent estimators. This is the difference between CAR and PCAR. While the PCAR set-up allows researchers the choice to consider a subset of predictors as distorted, the CAR set-up requires that all predictors are considered as distorted in order to yield consistent estimates.

As the PCAR set-up allows researchers to consider a subset of predictors as distorted, an important issue is how to determine whether a predictor should be considered as distorted or undistorted. Note that this distinction cannot be made using a statistical/analytical approach via studying the dependence or the relations between the observed variables and $U$. This is because all observed variables are dependent on $U$ in the model (correlate with $U$) *whether they are distorted or undistorted*. Hence, this decision instead should be made by considering the duality between (a) the decision/assumption on undistorted (distorted) variables and (b) the specification of the underlying model; more specifically, the consideration of the predictor choice in the underlying model.

More precisely, determining to include a predictor as distorted or undistorted corresponds to two completely different underlying models of interest, one involving $Z_s$ and the other involving $Z_s^*$, respectively. Specification of the underlying model will depend on the specific interest of the researcher and the specific context of the application. For example, in the data analysis presented in Section 5, the goal is to uncover the relation between a diabetes marker and diastolic blood pressure adjusted for body mass index. Hence, the diabetes marker and diastolic blood pressure are considered as adjusted/distorted variables. On the other hand, age and triceps skin fold thickness are considered unadjusted/undistorted, as we are interested in the direct effects of age and triceps skin fold thickness on the *BMI*-adjusted diabetes markers.

We also note here that the PCAR estimators given in (8) are consistent for the parameters of the underlying model (2) also under additive distortion. More precisely, consider the simple case of one distorted and one undistorted predictor. The regression model in (2) simplifies to $Y = \gamma_0+\gamma_1 X + \delta_1 Z + e$, where the multiplicative error structure is replaced by the additive error structure, given by $\tilde{Y} = Y + \psi_a(U)$ and $\tilde{X} = X + \phi_a(U)$. The proposed PCAR estimators given in (8) are consistent even when the multiplicative error is replaced by additive error as described

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

above. The above additive error model leads to the following specific varying coefficient model: $E(\tilde{Y}|\tilde{X},Z,U) = \beta_0(U) + \beta_1(U)\tilde{X} + \eta_1(U)Z$, where $\beta_0(U) = \gamma_0 - \gamma_1\phi_a(U) + \psi_a(U)$, $\beta_1(U) = \gamma_1$ and $\eta_1(U) = \delta_1$. The PCAR estimators given in (8), namely $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\delta}_1$, target

$$E\{\beta_0(U)\}, \quad E\{\beta_1(U)\tilde{X}\}/E(\tilde{X}), \quad \text{and} \quad E\{\eta_1(U)\},$$

respectively. This holds regardless of the specific error structure, whether it be additive or multiplicative. Furthermore, under the additive distortion model, we have that

$$E\{\beta_0(U)\} = \gamma_0, \quad E\left\{\beta_1(U)\tilde{X}\right\}/E(\tilde{X}) = \gamma_1 \quad \text{and} \quad E\{\eta_1(U)\} = \delta_1.$$

This follows since $E\{\psi_a(U)\} = E\{\phi_a(U)\} = 0$ in the additive distortion model, under the identifiability condition of no average distortion, i.e. $E(\tilde{Y}) = E(Y)$ and $E(\tilde{X}) = E(X)$. Thus, the PCAR estimators proposed in (8) are consistent for parameters of the underlying model also under additive distortion structure.

## 4 Asymptotic properties

We present the asymptotic distribution of the estimators $\hat{\gamma}_{n0}$, $\hat{\gamma}_{nr}$ and $\hat{\delta}_{ns}$ in (8) when the number of subjects $n$ tends to infinity. 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 $\xrightarrow{\mathscr{D}}$ and convergence in probability by $\xrightarrow{p}$. Also let

$$\left(\tilde{\gamma}_{n0j}, \ldots, \tilde{\gamma}_{npj}, \tilde{\delta}_{n1j}, \ldots, \tilde{\delta}_{nqj}\right)^{\mathrm{T}} = \left(\mathcal{X}'^{\mathrm{T}}_{nj}\mathcal{X}'_{nj}\right)^{-1}\mathcal{X}'^{\mathrm{T}}_{nj}Y'^{\mathrm{T}}_{nj} \tag{9}$$

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

For the PCAR estimators given in (8) to be well defined, the least squares estimators given in (7) must exist for each bin $B_{nj}$, i.e. $\det\left(\tilde{\mathcal{X}}'^{\mathrm{T}}_{nj}\tilde{\mathcal{X}}'_{nj}\right) \neq 0$. Correspondingly, the estimators in (9) will exist under the condition that $\det\left(\mathcal{X}'^{\mathrm{T}}_{nj}\mathcal{X}'_{nj}\right) \neq 0$. The following theorems are given under event $E_n$, explicitly defined in the Appendix, which summarizes the above outlined conditions for the PCAR estimators to be well defined.

For the following theorems, we define the following notations: $\lambda_\psi = E\{\psi^2(U)\}$, $\lambda_\phi = E\{\phi^2(U)\}$, $\lambda_{\psi\phi r} = E\{\psi(U)\psi_r(U)\}$, $\sigma^2_\psi = \mathrm{var}\{\psi(U)\}$, $m_{r,k} = E(X^k_r)$, $\chi^{\mathrm{T}} = (1, X_1, \ldots, X_p, Z_1, \ldots, Z_q)$, $\mathbf{\Gamma} = E(\chi\chi^{\mathrm{T}}|U)$, $\mathcal{M} = \mathbf{\Gamma}^{-1}\chi\chi^{\mathrm{T}}\mathbf{\Gamma}^{-1}$, $\tilde{\mathbf{\Theta}}_{nj} = L^{-1}_{nj}\tilde{\mathcal{X}}'^{\mathrm{T}}_{nj}\tilde{\mathcal{X}}'_{nj}$ and $\omega_{n,l,k} = (\tilde{\mathbf{\Theta}}_{nj})^{-1}_{l1} + (\tilde{\mathbf{\Theta}}_{nj})^{-1}_{l2}\tilde{X}'_{n1jk} + \ldots + (\tilde{\mathbf{\Theta}}_{nj})^{-1}_{l,p+q+1}Z'_{nqjk}$ for $l = 1, \ldots, p + s + 1$ 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 $n \to \infty$,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\sqrt{n}(\widehat{\gamma}_{nr} - \gamma_r) \xrightarrow{\mathscr{D}} \mathbb{N}(0, \sigma^2_{\gamma r}),\ 0 \le r \le p,$$

$$\sqrt{n}(\widehat{\delta}_{ns} - \delta_s) \xrightarrow{\mathscr{D}} \mathbb{N}(0, \sigma^2_{\delta s}),\ 1 \le s \le q,$$

where

$$\begin{aligned}
\sigma^2_{\gamma 0} &= \gamma_0^2 \sigma^2_{\psi} + \sigma^2 E\left\{\psi^2(U).\mathscr{M}_{11}\right\}, \\
\sigma^2_{\gamma r} &= \frac{\gamma_r^2(\lambda_{\psi} m_{r,2} - m^2_{r,1}) + \sigma^2 m^2_{r,1} E\left\{\psi^2(U).\mathscr{M}_{r+1,r+1}\right\} - 2\gamma_r^2(\lambda_{\psi\phi_r} m_{r,2} - m^2_{r,1}) - \gamma_r^2 \mathrm{var}(\tilde{X}_r)}{m^2_{r,1}}, \\
\sigma^2_{\delta s} &= \delta_s^2 \sigma^2_{\psi} + \sigma^2 E\left\{\psi^2(U).\mathscr{M}_{p+s+1,p+s+1}\right\} \textit{ for } 1 \le r \le p \textit{ and } 1 \le s \le q.
\end{aligned}$$

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 $n \to \infty$,

$$\begin{aligned}
\widehat{\sigma}^2_{n\gamma r} &\xrightarrow{p} \sigma^2_{\gamma r}, \quad 0 \le r \le p, \\
\widehat{\sigma}^2_{n\delta s} &\xrightarrow{p} \sigma^2_{\delta s}, \quad 1 \le s \le q,
\end{aligned}$$

where

$$\widehat{\sigma}^2_{n\gamma 0} = \sum_{j=1}^{m} \frac{L_{nj}}{n} \widehat{\beta}^2_{n0j} - \widehat{\gamma}^2_{n0} + \widehat{\sigma}^2_n \frac{n^{-1} \sum_{j=1}^{m} \widehat{\beta}^2_{n0j} \sum_{k=1}^{L_{nj}} \omega^2_{n,1,k}}{\sum_{j=1}^{m} n^{-1} L_{nj} \widehat{\beta}^2_{n0j}},$$

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

Normalizing by the above consistent variance estimators, it holds that

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

Therefore, the approximate (1−α)100% asymptotic confidence intervals for $\gamma_r$ and $\delta_s$ have the endpoints

$$\widehat{\gamma}_{nr}\pm z_{\alpha/2}\frac{\widehat{\sigma}_{n\gamma r}}{\sqrt{n}},\quad\text{and}\quad\widehat{\delta}_{ns}\pm z_{\alpha/2}\frac{\widehat{\sigma}_{n\delta s}}{\sqrt{n}},\tag{10}$$

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

**Remark—**These proposed variance estimators are motivated by the identifiability conditions, the definition of the smooth varying coefficients functions given in (6), Lemma 3 and Lemma 4 (a.). Using the consistency of $\hat{\beta}_{nrj}$ and $\hat{\eta}_{nsj}$ for the values of the functions $\beta_r$ and $\eta_s$ at the midpoint of the *j*th bin and the definitions of $\tilde{Y}'_{njk}$ and $\tilde{X}'_{nrjk}$, 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 $\widehat{\sigma}_n^2,\sum_{j=1}^m n^{-1}L_{nj}\widehat{\beta}_{n0j}^2,\sum_{j=1}^m n^{-1}L_{nj}\widehat{\eta}_{nsj}^2,n^{-1}\sum_{j=1}^m\widehat{\beta}_{nrj}^2\sum_{k=1}^{L_{nj}}\tilde{X}'^2_{nrjk}$ and $n^{-1}\widehat{\gamma}_{nr}\sum_{j=1}^m\widehat{\beta}_{nrj},\sum_{k=1}^{L_{nj}}\tilde{X}'^2_{nrjk}$, respectively. Furthermore, relying mainly on Lemma 3 and Lemma 4 (a.), we target $\gamma_0^2E\left\{\psi^2(U)\mathscr{M}_{11}\right\},\gamma_0^2m_{r,1}^2E\left\{\psi^2(U)\mathscr{M}_{r+1,r+1}\right\}$ and $\gamma_0^2E\left\{\psi^2(U)\mathscr{M}_{p+s+1,p+s+1}\right\}$ with $n^{-1}\sum_{j=1}^m\widehat{\beta}_{n0j}^2\sum_{k=1}^{L_{nj}}\omega_{n,1,k}^2,n^{-1}\sum_{j=1}^m\widehat{\beta}$ and $n^{-1}\sum_{j=1}^m\widehat{\beta}_{n0j}^2\sum_{k=1}^{L_{nj}}\omega_{n,p+s+1,k}^2$, respectively.

## 5 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 $n = 524$ 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. The goal is to uncover the underlying, *BMI* adjusted, regression relation, $PGC = \gamma_0 + \gamma_1 DBP + \delta_1 Age + \delta_2 TSFT + e$, based on the observed plasma glucose concentration ($\tilde{PGC}$; from a oral glucose tolerance test), diastolic blood pressure ($\tilde{DBP}$), triceps skin fold thickness (*TSFT*), age and body mass index. We chose to adjust only the main relation of interest, namely the one between plasma glucose concentration (the response) and diastolic blood pressure for body mass index, and included age and triceps skin fold thickness as unadjusted predictors as they are commonly accounted factors in studies on diabetes.

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 $\tilde{PGC}$ on ($\tilde{DBP}$, *Age*, *TSFT*) without adjusting for the confounder *BMI*, and adjustment via division, i.e. regressing $\tilde{PGC}/BMI$ on ($\tilde{DBP}/BMI$, Age, *TSFT*). The approximate 95% asymptotic confidence intervals for the regression parameters obtained through all three

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

methods are also displayed. The approximate confidence intervals for PCAR estimates were obtained as proposed in (10).

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. The merging algorithm is randomized to avoid the introduction of any additional bias. It starts by merging bins with no points. If there are more than one such bin, it randomly picks one and merges it with its neighbor of smallest number of points. After merging all the bins with no points, the bins with one point and eventually bins with $p + q$ points are merged. For this example with $n = 524$ (after the removal of outliers), 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, where the range of $m$ values yielding robust estimates for different sample sizes is given explicitly in the next section.

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 the undistorted predictors.

## 6 Numerical studies

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

$$Y=\gamma_0+\gamma_1 X_1+\gamma_2 X_2+\delta Z+e, \tag{11}$$

where the parameters of interest are $(\gamma_0, \gamma_1, \gamma_2, \delta)^T = (4,-1, 0.3, 3)$. 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: $(X_1,X_2,Z)^T \sim N_3(\boldsymbol{\mu},\boldsymbol{\Sigma})$, with a general covariance structure

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

The mean vector is $\boldsymbol{\mu} = (0.7, 1.2, |U| - 3.5)^T$, 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 $\tilde{Y} = \psi(U)Y$, $\tilde{X}_1 = \phi_1(U)X_1$, and $\tilde{X}_2 = \phi_2(U)X_2$. Under this simulation setting, we examine (1) the confidence interval coverage levels based on the asymptotic results and (2) the finite sample bias of the estimators, as well as comparing CAR and PCAR estimators in terms of variance and MSE.

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 (10). For the sample sizes $n$ = 100, 150, 350, 800 and 1400, the total number of bins formed were $m$ = 16, 27, 32, 50 and 70. 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 $(\gamma_0, \gamma_1, \gamma_2, \delta)^T = (4, -1, .3, 3)$. 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 in comparison to the CAR estimators. For example, the estimated (absolute bias, variance, MSE) values for PCAR estimators 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 (bias, variance, MSE) values for the CAR estimator for $\delta$ (even though their asymptotic distributions are different, the three other point estimates for $\gamma_0$, $\gamma_1$ and $\gamma_2$ are the same for the two methods), $\hat{\delta}^*$, are (1.294, 1.694, 3.368) at the same sample size of $n$ = 100. The multiplicative bias factor of the CAR estimate for $\delta$, 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 $\hat{\delta}^*$ is off target for $\delta$ = 3 (with a mean of 4.294 at $n$ = 100, and 4.146 at $n$ = 1400). In addition to being biased, note that the CAR estimator $\hat{\delta}^*$ has substantially larger variance relative to the PCAR estimator.

As stated above, for sample sizes $n$ = 100, 150, 350, 800 and 1400, the total number of bins formed were $m$ = 16, 27, 32, 50 and 70, respectively. We carried out additional simulation studies to examine the affect of the number of bins $m$. The results suggest that the estimators are robust, based on estimated MSE, when $m$ is chosen in the intervals [13, 18], [18, 27], [25, 45], [35, 65] and [60, 90] corresponding to sample sizes $n$ = 100, 150, 350, 800 and 1400. In a given application, the above intervals can give rough guidelines on how the choice of $m$ may change with sample size, although a sensitivity analysis for the choice of $m$ specific to the data would also be informative.

Finally, we note that even though the smallest sample size at which the proposed asymptotic confidence intervals attain reasonable coverage in our simulation study is $n$ = 100 (the coverage is around 5% off the targeted level for $n$ = 100), the proposed PCAR point estimators still yield reasonable (bias, variance and MSE) values: (0.0515, 0.5748, 0.5775), (0.0026, 0.3552, 0.3553), (0.0256, 0.0916, 0.0922) and (0.0205, 0.3328, 0.3332) for $\gamma_0 = 4$, $\gamma 1 = -1$, $\gamma 2 = 0.3$, $\delta = 3$, respectively, at $n$ = 50. These values are estimated under the current simulation set-up

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

with 3 predictor variables for $n = 50$ and $m = 8$. For the case of simple linear regression, roughly the same number of bins can be attained with $n = 30$, and hence $n = 30$ would be the smallest sample size where CAR and PCAR give reasonable point estimates. For sample sizes smaller than 30, systematic localization via binning may not be fully feasible so the PCAR estimates should be taken with caution. In this case a very rough localization by stratification (by $U$) into 2–3 groups for crude comparisons is possible.

## 7 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 lemma 1 to lemma 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 different-ttiable, satisfying $E\psi(U) = 1$, $E\phi_r(U) = 1$, and $\phi_r(\cdot) > 0$, $1 \leq r \leq p$.
- (C5) The matrices $\mathbf{\Gamma}_{nj}$, $j = 1,\ldots, m$ are nonsingular, i.e. $\rho = |\inf j \det(\mathbf{\Gamma}_{nj})| > 0$, where $\mathbf{\Gamma}_{nj}=E(L_{nj}^{-1}\mathcal{X}_{nj}^{\mathrm{T}}\mathcal{X}'_{nj}|U'^{*}_{nj}), \mathcal{X}'_{nj}=(\mathcal{X}'_{nj1},\ldots,\mathcal{X}'_{njL_{nj}})^{\mathrm{T}}$ is a $L_{nj} \times (p+q+1)$ undistorted data matrix in bin $j$, and $\mathcal{X}'_{njk}=(1, X'_{n1jk},\ldots, X'_{npjk}, Z'_{n1jk},\ldots, Z'_{nqjk})^{\mathrm{T}}$ denotes the $k$th 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 $E(\tilde{Y}|X) = E(Y|X)$ and $E(\tilde{X}_r|X_r) = X_r$.

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. $\hat{\boldsymbol{\theta}}_{nj} = (\hat{\beta}_{n0j}\, \hat{\beta}_{n1j},\ldots, \hat{\beta}_{npj}, \hat{\eta}_{n1j},\ldots, \hat{\eta}_{npj})^{\mathrm{T}}$.
4. $\tilde{\boldsymbol{\theta}}_{nj}=(\tilde{\gamma}_{n0j}\psi(U'^{*}_{nj}), \tilde{\gamma}_{n1j}\psi(U'^{*}_{nj})/\phi_1(U'^{*}_{nj}),\ldots, \tilde{\gamma}_{npj}\psi(U'^{*}_{nj})/\phi_p(U'^{*}_{nj}), \tilde{\delta}_{n1j}\psi(U'^{*}_{nj}),\ldots, \tilde{\delta}_{nqj}\psi(U'^{*}_{nj}))^{\mathrm{T}}$
5. $\boldsymbol{\theta}_{nj}=(\gamma_0\psi(U'^{*}_{nj}), \gamma_1\psi(U'^{*}_{nj})/\phi_1(U'^{*}_{nj}),\ldots, \tilde{\gamma}_p\psi(U'^{*}_{nj})/\phi_p(U'^{*}_{nj}), \delta_1\psi(U'^{*}_{nj}),\ldots, \delta_q\psi(U'^{*}_{nj}))^{\mathrm{T}}$
6. We use $\mathcal{X}'_{nj(i)}$ to denote the matrix $\mathcal{X}'_{nj}$ and $L_{nj(i)}$ to denote the number of points in the $j$th bin such that $U_{ni} \in B_{nj}$, and $\kappa_{rk(i)} \equiv \{(L_{nj(i)}^{-1}\mathcal{X}'^{\mathrm{T}}_{nj(i)}\mathcal{X}'_{nj(i)})^{-1}\mathcal{X}'^{\mathrm{T}}_{nj(i)}\}_{rk(i)}$ is the $(r, k)$th

element of the matrix $\{(L_{nj}^{-1}\mathcal{X}_{nj}'^{\mathrm{T}}\mathcal{X}_{nj}')^{-1}\mathcal{X}_{nj}'^{\mathrm{T}}\}$ for $1 \le r \le p + q + 1$, where $U_{ni}=U_{njk}'$ is the $k$th element in the ordered sample $(U_{nj1}', \ldots, U_{njL_{nj}}') \in B_{nj}$.

7. $\mathbf{\Theta}_{nj}=L_{nj}^{-1}\mathcal{X}_{nj}'^{\mathrm{T}}\mathcal{X}_{nj}'$.

## Proof of Theorem 1

From Lemma 4 (b.), we have that

$$\sup_j |(L_{nj}^{-1}\tilde{\mathcal{X}}_{nj}'^{\mathrm{T}}\tilde{Y}_{nj}') - \{\Delta \boxdot (L_{nj}^{-1}\mathcal{X}_{nj}'^{\mathrm{T}}Y_{nj}')\}| = O_p(m^{-1})\mathbf{1}_{(p+q+1)\times 1}, \tag{12}$$

where $\Delta=\left\{\psi(U_{nj}'^{*}), \psi(U_{nj}'^{*})\phi_1(U_{nj}'^{*}), \ldots, \psi(U_{nj}'^{*})\phi_p(U_{nj}'^{*}), \psi(U_{nj}'^{*}), \ldots, \psi(U_{nj}'^{*})\right\}^{\mathrm{T}}$

Lemma 3 together with (12) implies that, on event $E_n$,

$$\sup_j \left|\widehat{\boldsymbol{\theta}}_{nj} - \widehat{\boldsymbol{\theta}}_{nj}\right| = O_p(m^{-1})\mathbf{1}_{(p+q+1)\times 1}. \tag{13}$$

We first consider the case of $r = 0$ and show that $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$ is asymptotically normal. Using Lemma 4, (13), 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

$$\sum_{j=1}^{m}\sum_{k=l}^{L_{nj}}\left[\frac{\gamma_0\psi(U_{njk}')}{\sqrt{n}} + \frac{\psi(U_{njk}')e_{njk}'}{\sqrt{n}}\{(L_{nj}^{-1}\mathcal{X}_{nj}'^{\mathrm{T}}\mathcal{X}_{nj}')^{-1}\mathcal{X}_{nj}'^{\mathrm{T}}\}_{1k}\right] - \sqrt{n}\gamma_0 + O_p(\sqrt{n}/m).$$

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

$$\sum_{\substack{i=1\\ j,k}}^{t}\left[\frac{\gamma_0\psi(U_{ni})}{\sqrt{n}} + \frac{\psi(U_{ni})e_{ni}\kappa_{1k(i)}}{\sqrt{n}} - \frac{\gamma_0}{\sqrt{n}}\right] + O_p(\sqrt{n}/m) \equiv \sum_{i=1}^{t} W_{n0i} + O_p(\sqrt{n}/m). \tag{14}$$

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 $\left\{e_{n1}, \ldots, e_{nt}, U_{n1}, \ldots, U_{nt}, L_{nj(1)}, \ldots, L_{nj(t)}, \mathcal{X}_{nj(1)}', \ldots, \mathcal{X}_{nj(t)}'\right\}$. 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 $S_{n0n} \to \mathbb{N}(0, \sigma_{\gamma 0}^2)$ in distribution (McLeish, 1974, Theorem 2.3 and subsequent discussion). This establishes the asymptotic normality of $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$.



NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

$$\sqrt{n}\begin{pmatrix} \widehat{\nu}_{nr} - \gamma_r E(X_r) \\ \bar{\nu}_{nr} - E(X_r) \end{pmatrix} \xrightarrow{\mathscr{D}} \mathbb{N}_2(\underline{0}, \textstyle\sum_r). \tag{15}$$

For (15) to hold, by the Cramér-Wold device, it is enough to show the asymptotic normality of

$$\sqrt{n}\left[a\{\widehat{\nu}_{nr} - \gamma_r E(X_r)\} + b\{\bar{\nu}_{nr} - E(X_r)\}\right] \tag{16}$$

for real $a$, $b$. The asymptotic normality of $\sqrt{n}(\widehat{\gamma}_{nr} - \gamma_r)$ $(1 \le r \le p)$ will follow from (15) by applying the δ-method with $\hat{\gamma}_{nr} = \hat{\nu}_{nr}/\hat{\nu}_{nr}$. Again applying Lemma 4 together with (13) and some simple algebra, we can express $\hat{\nu}_{nr}$ and $\hat{\nu}_{nr}$ as

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

$$\bar{\nu}_{nr}= \sum_{j=l}^{m}\sum_{k=1}^{L_{nj}}\frac{1}{n}\phi_r(U'_{njk})X'_{nrjk}+O_p(m^{-1}).$$

Thus, using similar simplifications as was done for the case of $r = 0$ in (14), the linear combination (16), namely $\sqrt{n}\left[a\{\widehat{\nu}_{nr} - \gamma_r E(X_r)\} + b\{\bar{\nu}_{nr} - E(X_r)\}\right]$ can be expressed as

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

The second term $O_p(\sqrt{n}/m)$ is asymptotically negligible and it is straightforward to verify that $\left\{S_{nrt} \equiv \sum_{i=1}^{t} W_{nri}, F_{n0t}, 1 \le t \le n\right\}$ 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 $S_{nrn} \xrightarrow{\mathscr{D}} \mathbb{N}(0, (a,b)\sum_r (a,b)^{\mathrm{T}})$. 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 $\sigma_{\gamma r}^2$ is explicitly given in Theorem 1.

The asymptotic normality of $\sqrt{n}(\widehat{\delta}_{ns} - \delta_s) \to \mathbb{N}(0, \sigma_{\delta s}^2)$ follows similarly to the case of $\sqrt{n}(\widehat{\gamma}_{n0} - \gamma_0)$, since they have similar forms in (13). (See also definition/notation 4.) The asymptotic variance $\sigma_{\delta s}^2$ which has a similar form as $\sigma_{\gamma 0}^2$, is given explicitly in Theorem 1. This completes the proof of Theorem 1.

## Proof of Theorem 2

The following relation holds on event $A_n$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\sup_j \left| (\tilde{\gamma}_{n0j} - \gamma_0, \ldots, \tilde{\gamma}_{npj} - \gamma_p, \tilde{\delta}_{n1j} - \delta_1, \ldots, \tilde{\delta}_{nqj} - \delta_q)^{\mathrm{T}} \right| = o_p(1)\mathbf{1}_{(p+q+1)\times 1}. \tag{17}$$

It follows from Lemma 4 (a.) and (b.). Utilizing (17) together with (13) gives

$$\sup_j \left| \widehat{\boldsymbol{\theta}}_{nj} - \boldsymbol{\theta}_{nj} \right| = o_p(1)\mathbf{1}_{(p+q+1)\times 1}. \tag{18}$$

By the Law of Large Numbers, (18), and boundedness considerations

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

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

$$\begin{aligned} A_1 &\equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{nrj}^2\sum_{k=1}^{L_{nj}}\tilde{X}'^2_{nrjk} = \gamma_r^2\lambda_\psi m_{r,2} + o_p(1), \\ A_2 &\equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{nrj}\sum_{k=1}^{L_{nj}}\tilde{X}'^2_{nrjk} = \gamma_r\lambda_{\psi\phi_r} m_{r,2} + o_p(1), \\ A_3 &\equiv \sum_{j=1}^{m}\frac{L_{nj}}{n}\widehat{\beta}_{n0j}^2 = \gamma_0^2\lambda_\psi + o_p(1), \end{aligned}$$

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

$$\begin{aligned} A_5 &\equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{n0j}^2\sum_{k=1}^{L_{nj}}\omega_{n,1,k}^2 \xrightarrow{p} \gamma_0^2 E\left\{\psi^2(U)\mathscr{M}_{11}\right\}, \\ A_6 &\equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{n0j}^2\overline{\tilde{X}'}^2_{nrj}\sum_{k=1}^{L_{nj}}\omega_{n,r+1,k}^2 \xrightarrow{p} \gamma_0^2\{E(X_r)\}^2 E\left\{\psi^2(U)\mathscr{M}_{r+1,r+1}\right\}, \textit{and} \\ A_7 &\equiv \frac{1}{n}\sum_{j=1}^{m}\widehat{\beta}_{n0j}^2\sum_{k=1}^{L_{nj}}\omega_{n,p+s+1,k}^2 \xrightarrow{p} \gamma_0^2 E\left\{\psi^2(U)\mathscr{M}_{p+s+1,p+s+1}\right\}. \end{aligned}$$

The estimators of the asymptotic variances given in Theorem 2, in terms of the above quantities, are: (1) $\widehat{\sigma}_{n\gamma 0}^2 = A_3 - \widehat{\gamma}_{n0}^2 + \widehat{\sigma}_n^2 A_5/A_3$, (2) $\widehat{\sigma}_{n\gamma r}^2 = (A_1 + \widehat{\gamma}_{nr}^2\overline{\tilde{X}}_{nr}^2 + 2\widehat{\gamma}_{nr}A_2 + \widehat{\gamma}_{nr}^2 s_{\tilde{X}_r}^2)/\overline{\tilde{X}}_{nr}^2 + (\widehat{\sigma}_n^2 A_6)/(\overline{\tilde{X}}_{nr}^2 A_3)$, and (3) $\widehat{\sigma}_{n\delta s}^2 = A_4 - \widehat{\delta}_{ns}^2 + (\widehat{\sigma}_n^2 A_7)/A_3$. Thus, the first part of Theorem 2 follows by noting that $\widehat{\gamma}_{n0} \xrightarrow{p} \gamma_0, \widehat{\gamma}_{nr} \xrightarrow{p} \gamma_r, \widehat{\delta}_{ns} \xrightarrow{p} \delta_s, s_{\tilde{X}_r}^2 \xrightarrow{p} \mathrm{var}(\tilde{X}_r)$ and $\overline{\tilde{X}}_{nr} \xrightarrow{p} E(X_r)$. Asymptotic confidence

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

intervals given in the second part of the Theorem follow immediately from Theorem 1 and Slutsky's theorem using the consistent variance estimators.

## 8 Discussion

In this work we extend covariate adjusted regression (CAR) models to partial covariate adjusted regression (PCAR) models that allow for the specification of the effects of un-adjusted predictors. Asymptotic normality of the proposed estimators are derived. The PCAR (and CAR) estimation approach was designed to estimate the underlying regression relationship directly, bypassing the estimation of the exact distortion forms. Although the current approach leads to estimators that are simple to implement with known asymptotic properties and good finite sample performance, it does not provide direct estimates of the distorting functions. If the primary interest is in the distorting functions then alternative approaches are needed. A potential alternative approach would involve considering refined estimators for the varying coefficient functions to be used in targeting the distorting functions. However, nontrivial work is required and this remains largely an open problem.

Another interesting and relevant issue, brought to light by a reviewer, is an alternative analysis of the diabetes data given in Section 5. The analysis proceeds by considering the observed variable $Z_s = Age$ as $Z_s = \phi_r^*(U) Z_s^*$, where $U = BMI$ (body mass index) and $Z_s^*$ is a latent variable that can be interpreted as "core biological wear". This implies that the underlying model of interest would involve the latent construct of core biological wear instead of age (defined to be the length of time that a person has lived, which is observable). Such an application is an interesting extension of CAR in the presence of demographic variables, such as age or gender, and may be of key interest to the area of latent variable modeling in general. However, if one is interested in the direct effect of age on the response $Y$, then the specific model of interest would include age as a predictor. In this case, the PCAR methodology developed here would be appropriate, as the application of CAR estimation would result in inconsistent estimates. In either case, the specific area of application and relevant body of scientific literature can provide guidance to the researcher in choosing the relevant model. For example, if a researcher is unsure of whether to enter age into the model as actually age or as 'core biological wear', then the statistician can provide guidance on focusing/clarifying the specific aims of the research hypotheses with the researcher and the available relevant literature. After the specific research hypothesis is defined, which may involve illiciting the relationship (between the predictor and response) of interest to the re-searcher, then a suitable/reasonable PCAR model can be chosen/entertained. Hence, once the hypothesis (relationship of interest) is identified, then the (underlying) model can be specified (containing predictor $Z_s$ or $Z_s^*$) and the appropriate estimators (PCAR or CAR) can be applied accordingly.

Finally, we note the following regarding the implementation of the binning procedure, in the context of the data analysis. For the theory, we assumed that the support of $U$ is the interval $[a, b]$. To be able to bin the data with respect to $U = BMI$ and compute the estimators, we take $a$ and $b$ to be the min/max of the data. For any given population under study, a reasonable range can be inferred to define the limits $a$ and $b$. For adults, we can reasonably set the limits to 14 and 65 BMI, for instance. In our data, the observed min and max are 18.2 and 49.7 and the intervals/bins are between these observed limits. However, if one is applying the binning using the limits 14 and 65, for instance, our estimator weighted by $L_j/n$ will assign zero weight to bins/intervals with no data (as it should), e.g. bins with $U$ between 14 to 18.2 and 49.7 to 65 BMI. Thus, starting the binning at the min/max of the data is approximately equivalent to starting at [a, b].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

## Acknowledgments

We are grateful to the reviewers for many detailed suggestions which substantially improved the paper as well as to the Editor for careful review. This work is supported by Grant Number UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research, NIEHS grant P01-ES011269-06, NIH grants UL1RR024922 and RL1AG032115, and National Institute of Child Health and Human Development grant HD036071.

## Appendix

In this section we provide the additional lemmas and their proofs utilized earlier for the main results. We begin by formally defining the events under which the two main theorems of Section 4 are given. Summarizing the existence conditions for the PCAR estimators, define the events

$$\begin{aligned}\tilde{A}_n &= \{\omega \in \Omega : \inf_j \left|\det(L_{nj}^{-1}\tilde{\mathcal{X}}_{nj}'^{\mathrm{T}}\tilde{\mathcal{X}}_{nj}'\right| > \zeta \text{ and } \min_j L_{nj} > p+q\},\\ A_n &= \{\omega \in \Omega : \inf_j \left|\det(L_{nj}^{-1}\tilde{\mathcal{X}}_{nj}'^{\mathrm{T}}\tilde{\mathcal{X}}_{nj}'\right| > \zeta \text{ and } \min_j L_{nj} > p+q\},\end{aligned} \tag{19}$$

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 (8) and (9) are well defined on events $\tilde{A}_n$ and $A_n$, respectively. The event $E_n$ in Theorem 1 and Theorem 2 is defined to be the intersection of $A_n$ and $\tilde{A}_n$, $E_n = A \cap \tilde{A}_n$. It is shown following Lemma 4 that $\mathrm{pr}(E_n) \to 1$ as $n \to \infty$.

We next 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 $(\chi, U)$ and $(\chi e, U)$, respectively.
- (C7) The error term satisfies $E|e^\tau| < \infty$ for $\tau > 4$.

## Lemma 1

Under the technical conditions *(C1)–(C7)*, on event $A_n$ (19), the martingale differences $W_{n0t}$ satisfy the conditions

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

$$\Delta_{n0}^2 = \sum_{t=1}^{n} W_{n0t}^2 \xrightarrow{p} \sigma_{\gamma 0}^2 \text{ for } \sigma_{\gamma 0}^2 > 0. \tag{b.}$$

.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

### Proof

Let $W_{n0t} = w_{n0t}\upsilon_{n0t}$, where $w_{n0t}=1/\sqrt{n}$, $\upsilon_{n0t} = \gamma_0\psi(U_{nt}) + \psi(U_{nt})e_{nt}\kappa_{1k(t)}-\gamma_0 \equiv \alpha_{1nt}+\alpha_{2nt}e_{nt}$, $\alpha_{1nt}$. 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

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

Now, $E(\upsilon_{n0t}^4)$ is bounded uniformly in $n$ and $t$, since $e_{nt}$ has finite fourth moment by (C7). Also note that

$\Pr(\upsilon_{n0t}^2>n\epsilon^2)=\Pr((\alpha_{1nt}+\alpha_{2nt}e_{nt})^2>n\epsilon^2) \le \Pr(\alpha_{1nt}^2+\alpha_{2nt}^2e_{nt}^2+2|\alpha_{1nt}\alpha_{2nt}e_{nt}|>n\epsilon^2) \le \Pr(c_1^2+c_2^2e_{nt}^2+2c_1c_2|e_{nt}|>n\epsilon^2)$. Lemma 1 (a.) follows, since $\Pr\left(c_1^2+c_2^2e_{nt}^2+2c_1c_2|e_{nt}|>n\epsilon^2\right)\to 0$ uniformly in $n$ and $t$, $e_{nt}^2$ and $|e_{nt}|$ being i.i.d. with finite fourth moments.

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

$$\begin{aligned}
\Delta_{n0}^2 &= \gamma_0^2\left\{n^{-1}\sum_t \psi^2(U_{nt})\right\} + \gamma_0^2 - 2\gamma_0^2\left\{n^{-1}\sum_t \psi(U_{nt})\right\} + 2\gamma_0 n^{-1}\sum_t \psi^2(U_{nt})e_{nt}\kappa_{1k(t)}\\
&-2\gamma_0 n^{-1}\sum_t \psi(U_{nt})e_{nt}\kappa_{1k(t)} + n^{-1}\sum_{j=1}^{m}\sum_{k=1}^{L_{nj}}\left\{(\mathbf{\Theta}_{nj})_{11}^{-1}e'_{njk}\psi(U'_{njk})\right.\\
&\left.+(\mathbf{\Theta}_{nj})_{12}^{-1}e'_{njk}\psi(U'_{njk})X'_{n1jk}+\ldots+(\mathbf{\Theta}_{nj})_{1,p+q+1}^{-1}e'_{njk}\psi(U'_{njk})Z'_{nqjk}\right\}^2 \equiv \mathrm{T}_1+\ldots+\mathrm{T}_6.
\end{aligned}$$

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

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

where $\mathbf{\Gamma}$ and $\mathscr{M}$ are as defined prior to Theorem 1. Thus, $\Delta_{n0}^2 \xrightarrow{p} \gamma_0^2\sigma_\psi^2+\sigma^2 E\left\{\psi^2(U)\mathscr{M}_{11}\right\} \equiv \sigma_{\gamma 0}^2$ and Lemma 1 (b.) follows.

## Lemma 2

*Under the technical conditions (C1)–(C7), on event $A_n$(19), the martingale differences $W_{nrt}$ satisfy the conditions*

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

$$\Delta_{nr}^2 = \sum_{t=1}^{n} W_{nrt}^2 \xrightarrow{p} (a,b)\sum\nolimits_r (a,b)^{\mathrm{T}} \text{ for } (a,b)\sum\nolimits_r (a,b)^{\mathrm{T}} > 0. \tag{b.}$$

## 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 $\Delta_{nr}^2$ in Lemma 2 (b.) is equal to

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

Using the Law of Large Numbers, it holds that $T_1 + \ldots + T_{10} \xrightarrow{p} \alpha^2\gamma_r^2[m_{r,1}^2\sigma_\psi^2 + \mathrm{var}(\mathrm{X}_r)\lambda_\psi] + 2ab\gamma_r\,[\lambda_{\psi\phi_r}m_{r,2} - m_{r,1}^2] + b^2\mathrm{var}(\tilde{X}_r)$. 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

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

Thus

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



NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

## Lemma 3

*Under the technical conditions (C1)–(C6), it holds on event $E_n$ that,*

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

where

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

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

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

## Lemma 4

Under the technical conditions *(C1)–(C7)*, for a sequence $r_n$ such that $r_n=O_p\left\{\sqrt{(m\log n)/n}\right\}$, on event *An*

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

$$\sup_j\left|L_{nj}^{-1}\chi_{nj}'^{\mathrm{T}}e_{nj}'\right|=O_p(r_n)\mathbf{1}_{(p+q+1)\times 1}, \tag{b.}$$

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

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 $\mathrm{pr}(\inf_j\det(\mathbf{\Theta}_{nj})>\zeta)\to 1$ as $n\to\infty$, where

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$\zeta = \min\{\rho/2, [\inf_j(\phi_1^2(U_{nj}'^*), \ldots, \phi_p^2(U_{nj}'^*))]^p \rho/2\}$ and ρ is as defined in (C5). Similarly, it can be shown that $\mathrm{pr}(\min_j L_{nj} \leq p + q) \to 0$ as $n \to \infty$, where $p + q$ denotes the number of predictors. Thus, $\mathrm{pr}(A) \to 1$ as $n \to \infty$. Furthermore, Lemma 3 implies that

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

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

## References


Archie JP. Mathematical coupling of data: a common source of error. Annals of Surgery 1981;193:296–303. [PubMed: 7212790]

Berkson J. Are there two regressions? J. Am. Statist. Assoc 1950;45:164–180.

Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, C. Measurement Error in Nonlinear Models: A Modern Perspective. Second Edition. Boca Raton, FL: Chapman & Hall; 2006.

Cleveland, WS.; Grosse, E.; Shyu, WM. Local regression models. In: Chambers, JM.; Hastie, TJ., editors. Statistical Models in S. Pacific Grove: Wadsworth & Brooks; 1991. p. 309-376.

Hanson RL, Ehm MG, Pettitt DJ, Prochazka M, Thompson DB, Timberlake D, Foroud T, Kobes S, Baier L, Burns DL, Almasy L, Blangero J, Garvey WT, Bennett PH, Knowler WC. An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. Am. J.Hum. Genet 1998;63:1130–1138. [PubMed: 9758619]

Hastie T, Tibshirani R. Varying coefficient models. J. R. Statist. Soc. B 1993;55:757–796.

Hwang JT. Multiplicative errors-in-variables models with applications to recent data released by the U.S. department of energy. J. Am. Statist. Assoc 1986;81:680–688.

Iturria S, Carroll RJ, Firth D. Polynomial regression and estimating functions in the presence of multiplicative measurement error. J. R. Statist. Soc. B 1999;61:547–561.

Kaysen GA, Dubin JA, Müller HG, Mitch WE, Rosales LM, Levin NW. the Hemo Study Group. Relationship among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. Kidney Int 2003;61:2240–2249. [PubMed: 12028466]

Knowler WC, Pettitt DJ, Saad MF, Charles MA, Nelson RG, Howard BV, Bogardus C, Bennett PH. Obesity in the Pima Indians: its magnitude and relationship with diabetes. Am. J. Clin. Nutr 1991;53:1543S–1551S. [PubMed: 2031485]

Lai TL, Robbins H, Wei CZ. Strong consistency of least-squares estimates in multiple regression 2. J. Mult. Anal 1979;9:343–361.

McLeish DL. Dependent central limit theorems and invariance principles. Ann. Statist 1974;2:620–628.

Pinter JD, Brown WE, Eliez S, Schmitt JE, Capone GT, Reiss AL. Amygdala and hippocampal volumes in children with Down syndrome: A high-resolution MRI study. Neurology 2001;56:972–974. [PubMed: 11294940]

Şentürk D, Müller HG. Covariate adjusted regression. Biometrika 2005;92:75–89.

Şentürk D, Müller HG. Inference for covariate adjusted reg ression via varying coefficient models. Ann. Statist 2006;34:654–679.

Smith, JW.; Everhart, JE.; Dickson, WC.; Knowler, WC.; Johannes, RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus; Proceedings of the Symposium on Computer Applications and Medical Care; 1988. p. 261-265.

Wu CO, Yu KF. Nonparametric varying coefficient models for the analysis of longitudinal data. Int. Statist. Rev 2002;70:373–393.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

**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 $\tilde{Y} = \tilde{Y}_{=P\tilde{G}C}$ (plasma glucose concentration) on $= \tilde{X}_{1=D\tilde{B}P}$ (diastolic blood pressure), $Z_1 = Age$ and $Z_2 = TSFT$ (triceps skin fold thickness), and by adjustment through division, i.e. regressing $P\tilde{G}C/BMI$ on $D\tilde{B}P/BMI$, *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 (10) for PCAR and by asymptotic intervals given in Şentürk and Müller (2006) for CAR.

| | Least sq. reg. | | Adj. by division | | PCAR | | PCAR | |
|---|---|---|---|---|---|---|---|---|
| **Coeff.** | **Est.** | **95% CI** | **Est.** | **95% CI** | **Est.** | **95% CI** | **Est.** | **95% CI** |
| $\gamma_0$ | 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) |

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

**Table 2**

Estimated coverage (in percent) and average interval lengths for the approximate 95% confidence intervals formed for the parameters of the regression model (11), corresponding to sample sizes $n$ = 100, 150, 350, 800, and 1400. Results are based on 1000 Monte Carlo simulation runs for each sample size.

| | $\gamma_0$ | | $\gamma_1$ | | $\gamma_2$ | | $\delta$ | |
|---|---|---|---|---|---|---|---|---|
| ***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 |