

Covariate-Adjusted Generalized Linear Models

Damla Şentürk¹

Department of Statistics
Pennsylvania State University
University Park, PA 16802
U.S.A.

Hans-Georg Müller

Department of Statistics
University of California, Davis
One Shields Avenue
Davis, CA 95616
U.S.A.

September 2008

¹ Corresponding author, email: dsenturk@stat.psu.edu

SUMMARY

We propose covariate adjustment methodology for a situation where one wishes to study the dependency of a generalized response on predictors while both predictors and response are distorted by an observable covariate. The distorting covariate is thought of as a size measurement that affects predictors in a multiplicative fashion. The generalized response is modeled by means of a random threshold, where the subject-specific thresholds are affected by a multiplicative factor that is a function of the distorting covariate. While the various factors are modeled as smooth unknown functions of the distorting covariate, the underlying relationship between response and covariates is assumed to be governed by a generalized linear model with known link function. This model provides an extension of a covariate-adjusted regression approach to the case of a generalized linear model. We demonstrate that this contamination model leads to a semiparametric varying-coefficient model. Numerical implementation is straightforward by combining binning, quasi-likelihood and smoothing steps. The asymptotic distribution of the proposed estimators for the regression coefficients of the latent generalized linear model is derived by means of a martingale central limit theorem. Combining this result with consistent estimators for the asymptotic variance makes it then possible to obtain asymptotic inference for the targeted parameters. Both real and simulated data are used in illustrating the proposed methodology.

Keywords: Asymptotic Inference; Confounding; Covariate-adjusted Regression; Multiplicative Effects; Quasi-likelihood; Semiparametric Modeling.

1. INTRODUCTION

Covariate-adjusted regression (CAR) is a recent method to adjust for general multiplicative confounding effects of an observable covariate in the regression setting (Şentürk & Müller 2005, 2006). The main goal is the estimation of the regression coefficients in the latent linear regression model $E(Y_i|X_{1i}, \dots, X_{pi}) = \gamma_0 + \sum_{r=1}^p \gamma_r X_{ri}$, $i = 1, \dots, n$, where n is the sample size and the only observed variables are the distorted response \tilde{Y}_i , predictors \tilde{X}_{ri} and the confounding covariate U_i that needs to be adjusted for. The response and the predictors are multiplicatively distorted according to smooth unknown functions $\psi(\cdot)$ and $\phi_r(\cdot)$ such that $\tilde{Y}_i = \psi(U_i)Y_i$, $\tilde{X}_{ri} = \phi_r(U_i)X_{ri}$. CAR was motivated by the analysis of health sciences data where adjustments are often needed for the effects of body size measures such as body mass index, body surface area, height or weight. These effects are usually thought of as multiplicative, and common normalization proceeds by dividing both responses and predictors by the covariate U (see, e.g., Kaysen et al., 2002 for an example where U is body surface area). Other examples can be found in studies on environmental contaminants and human health risks: Schisterman et al. (2005) normalized polychlorinated biphenyl (PCB, a lipophilic contaminant) measurements through division by serum lipid levels. Such normalizations via division imply that the effect of the variable to be adjusted for is believed to be multiplicative. CAR provides a more flexible adjustment for such “size confounders” compared to simple division, as the distortion functions $\psi(\cdot)$ and $\phi_r(\cdot)$ are allowed to be general smooth functions of U . The naive adjustment through division corresponds to the special case where the distorting functions are equal to the identity function.

The purpose of this paper is to extend adjustment for size confounders to generalized linear models (GLM). A difficulty is the modeling of the confounding of

the response. For example, in the most important binary response case, it is not feasible to consider the response to be directly altered by additive or multiplicative confounding. Our solution to this basic problem of modeling response confounding in generalized response models is to reinterpret these models as governed by subject-specific random thresholds such that the size of the linear predictor relative to the threshold determines the (binary) response. These (unobserved) thresholds are then assumed to be subject to multiplicative confounding effects. We demonstrate that this approach gives rise to a GLM with varying coefficients in which the parameters that define the linear predictor of the GLM depend on the observed confounder U .

Adjustment methods for size confounders in GLMs are in demand for data analysis. For example, one goal of the third National Health and Nutrition Examination Survey (NHANES III) is to model hypertension as a response (Hosmer & Lemeshow, 2000). Numerous covariates, including body height, weight, body mass index (BMI) and serum cholesterol (SC) were measured for 17,030 individuals. The response variable hypertension (HP) is coded as a binary variable. We propose to model the $BMI \equiv U$ adjusted regression relation between the underlying unobserved $HP \equiv Y$ and $SC \equiv X$ by a random threshold model,

$$\mu_i = E(Y_i|X_i) = P(Z_i < \eta_i|X_i) = P(Z_i < \gamma_0 + \gamma_1 X_i|X_i), \quad (1)$$

where Z is a non-negative random variable, interpreted as a subject-specific random threshold. If the subject-specific linear predictor $\eta_i = \gamma_0 + \gamma_1 X_i$ exceeds the subject's threshold Z_i , then the response $Y_i = 1$ (hypertension is present) is observed, otherwise $Y_i = 0$ (no hypertension). The odds of hypertension for subject i with cholesterol level X_i is then $\mu_i/(1 - \mu_i)$. If e.g. $Z|X$ has a standard logistic distribution, $\mu_i/(1 - \mu_i) = \exp(\gamma_0 + \gamma_1 X_i)$. In this case, the odds ratio of the odds at level

$X_i + 1$ over the odds at level X_i , i.e., the odds ratio for increasing X_i by one unit takes the familiar form $\exp(\gamma_1)$. We discuss at the end of Section 4.1 how the odds of hypertension are altered for the covariate-adjusted model.

The main goal is estimation and inference for the regression coefficients γ_0 and γ_1 , based on the observed (and confounded) data consisting of observed hypertension status \tilde{Y}_i , observed serum cholesterol \tilde{X}_i and the *BMI* measurement U_i for each subject. The latent variables (Y and X) are related to their observed counterparts through the equations

$$\tilde{X}_i = \phi(U_i)X_i, \quad E(\tilde{Y}_i|\tilde{X}_i, U_i) = P(\tilde{Z}_i < \eta_i|\tilde{X}_i, U_i) = P\left\{\tilde{Z}_i < \gamma_0 + \frac{\gamma_1}{\phi(U_i)}\tilde{X}_i|\tilde{X}_i, U_i\right\},$$

where $\tilde{Z}_i = \psi(U_i)Z_i$ is the threshold that is distorted by U . One interpretation of this threshold distortion is that the strength of the “defense mechanism” against developing hypertension depends on the *BMI* of the patient. For the unknown smooth distortion functions $\psi(\cdot)$ and $\phi(\cdot)$ we require that $0 < \inf \psi(\cdot) \leq \sup \psi(\cdot) < \infty$, $0 < \inf \phi(\cdot) \leq \sup \phi(\cdot) < \infty$. In this distortion model, the underlying threshold Z and predictor X are independent of U , while their observed versions \tilde{Z} and \tilde{X} have been multiplicatively distorted with factors $\phi(U)$ and $\psi(U)$.

The distortion problem described here has connections to measurement error models in generalized regression models, since the main model of interest in (1) involves latent variables that cannot be observed directly. The measurement error literature in generalized linear models is by and large limited to additive errors in the predictors (Carroll, Ruppert & Stefanski, 1995; Stefanski & Carroll, 1985; Armstrong, 1985; Stefanski, 1989; Carroll, 1989; Carroll & Stefanski, 1990; Wang, Carroll & Liang, 1996) and ignores distortions in responses. The proposed model is quite different from the models considered in the literature, in that (i) it adjusts

for multiplicative error as is adequate for size covariates; (ii) the distortion is a function of an observable covariate; (iii) the distortion affects both predictors and responses, as expected for size confounders and many other confounders; (iv) the underlying model is not a varying-coefficient model and such models are only used as an auxiliary tool to obtain estimation and inference from the observed distorted data.

The above adjustment model, described for binomial regression, can be easily generalized to other GLMs such as Poisson regression or other models with generalized distributions of the responses as they are considered under the GLM or quasi-likelihood framework (see Section 2). Any choice of link function can be accommodated. Consistent estimators for the latent regression coefficients in (1) are proposed in Section 3.1 and the asymptotic distribution of the proposed estimators for the quasi-likelihood framework is derived in Section 3.2, under virtually no distributional assumptions for the responses. Outlines of proofs and additional details are provided in an Appendix. Consistent estimators of the asymptotic variance are used to gain asymptotic inference for the regression parameters. Finite sample performance is investigated through simulations in Section 4, where we also study the effect of serum cholesterol on hypertension in the light of *BMI*.

2. COVARIATE-ADJUSTED RANDOM THRESHOLDS

Consider the following generalization of random threshold models

$$\mu_i = E(Y_i|X_i) = h\left\{P\left(Z_i < \sum_{r=0}^p \gamma_r X_{ri}|X_i\right)\right\} = h\{F_{Z|X}(\eta_i)\} = g^{-1}(\eta_i) \quad (2)$$

for $i = 1, \dots, n$ subjects and p predictors with $X_{0i} = 1$, where $X_i = (X_{1i}, \dots, X_{pi})^T$, Z_i is the subject-specific threshold, $F_{Z|X}(\cdot)$ denotes the conditional cumulative distribution function (cdf) of Z given X , which is assumed to be strictly monotone

on its domain, and $g(\cdot)$ is a strictly monotone increasing continuous link function. We do not assume that the conditional distribution of Y is known, but do assume that $\text{var}(Y_i|X_i) = V(\mu_i) \equiv v_i$ for a known variance function $V(\cdot)$. The function $h(\cdot)$ corresponds to a strictly monotone and continuous transformation, and for example in the case of a binary response is the identity function. In this case, the conditional distribution of Z given X determines the link function g , and when $Z|X$ has a standard normal or standard logistic distribution, the corresponding link functions are probit and logit links, respectively; compare Brockhoff & Müller (1997).

The choice $h(x) = -\log(1 - x)$ is suitable for a Poisson distributed response, as then the inverse link function $g^{-1} \equiv h \circ F_{Z|X}(\cdot)$ maps from \mathbb{R} to \mathbb{R}^+ . In fact, for any given link function g , when using the above transformation, one can always find a cdf $F_{Z|X}(\cdot)$ such that $g^{-1} \equiv h \circ F_{Z|X}(\cdot)$ in (2). To see this, note that combining $h(x) = -\log(1 - x)$ and $g^{-1} \equiv h \circ F_{Z|X}(\cdot)$ gives $F_{Z|X}(x) = 1 - \exp\{-g^{-1}(x)\}$, which is a valid conditional cdf. For example the log link function $g(\mu) = \log(\mu)$ or the complementary log-log link function $g(\mu) = \log\{-\log(1 - \mu)\}$ yield $F_{Z|X}(x) = 1 - \exp\{-g^{-1}(x)\} = 1 - e^{-e^x}$ and $F_{Z|X}(z) = 1 - \exp\{-g^{-1}(z)\} = 1 - e^{-1+e^{-z}}$, respectively. The combination of h and this $F_{Z|X}$ then generates the desired link function g . While the two cases of identity and $-\log(1 - x)$ transformations cover all situations of interest, we note that we do not require knowledge of the precise nature of h or $F_{Z|X}$ for our methodology, as long as the link they generate is known.

Our main goal is the estimation of and inference for the regression parameters γ in model (2), based only on the distorted predictors $\tilde{X}_{ri} = \phi_r(U_i)X_{ri}$, $1 \leq r \leq p$, the distorted response \tilde{Y}_i and the covariate U_i that is to be adjusted for. While the latent predictors X_r and the latent threshold Z and thus the latent model in (2) are not subject to the distortion effects of U , (i.e. $Z \perp U$ and $X_r \perp U$), their distorted

versions \tilde{X}_r and \tilde{Z} are multiplicatively confounded by U , i.e. $\tilde{X}_{ri} = \phi_r(U_i)X_{ri}$ and $\tilde{Z}_i = \psi(U_i)Z_i$, resulting in the observed generalized linear model

$$\begin{aligned}
\tilde{\mu}_i &= E(\tilde{Y}_i|\tilde{X}_i, U_i) = h\left\{P\left(\tilde{Z}_i < \sum_{r=0}^p \gamma_r X_{ri}|\tilde{X}_i, U_i\right)\right\} \\
&= h\left[P\left\{Z_i < \sum_{r=0}^p \frac{\gamma_r}{\psi(U_i)\phi_r(U_i)} \tilde{X}_{ri}|X_i\right\}\right] \\
&= h\left[F_{Z|X}\left\{\sum_{r=0}^p \beta_r(U_i)\tilde{X}_{ri}\right\}\right] = h\{F_{Z|X}(\tilde{\eta}_i)\} = g^{-1}(\tilde{\eta}_i), \tag{3}
\end{aligned}$$

where

$$\beta_0(U_i) = \frac{\gamma_0}{\psi(U_i)}, \quad \beta_r(U_i) = \frac{\gamma_r}{\psi(U_i)\phi_r(U_i)}, \quad \text{for } r = 1, \dots, p, \tag{4}$$

and $\phi_0(\cdot) \equiv 1$. Here, $\tilde{X}_i = (\tilde{X}_{1i}, \dots, \tilde{X}_{pi})^T$, $\psi(\cdot)$ and $\phi_r(\cdot)$, $r = 1, \dots, p$, denote the unknown smooth distorting functions of U , and $\text{var}(\tilde{Y}_i|\tilde{X}_i, U_i) = V(\tilde{\mu}_i) \equiv \tilde{v}_i$.

The observed model in (3) that is subject to confounding is seen to correspond to a generalized varying-coefficient model (Cai et al., 2000, extending Hastie & Tibshirani, 1993) where the observed linear predictor $\tilde{\eta}$ depends on coefficient functions varying in U and the underlying predictors \tilde{X} . While we make use of the connection to the generalized varying-coefficient model in the proposed estimation procedure for covariate adjusted generalized linear models, we note that the two models are clearly distinct. In the proposed covariate adjusted generalized linear model, U is viewed as a “nuisance” parameter and the goal is estimation in a model that is free from or adjusted for the effects of U . In contrast, in the varying coefficient model, U is an integral and important predictor for the outcome and one of the main goals is to recover the (nonlinear) effects of U via the varying coefficient functions. To obtain γ from the varying coefficients of the observed model, we need to make the identifiability assumption that there is no mean distortion. This means

that $E(\tilde{Z}) = E(Z)$ and $E(\tilde{X}_r) = E(X_r)$, i.e. $E\{\psi(U)\} = 1$ and $E\{\phi_r(U)\} = 1$, due to the independence of Z and X_r from U . A re-parametrization of the observed model in (3) then lead to the desired estimates for γ .

3. ESTIMATION AND ASYMPTOTIC PROPERTIES

3.1 Proposed estimation methods

We reparametrize $\tilde{\eta}_i$ by

$$\tilde{\eta}_i = \sum_{r=0}^p \beta_r(U_i) \tilde{X}_{ri} = \frac{1}{\alpha_0(U_i)} + \sum_{r=1}^p \frac{\tilde{X}_{ri}}{\alpha_0(U_i) \alpha_r(U_i)},$$

where

$$\alpha_0(U_i) = \frac{1}{\beta_0(U_i)} = \frac{\psi(U_i)}{\gamma_0} \quad \text{and} \quad \alpha_r(U_i) = \frac{\beta_0(U_i)}{\beta_r(U_i)} = \frac{\phi_r(U_i) \gamma_0}{\gamma_r} \quad \text{for } r = 1, \dots, p. \quad (5)$$

It is assumed here that the vector of targeted parameters $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_r)^\text{T}$ has no zero elements. Our estimation procedure therefore includes a first step to check this assumption with a version of the bootstrap test of Şentürk & Müller (2005). If a predictor or the intercept is found insignificant in this test, the corresponding parameter is set to zero. The estimation procedure is then only applied to the remaining parameters. We include further details on the implementation of this bootstrap test in Section 4.1.

The new parametrization is useful in that once the functions $\alpha(\cdot)$ are estimated, we will be able to target $1/\gamma_0$ and γ_0/γ_r by simply averaging the estimators of $\alpha_0(\cdot)$ and $\alpha_r(\cdot)$, respectively. This follows from the identifiability conditions, which imply $E\{\alpha_0(U)\} = 1/\gamma_0$ and $E\{\alpha_r(U)\} = \gamma_0/\gamma_r$. A binning approach is used to obtain the varying coefficient functions $\alpha(\cdot)$. We assume that the covariate U is bounded below and above, $-\infty < a \leq U \leq b < \infty$, for real numbers $a < b$. The support

$[a, b]$ is divided into m equidistant bins (intervals) denoted by B_1, \dots, B_m . Given the number of bins m , the bins are fixed while the number of U_i falling into each bin is random, where the number falling into bin j is denoted by L_j , $j = 1, \dots, m$.

Denoting the observations falling into bins or quantities within a bin by a prime $\{(U'_{jk}, \tilde{X}'_{rjk}, \tilde{Y}'_{jk}), k = 1, \dots, L_j, r = 1, \dots, p\} = \{(U_i, \tilde{X}_{ri}, \tilde{Y}_i), i = 1, \dots, n, r = 1, \dots, p : U_i \in B_j\}$, we approximate the targeted varying coefficient function values, $\alpha_r(U'_{jk})$, within a bin by $\alpha_r(U_j^*) \equiv \alpha_{rj}$ for $r = 1, \dots, p$, $k = 1, \dots, L_j$ and $j = 1, \dots, m$, where $U_j^* = L_j^{-1} \sum_{k=1}^{L_j} U_{jk}$ denotes the sample average of the U 's falling into bin B_j . To obtain estimators for $\alpha_j \equiv (\alpha_{0j}, \dots, \alpha_{pj})^T$, we maximize the local likelihood

$$\ell_j(\alpha_j) = \frac{1}{L_j} \sum_{k=1}^{L_j} \ell \left\{ g^{-1} \left(\frac{1}{\alpha_{0j}} + \sum_{r=1}^p \frac{\tilde{X}'_{rjk}}{\alpha_{0j} \alpha_{rj}} \right), \tilde{Y}'_{jk} \right\} = \frac{1}{L_j} \sum_{k=1}^{L_j} \ell(\tilde{\mu}_{jk}^*, \tilde{Y}'_{jk}),$$

where $\tilde{\mu}_{jk}^* = g^{-1} \{1/\alpha_{0j} + \sum_{r=1}^p \tilde{X}'_{rjk}/(\alpha_{0j} \alpha_{rj})\}$.

If the conditional log likelihood function cannot be fully specified, it may be replaced with the quasi-likelihood function $Q_j(\cdot)$ defined by $(\partial/\partial \tilde{\mu}_{jk}^*)Q_j(\alpha_j) = L_j^{-1} \sum_{k=1}^{L_j} (\tilde{Y}'_{jk} - \tilde{\mu}_{jk}^*)/\tilde{v}_{jk}^*$, where $\tilde{v}_{jk}^* = V(\tilde{\mu}_{jk}^*)$. Let $W_j^*(\alpha_j)$ denote the $(p+1) \times 1$ score vector

$$W_j^*(\alpha_j) = \frac{\partial Q_j(\alpha_j)}{\partial \alpha_j} = \frac{1}{L_j} \sum_{k=1}^{L_j} \frac{D_{jk}^*}{\tilde{v}_{jk}^*} (\tilde{Y}'_{jk} - \tilde{\mu}_{jk}^*),$$

with $D_{jk}^* = (\partial \tilde{\mu}_{jk}^*/\partial \alpha_{0j}, \dots, \partial \tilde{\mu}_{jk}^*/\partial \alpha_{pj})^T$, and $I_j^*(\alpha_j) = -\partial^2 Q_j(\alpha_j)/\partial \alpha_j^2$ denoting the $(p+1) \times (p+1)$ Hessian matrix. Also define the information matrix $i_j^*(\alpha_j) = E\{I_j^*(\alpha_j) | \tilde{X}, U\}$. The maximum (quasi-)likelihood estimators $\hat{\alpha}_j = (\hat{\alpha}_{0j}, \hat{\alpha}_{1j}, \dots, \hat{\alpha}_{pj})^T$ can be obtained by standard statistical software using Newton-Raphson or Fisher Scoring iteration where the estimators are updated in the s th iteration according to

$$\hat{\alpha}_j^s = \hat{\alpha}_j^{s-1} + \{I_j^*(\hat{\alpha}_j^{s-1})\}^{-1} W_j^*(\hat{\alpha}_j^{s-1}) \quad \text{or} \quad \hat{\alpha}_j^s = \hat{\alpha}_j^{s-1} + \{i_j^*(\hat{\alpha}_j^{s-1})\}^{-1} W_j^*(\hat{\alpha}_j^{s-1}).$$

Once the bin estimators are obtained, we simply take a weighted average to target the vector $\theta = (\theta_0, \theta_1, \dots, \theta_p)^\top \equiv (1/\gamma_0, \gamma_0/\gamma_1, \dots, \gamma_0/\gamma_p)^\top$:

$$\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_p)^\top = \frac{1}{n} \sum_{j=1}^m L_j \hat{\alpha}_j. \quad (6)$$

The bin estimates are weighted according to the number of points L_j that fall into each bin, $1 \leq j \leq m$. The form of the estimator in (6) is motivated by the identifiability conditions and the relations given in (5).

3.2 Asymptotic inference

Our main result (Theorem 1) provides the asymptotic distribution of $\hat{\theta}$ as $n \rightarrow \infty$ and the total number of bins m is such that $\sqrt{n}/m^2 \rightarrow 0$ and $m/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$ under the assumption that the vector of targeted parameters $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_r)^\top$ is different from zero. Under this assumption, the asymptotic distribution of γ follows from Theorem 1 by a straightforward application of the delta method. Theorem 2 gives consistent estimators for the asymptotic variance of $\hat{\theta}$, from which consistent estimators can also be obtained for the asymptotic variance of the estimators of γ . Thus, the two following theorems provide the tools for asymptotic inference for γ . The necessary conditions and outlines of the proofs can be found in the Appendix.

THEOREM 1. *Under the technical conditions given in the Appendix,*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} \mathbb{N}_{p+1}(0_{(p+1) \times 1}, \Sigma),$$

where $0_{(p+1) \times 1}$ denotes the $(p+1) \times 1$ vector of zeros and Σ is the $(p+1) \times (p+1)$ limiting covariance matrix.

The asymptotic normality of $\sqrt{n}(\hat{\gamma} - \gamma)$ follows from Theorem 1 by a simple application of the delta method when θ does not contain zeros, using the transfor-

mation $h(x_0, x_1, \dots, x_p)^T = \{1/x_0, 1/(x_0x_1), \dots, 1/(x_0x_p)\}^T$. Note that θ will not contain zeros since γ does not contain zeros. Even though the explicit form of Σ is not given, the following Theorem demonstrates consistent estimation which is sufficient for asymptotic inference.

THEOREM 2. *Under the technical conditions given in the Appendix,*

$$(\Xi)_{(p+1) \times (p+1)} + \frac{1}{n} \sum_{j=1}^m \sum_{k=1}^{L_j} \{v_j^*(\hat{\alpha}_j)\}^{-1} (\hat{D}_{jk}^*) (\hat{D}_{jk}^*)^T \{v_j^*(\hat{\alpha}_j)\}^{-1} \frac{(\tilde{Y}_{jk}' - \hat{\mu}_{jk}^*)^2}{(\hat{v}_{jk}^*)^2} \xrightarrow{\mathcal{P}} \Sigma,$$

where the (r, r') element of Ξ is $n^{-1} \sum_{j=1}^m L_j \hat{\alpha}_{rj} \hat{\alpha}_{r'j} - \hat{\theta}_r \hat{\theta}_{r'}$, for $r, r' = 0, 1, \dots, p$, $\hat{\mu}_{jk}^* = g^{-1}\{1/\hat{\alpha}_{0j} + \sum_{r=1}^p \tilde{X}'_{rjk}/(\hat{\alpha}_{0j} \hat{\alpha}_{rj})\}$, $\hat{v}_{jk}^* = V(\hat{\mu}_{jk}^*)$ and $(\hat{D}_{jk}^*)_{(p+1) \times 1}$ denotes D_{jk}^* evaluated at $\hat{\alpha}_j$ instead of α_j .

4. APPLICATIONS

4.1 Data analysis

Hypertension (*HP*) or high blood pressure, categorized as a cardiovascular disease, occurs when the force of blood passing through blood vessels is above normal. For the NHANES III data set analyzed here, subjects with a systolic blood pressure above 140 are coded as 1, indicating hypertension, and the others as 0, so that the response is binary. Obesity and high cholesterol are well-known risk factors for hypertension and cardiovascular disease. We address the relationship between *HP* and *SC* adjusted for body mass *BMI* in the NHANES III data set with sample size $n = 17,030$. NHANES III is a survey conducted between 1988 and 1994, designed to provide national estimates of the health and nutritional status of the U.S. population (National Center for Health Statistics, 1994). The data can be obtained at <http://www.cdc.gov/nchs/about/major/nhanes/nh3data.htm>. In our analysis, the clustering and weighting in the design of the study is ignored and this

may potentially have an impact on the interpretations drawn from the inference.

The goal is to uncover the underlying relationship

$$E(HP_i|SC_i) = P(Z_i < \gamma_0 + \gamma_1 SC_i | SC_i),$$

based on the observed presence of hypertension, denoted by \widetilde{HP} , observed serum cholesterol \widetilde{SC} and the *BMI* measurements. \widetilde{SC} and *BMI* are both known to correlate with \widetilde{HP} . The current analysis will help in understanding this three-way relationship better. A pertinent question is whether \widetilde{SC} still has a marginal significant effect on \widetilde{HP} , once the effect of *BMI* on both \widetilde{SC} and \widetilde{HP} is removed by adjusting for it. We address this question by using the proposed model, which will also include the bootstrap test to check the significance of included predictors.

After removing subjects with missing values and outliers in *BMI* and \widetilde{SC} (such that $15 < BMI < 39$ and $100 < \widetilde{SC} < 310$), the sample size is $n = 15,073$, and the resulting subset of the data used is posted at <http://www.stat.psu.edu/dsenturk/Research/supplements.html>. The parameters γ_0 and γ_1 are estimated by the proposed covariate-adjusted random threshold model. We report analysis results for estimates obtained from (a) the proposed covariate-adjusted logistic regression; (b) standard logistic regression of the observed *HP* on observed *SC*, without any adjustment; (c) logistic regression with classical adjustment, obtained by including *BMI* as an additional predictor in the logistic regression model, regressing observed *HP* on observed *SC* and *BMI*; (d) logistic regression with a modified adjustment by also including the interaction term between *SC* and *BMI* in the model used in (c), evaluated at the sample mean level of *BMI*. Estimates and approximate 90% asymptotic confidence intervals for the regression parameters are displayed in Table 1. Approximate confidence intervals for the proposed covariate-adjusted es-

estimates were obtained from the asymptotic results in Theorems 1 and 2, applying appropriate transformations.

The implementation of the binning algorithm utilizes merging of sparsely populated bins. Bin widths are chosen such that there are at least $(p + 1)$ data points in each bin, enough to fit generalized linear regression with $(p - 1)$ predictors. Bins with initially less than $(p + 1)$ elements were merged with neighboring bins. An additional consideration in choosing the bin widths for the case of binary responses is the abundance of response values at 1 and 0, since bins with total or quasi separation needs to be avoided or merged with adjacent bins to obtain stable estimates. For this example, the average number of data points per bin is about 250, yielding a total of 60 bins after merging. To facilitate replication of the analysis, the exact number of points falling in each of the 60 bins is available at <http://www.stat.psu.edu/dsenturk/Research/supplements.html>.

To establish the validity of the proposed estimates and inference, we carry out a check of the significance of the predictors, testing whether the elements of the parameter vector γ are different from zero. Equation (4) shows that $\gamma_r = 0$ is equivalent to $\beta_r(\cdot) = 0$ for $r = 0, \dots, p$, whenever $\psi(\cdot)$ and $\phi_r(\cdot)$ satisfy the identifiability conditions. Thus, testing $H_0 : \beta_r(\cdot) = 0$ is equivalent to testing $H_0 : \gamma_r = 0$, for which we adopt the wild bootstrap as in Şentürk & Müller (2005) and in a first step we obtain raw estimates $(\hat{\beta}_{r1}, \dots, \hat{\beta}_{rm})$ by fitting logistic regressions to the observed data $\{\tilde{Y}_{jk}, \tilde{X}_{1jk}, \dots, \tilde{X}_{pjk}\}_{k=1}^{L_j}$ within each bin. These are plotted against the midpoints of the corresponding bins (B_1, \dots, B_m) in the top panels of Figure 1 for $r = 0, 1$, which also contain local linear fits applied to these scatterplots, providing smooth function estimates for functions β_0 and β_1 within the model $E(\widetilde{HP}_i | \widetilde{SC}_i, BMI_i) = P\{Z_i < \beta_0(BMI_i) + \beta_1(BMI_i)\widetilde{SC}_i\}$. The test procedure is

then based on a quantification of departures of the smooth fits from zero as the test statistic. The tests for $\gamma_0 = 0$ and $\gamma_1 = 0$ reject these null hypotheses with a p-value of 0, confirming that the underlying regression parameters are different from zero and the assumptions are met.

Given the raw estimates $(\hat{\beta}_{r1}, \dots, \hat{\beta}_{rm})$ of the varying coefficients, equation (4) suggests corresponding raw estimates of the distortion factors $\hat{\psi}_j = \hat{\gamma}_0 / \hat{\beta}_{0j}$ and $\hat{\phi}_j = \hat{\gamma}_r \hat{\beta}_{0j} / (\hat{\gamma}_0 \hat{\beta}_{1j})$ for $j = 1, \dots, m$. The scatterplots of these raw distortions against the bin midpoints are included in the lower panels of Figure 1, along with smoothed versions, targeting the underlying distorting functions ψ and ϕ .

In all four analyses, the positive association between *SC* and *HP* is found significant. Adjusted for *BMI*, *SC* has a less significant effect on *HP*, where adjustment by including *BMI* as an extra predictor yields a slightly lower slope estimate than no adjustment. The adjusted slope estimate obtained from the proposed method is the lowest among all four methods and it is noteworthy that this estimate is not included in the 90% asymptotic confidence intervals obtained for the other three established types of analysis. Hence it appears that *BMI* does explain some of the marginal effect of *SC*, but not to the extent that *SC* would become insignificant after this adjustment. As shown in Figure 1, for normal weight subjects ($BMI < 25$), the effect of *SC* is clearly declining as *BMI* increases. This decline starts levelling off for the overweight group ($25 < BMI < 30$), and the effect is more or less constant for obese subjects ($BMI > 30$).

Figure 2 provides the odds of hypertension as a function of serum cholesterol for subjects with *BMI* values 20, 27 and 35, obtained using the smooth varying coefficient function estimates given in Figure 1. Consistent with the levelling off of the slope function for larger *BMI* values in Figure 1, the odds for overweight

and obese subjects behave similarly, while the odds for the normal weight subjects have a sharper increase with the increase of serum cholesterol. Hence the positive association of SC and HP is more pronounced in subjects with lower BMI values and is weaker for overweight or obese subjects, regardless of the level of obesity.

The pronounced positive association of SC and HP for lower BMI values is consistent with the shape of $\hat{\psi}(\cdot)$ in Figure 1. Lower values of $\hat{\psi}(\cdot)$ which are seen to correspond to lower BMI values, imply a lower threshold \tilde{Z} and hence a more pronounced positive association of SC and hypertension (since it is easier to exceed the threshold, an event that corresponds to hypertension). The shape of $\hat{\phi}(\cdot)$ reflects the enhancing effect of BMI on SC . We note that for overweight and obese subjects, raw estimates for the distorting functions have large variances and hence these functions are less well determined for these subjects. The proposed method provides insights into the nature of the effect of risk factors in the presence of confounders that go beyond the findings that are possible with traditional adjustment methods.

4.2 Simulation studies

We carried out two simulation studies for binomial and Poisson distributed responses, choosing the distorting covariate U from Uniform(2, 6) and the distortion functions as $\psi(U) = (U + 3)/7$, $\phi_1(U) = 3(U + 1)^2/79$, satisfying the identifiability conditions for both simulations. The underlying unobserved model is

$$E(Y_i|X_i) = h\{P(Z_i < \gamma_0 + \gamma_1 X_i|X_i)\} = h\{F_{Z|X}(\eta_i)\} = g^{-1}(\eta_i),$$

where in the first study $h(x) = x$ and $F_{Z|X}(z) = e^z/(1 + e^z)$, yielding the logit link function $g(\mu) = \log\{\mu/(1 - \mu)\}$. The predictor is simulated from $X \sim \mathcal{N}(1, 1)$ and the targeted parameters are $\gamma_0 = -3$ and $\gamma_1 = 3$. In the second study $h(x) = -\log(1 - x)$ for Poisson distributed data and $F_{Z|X}(z) = 1 - e^{-e^z}$, hence the link

function is $g(\mu) = \log(\mu)$. The predictor is simulated from $X \sim \text{Uniform}(1, 4)$ and the targeted parameters are $\gamma_0 = 3$ and $\gamma_1 = 1$.

We conducted 1000 Monte Carlo runs for both simulations with sample sizes 200, 500 and 10000. For each run approximate 90% asymptotic confidence intervals were formed for the regression parameters by applying the asymptotic results. Estimated mean squared errors for the estimators for both simulations are listed in Table 2. Reported values are obtained after removing (5, 2, 0.5) percent outliers in the first study and (1, 0.4, 0.1) percent outliers in the second study for each sample size $n = (200, 500, 10000)$, respectively. The estimated coverage fractions and mean interval lengths for these confidence intervals are also given in Table 2 for both the binomial and Poisson cases. The estimated non-coverage fractions are seen to get very close to the target value 0.10, as sample size increases, while estimated interval lengths are sharply decreasing.

Additional simulation studies (not reported) indicate that the estimated mean squared errors of the proposed estimates are sufficiently robust regarding the choice of total number of bins (m) within the ranges [8, 12], [10, 20], [15, 40] and [50, 140] for sample sizes $n = 200, 500, 1000$ and 10000, respectively. Estimated mean squared errors for all sample sizes for different choices of m can be found in the supplemental material at <http://www.stat.psu.edu/dsenturk/Research/supplements.html>, where also simulation results for normally distributed U are documented, with similar results as those presented here for uniform U . Also documented there are the MSEs of the three alternate adjustment methods as described in Section 4.1. These are found not to decrease with increasing sample size, suggesting that these MSEs are dominated by the bias of these methods. This indicates that these methods indeed do not target the right parameters under the multiplicative distortion setting for

which they are not designed for and demonstrates that there is a need for a new adjustment procedure for the type of data distortion we discuss here.

5. CONCLUDING REMARKS

Within the framework of generalized linear models, we propose a flexible method to adjust for the effect of “size type” confounders that exercise multiplicative effects on both predictors and responses. Upon rewriting generalized linear models as random threshold models, the response effect can be modeled as a multiplicative effect on the random threshold. Implementation of this method is straightforward by employing varying coefficient generalized linear models as an auxiliary tool. Preliminary estimators for the distorting functions ψ and ϕ_r are developed in Section 4.1; developing refined estimates and inference procedures for these distorting functions will be interesting topics for future research. We have demonstrated in simulations that the proposed method works well in a variety of situations. In the data example, it leads to insights that classical adjustment methods do not provide. This is especially useful for the complex effects of risk factor levels in the presence of additional influential factors, as demonstrated for the risk of serum cholesterol levels for hypertension in the presence of various levels of body mass index.

ACKNOWLEDGEMENT

We are extremely grateful to two anonymous referees, the associate editor and editor for helpful remarks that improved the paper. This research was supported by the National Science Foundation grants DMS05-05537 and DMS08-06199.

APPENDIX

A1. Technical conditions.

Condition 1. The covariate U satisfies $-\infty < a \leq U \leq b < \infty$, for real $a < b$ and

its density $f(u)$ satisfies $\inf_{a \leq u \leq b} f(u) > 0$, $\sup_{a \leq u \leq b} f(u) < \infty$.

Condition 2. The predictors $\{X_r\}_{r=1}^p$ and the threshold Z are independent of U .

Condition 3. The quasi-likelihood $\{Q_j(\alpha_j)\}_{j=1}^m$ is four times continuously differentiable with respect to α_j . The same also holds for the link function $g^{-1}(\cdot)$.

Condition 4. Contamination functions $\psi(\cdot)$ and $\{\phi_r(\cdot)\}_{r=1}^p$ are twice continuously differentiable, satisfying $E\psi(U) = 1$, $E\phi_r(U) = 1$; and are bounded away from zero, i.e. $\inf_{1 \leq i \leq n} |\psi(U_i)| \geq \Delta$, $\inf_{1 \leq i \leq n, 1 \leq r \leq p} |\phi_r(U_i)| \geq \Delta$ for a $\Delta > 0$.

Condition 5. For the predictors, $\sup_{1 \leq i \leq n, 1 \leq r \leq p} |X_{ri}| \leq C$ for some bound $C > 0$.

Condition 6. The parameters γ_r satisfy $\min_{1 \leq r \leq p} |\gamma_r| \geq \Delta$ for a $\Delta > 0$.

Condition 7. The variance function $V(\cdot)$ is continuously differentiable such that \tilde{v}_i is bounded away from zero uniformly in i , i.e. $\inf_{1 \leq i \leq n} |\tilde{v}_i| \geq \Delta$ for a $\Delta > 0$.

Condition 8. The response variable \tilde{Y} has a finite fourth moment, i.e. $E|\tilde{Y}^4| < \infty$.

Condition 9. The matrices i_j defined below satisfy $\inf_{1 \leq j \leq m} \det(i_j) \geq \Delta$ for a $\Delta > 0$.

A2. Proofs of main results.

Proof of Theorem 1. The equation used for determining the maximum likelihood estimator $\hat{\alpha}_j$ within bin B_j is $W_j^*(\hat{\alpha}_j) = 0$. By the mean value theorem $0 = W_j^*(\alpha_j) - I_j^*(\alpha'_j)(\hat{\alpha}_j - \alpha_j)$, where α'_j lies on the line segment joining $\hat{\alpha}_j$ and α_j . By Condition 1 and a Taylor expansion $L_j^{-1} \sum_{k=1}^{L_j} f(U_j^*) = L_j^{-1} \sum_{k=1}^{L_j} f(U'_{jk}) + O(m^{-2})$ uniformly in j . Note that $W_j^*(\alpha_j)$ and $I_j^*(\alpha_j)$ are functions of $(\tilde{Y}'_{jk}, \tilde{X}'_{rjk}, \alpha_j)$, $k = 1, \dots, L_j$, $r = 1, \dots, p$ and $i_j^*(\alpha_j)$ is a function of $(\tilde{X}'_{rjk}, \alpha_j)$. Using *Conditions 3 and 4* we may replace $\alpha_j = \alpha(U_j^*)$ in these quantities by $\alpha(U'_{jk})$, resulting in asymptotically equivalent quantities W_j , I_j and i_j . Specifically, for $\tilde{\mu}'_{jk} = g^{-1}[1/\alpha_0(U'_{jk}) +$

$\sum_{r=0}^p \tilde{X}'_{rjk}/\{\alpha_0(U'_{jk})\alpha_r(U'_{jk})\}$], $D_{jk} = \{\partial\tilde{\mu}'_{jk}/\partial\alpha_0(U'_{jk}), \dots, \partial\tilde{\mu}'_{jk}/\partial\alpha_p(U'_{jk})\}^T$ and $\tilde{v}'_{jk} = V(\tilde{\mu}'_{jk})$, $W_j = L_j^{-1} \sum_{k=1}^{L_j} D_{jk}(\tilde{Y}'_{jk} - \tilde{\mu}'_{jk})/\tilde{v}'_{jk}$ and $i_j = L_j^{-1} \sum_{k=1}^{L_j} D_{jk}D_{jk}^T/\tilde{v}'_{jk}$, where $W_j^*(\alpha_j) = W_j + O_p(m^{-2})$, $I_j^*(\alpha_j) = I_j + O_p(m^{-2})$ and $i_j^*(\alpha_j) = i_j + O(m^{-2})$. By Conditions 1 and 5 and $E(W_j) = 0$ (see also McCullagh, 1983), $\hat{\alpha}_j - \alpha_j = \{i_j^*(\alpha_j)\}^{-1}W_j^*(\alpha_j) + O_p(m/n) = i_j^{-1}W_j + O_p(m/n) + O_p(m^{-2})$. Note that this result holds uniformly in j , for $j = 1, \dots, m$, as the relevant bounds hold uniformly over all bins. Hence,

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta) &= \sqrt{n} \left[\frac{1}{n} \sum_{j=1}^m \sum_{k=1}^{L_j} \left\{ \frac{\psi(U'_{jk})}{\gamma_0} - \frac{1}{\gamma_0}, \frac{\phi_1(U'_{jk})\gamma_0}{\gamma_1} - \frac{\gamma_0}{\gamma_1}, \dots, \frac{\phi_p(U'_{jk})\gamma_0}{\gamma_p} - \frac{\gamma_0}{\gamma_p} \right\}^T \right. \\ &\quad \left. + \frac{1}{n} \sum_{j=1}^m i_j^{-1} \sum_{k=1}^{L_j} \frac{D_{jk}}{\tilde{v}'_{jk}} (\tilde{Y}'_{jk} - \tilde{\mu}'_{jk}) \right] + O_p(m/\sqrt{n}) + O_p(\sqrt{n}/m^2), \end{aligned}$$

so that $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\left\{ \frac{\psi(U_i)}{\gamma_0} - \frac{1}{\gamma_0}, \frac{\phi_1(U_i)\gamma_0}{\gamma_1} - \frac{\gamma_0}{\gamma_1}, \dots, \frac{\phi_p(U_i)\gamma_0}{\gamma_p} - \frac{\gamma_0}{\gamma_p} \right\}^T + i_{j(i)}^{-1} \frac{D_i}{\tilde{v}_i} (\tilde{Y}_i - \tilde{\mu}_i) \right],$$

where $i_{j(i)}$ denotes the information matrix of the j th bin such that $U_i \in B_j$, $D_i = \{\partial\tilde{\mu}_i/\partial\alpha_0(U_i), \dots, \partial\tilde{\mu}_i/\partial\alpha_p(U_i)\}^T$ and $\tilde{v}_i = V(\tilde{\mu}_i)$ for $\tilde{\mu}_i = g^{-1}[1/\alpha_0(U_i) + \sum_{r=1}^p \tilde{X}_{ri}/\{\alpha_0(U_i)\alpha_r(U_i)\}]$. By the Crámer Wold device, in order to show the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta)$, it is enough to show the asymptotic normality of

$$\begin{aligned} &\sum_{i=1}^n \frac{1}{\sqrt{n}} \left[a_1 \left\{ \frac{\psi(U_i)}{\gamma_0} - \frac{1}{\gamma_0} \right\} + a_2 \left\{ \frac{\phi_1(U_i)\gamma_0}{\gamma_1} - \frac{\gamma_0}{\gamma_1} \right\} + \dots + a_{p+1} \left\{ \frac{\phi_p(U_i)\gamma_0}{\gamma_p} - \frac{\gamma_0}{\gamma_p} \right\} \right. \\ &\quad \left. + \frac{(\tilde{Y}_i - \tilde{\mu}_i)}{\tilde{v}_i} \sum_{r=1}^{p+1} \sum_{r'=1}^{p+1} a_r (i_{j(i)})_{rr'}^{-1} (D_i)_{r'} \right] \equiv \sum_{i=1}^n T_i \end{aligned} \quad (7)$$

for real a_1, \dots, a_{p+1} , where $(i_{j(i)})_{rr'}^{-1}$ denotes the element of $(i_{j(i)})^{-1}$ in the r th row and r' th column and $(D_i)_{r'}$ denotes the r' th element of the $(p+1) \times 1$ vector D_i . Let $S_t = \sum_{i=1}^t T_i$ and let F_t be the σ -field generated by S_t . Note that $\{S_t, F_t, 1 \leq t \leq n\}$ is a mean zero martingale for $n \geq 1$, since $E(\tilde{Y}_i - \tilde{\mu}_i | \tilde{X}_i, U_i) = 0$, $E\psi(U) = 1$

and $E\phi_r(U) = 1$, $r = 1, \dots, p$, by the identifiability conditions. Using Lemma 1, $S_n \xrightarrow{\mathcal{D}} \mathbb{N}\{0, (a_1, \dots, a_{p+1})\Sigma(a_1, \dots, a_{p+1})^T\}$ [McLeish (1974), Theorem 2.3].

Proof of Theorem 2. By a Taylor expansion and Condition 9, $\sup_j |\hat{\alpha}_j - \alpha_j| = o_p(1)$, $\alpha_{rj} = L_j^{-1} \sum_{k=1}^{L_j} \alpha_r(U'_{jk}) + O(m^{-2})$ and $\alpha_{rj}\alpha_{r'j} = L_j^{-1} \sum_{k=1}^{L_j} \alpha_r(U'_{jk})\alpha_{r'}(U'_{jk}) + O(m^{-2})$, uniformly in j for $r, r' = 0, \dots, p$. Therefore,

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^m L_j \hat{\alpha}_{0j}^2 &= \frac{1}{n} \sum_{i=1}^n \alpha_0^2(U_i) + o_p(1) = \frac{E\{\psi^2(U)\}}{\gamma_0^2} + o_p(1), \\ \frac{1}{n} \sum_{j=1}^m L_j \hat{\alpha}_{0j} \hat{\alpha}_{rj} &= \frac{1}{n} \sum_{i=1}^n \alpha_0(U_i) \alpha_r(U_i) + o_p(1) = \frac{E\{\psi(U)\phi_r(U)\}}{\gamma_r} + o_p(1), \\ \frac{1}{n} \sum_{j=1}^m L_j \hat{\alpha}_{rj} \hat{\alpha}_{r'j} &= \frac{1}{n} \sum_{i=1}^n \alpha_r(U_i) \alpha_{r'}(U_i) + o_p(1) = \frac{\gamma_0^2 E\{\phi_r(U)\phi_{r'}(U)\}}{\gamma_r \gamma_{r'}} + o_p(1), \end{aligned}$$

for $r, r' = 1, \dots, p$. Since $|\hat{\theta}_r - \theta_r| = o_p(1)$ for $r = 1, \dots, p$, the $(p+1) \times (p+1)$ matrix Ξ in Theorem 2 converges in probability to a $(p+1) \times (p+1)$ limiting matrix Σ_1 with $(1,1)$ th element equal to $\text{var}\{\psi(U)\}/\gamma_0^2$, $(1,r)$ th element equal to $\text{cov}\{\psi(U), \phi_r(U)\}/\gamma_r$ and (r,r') th element equal to $\gamma_0^2 \text{cov}\{\phi_r(U), \phi_{r'}(U)\}/(\gamma_r \gamma_{r'})$, for $r, r' = 1, \dots, p$. The limit of the terms $A_1 + A_2 + A_3$ (defined in (9)) is $(a_1, \dots, a_{p+1})\Sigma_1(a_1, \dots, a_{p+1})^T$ by (10). The second term of the estimator given in Theorem 2 is now seen to be asymptotically equivalent to

$$\frac{1}{n} \sum_{j=1}^m \sum_{k=1}^{L_j} \{i_j^*(\alpha_j)\}^{-1} (D_j^*) (D_j^*)^T \{i_j^*(\alpha_j)\}^{-1} \frac{(\tilde{Y}'_{jk} - \tilde{\mu}'_{jk})^2}{(\tilde{v}'_{jk})^2}. \quad (8)$$

Using $\sup_{j,k} |U_j^* - U'_{jk}| = O(m^{-1})$ and Taylor expansions yields the following asymptotically equivalent form of (8),

$$\frac{1}{n} \sum_{j=1}^m \sum_{k=1}^{L_j} (i_j)^{-1} (D'_j) (D'_j)^T (i_j)^{-1} \frac{(\tilde{Y}'_{jk} - \tilde{\mu}'_{jk})^2}{(\tilde{v}'_{jk})^2} = \frac{1}{n} \sum_{i=1}^n (i_{j(i)})^{-1} (D_i) (D_i)^T (i_{j(i)})^{-1} \frac{(\tilde{Y}_i - \tilde{\mu}_i)^2}{(\tilde{v}_i)^2},$$

where $D'_j = (D'_{j1}, \dots, D'_{jL_j})$. Hence if we denote the limit of (8) by Σ_2 , term A_6 given

in Lemma 1 converges to $(a_1, \dots, a_{p+1})\Sigma_2(a_1, \dots, a_{p+1})^T$ so that Σ_2 is the remaining part of Σ and Theorem 2 follows.

A3. Auxiliary results and proofs. We assume in the following Conditions 1-9.

LEMMA 1. *The martingale differences T_i defined in (7) satisfy*

$$(a.) \quad \sum_{i=1}^n E\{T_i^2 I(|T_i| > \epsilon)\} \rightarrow 0 \quad \text{for all } \epsilon > 0,$$

$$(b.) \quad \sum_{i=1}^n T_i^2 \xrightarrow{p} (a_1, \dots, a_{p+1})\Sigma(a_1, \dots, a_{p+1})^T \quad \text{for } (a_1, \dots, a_{p+1})\Sigma(a_1, \dots, a_{p+1})^T > 0.$$

Proof. Part (a.) follows from uniform boundedness of the crucial terms which is implied by the assumptions. For part (b.), consider

$$\begin{aligned} \sum_{i=1}^n T_i^2 &= \frac{a_1^2}{\gamma_0^2} \frac{1}{n} \sum_{i=1}^n \{\psi(U_i) - 1\}^2 + \sum_{r=1}^p \frac{2a_1 a_{r+1}}{\gamma_r} \frac{1}{n} \sum_{i=1}^n \{\psi(U_i) - 1\} \{\phi_r(U_i) - 1\} \\ &+ \sum_{r=1}^p \sum_{r'=1}^p \frac{a_{r+1} a_{r'+1} \gamma_0^2}{\gamma_r \gamma_{r'}} \frac{1}{n} \sum_{i=1}^n \{\phi_r(U_i) - 1\} \{\phi_{r'}(U_i) - 1\} \\ &+ \sum_{r=1}^{p+1} \sum_{r'=1}^{p+1} \frac{2a_1 a_r}{\gamma_0} \frac{1}{n} \sum_{i=1}^n \frac{\tilde{Y}_i - \tilde{\mu}_i}{\tilde{v}_i} \{\psi(U_i) - 1\} (i_{j(i)})_{rr'}^{-1} (D_i)_{r'} \\ &+ \sum_{s=1}^p \sum_{r=1}^{p+1} \sum_{r'=1}^{p+1} \frac{2a_{s+1} a_r \gamma_0}{\gamma_s} \frac{1}{n} \sum_{i=1}^n \frac{\tilde{Y}_i - \tilde{\mu}_i}{\tilde{v}_i} \{\phi_s(U_i) - 1\} (i_{j(i)})_{rr'}^{-1} (D_i)_{r'} \quad (9) \\ &+ \sum_{s=1}^{p+1} \sum_{s'=1}^{p+1} \sum_{r=1}^{p+1} \sum_{r'=1}^{p+1} a_s a_r \frac{1}{n} \sum_{i=1}^n \frac{(\tilde{Y}_i - \tilde{\mu}_i)^2}{\tilde{v}_i^2} (i_{j(i)})_{ss'}^{-1} (D_i)_{s'} (i_{j(i)})_{rr'}^{-1} (D_i)_{r'} \equiv A_1 + \dots + A_6. \end{aligned}$$

By the weak law of large numbers, as $n \rightarrow \infty$,

$$\begin{aligned} A_1 + A_2 + A_3 &\xrightarrow{p} \frac{a_1^2 \text{var}\{\psi(U)\}}{\gamma_0^2} + \sum_{r=1}^p \frac{2a_1 a_{r+1} \text{cov}\{\psi(U), \phi_r(U)\}}{\gamma_r} \\ &+ \sum_{r=1}^p \sum_{r'=1}^p \frac{a_{r+1} a_{r'+1} \gamma_0^2 \text{cov}\{\phi_r(U), \phi_{r'}(U)\}}{\gamma_r \gamma_{r'}}. \quad (10) \end{aligned}$$

Using bounds implied by the assumptions, $A_4 = O_p(n^{-1/2})$ and $A_5 = O_p(n^{-1/2})$. By Lemma 2, using the notation introduced there, and a Taylor expansion,

$$A_6 = (1 + o_p(1)) \frac{1}{n} \sum_{i=1}^n \sum_{s=1}^{p+1} \sum_{s'=1}^{p+1} \sum_{r=1}^{p+1} \sum_{r'=1}^{p+1} a_s a_r \frac{(\tilde{Y}_i - \tilde{\mu}_i)^2}{\tilde{v}_i^2} \{\iota(U_i)\}_{ss'}^{-1} (D_i)_{s'} \{\iota(U_i)\}_{rr'}^{-1} (D_i)_{r'},$$

where the first sum taken over i is over i.i.d random variables with finite variance. The weak law of large numbers implies $A_6 = O_p(1)$ and hence part (b.) follows.

LEMMA 2. Let $\mathbf{1}_{(p+1) \times (p+1)}$ denote a $(p+1) \times (p+1)$ matrix of ones, $u_j^M = a + (2j - 1)\{(b-a)/(2m)\}$ the midpoints of the bins B_j and $\{\iota(u)\}_{r,r'} = E\{(D_i)_r(D_i)_{r'}/\tilde{v}_i|U = u\}$. Then

$$\sup_j |(i_j)^{-1} - \{\iota(u_j^M)\}^{-1}| = o_p(1)\mathbf{1}_{(p+1) \times (p+1)}.$$

Proof. The (r, r') th element of i_j for $r, r' = 1, \dots, p$, can be viewed as a Nadaraya-Watson kernel estimator equaling $\{\hat{\iota}_n(u_j^M)\}_{r,r'} \equiv \sum_{i=1}^n \{(D_i)_r(D_i)_{r'}/\tilde{v}_i\} K\{(U_i - u_j^M)/h\} / \sum_{i=1}^n K\{(U_i - u_j^M)/h\}$ with kernel $K(\cdot) = (1/2)\mathbf{1}_{[-1,1]}$ and $h = (b-a)/m$. Uniform consistency of Nadaraya-Watson estimators implies $\sup_{a \leq u \leq b} |\{\hat{\iota}_n(u)\}_{r,r'} - \{\iota(u)\}_{r,r'}| = O_p(r_n)$, where $r_n = O_p\{\sqrt{(m \log n)/n}\}$ and $\sup_j |\{\hat{\iota}_n(u_j^M)\}_{r,r'} - \{\iota(u_j^M)\}_{r,r'}| = O_p(r_n)$, whence $\sup_j |i_j - \iota(u_j^M)| = O_p(r_n)\mathbf{1}_{(p+1) \times (p+1)} = o_p(1)\mathbf{1}_{(p+1) \times (p+1)}$. For the convergence of the inverse of i_j , one observes that the convergence of the determinant and of the cofactors in the matrix is uniform with rate r_n which implies the result.

REFERENCES

- ARMSTRONG, B. (1985). Measurement error in the generalized linear model. *Commun. Stat. Simulat.* **14**, 529-44.
- BROCKHOFF, P. & MÜLLER, H.G. (1997). Random effects threshold models for dose-response relations with repeated measurements. *J.R.S.S. B* **59**, 431-46.
- CAI, Z., FAN, J. & LI, R. (2000). Efficient estimation and inferences for varying coefficient models. *J. Am. Statist. Assoc.* **95**, 888-902.
- CARROLL, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Stat. Med.* **8**, 1075-93.
- CARROLL, R. J., RUPPERT, D. & STEFANSKI, L. A. (1995). Measurement error in non linear models. New York: Chapman and Hall.

- CARROLL, R. J. & STEFANSKI, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Am. Statist. Assoc.* **85**, 652-63.
- HASTIE, T. & TIBSHIRANI, R. (1993). Varying coefficient models. *J.R.S.S. B* **55**, 757-96.
- HOSMER, D. W. & LEMESHOW, S. (2000). *Applied Logistic Regression, Second Edition* Wiley Series in Probability and Statistics.
- KAYSEN, G. A., DUBIN, J. A., MÜLLER, H. G., MITCH, W. E., ROSALES, L. M., LEVIN, N. W. & THE HEMO STUDY GROUP. (2002). Relationship among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. *Kidney Int.* **61**, 2240-9.
- MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Stat.* **11**, 59-67.
- MCLEISH, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* **2**, 620-8.
- NATIONAL CENTER FOR HEALTH STATISTICS, US DEPARTMENT OF HEALTH AND HUMAN SERVICES, PUBLIC HEALTH SERVICE, CENTERS FOR DISEASE CONTROL PREVENTION (1994). Vital and health statistics: plan and operation of the Third National Health and Nutrition Examination Survey, 1988-1994.
- SCHISTERMAN, E. F., WHITCOMB, B. W., LOUIS, G. M. B. & LOUIS, T. A. (2005). Lipid adjustment in the analysis of environmental contaminants and human health risks. *Environ. Health Persp.* **113**, 853-7.
- ŞENTÜRK, D. & MÜLLER, H. G. (2005). Covariate-adjusted regression. *Biometrika* **92**, 75-89.
- ŞENTÜRK, D. & MÜLLER, H. G. (2006). Inference for covariate-adjusted regression via varying coefficient models. *Ann. Stat.* **34**, 654-79.
- STEFANSKI, L. A. (1989). Correcting data for measurement error in generalized linear-models. *Commun Stat-Theor. M.* **18**, 1715-33.
- STEFANSKI, L. A. & CARROLL, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Stat.* **13**, 1335-51.
- WANG, N., CARROLL, R. J. & LIANG, K. Y. (1996). Quasi-likelihood estimation in measurement error models with correlated replicates. *Biometrics* **52**, 401-11.

Table 1: Parameter estimates for the regression model $E(HP_i|SC_i) = P(Z_i < \gamma_0 + \gamma_1 SC_i | SC_i)$, obtained by (a) the proposed covariate-adjusted generalized linear model, adjusting for BMI , (b) standard logistic regression of \widetilde{HP} (hypertension) on \widetilde{SC} (serum cholesterol level), (c) logistic regression of \widetilde{HP} on \widetilde{SC} and BMI (body mass index), (d) of \widetilde{HP} on \widetilde{SC} , BMI and the interaction between \widetilde{SC} and BMI , evaluated at the sample mean level of BMI , for $n = 15,073$ subjects. Approximate confidence intervals at the 90% level are based on standard likelihood asymptotics for models (b)-(d) and on the asymptotic results in Section 3.2 for the proposed model (a).

Intercept			
Methods	Lower Bound	Estimate	Upper Bound
(a) proposed method	-4.235	-3.620	-3.005
(b) no adjustment	-4.200	-4.013	-3.826
(c) adding BMI as predictor	-4.835	-4.579	-4.323
(d) adding the interaction with BMI	-8.614	-7.543	-6.472

SC			
Methods	Lower Bound	Estimate	Upper Bound
(a) proposed method	0.0084	0.0108	0.0133
(b) no adjustment	0.0117	0.0125	0.0134
(c) adding BMI as predictor	0.0113	0.0121	0.0130
(d) adding the interaction with BMI	0.0115	0.0124	0.0132

Table 2: Estimated mean squared errors, coverage (in percent) and mean interval length for the approximate 90% asymptotic confidence intervals formed for the parameters of the regression models described in Section 4.2 for binomial and Poisson cases. The values were obtained from 1000 Monte Carlo runs. The average number of points per bin was 10, 15 and 100 for sample sizes 200, 500 and 10000.

Binomial						
		γ_0		γ_1		
n	MSE	Coverage	Length	MSE	Coverage	Length
200	.5573	87.5	2.49	.5210	86.6	2.43
500	.1619	88.2	1.27	.1297	88.0	1.15
10000	.0055	89.1	0.23	.0050	88.8	0.21
Poisson						
		γ_0		γ_1		
n	MSE	Coverage	Length	MSE	Coverage	Length
200	.0474	86.2	3.69	.0068	85.8	1.22
500	.0124	86.0	1.15	.0016	87.3	0.25
10000	.0000	89.3	0.02	.0000	90.0	0.02

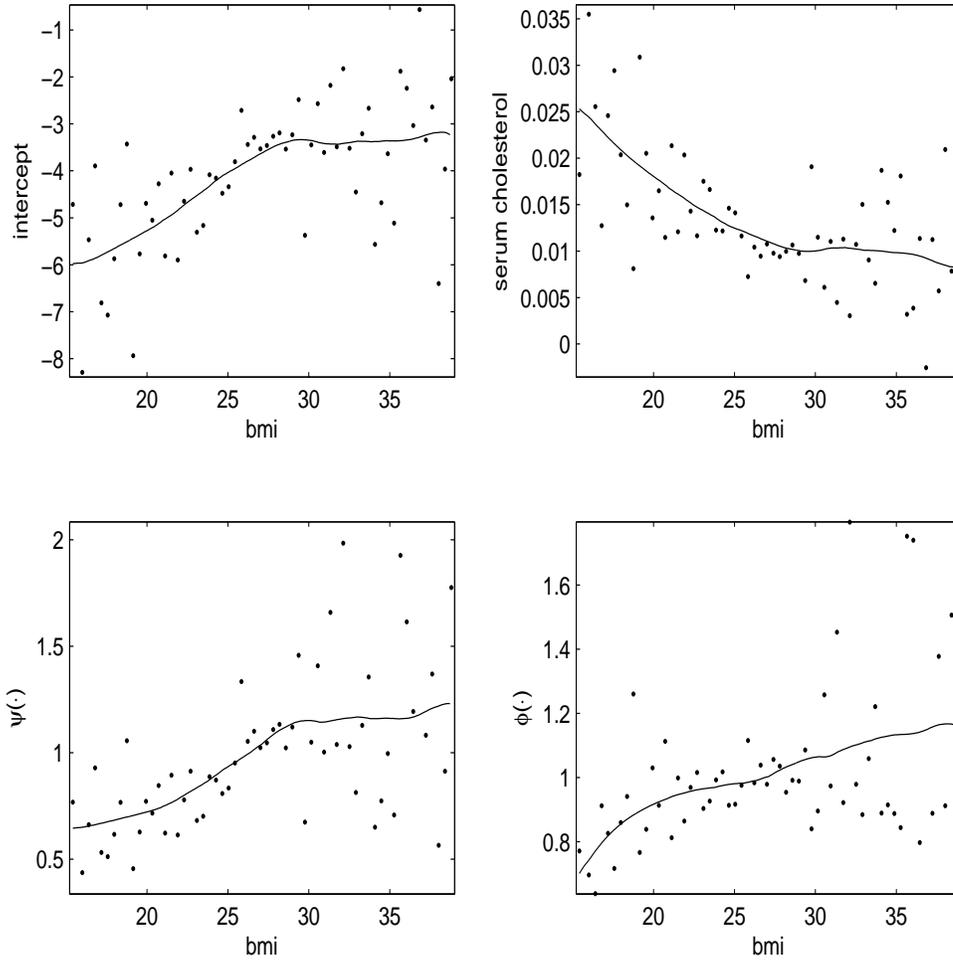


Figure 1: Top panels: Scatterplots of the raw varying coefficients $(\hat{\beta}_{r1}, \dots, \hat{\beta}_{rm})$ vs. midpoints of the bins (B_1, \dots, B_m) for *BMI* (body mass index) and the corresponding local linear fits for $r = 0$ (left) and $r = 1$ (right) for the NHANES III data (sample size $n = 15,073$) with 60 bins. Bottom panels: Corresponding scatterplots of the raw distortion coefficients $(\hat{\psi}_1, \dots, \hat{\psi}_m)$ (left), $(\hat{\phi}_1, \dots, \hat{\phi}_m)$ (right) vs. midpoints of the bins (B_1, \dots, B_m) and the corresponding local linear fits, obtained with cross-validation bandwidth choices (bandwidth=5).

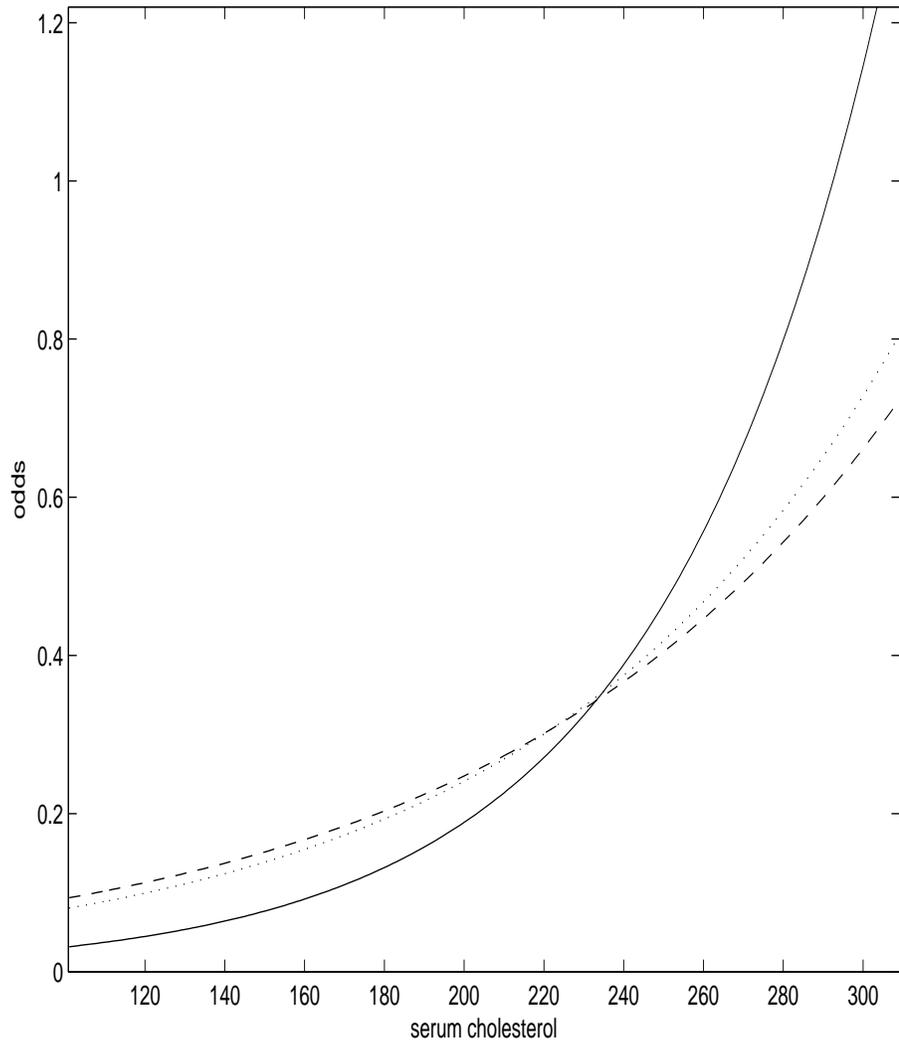


Figure 2: Odds of hypertension as a function of serum cholesterol given for subjects with $BMI = 20$ (solid), $BMI = 27$ (dotted) and $BMI = 35$ (dashed).