4-30-2008

Nonparametric Heteroscedastic Transformation Regression Models for Skewed Data with an Application to Health Care Costs

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

Huazhen Lin

Eric Johnson

Nonparametric heteroscedastic transformation regression models for skewed data with an application to health care costs

Xiao-Hua Zhou [*], [†] **, Huazhen Lin** [*] [‡]**, and Eric Johnson** [*]

SUMMARY

In this paper we develop a new non-parametric heteroscedastic transformation regression model for predicting the expected value of the outcome of a patient with given patient's covariates when the distribution of the outcome is highly skewed with a heteroscedastic variance. In our model, we allow both the transformation function and the error distribution function to be unknown. We show the estimators for regression parameters, the expected value of the original outcome, and the transformation function converge to their true values at the rate $n^{-1/2}$, and the convergent rate that one could expect for a parametric model. In a simulation study, we demonstrate that our proposed nonparametric method is robust with little loss of efficiency. Finally, we apply our model to a study on health care costs.

KEY WORDS: Nonparametric; heteroscedastic; transformation regression models; skewed; health care costs.

1 Introduction

In health services research, risk adjustment has been widely used for assessing provider efficiency, setting capitation rates, and examining resource allocation (Ash, et al.,2000). A key component of the risk adjustment scheme is to predict the health care cost of an individual, given certain demographic characteristics and a measure of the prior health status of that individual; that is, the mean function $\mu(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}]$. The main challenge for such

[*]HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA 98108
[†]Department of Biostatistics, University of Washington, Seattle, WA 98195
[‡]School of Mathematics, Sichuan University, Chengdu, Sichuan 610064, P. R. China
[*]Department of Biostatistics, University of Washington, Seattle, WA 98195

prediction is that the distribution of health care costs is highly skewed with non-constant variance and estimates of $\mu(\boldsymbol{x})$ may be quite sensitive to how estimators treat the skewness and heteroscedasticity (Manning, 1998; Mullahy, 1998; Blough et al., 1999; Manning et al., 2005).

In statistical literature, there are two well established ways of modeling skewed data. The first one is the traditional generalized linear model (GLM) and its semi-parametric and non-parametric extensions (Basu and Rathouz, 2005; Blough et al., 1999; Chiou and Muller, 1998), and the second is to assume that we can transform Y into a special type of the distribution that makes the analysis easier to perform (Carroll and Ruppert, 1988, page 116; Manning, 1998; Mullahy, 1998; Manning et al., 2005). Usually, GLM and transformation models lead to different non-linear regression relationships between $\mu(\boldsymbol{X})$ and $\boldsymbol{X}$. The adequacy of the assumed model depends on a particular application. One major potential advantage of a transformation method over the GLM is that when the expected value of Y has a complex relationship with a vector of covariates, often a transformation of Y simplifies this relationship by inducing a particular type of distribution, e.g. normal, homoscedastic or symmetric distribution so that more efficient estimators and more appropriate plotting can be obtained (Ruppert, 2001).

Most of the existing transformation models in health services research require a specification of the transformation function. The parametric assumption on the transformation function may not always be desirable, as the outcome variable may depend on covariates in a complicated manner. Since the estimates of $\mu(\boldsymbol{x})$ may be quite sensitive to the specification of the transformation function, in this paper, we propose the following nonparametric transformation model to analyze skewed and heteroscedastic variance data:

$$H(Y) = \boldsymbol{X}'\boldsymbol{\beta} + \sigma(\boldsymbol{X}'\gamma)\varepsilon. \tag{1.1}$$

Here $H(\cdot)$ is an unknown increasing transformation function, $\sigma(\cdot)$ is the known variance function, $\boldsymbol{X}$ is a $q \times 1$ vector of observed explanatory variables with the first element being 1, $\boldsymbol{\beta}$ and γ are vectors of unknown parameters, and ε is an error term with mean 0 and variance 1. The error term ε is assumed to be independent of $\boldsymbol{X}$. Let F denote the unknown cumulative distribution function of ε. In the model (1.1), we allow the effect of $\boldsymbol{X}$ on the mean and variance in the transformation scale to be different. With a known transformation function, Welsh and

Zhou (2006) proposed a heteroscedastic regression model for a skewed population.

In this paper, we focus on estimation of $\mu(x)$ under the model (1.1). We propose a new method for estimating the unknown transformation and regression parameters. We then use the resulting estimates to derive an estimator of the expected value of an outcome on the original scale of an individual given their covariates, $\mu(x)$. We show that all the proposed estimators are asymptotically normal and converge to their true values at the rate $n^{-1/2}$, which is the convergent rate that one can expect for a fully parametric model. This result implies that the robustness gained by allowing nonparametric transformation and error distribution functions, comes at little cost of efficiency for estimating the regression parameters and the mean of the original response. The conclusion also is confirmed by our simulation studies in Section 5.

The paper is organized as follows. In Section 2, we present the notations used in forming our model. In Section 3, we give the estimators of H, $\boldsymbol{\beta}$ and $\mu(x)$. We derive the asymptotic distribution properties for the proposed estimators in Section 4, and present the simulation results on the robustness and efficiency of the estimators in Section 5. Finally, we illustrate our methods using a real example in Section 6.

2 Estimation

2·1 *Estimation of regression parameters and transformation function*

To make the model (1.1) identifiable, we assume that $H(y_0) = 0$ for some finite y_0. We first derive the estimator of H. Let $\{Y_i, \boldsymbol{X}_i, i = 1, \cdots, n\}$ be a random sample of (Y, X) that satisfies the model (1.1). Denote $Z_1 = \boldsymbol{X}'\boldsymbol{\beta}$, $Z_2 = \boldsymbol{X}'\gamma$, $Z_{1i} = \boldsymbol{X}_i'\boldsymbol{\beta}$, and $Z_{2i} = \boldsymbol{X}_i'\gamma$. Let $G(\cdot|z_1, z_2)$ be the cumulative distribution function (CDF) of Y conditional on $Z_1 = z_1$ and $Z_2 = z_2$, and $p(\cdot, \cdot)$ be the probability density function of (Z_1, Z_2). Assume that H, F, and G are all differentiable. Define $h(y) = dH(y)/dy$, $f(y) = dF(y)/dy$, $p(y|z_1, z_2) = dG(y|z_1, z_2)/dy$, and $g_j(y|z_1, z_2) = dG(y|z_1, z_2)/dz_j, j = 1, 2$. Notice that under the model (1.1), Y depends on $\boldsymbol{X}$ only through the index Z_1 and Z_2, and the model (1.1) implies the expression for G: $G(y|z_1, z_2) = F(\frac{H(y)-z_1}{\sigma(z_2)})$. By differentiating $G(y \mid z_1, z_2)$ with respect to y and z_1, we obtain

the following expressions:

$$p(y|z_1,z_2)=f(\frac{H(y)-z_1}{\sigma(z_2)})\frac{h(y)}{\sigma(z_2)} \quad \text{and} \quad g_1(y|z_1,z_2)=-f(\frac{H(y)-z_1}{\sigma(z_2)})\frac{1}{\sigma(z_2)},$$

which give us the relationship between $p(y|z_1,z_2)$ and $g_1(y|z_1,z_2)$: $h(y)g_1(y \mid z_1,z_2)=-p(y \mid z_1,z_2)$. If we denote $g_1(y,z_1,z_2)=g_1(y|z_1,z_2)p(z_1,z_2)$ and $p(y,z_1,z_2)=p(y|z_1,z_2)p(z_1,z_2)$, we obtain the following expression:

$$g_1(y,z_1,z_2)h(y)=-p(y,z_1,z_2). \tag{2.1}$$

Replacing z_1 and z_2 with Z_{1i} and Z_{2i} in (2.1) and making a summation over all subjects, we obtain the following expression:

$$\sum_{i=1}^{n} g_1(y,Z_{1i},Z_{2i})h(y)=-\sum_{i=1}^{n} p(y,Z_{1i},Z_{2i}),$$

and solving for $h(.)$, we obtain the following expression:

$$h(y)=-\frac{\sum_{i=1}^{n} p(y,Z_{1i},Z_{2i})}{\sum_{i=1}^{n} g_1(y,Z_{1i},Z_{2i})}. \tag{2.2}$$

Integrating the both sides of (2.2) gives us the following expression of $H(.)$:

$$H(y)=-\int_{y_0}^{y}\frac{\sum_{i=1}^{n} p(u,Z_{1i},Z_{2i})}{\sum_{i=1}^{n} g_1(u,Z_{1i},Z_{2i})}du. \tag{2.3}$$

The expression (2.3) forms the basis for the estimator of H proposed here.

From (2.3), we see that to derive an estimator of $H(.)$, we need to estimate $p(z_1,z_2)$, $G(y|z_1,z_2)$, and derivatives of $G(y|z_1,z_2)$ when the values of $\boldsymbol{\beta}$ and γ are given. We estimate $G(y|z_1,z_2)$ by the following kernel estimator:

$$G_n(y|z_1,z_2)=\frac{1}{nh_1h_2p_n(z_1,z_2)}\sum_{i=1}^{n} I(Y_i\le y)K_1\left(\frac{Z_{1i}-z_1}{h_1}\right)K_2\left(\frac{Z_{2i}-z_2}{h_2}\right), \tag{2.4}$$

where K_1 and K_2 are bounded and symmetric kernel functions with the support $[-1,1]$, and h_1 and h_2 are their corresponding bandwidths. Here $p_n(z_1,z_2)$ is the kernel density estimate of $p(z_1,z_2)$, which is given by the following expression:

$$p_n(z_1,z_2)=\frac{1}{nh_1h_2}\sum_{i=1}^{n} K_1\left(\frac{Z_{1i}-z_1}{h_1}\right)K_2\left(\frac{Z_{2i}-z_2}{h_2}\right). \tag{2.5}$$

Since $g_1(y|z_1,z_2) = dG(y|z_1,z_2)/dz_1$, we obtain an estimator of $g_1(y|z_1,z_2)$ by differentiating $G_n(y|z_1,z_2)$ with respect to z_1:

$$g_{1n}(y|z_1,z_2) = \partial G_n(y|z_1,z_2)/\partial z_1. \tag{2.6}$$

Although $p(y|z_1,z_2)$ is the probability density function of Y conditional on $Z_1 = z_1, Z_2 = z_2$, it can not be directly estimated by $\partial G_n(y|z_1,z_2)/\partial y$ because $G_n(y|z_1,z_2)$ is a step function of y. Instead, we use the following kernel density estimator for $p(y|z_1,z_2)$:

$$p_n(y|z_1,z_2) = \frac{1}{nh_1h_2h_0p_n(z_1,z_2)} \sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0}) K_1\left(\frac{Z_{1i}-z_1}{h_1}\right) K_2\left(\frac{Z_{2i}-z_2}{h_2}\right), \tag{2.7}$$

where K_0 be a bounded and symmetric kernel function with the support $[-1,1]$, and h_0 is its bandwidth. Finally, by substituting (2.5), (2.6) and (2.7) into (2.3), we obtain the following estimator H_n of H:

$$H_n(y) = -\int_{y_0}^{y} \frac{\sum_{i=1}^{n} p_n(u|Z_{1i},Z_{2i}) p_n(Z_{1i},Z_{2i})}{\sum_{i=1}^{n} g_{1n}(u|Z_{1i},Z_{2i}) p_n(Z_{1i},Z_{2i})} du. \tag{2.8}$$

Since $E((H(Y) - Z_1)^2 \mid \boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\gamma)$, without imposing a parametric structure on F, when given H, we can use the following estimating equations to simultaneously estimate $\boldsymbol{\beta}$ and γ:

$$\sum_{i=1}^{n} \frac{(H(Y_i) - \boldsymbol{X}_i'\boldsymbol{\beta})\boldsymbol{X}_i}{\sigma^2(\boldsymbol{X}_i'\gamma)} = 0, \tag{2.9}$$

and

$$\sum_{i=1}^{n} \left\{(H(Y_i) - \boldsymbol{X}_i'\boldsymbol{\beta})^2 - \sigma^2(\boldsymbol{X}_i'\gamma)\right\} \boldsymbol{X}_i = 0. \tag{2.10}$$

From the equation (2.9), we obtain a closed-form estimator of $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_n = \left(\sum_{i=1}^{n} \frac{\boldsymbol{X}_i\boldsymbol{X}_i'}{\sigma^2(\boldsymbol{X}_i'\gamma)}\right)^{-1} \sum_{i=1}^{n} \frac{\boldsymbol{X}_i H(Y_i)}{\sigma^2(\boldsymbol{X}_i'\gamma)}. \tag{2.11}$$

2·2 *Implementation*

We now outline an algorithm for computing $\boldsymbol{\beta}, \gamma$ and $H(\cdot)$.

1. Selection of initial values.

(a) Initial values for $\boldsymbol{\beta}$ and H. We can still obtain consistent estimates for $\boldsymbol{\beta}$ and H even if we misspecify the variance function; hence, we can obtain reasonable starting values for $\boldsymbol{\beta}$ and H with estimates obtained under the homoscedasticity model,

$$H(Y) = \boldsymbol{X}'\boldsymbol{\beta} + \sigma\varepsilon. \tag{2.12}$$

Under this homoscedastic model with the unknown transformation function $H(.)$, we can estimate $\boldsymbol{\beta}$ by the maximum rank correlation (MRC) method proposed by Han (1987); that is, we estimate $\boldsymbol{\beta}$ with $\widetilde{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} W_n(\boldsymbol{\beta})$, where $W_n(\boldsymbol{\beta}) = \sum_{i \neq j}\{Y_i > Y_j\}\{\boldsymbol{X}_i'\boldsymbol{\beta} > X_j'\boldsymbol{\beta}\}$. Then we can estimate H using the proposed method with h_2 so that $K_2(\frac{Z_{2i}-z_2}{h_2}) = 1$ for any z_2 and $i = 1, \cdots, n$.

(b) Initial values for γ. Given $\boldsymbol{\beta}$ and H, we estimate γ by the equation (2.10).

2. Estimation of $H(.)$. Given $\boldsymbol{\beta}$ and γ, we estimate H by (2.8).
3. Estimation of $\boldsymbol{\beta}$ and γ. Given H, we estimate $\boldsymbol{\beta}$ and γ by solving the estimating equations (2.11) and (2.10).
4. Iteration. Repeat Steps 2 and 3 until two successive values of $\boldsymbol{\beta}$, γ, and $H(.)$ don't differ significantly.

In practice, the function H may not be estimated well in the area of extreme observations because of sparsity. It may be necessary to specify a parametric form of H for the area of extreme observations. In general, the parametric form for the area of extreme observations can be inducted according to the estimated function in the interiors of the observations. In the simulations and the example, we assume that the linear form for H at the tail of observations.

2·3 *Prediction of the expected value of Y given covariates $\boldsymbol{X}$*

For given covariates $\boldsymbol{x}$ of a patient, we are interested in predicting $\mu(\boldsymbol{x})$. Under the model (1.1), we can write $\mu(\boldsymbol{x})$ as follows:

$$\mu(\boldsymbol{x}) = \int H^{-1}(\boldsymbol{x}^T\boldsymbol{\beta} + \sigma(\boldsymbol{x}^T\gamma)u)dF(u). \tag{2.13}$$

We propose to estimate F by the empirical distribution $\widehat{F}$ of the standardized residuals, $\widehat{e}_i = \frac{\widehat{H}(Y_i)-\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}}{\sigma(\boldsymbol{X}_i^T\widehat{\gamma})}$, where $\widehat{H}, \widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$ are the estimators of $H, \boldsymbol{\beta}$ and γ, given in Sections 2.1 and 2.2. Therefore, replacing H, $\boldsymbol{\beta}$, γ, and F by $\widehat{H}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\gamma}$, and $\widehat{F}$ in (2.13), we obtain the following estimator of $\mu(\boldsymbol{x})$:

$$\widehat{\mu}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{H}^{-1}\left(\boldsymbol{x}'\widehat{\boldsymbol{\beta}} + \sigma(\boldsymbol{x}'\widehat{\gamma})\frac{\widehat{H}(Y_i) - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}}{\sigma(\boldsymbol{X}_i'\widehat{\gamma})}\right). \tag{2.14}$$

This estimator can be considered as an extension of Duan's smearing estimator (Duan, 1983) to the heteroscedastic transformation model with the unknown transformation and error distribution functions.

3 Large sample properties

In this section, we present the large sample properties of all estimators. In the rest of the paper, we denote the $(k_1 + k_2 + \cdots)$-th order partial derivative of a function $f(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots)$ by $f^{(k_1,k_2,\cdots)}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots)$; that is, $f^{(k_1,k_2,\cdots)}(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots) = \frac{d^{(k_1+k_2+\cdots)}f(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots)}{dx_1^{k_1}dx_2^{k_2}\cdots}$.

To present Theorems 1 to 3 below, we need the following notations. Define

$$\varrho(y) = Eg_1(y, Z_1, Z_2),\ \boldsymbol{\pi}_1(y) = -\int_{y_0}^{y} h(u)\frac{E\left[g_1(u, Z_1, Z_2)q^{(10)}(Z_1, Z_2)\right]}{\varrho(u)}du,$$

$$\mathbf{A}_1 = E\frac{\boldsymbol{X}\left(\boldsymbol{X} + \boldsymbol{\pi}_1(Y)\right)'}{\sigma^2(\boldsymbol{X}'\gamma)},\quad \mathbf{A}_2 = E\left[\frac{\boldsymbol{X}\boldsymbol{\pi}_2(Y)'}{\sigma^2(\boldsymbol{X}'\gamma)}\right],\quad \mathbf{A}_3 = E\left[\boldsymbol{X}\sigma(Z_2)\varepsilon\boldsymbol{\pi}_1'(Y)\right],$$

$$\mathbf{A}_4 = E\left[\left(\sigma^2\right)^{(1)}(Z_2)\boldsymbol{X}\boldsymbol{X}'\right] + 2E\left[\boldsymbol{X}\sigma(Z_2)\varepsilon\boldsymbol{\pi}_2'(Y)\right] - 2\mathbf{A}_3\mathbf{A}_1^{-1}\mathbf{A}_2,$$

$$\boldsymbol{\pi}_2(y) = -\int_{y_0}^{y} h(u)\frac{E\left[(H(u) - Z_1)g_1(u, Z_1, Z_2)q^{(10)}(Z_1, Z_2)\sigma^{(1)}(Z_2)/\sigma(Z_2)\right]}{\varrho(u)}du,$$

$$\boldsymbol{\pi}(y) = \left(\mathbf{A}_1^{-1}\mathbf{A}_2\right)'\boldsymbol{\pi}_1(y) - \boldsymbol{\pi}_2(y),\ \ \eta(y, z_1, z_2) = \frac{2h(y)p^{(10)}(z_1, z_2)}{\varrho(y)},$$

and $q(z_1, z_2) = E[\boldsymbol{X}|Z_1 = z_1, Z_2 = z_2]$. Then, we have the following results.

Lemma A. Under the conditions given in Appendix A, we have the following linear expansion of $\widehat{H}(y) - H(y)$:

$$\begin{aligned}\widehat{H}(y) - H(y) &= \frac{1}{n}\sum_{i=1}^{n}\delta_i(y) - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{\pi}_1(y) - (\widehat{\gamma} - \gamma)'\boldsymbol{\pi}_2(y) \\ &\quad + o_p(n^{-1/2}) + o_p(\widehat{\gamma} - \gamma) + o_p(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \end{aligned} \tag{3.1}$$

where $\delta_i(y) = \int_{y_0}^{y} \eta(u, Z_{1i}, Z_{2i}) \left[I(Y_i \le u) - G(u|Z_{1i}, Z_{2i})\right] du$

$$- \left[\frac{I(y_0 \le Y_i \le y)}{\varrho(Y_i)} - \int_{y_0}^{y} \frac{p(u|Z_{1i}, Z_{2i})}{\varrho(u)} du\right] p(Z_{1i}, Z_{2i}).$$

We give a proof of Lemma A in Appendix B. Lemma A is a key to establish the asymptotic properties of all the estimators given in Theorems 1 to 4 below. We present a proof of Theorems 1 to 3 in Appendix C and a proof of Theorem 4 in Appendix D.

Theorem 1. Under the conditions given in Appendix A, we have the following asymptotic expansion of $\widehat{\beta} - \beta$:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n}\varphi_i^{\boldsymbol{\beta}} + o_p(n^{-1/2}), \tag{3.2}$$

where $\varphi_i^{\boldsymbol{\beta}} = \left(\mathbf{A}_1^{-1} + 2\mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{A}_4^{-1}\mathbf{A}_3\mathbf{A}_1^{-1}\right)\boldsymbol{\varpi}_i - \mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{A}_4^{-1}\boldsymbol{\zeta}_i - 2\mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{A}_4^{-1}\boldsymbol{\varsigma}_i$, $\boldsymbol{\varpi}_i = E\frac{\delta_i(Y)\boldsymbol{X}}{\sigma^2(Z_2)} + \frac{\varepsilon_i \boldsymbol{X}_i}{\sigma(Z_{2i})}$, $\boldsymbol{\zeta}_i = \sigma^2(Z_{2i})(\varepsilon_i^2 - 1)\boldsymbol{X}_i$, and $\boldsymbol{\varsigma}_i = E\left\{\delta_i(Y)\sigma(Z_2)\varepsilon\boldsymbol{X}\right\}$.

Theorem 2. Under the conditions given in Theorem 1, we have the following asymptotic expansion of $\widehat{\gamma} - \gamma$:

$$\widehat{\gamma} - \gamma = \frac{1}{n}\sum_{i=1}^{n}\varphi_i^{\gamma} + o_p(n^{-1/2}), \tag{3.3}$$

where $\varphi_i^{\gamma} = -2\mathbf{A}_4^{-1}\mathbf{A}_3\mathbf{A}_1^{-1}\boldsymbol{\varpi}_i + \mathbf{A}_4^{-1}\boldsymbol{\zeta}_i + 2\mathbf{A}_4^{-1}\boldsymbol{\varsigma}_i$.

So, $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ and $\sqrt{n}\,(\widehat{\gamma} - \gamma)$ have an asymptotically normal distribution with mean vector 0 and covariance matrix $E[\varphi_i^{\boldsymbol{\beta}}(\varphi_i^{\boldsymbol{\beta}})']$ and $E[\varphi_i^{\gamma}(\varphi_i^{\gamma})']$, respectively, which can be estimated by replacing all theoretical quantities in $\frac{1}{n}\sum_{i=1}^{n}\varphi_i^{\boldsymbol{\beta}}(\varphi_i^{\boldsymbol{\beta}})'$ and $\frac{1}{n}\sum_{i=1}^{n}\varphi_i^{\gamma}(\varphi_i^{\gamma})'$ by their sample counterparts.

Substituting the results of Theorems 1 and 2 into (3.1) of Lemma A, we obtain the following theorem.

Theorem 3. Under the conditions given in Theorem 1, we have the following asymptotic expansion of $\widehat{H}$:

$$\widehat{H}(y) - H(y) \;=\; \frac{1}{n}\sum_{i=1}^{n} \varphi_i^H(y) + o_p(n^{-1/2}), \tag{3.4}$$

where $\varphi_i^H(y) = \delta_i(y) - \left\{\boldsymbol{\pi}_1'(y) + 2\boldsymbol{\pi}'(y)\mathbf{A}_4^{-1}\mathbf{A}_3\right\}\mathbf{A}_1^{-1}\boldsymbol{\varpi}_i + \boldsymbol{\pi}'(y)\mathbf{A}_4^{-1}\boldsymbol{\zeta}_i + 2\boldsymbol{\pi}'(y)\mathbf{A}_4^{-1}\boldsymbol{\varsigma}_i$.

So, by the central limit theorem, it is easy to show that $\sqrt{n}\left(\widehat{H}(y) - H(y)\right)$ has an asymptotically normal distribution with mean 0 and variance $E[\varphi_i^H(y)]^2$, which can be estimated by replacing all theoretical quantities in $\frac{1}{n}\sum_{i=1}^{n}\left[\varphi_i^H(y)\right]^2$ with their sample counterparts.

Hence, $\widehat{H}(y)$ converges to $H(y)$ at rate of $n^{-1/2}$; this result shows that we can estimate the nonparametric function $H(.)$ with a parametric convergent rate. The result also assures that we can estimate $\mu(\boldsymbol{x})$ at the rate of $n^{-1/2}$, which is presented in Theorem 4 below. Denote $\kappa(\boldsymbol{x}) = H^{-1}\left(\boldsymbol{x}'\boldsymbol{\beta} + \sigma(\boldsymbol{x}'\gamma)\varepsilon\right)$, $\kappa_i(\boldsymbol{x}) = H^{-1}\left(\boldsymbol{x}'\boldsymbol{\beta} + \sigma(\boldsymbol{x}'\gamma)\varepsilon_i\right)$, for $\boldsymbol{x}$ in the interior of the support of $\boldsymbol{X}$,

$$\mathbf{B}_1(\boldsymbol{x}) = E\left[\frac{1}{h\left(\kappa(\boldsymbol{x})\right)}\left\{x - \frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(Z_2)}\boldsymbol{X} - \frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(Z_2)}\boldsymbol{\pi}_1(Y) + \boldsymbol{\pi}_1(Y)\right\}\right],$$

$$\mathbf{B}_2(\boldsymbol{x}) = E\left[\frac{1}{h\left(\kappa(\boldsymbol{x})\right)}\left\{\left(\sigma^{(1)}(\boldsymbol{x}'\gamma)x - \frac{\sigma(\boldsymbol{x}'\gamma)\sigma^{(1)}(Z_2)}{\sigma(Z_2)}\boldsymbol{X}\right)\varepsilon - \left(\frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(Z_2)} - 1\right)\boldsymbol{\pi}_2(Y)\right\}\right],$$

and

$$B_3'(\boldsymbol{x}) = \mathbf{B}_2(\boldsymbol{x})'\mathbf{A}_4^{-1} - \mathbf{B}_1(\boldsymbol{x})'\mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{A}_4^{-1}.$$

Theorem 4. Under the conditions given in Theorem 1, we have the following asymptotic expansion of $\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})$:

$$\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\varphi_i^{\mu}(\boldsymbol{x}) + o_p(n^{-1/2}), \tag{3.5}$$

where $\varphi_i^{\mu}(\boldsymbol{x}) = \kappa_i(\boldsymbol{x}) + \left[\mathbf{B}_1(\boldsymbol{x})'\mathbf{A}_1^{-1} - 2B_3(\boldsymbol{x})'\mathbf{A}_3\mathbf{A}_1^{-1}\right]\boldsymbol{\varpi}_i + B_3'(\boldsymbol{x})\boldsymbol{\zeta}_i + 2B_3'(\boldsymbol{x})\boldsymbol{\varsigma}_i + \tau_i(\boldsymbol{x})$, $\tau_i(\boldsymbol{x}) = E\left[\frac{1}{h(\kappa(\boldsymbol{x}))}\left\{\frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(Z_2)}\delta_i(Y) - \delta_i\left(\kappa(\boldsymbol{x})\right)\right\}\right]$, $\varepsilon = \frac{H(Y) - \boldsymbol{X}'\boldsymbol{\beta}}{\sigma(\boldsymbol{X}'\gamma)}$, and $\varepsilon_i = \frac{H(Y_i) - \boldsymbol{X}_i'\boldsymbol{\beta}}{\sigma(\boldsymbol{X}_i'\gamma)}$. Throughout the paper we let E denote the expectation with respect to $(\boldsymbol{X}, Y, \varepsilon)$.

Hence $\sqrt{n}\left(\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})\right)$ has an asymptotically normal distribution with mean 0 and variance $E[\varphi_i^{\mu}(\boldsymbol{x})]^2$, which can be estimated by replacing all theoretical quantities in $\frac{1}{n}\sum_{i=1}^{n}\left[\varphi_i^{\mu}(\boldsymbol{x})\right]^2$ with their sample counterparts.

Since the leading terms in Theorems 1 to 4 do not depend on the bandwidths h_0, h_1 and h_2, we can conclude that the bandwidths h_0, h_1 and h_2 are not crucial for the asymptotic performance of the estimates; this conclusion has also has been confirmed in our simulation studies. A practical implication of this result is that our estimates are not sensitive to the bandwidths h_0, h_1 and h_2, which makes practical implementation of our method much easier. A rough selecting method for h_0, h_1 and h_2 is enough. Next, we give a practical method for selecting h_0, h_1 and h_2. Since $h(y) = -\frac{p(y,z_1,z_2)}{g_1(y,z_1,z_2)}$ is independent of z_1 and z_2 for any given y, $\frac{p(y,Z_{1i},Z_{2i})}{g_1(y,Z_{1i},Z_{2i})}$ is a constant for any $i = 1, \cdots, n$, which means that $Var[\frac{p(y,Z_{1i},Z_{2i})}{g_1(y,Z_{1i},Z_{2i})}] = 0$ for any given y. The data-based bandwidths for h_0, h_1 and h_2 are then chosen to minimize the following sample variance of $\frac{p(y,Z_{1i},Z_{2i})}{g_1(y,Z_{1i},Z_{2i})}$:

$$\frac{1}{(n-1)R}\sum_{r=1}^{R}\sum_{i=1}^{n}\left(\frac{\widehat{p}(y_r, Z_{1i}, Z_{2i})}{\widehat{g}_1(y_r, Z_{1i}, Z_{2i})} - \overline{h}(y_r)\right)^2,$$

where $\overline{h}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{p}(y,Z_{1i},Z_{2i})}{\widehat{g}_1(y,Z_{1i},Z_{2i})}$, and $y_1, \cdots, y_R$ are chosen to evaluate the variance of $\frac{p(y,Z_{1i},Z_{2i})}{g_1(y,Z_{1i},Z_{2i})}$.

Although we have derived estimators of the variances of $\widehat{\boldsymbol{\beta}}, \widehat{\gamma}, \widehat{H}(y)$ and $\widehat{\mu}(\boldsymbol{x})$, their computation involves the unknown function $q(z_1, z_2)$, its derivative $q^{(10)}(z_1, z_2)$, and the derivative $p^{(10)}(z_1, z_2)$. Hence, the performance of the resulting variance estimates in finite-sample sizes may be unstable because it may be difficult to get a good estimate for a derivative. Alternatively, we could use a resampling scheme, for example, a bootstrap method, to approximate the variances or covariance matrices.

4 Simulation studies

We conducted simulations studies to assess the finite-sample performance of the proposed method. Since the validity of our method does not rely on parametric specifications for the transformation functions, we expect our estimators of the untransformed scale expectation and regression parameters to be more robust than the ones derived under the assumed parametric transformation methods. In addition, since our method requires specifying the variance function, we also want to know how the misspecified variance function can affect our new estimators. To investigate these issues, we compared the performance of the proposed method with

the following models: (1) the CTCV model, where the transformation and variance functions were correctly specified, the case that served as the gold standard; (2) the CTMV model, where the transformation function was correctly specified and the variance function was misspecified; (3) the MTCV model, where the transformation function was misspecified and the variance function was correctly specified. The MTCV models were used to investigate the robustness of the proposed method. In addition, as suggested by referees, we also compared our approach to Chiou and Muller's non-parametric GLM approach and Basu and Rathouz's semi-parametric GLM approach that used data to determine appropriate link and variance functions. We are interested in assessing whether the proposed method has any advantage over these competing models both in terms of bias and efficiency.

From Theorems 1 to 4, described in Section 3, we saw that bandwidth selection was not a vital issue for estimating the parameters and non-parametric functions. That is, we could just select any bandwidths that satisfied the technical assumption of undersmoothing but were not too ridiculously small to get consistent estimates of the parameters and non-parametric functions. We would confirm this conclusion in our simulation studies.

We generated data from a non-logarithm transformation model with one binary covariate and one continuous covariate. For $n = 2000$ subjects, we generated covariates X_1 and X_2 from the binomial distribution with $p = 0.5$ and the uniform distribution on $[0, 2]$, respectively, while generating the random error ε from the standard normal distribution. Let us denote $X = (X_1, X_2)$. We generated our outcome by the following transformation model with heteroscedastic variance: $H(Y) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \sqrt{0.4 + \gamma X_1}\varepsilon$, where $H(y) = \Phi^{-1}\{\exp(y - 10)\}$, $\beta_0 = -1.8$, $\beta_1 = 1.4$, $\beta_2 = 1.4$, and $\gamma = -0.35$. The coefficients of skewness and kurtosis of Y were -2.2 and 8.5, respectively. We assessed the performance of the various estimators of the untransformed scale expectation in terms of standard deviation (SD), bias and square root of mean squared error (RMSE).

To assess the effect of the chosen bandwidths on the estimators, we first chose the values for h_1, h_2 and h_0, based on the data-based method, described in Section 3 for a single randomly selected data set. The chosen values were $0.37, 0.7$, and 0.1, respectively. We then chose

five additional values for each of h_1, h_2, and h_0 by independently drawing from the normal distributions with mean $0.37, 0.7$, and 0.1, respectively, with the common variance 0.2. The results showed that although the estimation of $\boldsymbol{\beta}$ was affected by the selection of h_1, h_2, and h_0, the bandwidth selection had little effect on estimation of the average cost on the original scale, which was our focus. The reason was that the extra amount of smoothing inherent in the summation in (2.14) should mean that $\widehat{\mu}(\boldsymbol{x})$, our estimate of $E(Y \mid X = x)$, would be even less sensitive to the bandwidth, the so-called double-smoothing phenomenon (Maity, Ma and Carroll, 2007).

Based on our sensitivity result of bandwidth selection on $\boldsymbol{\beta}$, we chose the following method for selecting the bandwidth for each generated data set. For a randomly generated data set with a given sample size, we would use the data-based method, described in Section 3, to select the values for the bandwidths, h_0, h_1, and h_2, denoted by $\widehat{h}_0$, $\widehat{h}_1$, and $\widehat{h}_2$, respectively. Unfortunately, the estimation algorithm failed to converge using the selected bandwidths for some of the 200 simulated data sets. If that happen, we just used the absolute value of a randomly generated number from the normal distribution with mean $\widehat{h}_0$ and variance 0.01 until the algorithm converged. For the 200 simulated data sets, the chosen h_0 values varied from 0.01 to 0.36. Similarly, we applied the same way for h_1 and h_2 to assure the convergency.

In Table 1, we reported bias and standard deviation (SD) of the various estimators for the untransformed scale expectation at the combination of $X_1 = 0, 1$ and $X_2 = 0, 1, 2$. In the MTCV model, the transformation function was misspecified as a function $H(y) = exp(y - 10)$. By comparing results among the parametric CTCV, CTMV, and MTCV estimates, we can conclude that misspecification of the transformation function can lead to large bias and large RMSE while misspecification of variance function has minimal effect on bias and RMSE. Among the two existing GLM-based estimators, we found that Basu and Rathouz's procedure failed to converge for all of 200 simulated samples, suggesting that the Basu and Rathouz's estimator is not stable. Chiou and Muller's estimator has much larger bias and SD than our newly proposed estimator. Our newly proposed estimator is essentially unbiased. Comparing our new estimator with the gold standard estimator, the CTCV estimator, derived under correctly

specified transformation and variance function, the empirical efficiency of our new estimator is around 60% on average.

For each simulated data set, we also obtained estimates of the transformation H using the proposed approach. Figure 1(a) displays the averaged estimated transformation function and their 95% empirical pointwise confidence limits, based on the 200 simulated data sets. Figure 1(a) shows that our proposed estimate of the transformation function is very close to the true transformation function.

Table 1 Simulation results for the average cost over the 200 replications

$\boldsymbol{x}$	$\mu(\boldsymbol{x})$	Method	Bias	SD	$\boldsymbol{x}$	$\mu(\boldsymbol{x})$	Bias	SD
(0,0)	6.502	Prop.	0.1362	0.217	(0,1)	8.795	-0.0062	0.042
		CTCV	0.0015	0.088			-0.0030	0.028
		CTMV	0.0022	0.087			-0.0015	0.025
		MTCV	2.8103	0.018			0.7702	0.023
		CHIOU	0.8027	0.226			-0.3023	0.063
		BASU	Failed to converge					
(0,2)	9.753	Prop.	-0.0254	0.029	(1,0)	8.917	0.0384	0.033
		CTCV	-0.0023	0.015			0.0015	0.036
		CTMV	-0.0013	0.015			0.0019	0.035
		MTCV	-0.0136	0.026			0.6480	0.005
		CHIOU	-0.4317	0.173			0.1796	0.189
(1,1)	9.818	Prop.	-0.0057	0.009	(1,2)	9.990	0.0043	0.004
		CTCV	-0.0001	0.002			-0.0001	0.001
		CTMV	-0.0001	0.002			-0.0001	0.001
		MTCV	-0.0724	0.005			-0.1297	0.009
		CHIOU	-0.0637	0.023			0.2003	0.051

5 An Example

The sample used here was from a study on the effectiveness of the Improving Mood-Promoting Access to Collaborative Treatment (IMPACT) collaborative care management program for

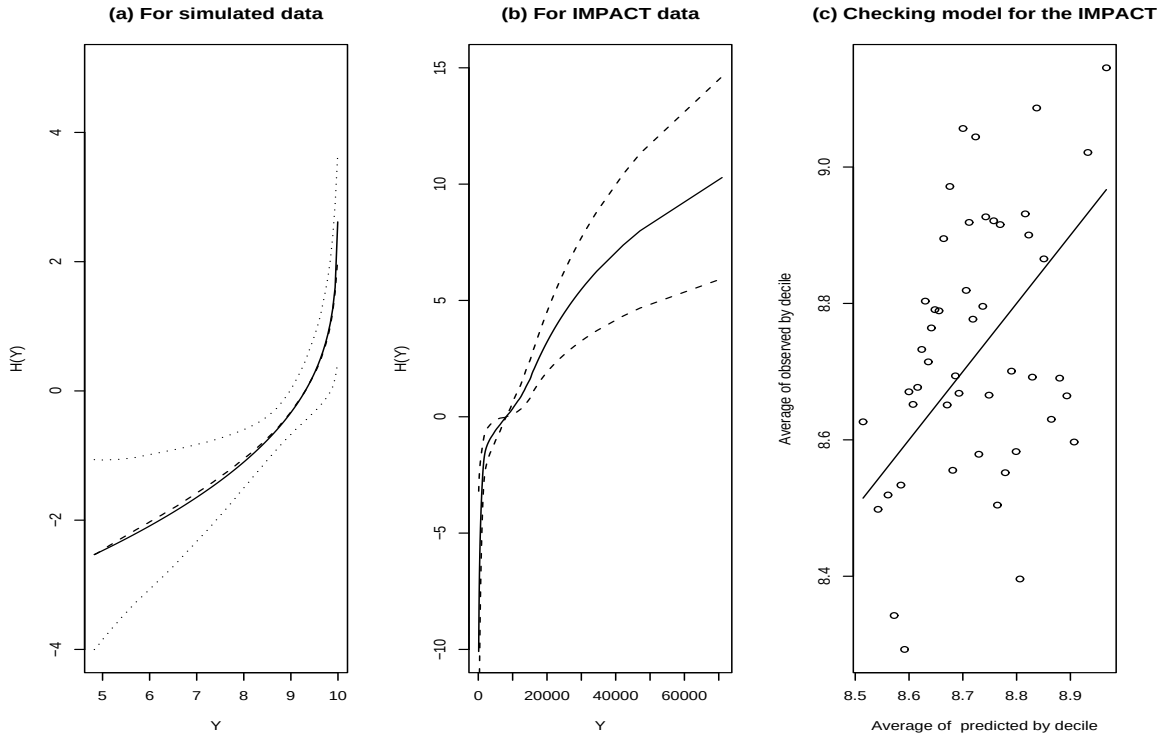

Figure 1: (a) The averaged estimates of transformation curve (Solid — true functions; dashed—estimated; dotted-linear—confidence limit). (b) The estimated transformation and its 95% confidence limits for IMPACT data. (c) Prediction against actual for logarithm of cost with bandwidth $h_1 = 4, h_2 = 10, h_y = 90$, the solid line is diagonal.

late-life depression (Unutzer et al., 2002). A total of 1801 patients aged 60 years or older with major depression (17%), dysthymic disorder (30%), or both (53%) were randomly assigned to the IMPACT intervention ($n = 906$) or usual care ($n = 895$). Intervention patients had up to 12 months access to a depression care manager who offered education, care management, and support of antidepressant management by the patient's primary care physician. Primary outcomes were collected at baseline, 6, 12, 18 and 24 months. In the paper, we focus on the cost in the first year (Y), the mean and standard deviations of Y were \$6258.442 and \$5065.507, respectively, and the coefficients of skewness and kurtosis of Y are 3.36 and 26.94, respectively. We fit the data using the model (1.1) with the outcome variable being outpatient costs in the first year, and the two independent variable, X_1 and X_2. Here X_1 was the binary treatment indicator, and X_2 was the mean score of the 20 depression items from the Symptom Checklist. To reduce the computational time, we applied a log transformation to the outcome variable

before performing our analysis.

We set the variance function to be a polynomial function $\sigma^2(x;p) = \sum_{k=0}^{p} \alpha_k x^k$, $p = 1, 2, \cdots$, where p was chosen to minimize the following $GF(p)$:

$$GF(p) = \min_{\gamma} \sum_{i=1}^{n} \left\{ \left(\tilde{H}(Y_i) - \boldsymbol{X}_i' \tilde{\boldsymbol{\beta}} \right)^2 - \sigma^2(\boldsymbol{X}_i' \gamma; p) \right\}^2,$$

where $\widetilde{H}$ and $\tilde{\boldsymbol{\beta}}$ were the initial values of H and $\boldsymbol{\beta}$, respectively, obtained using the method in Section 3.2. The results showed that $GF(p)$ did not substantially change with p. Hence to assure $\sigma^2(x) \geq 0$, here we selected $\sigma^2(x;p) = x^2$. We then applied our newly proposed method to the data set. For a comparison purpose, we also applied Chiou and Muller' method and Basu and Rathouz's methods to the data set. In Table 2, we reported the three estimates of the average cost of a patient with the given covariate values, X_1 and X_2, where $X_1 = 0, 1$ and $X_2 = 0.04, 1.5, 3.2$, and their corresponding standard errors. Here we used a bootstrap method to estimate the standard errors of the proposed estimator and Chiou and Muller's estimator. Table 2 suggested that the three methods could give different estimates of the average cost of a patient with the given X_1 and X_2. However, all three methods reached the same conclusion that patients in the treatment group might incur higher costs than those in the control group. Figure 1(b) displayed the estimated transformation function and their 95% bootstrap pointwise confidence limits.

Finally, we proposed a procedure to check validity of the assumed heteroscedastic nonparametric transformation model (1.1). First, we randomly divided the data into two subsets with equal sizes, called the training set and validation set. Then, we fit a model using the training set. For each subject in the validation set, we predicted the subject's cost using the fitted model from the training set. We investigated the performance of the model by examining how well observed mean costs agreed with averages of predicted costs in each of the ten groups formed by decile of predicted costs. To better visualize this result, we plotted the mean of logarithms of predicted costs by decile of logarithms of predicted costs against the logarithms of actual mean spending for the individuals in that decile. We reported this plot in Figure 1(c). This provides a graphical depiction of the degree to which these models can estimate actual costs across the span of the data (Buntin and Zaslavsky, 2004). A perfect fit corresponds to the solid

diagonal line. Points above or below this diagonal line indicate over-predict or under-predict of the model. Figure 1(c) suggested that the goodness-of-fit of the proposed model (1.1) was reasonable.

Table 2 The average cost estimates of a patient in the IMPACT study

		Proposed	CHIOU	BASU
Rand	SCL	Average cost(se)	Average cost(se)	Average cost(se)
1	0.04	5008.6(444.4)	5239.9(74.7)	5334.8(213.0)
0	0.04	4424.0(500.6)	4639.4(10.5)	4991.9(390.1)
1	1.50	6916.1(392.9)	6779.1(26.7)	6717.8 (57.6)
0	1.50	6172.1(216.4)	6177.7(40.7)	6167.4(2.2)
1	3.20	9177.3(1156.9)	8574.4(31.9)	9802.8(3921.1)
0	3.20	8349.5(816.3)	7971.8(97.5)	8622.3(1668.5)

6 Discussion

In this paper, we have extended the traditional parametric transformation model with a known transformation function to a nonparametric heteroscedastic transformation model with unknown transformation function and error distribution. The theoretical studies show that our estimators are asymptotically normal with convergent rate $n^{1/2}$, which is the rate for a fully parametric regression model. In our simulation comparisons with two existing generalized linear model (GLM) based methods, Basu and Rathouz's semi-parametric GLM method and Chiou and Muller's non-parametric GLM method, we have found that Basu and Rathouz's procedure is unstable, failing to converge in most of our simulated data sets. Between our newly proposed method and Chiou and Muller's procedure, our simulation results show that our new method greatly outperforms Chiou and Muller's procedure for estimating the expected value of a skewed outcome, although Chiou-Muller's procedure is not originally designed for such situations.

Modeling heteroscedasticity is a complicated matter. If we fit a non-parametric function to the variance (e.g. allowing an unknown function σ), it may be difficult to get an good estimate of the variance. In this paper, we fit the heteroscedasticity through a known function σ, which

depends on unknown parameters γ. Another possibility to model the heteroscedasticity is setting $\gamma = \boldsymbol{\beta}$, which assumes that the variance depends only on the mean, as in the GLM literature, and leave the variance function unknown.

There are some potential limitations to the proposed method. First, our method can not be easily extended to the case $\gamma = \boldsymbol{\beta}$. Second, the validity of the proposed model requires the assumption that ϵ is independent of X and $E(\epsilon) = 0$. In contrast, the GLM and Basu's method only requires mean independence of ϵ with covariates X , $E(\epsilon \mid X = x) = 0$. It is obvious that the effect of our independence assumption is stronger from a mathematical point of view, however, it is unclear how much practical difference it makes. For example, in the content of causal inferences with instrument variables, Imbens and Rubin (1997) argued against the mean independence assumption and favored the full independence assumption. Finally, the heteroscedastic term in our model (1.1) only attempts to model differences in the second order moments of $H(Y)$, and does not address variations of even higher order moments of $H(Y)$ with respect to X.

Appendix A. Conditions

1. Functions K_0, K_1, and K_2 are one-dimensional bounded and symmetric density functions around zero with compact supports, and without loss of generality, we assume that their supports are $[-1, 1]$. We assume that K_0, K_1 and K_2 have orders of r_0, r_1 and r_2, respectively; that is, $\int_{-1}^{1} \mu^j K_p(\mu) d\mu$ is 1 when $j = 0$, is 0 when $1 \leq j \leq r_p - 1$, and is not zero when $j = r_p$, for $p = 0, 1, 2$. We also assume that K_0 has bounded variation and that K_1 and K_2 are everywhere twice differentiable. The derivatives are bounded and have bounded variation. We further assume that the second derivative of K_j satisfies $|K_j^{(2)}(x_1) - K_j^{(2)}(x_2)| \leq M|x_1 - x_2|$ for some $M < \infty$ and $j = 1, 2$.

2. The interval $[y_0, y_1]$ is the domain of H. In practice, this would be the range of the observed Y's. The function H is strictly increasing, and the derivatives $H^{(k)}(y)(k = 1, \cdots, r_0 + 1)$ exist and are uniformly bounded over $y \in [y_0, y_1]$.

3. There exists a sequence $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_n$ and $\widehat{\gamma} = \widehat{\gamma}_n$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = o_p(1)$ and $\|\widehat{\gamma} - \gamma\| = o_p(1)$.

4. The derivatives $p^{(k_1,k_2)}(z_1,z_2)$ and $p^{(k_0,k_1,k_2)}(y,z_1,z_2)$, for $k_0 = 1,\cdots,r_0+1, k_1 = 1,\cdots,r_1+1, k_2 = 1,\cdots,r_2+1$, exist and are uniformly bounded over $y \in [y_0,y_1]$ and (z_1,z_2) in the support of (Z_1,Z_2).

5. As $n \to \infty$, we have $\sqrt{n}h_1^k h_2^{4-k} \to \infty$ when $2 \le k \le 4$, $\frac{\log n}{\sqrt{nh_1^k h_2^{7-k}h_0}} \to 0$ when $k = 4, 6$, and $nh_1^{2r_1} \to 0$, $nh_2^{2r_2} \to 0$, $nh_0^{2r_0} \to 0$.

6. X is bounded.

The assumption on h_0, h_1, and h_2 can be satisfied, for example, if K_0 is a second-order kernel, K_1 and K_2 are sixth-order kernel functions with $h_0 \propto n^{-1/3}$ and $h_1 = h_2 \propto n^{-1/11}$. The second-order and sixth-order kernel functions can be taken from Muller (1984). Since $g_{1n}(y|z_1,z_2)$ is a function of derivatives of K_1, and derivative functional converge relatively slowly, the higher-order kernel for K_1 is needed to insure sufficiently rapid convergence. Note that our model is a special case of the single-index model, and the existence of the consistency estimators for the index parameters has been proved by Yin and Cook (2005); hence the condition 3 can be satisfied.

Our proof on the asymptotic expansion for all the estimators relies on three steps. The first step consists of an expansion of $\widehat{H}$ at an argument y, which is included in the proof of Lemma A. The second step consists of the expansions of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$, which is included in the proof of Theorem 1 and 2. Finally, based on the expansions of $\widehat{\boldsymbol{\beta}},\widehat{\gamma}$ and $\widehat{H}$, we obtain the asymptotic expression of $\widehat{\mu}(\boldsymbol{x})$. Because the argument used to prove Lemma 3 is essentially the same as that in Lemma 2, Lemma 3 is stated without a proof.

Appendix B. Proof of Lemma A.

To prove Lemma A, we first need to prove the following four lemmas.

Lemma 1: Under the conditions on h_0, h_1 and h_2 given in Appendix A, we have

$$\frac{(-1)^k}{h_0 h_1^{k+1} h_2} ES(Y,Z_1,Z_2)K_0(\frac{Y-y}{h_0})K_1^{(k)}(\frac{Z_1-z_1}{h_1})K_2(\frac{Z_2-z_2}{h_2})$$
$$= \{S(y,z_1,z_2)p(y,z_1,z_2)\}^{(0k0)} + O_p(h_0^{r_0} + h_1^{r_1} + h_2^{r_2}),$$

where $k = 0, 1$, the derivatives $S^{(r_0+1,r_1+1,r_2+1)}(y, z_1, z_2)$ exist and are uniformly bounded over the support of Y, Z_1 and Z_2.

Proof of Lemma 1. See Horowitz (1996).

Lemma 2: Define

$$\begin{aligned}
\widehat{p}(y, z_1, z_2) &= \frac{1}{nh_1h_2h_0}\sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0})K_1\left(\frac{\boldsymbol{X}_i'\widehat{\boldsymbol{\beta}} - z_1}{h_1}\right)K_2\left(\frac{\boldsymbol{X}_i'\widehat{\gamma} - z_2}{h_2}\right), \\
p_n(y, z_1, z_2) &= \frac{1}{nh_1h_2h_0}\sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0})K_1\left(\frac{\boldsymbol{X}_i'\boldsymbol{\beta} - z_1}{h_1}\right)K_2\left(\frac{\boldsymbol{X}_i'\gamma - z_2}{h_2}\right), \\
&\text{and} \quad \boldsymbol{\Gamma}(y, z_1, z_2, \boldsymbol{x}) = -p(y, z_1, z_2)\left[q(z_1, z_2) - \boldsymbol{x}\right],
\end{aligned}$$

where $q(z_1, z_2) = E[\boldsymbol{X}|Z_1 = z_1, Z_2 = z_2]$ and $p(y, z_1, z_2)$ is the joint density function of (Y, Z_1, Z_2). As $n \to \infty$, we have

$$\begin{aligned}
\widehat{p}(y, \boldsymbol{x}'\widehat{\boldsymbol{\beta}}, \boldsymbol{x}'\widehat{\gamma}) &= p_n(y, \boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{x}'\gamma) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{\Gamma}^{(0100)}(y, \boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{x}'\gamma, \boldsymbol{x}) \\
&\quad + (\widehat{\gamma} - \gamma)'\boldsymbol{\Gamma}^{(0010)}(y, \boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{x}'\gamma, \boldsymbol{x}) + o_p(\widehat{\gamma} - \gamma) + o_p(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).
\end{aligned}$$

Proof of Lemma 2. By the Taylor series expansion and the conditions on $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$, we obtain that

$$\begin{aligned}
&\widehat{p}(y, \boldsymbol{x}'\widehat{\boldsymbol{\beta}}, \boldsymbol{x}'\widehat{\gamma}) \\
&= \frac{1}{nh_1h_2h_0}\sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0})K_1\left(\frac{\boldsymbol{X}_i'\boldsymbol{\beta} - \boldsymbol{x}'\boldsymbol{\beta}}{h_1}\right)K_2\left(\frac{\boldsymbol{X}_i'\gamma - \boldsymbol{x}'\gamma}{h_2}\right) \\
&\quad + \frac{1}{nh_1^2h_2h_0}\sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0})K_1^{(1)}\left(\frac{(\boldsymbol{X}_i - \boldsymbol{x})'\boldsymbol{\beta}}{h_1}\right)K_2\left(\frac{(\boldsymbol{X}_i - \boldsymbol{x})'\gamma}{h_2}\right)(\boldsymbol{X}_i - \boldsymbol{x})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\quad + \frac{1}{nh_1h_2^2h_0}\sum_{i=1}^{n} K_0(\frac{Y_i - y}{h_0})K_1\left(\frac{(\boldsymbol{X}_i - \boldsymbol{x})'\boldsymbol{\beta}}{h_1}\right)K_2^{(1)}\left(\frac{(\boldsymbol{X}_i - \boldsymbol{x})'\gamma}{h_2}\right)(\boldsymbol{X}_i - \boldsymbol{x})'(\widehat{\gamma} - \gamma) \\
&\quad + o_p(\widehat{\gamma} - \gamma) + o_p(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= p_n(y, \boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{x}'\boldsymbol{\beta}) + p_{n1}(y, \boldsymbol{x})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + p_{n2}(y, \boldsymbol{x})'(\widehat{\gamma} - \gamma) + o_p(\widehat{\gamma} - \gamma) + o_p(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \qquad (6.1)
\end{aligned}$$

By Theorem 2.37 of Pollard (1984), we obtain $sup_{y,\boldsymbol{x}}|p_{n1}(y, \boldsymbol{x}) - Ep_{n1}(y, \boldsymbol{x})| = o[(\log n)/(nh_1^3h_2h_0)^{1/2}] \to 0$; by Lemma 1, we obtain that $Ep_{n1}(y, \boldsymbol{x}) - \boldsymbol{\Gamma}^{(0100)}(y, \boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{x}'\gamma, \boldsymbol{x}) = O(h_0^{r_0} + h_1^{r_1} + h_2^{r_2})$. Hence, it follows from the conditions on h_0, h_1 and h_2 that

$$p_{n1}(y, \boldsymbol{x}) - \boldsymbol{\Gamma}^{(0100)}(y, \boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{x}'\gamma, \boldsymbol{x}) = o_p(1). \qquad (6.2)$$

Similarly, we can obtain that

$$p_{n2}(y,\boldsymbol{x}) - \boldsymbol{\Gamma}^{(0010)}(y,\boldsymbol{x}'\boldsymbol{\beta},\boldsymbol{x}'\gamma,\boldsymbol{x}) \;=\; o_p(1). \tag{6.3}$$

The lemma 2 follows from (6.1), (6.2) and (6.3).

Lemma 3. Let $\Delta(z_1,z_2,\boldsymbol{x}) = -p(z_1,z_2)[q(z_1,z_2)-\boldsymbol{x}], \Lambda(y,z_1,z_2,\boldsymbol{x}) = -G(y|z_1,z_2)\Delta(z_1,z_2,\boldsymbol{x})$,

$$\begin{aligned}
\widehat{p}(z_1,z_2) &= \frac{1}{nh_1h_2}\sum_{i=1}^{n} K_1\left(\frac{\boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}-z_1}{h_1}\right)K_2\left(\frac{\boldsymbol{X}_i'\widehat{\gamma}-z_2}{h_2}\right),\\
p_n(z_1,z_2) &= \frac{1}{nh_1h_2}\sum_{i=1}^{n} K_1\left(\frac{\boldsymbol{X}_i'\boldsymbol{\beta}-z_1}{h_1}\right)K_2\left(\frac{\boldsymbol{X}_i'\gamma-z_2}{h_2}\right),\\
\widehat{G}(y,z_1,z_2) &= \frac{1}{nh_1h_2}\sum_{i=1}^{n} I(Y_i\le y)K_1\left(\frac{\boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}-z_1}{h_1}\right)K_2\left(\frac{\boldsymbol{X}_i'\widehat{\gamma}-z_2}{h_2}\right),\\
G_n(y,z_1,z_2) &= \frac{1}{nh_1h_2}\sum_{i=1}^{n} I(Y_i\le y)K_1\left(\frac{\boldsymbol{X}_i'\boldsymbol{\beta}-z_1}{h_1}\right)K_2\left(\frac{\boldsymbol{X}_i'\gamma-z_2}{h_2}\right).
\end{aligned}$$

As $n\to\infty$, we have the following asymptotic expansions:

$$\begin{aligned}
\widehat{p}^{(k0)}(\boldsymbol{x}'\widehat{\boldsymbol{\beta}},\boldsymbol{x}'\widehat{\gamma}) &= p_n^{(k0)}(\boldsymbol{x}'\boldsymbol{\beta},\boldsymbol{x}'\gamma) + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\Delta^{(k+1,00)}(\boldsymbol{x}'\boldsymbol{\beta},\boldsymbol{x}'\gamma,\boldsymbol{x})\\
&\quad +(\widehat{\gamma}-\gamma)'\Delta^{(k10)}(\boldsymbol{x}'\boldsymbol{\beta},\boldsymbol{x}'\gamma,\boldsymbol{x}) + o_p(\widehat{\gamma}-\gamma) + o_p(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}),\\
\widehat{G}^{(0k0)}(y,\boldsymbol{x}'\widehat{\boldsymbol{\beta}},\boldsymbol{x}'\widehat{\gamma}) &= G_n^{(0k0)}(y,\boldsymbol{x}'\boldsymbol{\beta},\boldsymbol{x}'\gamma) - (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\Lambda^{(0,k+1,00)}(y,\boldsymbol{x}'\boldsymbol{\beta},\boldsymbol{x}'\gamma,\boldsymbol{x})\\
&\quad \text{and} \qquad -(\widehat{\gamma}-\gamma)'\Lambda^{(0k10)}(y,\boldsymbol{x}'\boldsymbol{\beta},\boldsymbol{x}'\gamma,\boldsymbol{x}) + o_p(\widehat{\gamma}-\gamma) + o_p(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}),
\end{aligned}$$

where $k=0,1$.

Lemma 4: Let $\varrho(y) = Eg_1(y,Z_1,Z_2), \quad G(y,z_1,z_2) = G(y|z_1,z_2)p(z_1,z_2)$,

$$\widehat{g}_1(y,z_1,z_2) = \frac{\partial\widehat{G}(y|z_1,z_2)}{\partial z_1}\widehat{p}(z_1,z_2), \quad \Sigma_1(y) = -h(y)\frac{E\left[g_1(y,Z_1,Z_2)q^{(10)}(Z_1,Z_2)\right]}{\varrho(y)},$$

$$\Sigma_2(y) = -h(y)\frac{E\left[(H(y)-Z_1)g_1(y,Z_1,Z_2)q^{(10)}(Z_1,Z_2)\sigma^{(1)}(Z_2)/\sigma(Z_2)\right]}{\varrho(y)},$$

$$\begin{aligned}
\Psi_{n1}(y) &= \frac{1}{n^2h_1h_2\varrho(y)}\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{h_0^{-1}K_0(\frac{Y_i-y}{h_0})\right.\\
&\quad \left.-\frac{h(y)p^{(10)}(Z_{1j},Z_{2j})}{p(Z_{1j},Z_{2j})}\left(I(Y_i\le y)-G(y|Z_{1j},Z_{2j})\right)\right\}K_1\left(\frac{Z_{1i}-Z_{1j}}{h_1}\right)K_2\left(\frac{Z_{2i}-Z_{2j}}{h_2}\right),
\end{aligned}$$

and $\Psi_{n2}(y) = -\dfrac{h(y)}{n^2 h_1^2 h_2 \varrho(y)} \sum_{j=1}^{n}\sum_{i=1}^{n} \left(I(Y_i \le y) - G(y|Z_{1j}, Z_{2j})\right) K_1^{(1)}\left(\dfrac{Z_{1i}-Z_{1j}}{h_1}\right) K_2\left(\dfrac{Z_{2i}-Z_{2j}}{h_2}\right)$.

Then we have the following asymptotic expansion:

$$\begin{aligned}\frac{\sum_{j=1}^n \widehat{p}(y,\widehat{Z}_{1j},\widehat{Z}_{2j})}{\sum_{j=1}^n \widehat{g}_1(y,\widehat{Z}_{1j},\widehat{Z}_{2j})} - \frac{\sum_{j=1}^n p(y,Z_{1j},Z_{2j})}{\sum_{j=1}^n g_1(y,Z_{1j},Z_{2j})} &= \Psi_{n1}(y) + \Psi_{n2}(y) + \Sigma_1(y)'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad + \Sigma_2(y)'(\widehat{\gamma} - \gamma) + o_p(\widehat{\gamma} - \gamma) + o_p(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}),\end{aligned}$$

where $\widehat{Z}_{1j} = \boldsymbol{X}'_j\widehat{\boldsymbol{\beta}}$, $\widehat{Z}_{2j} = \boldsymbol{X}'_j\widehat{\gamma}$.

Proof of Lemma 4. Since $\widehat{g}_1(y,z_1,z_2) = \widehat{G}^{(010)}(y,z_1,z_2) - \frac{\widehat{G}(y,z_1,z_2)\widehat{p}^{(10)}(z_1,z_2)}{\widehat{p}(z_1,z_2)}$, $g_1(y,z_1,z_2) = G^{(010)}(y,z_1,z_2) - \frac{G(y,z_1,z_2)p^{(10)}(z_1,z_2)}{p(z_1,z_2)}$, and $p(y,z_1,z_2) = -h(y)g_1(y,z_1,z_2)$, we obtain the following asymptotic expansion:

$$\begin{aligned}&\frac{\sum_{j=1}^n \widehat{p}(y,\widehat{Z}_{1j},\widehat{Z}_{2j})}{\sum_{j=1}^n \widehat{g}_1(y,\widehat{Z}_{1j},\widehat{Z}_{2j})} - \frac{\sum_{j=1}^n p(y,Z_{1j},Z_{2j})}{\sum_{j=1}^n g_1(y,Z_{1j},Z_{2j})} \approx \frac{1}{n\varrho(y)}\sum_{j=1}^n \left\{\widehat{p}(y,\widehat{Z}_{1j},\widehat{Z}_{2j}) - p(y,Z_{1j},Z_{2j})\right\} \\ &+\frac{h(y)}{n\varrho(y)}\sum_{j=1}^n \Big\{\widehat{G}^{(010)}(y,\widehat{Z}_{1j},\widehat{Z}_{2j}) - G^{(010)}(y,Z_{1j},Z_{2j}) \\ &+\frac{G(y,Z_{1j},Z_{2j})p^{(10)}(Z_{1j},Z_{2j})}{\left(p(Z_{1j},Z_{2j})\right)^2}\left(\widehat{p}(\widehat{Z}_{1j},\widehat{Z}_{2j}) - p(Z_{1j},Z_{2j})\right) \\ &-\frac{G(y,Z_{1j},Z_{2j})}{p(Z_{1j},Z_{2j})}(\widehat{p}^{(10)}(\widehat{Z}_{1j},\widehat{Z}_{2j}) - p^{(10)}(Z_{1j},Z_{2j})) - \frac{p^{(10)}(Z_{1j},Z_{2j})}{p(Z_{1j},Z_{2j})}(\widehat{G}(y,\widehat{Z}_{1j},\widehat{Z}_{2j}) - G(y,Z_{1j},Z_{2j}))\Big\}\end{aligned}$$

By combining Lemmas 2 and 3 with the conditions on h_0, h_1, h_2 and after some tedious computations, we obtain Lemma 4.

Proof of Lemma A. Note that $h(u) = -p(u,z_1,z_2)/g_1(u,z_1,z_2)$. Hence by the result in Lemma 4 and (2.8) in the paper, we have

$$\begin{aligned}\widehat{H}(y) - H(y) &= -\int_{y_0}^y \Psi_{n1}(u)du - \int_{y_0}^y \Psi_{n2}(u)du - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\int_{y_0}^y \Sigma_1(u)dy \\ &\quad -(\widehat{\gamma} - \gamma)'\int_{y_0}^y \Sigma_2(u)du + o_p(\widehat{\gamma} - \gamma) + o_p(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \qquad (6.4)\end{aligned}$$

Interchanging the summations, following the same line for (A.15) in Carroll et al., 1997) and

using Lemma 1 , we obtain the following asymptotic expansion:

$$\Psi_{n1}(y)+\Psi_{n2}(y)\approx\frac{1}{n\varrho(y)}\sum_{i=1}^{n}p(Z_{1i},Z_{2i})h_0^{-1}K_0(\frac{Y_i-y}{h_0})+\frac{h(y)}{n\varrho(y)}\sum_{i=1}^{n}g_1(y,Z_{1i},Z_{2i})$$
$$-\frac{2h(y)}{n\varrho(y)}\sum_{i=1}^{n}p^{(10)}(Z_{1i},Z_{2i})\left(I(Y_i\le y)-G(y|Z_{1i},Z_{2i})\right) \tag{6.5}$$

uniformly over $y\in[y_0,y_1]$. Define $\Upsilon_n(y)=\frac{1}{n}\sum_{i=1}^{n}\int_{y_0}^{y}\frac{1}{h_0\varrho(u)}K_0(\frac{Y_i-u}{h_0})p(Z_{1i},Z_{2i})du$ and $\vartheta_n(y)=\frac{1}{n}\sum_{i=1}^{n}I(y_0\le Y_i\le y)p(Z_{1i},Z_{2i})/\varrho(Y_i)$. It can be shown that

$$E_{Y|Z_1,Z_2}\left[\Upsilon_n(y)-\vartheta_n(y)\right]=O(h_0^s). \tag{6.6}$$

Since $E_{Y_i|Z_{1i},Z_{2i}}\left[\int_{y_0}^{y}\frac{1}{h_0\varrho(u)}K_0(\frac{Y_i-u}{h_0})p(Z_{1i},Z_{2i})du-I(y_0\le Y_i\le y)p(Z_{1i},Z_{2i})/\varrho(Y_i)\right]^2=O(h_0)$, by Theorem 2.37 of Pollard (1984), we have

$$\sup_{y\in[y_0,y_1]}\|\Upsilon_n(y)-\vartheta_n(y)-E[\Upsilon_n(y)-\vartheta_n(y)]\|=o(h_0^{1/2}(\log n)/n^{1/2}) \tag{6.7}$$

almost surely. Combining (6.6) and (6.7) with the condition on h_0, we obtain that $\Upsilon_n(y)-\vartheta_n(y)=o_p(n^{-1/2})$ uniformly over $y\in[y_0,y_1]$. Therefore, $\Upsilon_n(y)$ can be replaced by $\vartheta_n(y)$. Then, by (6.4) and (6.5), Lemma A follows.

Appendix C. Proof of Theorems 1 and 2

First, we consider the asymptotic expression form of $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}$. Using the expression of $\widehat{\boldsymbol{\beta}}$, given by (2.11) and the assumed model, given by (1.1), we can obtain the following expression:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\boldsymbol{X}_i'}{\sigma^2(\boldsymbol{X}_i'\widehat{\gamma})}\widehat{\boldsymbol{\beta}}-\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\boldsymbol{X}_i'}{\sigma^2(\boldsymbol{X}_i'\gamma)}\boldsymbol{\beta}$$
$$=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\widehat{H}(Y_i)}{\sigma^2(\boldsymbol{X}_i'\widehat{\gamma})}-\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_iH(Y_i)}{\sigma^2(\boldsymbol{X}_i'\gamma)}\right)+\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\varepsilon_i}{\sigma(\boldsymbol{X}_i'\gamma)}. \tag{6.8}$$

Using Taylor's expansion, and the results, $\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\boldsymbol{X}_i'(\sigma^2)^{(1)}(\boldsymbol{X}_i'\gamma)\varepsilon_i}{\sigma^3(\boldsymbol{X}_i'\gamma)}=O_p(n^{-1/2})$ and $\widehat{\gamma}-\gamma=o_p(1)$, we get the following asymptotic expansion:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\boldsymbol{X}_i'}{\sigma^2(\boldsymbol{X}_i'\gamma)}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i}{\sigma^2(\boldsymbol{X}_i'\gamma)}\left(\widehat{H}(Y_i)-H(Y_i)\right)+\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\varepsilon_i}{\sigma(\boldsymbol{X}_i'\gamma)}$$
$$+o_p(n^{-1/2})+o_p(\widehat{\gamma}-\gamma)+o_p(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}). \tag{6.9}$$

Substituting (3.1) into (6.9) and interchanging the summations, we get:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{A}_1^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\varpi}_i - \mathbf{A}_1^{-1} \mathbf{A}_2 (\widehat{\gamma} - \gamma) + o(n^{-1/2}) + o_p(\widehat{\gamma} - \gamma) + o_p(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{6.10}$$

where $\mathbf{A}_1 = E \frac{\boldsymbol{X}(\boldsymbol{X} + \boldsymbol{\pi}_1(Y))'}{\sigma^2(Z_2)}$, $\mathbf{A}_2 = E\left[\frac{\boldsymbol{X}\boldsymbol{\pi}_2(Y)'}{\sigma^2(Z_2)}\right]$, and $\boldsymbol{\varpi}_i$ is defined in Section 2.

Now we consider the asymptotic expression form of $\widehat{\gamma}$. Since the estimate of γ, $\widehat{\gamma}$, solves the following equation:

$$0 = \frac{1}{n} \sum_{i=1}^{n} \left\{ (\widehat{H}(Y_i) - \boldsymbol{X}_i' \widehat{\boldsymbol{\beta}})^2 - \sigma^2(\boldsymbol{X}_i' \widehat{\gamma}) \right\} \boldsymbol{X}_i,$$

using the Taylor's expansion, we can re-write the above expression as follows:

$$\begin{aligned} 0 = & \frac{1}{n} \sum_{i=1}^{n} \left\{ (H(Y_i) - \boldsymbol{X}_i' \boldsymbol{\beta})^2 - \sigma^2(\boldsymbol{X}_i' \gamma) \right\} \boldsymbol{X}_i + \frac{2}{n} \sum_{i=1}^{n} (H(Y_i) - \boldsymbol{X}_i' \boldsymbol{\beta})(\widehat{H}(Y_i) - H(Y_i)) \boldsymbol{X}_i \\ & - \frac{2}{n} \sum_{i=1}^{n} (H(Y_i) - \boldsymbol{X}_i' \boldsymbol{\beta}) \boldsymbol{X}_i \boldsymbol{X}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) - \frac{1}{n} \sum_{i=1}^{n} \left(\sigma^2\right)^{(1)} (\boldsymbol{X}_i' \gamma) \boldsymbol{X}_i \boldsymbol{X}_i' (\widehat{\gamma} - \gamma) + o(n^{-1/2}) \\ \equiv & D_{n1} + D_{n2} + D_{n3} + D_{n4} + o(n^{-1/2}). \end{aligned} \tag{6.11}$$

By the fact $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$ and $\frac{1}{n} \sum_{i=1}^{n} (H(Y_i) - \boldsymbol{X}_i' \boldsymbol{\beta}) \boldsymbol{X}_i \boldsymbol{X}_i' = O_p(n^{-1/2})$, we have the following result: $D_{n3} = o(n^{-1/2})$. Substituting (3.1) into D_{n2}, interchanging the summations and using the large number theorem, we have the result:

$$D_{n2} = \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\varsigma}_i - 2\mathbf{A}_3 (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - 2E\left[\sigma(Z_2) \varepsilon \boldsymbol{X} \boldsymbol{\pi}_2'(Y)\right] (\widehat{\gamma} - \gamma) + o(n^{-1/2}),$$

where $\boldsymbol{\varsigma}_i$, $\mathbf{A}_3$ and $\boldsymbol{\pi}_2(y)$ are defined in Section 2. Substituting the results on D_{n2} and D_{n3} into (6.11), we obtain the following asymptotic expansion:

$$\begin{aligned} & E\left[\left(\sigma^2\right)^{(1)} (Z_2) \boldsymbol{X} \boldsymbol{X}' + 2\boldsymbol{X} \sigma(Z_2) \varepsilon \boldsymbol{\pi}_2'(Y) \right] (\widehat{\gamma} - \gamma) \\ & \qquad = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\zeta}_i + \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\varsigma}_i - 2\mathbf{A}_3 (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(n^{-1/2}), \end{aligned} \tag{6.12}$$

where $\boldsymbol{\zeta}_i$ is defined in Section 2. Substituting (6.10) into (6.12), Theorem 3 follows. Theorem 2 follows by substituting (6.12) into (6.10).

Appendix D. Proof of Theorem 4

Let $\upsilon(\boldsymbol{x}) = (H^{-1})^{(1)}(\boldsymbol{x}'\boldsymbol{\beta} + \sigma(\boldsymbol{x}'\gamma)\varepsilon)$, $\upsilon_i(\boldsymbol{x}) = (H^{-1})^{(1)}(\boldsymbol{x}'\boldsymbol{\beta} + \sigma(\boldsymbol{x}'\gamma)\varepsilon_i)$, $\xi_i(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} + \sigma(\boldsymbol{x}'\gamma)\varepsilon_i$ and $\mu_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \upsilon_i(\boldsymbol{x})$. We have the expression of $\widehat{\mu}(\boldsymbol{x}) - \mu_n(\boldsymbol{x})$:

$$\begin{aligned}\widehat{\mu}(\boldsymbol{x}) - \mu_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \upsilon_i(\boldsymbol{x})\boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \frac{1}{n}\sum_{i=1}^n \upsilon_i(\boldsymbol{x})\left\{\sigma(\boldsymbol{x}'\widehat{\gamma})\frac{\widehat{H}(Y_i) - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}}{\sigma(\boldsymbol{X}_i'\widehat{\gamma})}\right. \\ \left. -\sigma(\boldsymbol{x}'\gamma)\varepsilon_i\right\} + \frac{1}{n}\sum_{i=1}^n\left\{\widehat{H}^{-1}(\xi_i(\boldsymbol{x})) - H^{-1}(\xi_i(\boldsymbol{x}))\right\} + o_p(n^{-1/2}) \\ \equiv EE_1 + EE_2 + EE_3 + o_p(n^{-1/2}). \qquad (6.13)\end{aligned}$$

By the large number theorem, it is easy to know $EE_1 = E\left[(\upsilon_i(\boldsymbol{x})\right]\boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$. Now we consider EE_2. Using the expansion,

$$\begin{aligned}\sigma(\boldsymbol{x}'\widehat{\gamma})\frac{\widehat{H}(Y_i) - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}}{\sigma(\boldsymbol{X}_i'\widehat{\gamma})} - \sigma(\boldsymbol{x}'\gamma)\varepsilon_i = \left[-\frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(\boldsymbol{X}_i'\gamma)}\sigma^{(1)}(\boldsymbol{X}_i'\gamma)\boldsymbol{X}_i' + \sigma^{(1)}(\boldsymbol{x}'\gamma)\boldsymbol{x}'\right](\widehat{\gamma} - \gamma)\,\varepsilon_i \\ +\frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(\boldsymbol{X}_i'\gamma)}\left(\widehat{H}(Y_i) - H(Y_i)\right) - \frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(\boldsymbol{X}_i'\gamma)}\boldsymbol{X}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + o_p(n^{-1/2}), \qquad (6.14)\end{aligned}$$

we get the following the result for EE_2:

$$\begin{aligned}EE_2 = E\left[\upsilon_i(\boldsymbol{x})\left(\sigma^{(1)}(\boldsymbol{x}'\gamma)x - \frac{\sigma(\boldsymbol{x}'\gamma)\sigma^{(1)}(\boldsymbol{X}_i'\gamma)}{\sigma(\boldsymbol{X}_i'\gamma)}\boldsymbol{X}_i\right)'\varepsilon_i\right](\widehat{\gamma} - \gamma) \\ -E\left[\frac{\upsilon_i(\boldsymbol{x})\sigma(\boldsymbol{x}'\gamma)}{\sigma(\boldsymbol{X}_i'\gamma)}\boldsymbol{X}_i'\right]\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \frac{1}{n}\sum_{i=1}^n \frac{\upsilon_i(\boldsymbol{x})\sigma(\boldsymbol{x}'\gamma)}{\sigma(\boldsymbol{X}_i'\gamma)}\left(\widehat{H}(Y_i) - H(Y_i)\right) + o_p(n^{-1/2}).\end{aligned}$$

Using the Taylor expansion and (3.4), we have the following expansion:

$$\begin{aligned}\widehat{H}^{-1}(y) - H^{-1}(y) &= (h\left(H^{-1}(y)\right))^{-1}\left(H\left(\widehat{H}^{-1}(y)\right) - H\left(H^{-1}(y)\right)\right) + o_p(n^{-1/2}) \\ &= (h\left(H^{-1}(y)\right))^{-1}\left(H\left(\widehat{H}^{-1}(y)\right) - \widehat{H}\left(\widehat{H}^{-1}(y)\right)\right) + o_p(n^{-1/2}) \\ &= -(h\left(H^{-1}(y)\right))^{-1}\frac{1}{n}\sum_{i=1}^n \varphi_i^H(H^{-1}(y)) + o_p(n^{-1/2}),\end{aligned}$$

where $\varphi_i^H(\cdot)$ is defined in Theorem 3. Hence we obtain the following expression of EE_3:

$$\begin{aligned}EE_3 &= -\frac{1}{n}\sum_{i=1}^n E\left[(h\left(\kappa(\boldsymbol{x})\right))^{-1}\delta_i\left(\kappa(\boldsymbol{x})\right)\right] + E\left[(h\left(\kappa(\boldsymbol{x})\right))^{-1}\boldsymbol{\pi}_1'(Y)\right]\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &\quad +E\left[(h\left(\kappa(\boldsymbol{x})\right))^{-1}\boldsymbol{\pi}_2'(Y)\right](\widehat{\gamma} - \gamma) + o_p(n^{-1/2}).\end{aligned}$$

By the results on EE_1, EE_2, EE_3 and (3.1), we re-write $\widehat{\mu}(\boldsymbol{x}) - \mu_n(\boldsymbol{x})$ as follows:

$$\widehat{\mu}(\boldsymbol{x}) - \mu_n(\boldsymbol{x}) = \mathbf{B}_1(\boldsymbol{x})' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \mathbf{B}_2(\boldsymbol{x})' \left(\widehat{\gamma} - \gamma\right)$$
$$+\frac{1}{n}\sum_{i=1}^{n} E\left[\upsilon(\boldsymbol{x})\frac{\sigma(\boldsymbol{x}'\gamma)}{\sigma(Z_2)}\delta_i(Y) - \left(h\left(\kappa(\boldsymbol{x})\right)\right)^{-1}\delta_i\left(\kappa(\boldsymbol{x})\right)\right],$$

where $\mathbf{B}_1(\boldsymbol{x})$ and $\mathbf{B}_2(\boldsymbol{x})$ are defined in Section 2. Then by the definition of $\mu_n(\boldsymbol{x})$, (3.2) and (3.3), Theorem 4 follows.

Acknowledgements

Zhou's work was supported in part by AHRQ grant R01HS013105 and U.S. Department of Veterans Affairs, Veterans Affairs Health Administration, HSR&D grant ECI-03-206. Lin's work was supported in part by the Fund of National Natural Science (Grant 10771148) of China. We also like to thank Dr. Jurgen Unutzer for providing us with the data. This paper presents the findings and conclusions of the authors. It does not necessarily represent those of VA HSR&D Service. We also would like to express our thanks to the editor, the associate editor, and the two referees for many helpful comments and suggestions that lead to the improvement of this manuscript.

References

Ash, A., Ellis, R., Pope, G., Ayanian, J., Bates, D., Burstin,H., Iezzoni, L., Mackay,E., Yu, W. (2000). Using diagnoses to describe populations and predict costs. *Health Care Financial Review*, **21**, 7-28.

Basu, A., and Rathouz, P.J. (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics*, **17**, 93-109.

Blough, D.K., Madden, C.W., Hornbrook, M.C. (1999). Modeling risk using generalized linear models. *Journal of Health Economics*, **18**, 153-171.

Buntin, M. B. and Zaslavsky, A. M. (2004). Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*, **23**, 525-542.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models. *Journal of American Statistical Association*, 92, 477-489.

Carroll, R. J. and Ruppert, D. (1988), Transformation and Weighting in Regression. *Chapman & Hall*.

Chiou, J. M., and Muller, H. (1998). Quasi-likelihood regression with unknown link and variance functions. *JASA*, **93**, 1376-1387.

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, **78**, 605-610.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model, *Journal of Econometrics*, **35**, 303-316.

Horowitz, J.L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, **64**, 103-137.

Imbens, G. and Rubin, D. B. (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *Review of Economic Studies*, **64**. 555-574.

Maity, A., Ma, Y. and Carroll, R. (2007). Efficient estimation of population-level summaries in general semiparametric regression models. *Journal of American medical association*, **477**, 123-139.

Manning, W.G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation, *Journal of Health economics*, **17**, 283-295.

Manning, W.G., Basu, A. and Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*.

Manning, W.G., and Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health economics*, **20**, 461-494.

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics, *Journal of Health economics*, **17**, 247-281.

Muller, H.G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes, *Annals of Statistics*, **12**, 766-774.

Unutzer, J., Katon, W., Callahan, C.M., and et al. (2002). Collaborative care management of late-life depression in the primary care setting. *Journal of American medical Association*, **288**, 2836-2845.

Pollard, D. (1984), Convergence of stochastic processes, *Springer-Verlag*, New York.

Ruppert, D. (2001). Transformations of data. *International Encyclopedia of social and Behavioral sciences*.

Welsh, A. and Zhou, X. H. (2006). Estimating the retransformed mean in a heteroscedastic two-part model. *Journal of Statistical Planning and Inferences*, **136**, 860-881.

Yin, X. and Cook, R. (2005). Direction estimation in single-index regressions. *Biometrika*, **92**, 371-384.