12-6-2004

Semi-parametric Single-index Two-Part Regression Models

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

Hua Liang
St. Jude Children's Research Hospital, hua.liang@stjude.org

1 Introduction

In data analysis, we sometimes encountered skewed data with additional zero values. Although it is common to use the median instead of the mean as the measure of central location in skewed data, many applications require mean as the parameter of interest. This is particularly true in the analysis of medical cost data (Zhou and Tu, 1999). For example, health care policymakers and managers are interested in the entire expenditure on health care in a given patient population, which can be measured by the total cost; only the mean, not median, can be used to recover the total cost. For example, to achieve fairness in the allocation of fixed assets to different veteran affairs (VA) hospitals, the federal VA administration is interested in the most accurate prediction of the total cost for each VA hospital.

For modeling the mean of skewed data with additional zero values, several parametric regression models and methods have been proposed (Ichimura, 1993). These models include the Tobit model (Tobin, 1958) and Heckman's selection model (Heckman, 1976). Duan et al. (1983) argued that these models may not be the best models for skewed data containing zeros, and proposed an alternative two-part parametric regression model. The two-part model is a generalization of the delta distribution model (Aitchison, 1955) and consists of two stages. The first stage uses a probit equation for the dichotomous event of having zero or positive values, and the second stage uses a linear model for non-zero values on the log-scale. Olsen and Schafer (2001) extended Duan et al.'s (1983) two-part parametric regression model to longitudinal data. If the parametric distribution assumption is true, regression estimators are usually $\sqrt{n}$ consistent. However, if the assumption of a parametric distribution is violated, the resulting estimators can be biased.

One alternative way of modeling non-zero values is to use a totally non-parametric regression model, but the convergence rate of nonparametric estimators to the true parameter decreases rapidly as the number of covariates increases. To make a trade-off between nonparametric and parametric models, we propose a semi-parametric single-index regression model for non-zero costs and use it for the analysis of skewed data with additional zeros. A single-index model is one of effective tools to avoid the curse of dimensionality occured in nonparametric multivariate regression. It generalizes linear regression by replacing $\alpha^{\mathrm{T}}x$ by $g(\alpha^{\mathrm{T}}x)$ but keeps feasibility of univariate nonparametric regression. Efforts were mainly focused on estimation

of α and $g(\cdot)$ when the covariates x are continuous. In many practical problems, covariates are mixed with continuous and binary/discrete variables, for instance in our example. How to estimate the coefficients of the continuous and discrete components forms the goal of this paper. The similar topics have been studied by Bonneu, Delecroix, and Malin (1993), Delecroix, Härdle, and Hristache (2003), and Horowitz and Härdle (1996).

The paper is organized as follows. In Section 2 we introduce a semi-parametric single-index two-part regression model. In Section 3 we propose an estimation procedure for the proposed semi-parametric single-index two-part regression model. In Section 4 we conduct a simulation study to assess the performance of the proposed method in finite-sample sizes. In Section 5 we illustrate the application of the proposed methods in a heath care cost study, in which the cost was the main outcome. We state the assumptions for our method in the Appendix.

2 The Models

Let Y_i be a random variable that represents the total inpatient cost of the ith patient, where $i = 1, \ldots, n$. We assume that $Y_1, \ldots, Y_n$ are independent. The proposed model consists of the following two parts. In the first part, we relate the probability of $(Y_i > 0)$ to a vector of known covariates W_i through a logistic link function so that

$$\text{logit}\{P(Y_i > 0|W_i)\} = W_i^{\mathrm{T}}\alpha, \tag{2.1}$$

where α is a vector of unknown parameters. In the second part, we relate the conditional mean of Y_i given $Y_i > 0$ to a vector of covariates, X_i and Z_i, by a semi-parametric single-index model,

$$E(Y_i \mid X_i, Z_i, Y_i > 0) = g(X_i^{\mathrm{T}}\beta + Z_i^{\mathrm{T}}\gamma), \tag{2.2}$$

where $X_i(k \times 1)$ and $Z_i(l \times 1)$ are, respectively, continuous and discrete covariates, $g(\cdot)$ is an unknown smooth function, and β and γ are the vectors of unknown parameters. Note that some elements in W_i may overlap with those in X_i and Z_i. In order to identify β and γ we require that the model (2.2) contains at least one continuous variable (Klein and Spady, 1993). See Bierens and Hartog (1988) for a detailed discussion of the case where $k = 0$. Because β

and γ are identified only up to sign and scale, sign and scale normalizations are required; we assume the coefficient of the first component of X, β_1, is 1.

3 Estimation Procedure

Let $\{(Y_i, X_i, Z_i, W_i, \delta_i), i = 1, \ldots, n\}$ be a sample of size n from models (2.1) and (2.2). Denote $\pi_i = 1 - P(Y_i = 0) = P(Y_i > 0)$. Then, $\pi_i = \{1 + \exp(-W_i^{\mathrm{T}}\alpha)\}^{-1}$. Our goal is to provide point estimates and confidence intervals of the parameters β, γ, and $E(Y_i|W_i = w_i, X_i = x_i, Z_i = z)$. Note that $E(Y_i|W_i = w_i, X_i = x_i, Z_i = z) = \pi_i g(x_i^{\mathrm{T}}\beta + z^{\mathrm{T}}\gamma)$. To estimate $E(Y_i|W_i = w_i, X_i = x_i, Z_i = z_i)$, we estimate α, β, γ and the nonparametric function $g(\cdot)$ first. Using a standard logistic technique, we obtain an estimator, say $\widehat{\alpha}_n$, of α. Denote $\widehat{\pi}_{in} = \{1 + \exp(-W_i^{\mathrm{T}}\widehat{\alpha}_n)\}^{-1}$. It is easy to show that $\sqrt{n}(\widehat{\alpha}_n - \alpha) = O_P(1)$, and then

$$\sqrt{n}(\widehat{\pi}_{in} - \pi_i) = O_P(1). \tag{3.1}$$

We next consider estimation of β, γ, and the nonparametric function $g(\cdot)$ in the single-index model using the data $(Y_i, X_i, Z_i, W_i, Y_i = 0)$. We first discuss estimation of β and $g(\cdot)$ and then discuss the method for estimating γ, the vector of discrete covariates.

For estimating β and $g(\cdot)$, several methods have already been proposed in the statistical literature, including average derivative estimation (ADE) (Härdle and Stoker, 1989), projection pursuit regression (Friedman and Stuetzle, 1981), and sliced inverse regression (Li, 1991). The method of ADE is most computationally efficient because it does not require iteration as other methods do. We extend the method of ADE to our two-part semi-parametric single index model. Define $\Omega_z = \{z^{(i)}, i = 1, \cdots, M\}$ to be the support of the discrete random vector Z. The estimation procedure for β and $g(\cdot)$, proposed by Härdle and Stoker (1989), are summarized as follows.

To estimate β, we denote $\{X_{iz}, Y_{iz}, Z_{iz}\}$ to be the subset of $\{X_i, Y_i, Z_i\}$ with $Z_i = z$ for each $z \in \Omega_z$. Let $M_z = \#\{i : Z_i = z, i = 1, \cdots, n\}$. Applying the method of ADE to this subset, we obtain an estimate of β using the data in this subset, denoted by β_{nz}. Combining data across the subsets, we obtain an estimate of β as

$$\widehat{\beta}_n = \sum_{z \in \Omega_z} M_z \beta_{nz}/n.$$

After estimating β by $\widehat{\beta}_n$, we next estimate $g(\cdot)$. Noting that Z is a discrete variable, we estimate $g(\cdot)$ for each of the z values. For each given $z \in \Omega_z$, we estimate the function $g(v + z^{\mathrm{T}}\gamma)$ by

$$\widehat{g}_{nz}(v) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{v - \widehat{\Lambda}_i}{h}\right) Y_i I(Z_i = z) \Big/ \widehat{f}_{nz}(v),$$

where $K(\cdot)$ is a kernel function, $\widehat{\Lambda}_i = X_i^{\mathrm{T}}\widehat{\beta}_n$, and $\widehat{f}_{nz}(v)$ is the density estimator of $X^{\mathrm{T}}\beta$ given $Z = z$, i.e.,

$$\widehat{f}_{nz}(v) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{v - \widehat{\Lambda}_i}{h}\right) I(Z_i = z).$$

Under the appropriate assumptions (listed in the Appendix), by modifying the proof in Theorem 3.3 of Härdle and Stoker (1989), we can show that the regression estimator $\widehat{g}_{nz}(v)$ is asymptotically normal and converges to $g(v + Z^T\gamma)$ (pointwise) with the optimal rate $n^{2/5}$. More specifically,

$$n^{2/5}\{\widehat{g}_{nz}(v) - g(\bullet)\} \to N\{M_{(v|z)}, \Sigma_{(v|z)}\}, \tag{3.2}$$

where

$$\begin{aligned} M_{(v|z)} &= \{g^{(2)}(\bullet)/2 + g'(\bullet)f'(\bullet|z)/f(\bullet|z)\}\int s^2 K(s)ds \\ \text{and} \quad \Sigma_{(v|z)} &= \{\text{var}(Y|x^{\mathrm{T}}\beta, z)/f(\bullet|z)\}\int K^2(s)ds. \end{aligned}$$

Here and in the sequel, $(\bullet)$ denotes $(v + z^{\mathrm{T}}\gamma)$.

There are some good alternative methods for estimating $\widehat{g}_{nz}(\nu)$ and $\widehat{f}_{nz}(v)$ including local linear or local polynomial kernel smoothing methods and regression spline methods. In this paper we chose the local constant smoothing method for its simple descriptions. The results still apply for any other kernel-based methods, as well as for spline methods. One critical concern in any kernel based method is the bandwidth selection. In the Appendix, we give the theoretical conditions for selecting appropriate bandwidths in our kernel smoothing method. In our practical implementation below, we compute the average error using a geometric sequence of 30 bandwidths ranging in $[0.1, 0.5]$. The optimal bandwidth is selected to minimize the average squared error among the 30 candidates.

Finally we estimate γ by employing the estimation procedure proposed by Horowitz and Härdle (1996). It can briefly be described as follows. Assume that there are finite numbers

v_0, v_1, c_0 and c_1 such that $v_0 < v_1$, $c_0 < c_1$ and $g(v + z\gamma) < c_0$ for each $z \in \Omega_z$ if $v < v_0$, and $g(v + z\gamma) > c_1$ for $z \in \Omega_z$ if $v > v_1$. For $z \in \Omega_z$, define

$$\begin{aligned} J(z) &= \int_{v_0}^{v_1} \Big[c_0 I\{g(v + z\gamma) < c_0\} + c_1 I\{g(v + z\gamma) > c_1\} \\ &\quad + g(v + z\gamma) I\{c_0 \le g(v + z\gamma) < c_1\} \Big] dv \end{aligned}$$

and let

$$\Delta J = \begin{bmatrix} J\{\boldsymbol{z}^{(2)}\} - J\{\boldsymbol{z}^{(1)}\} \\ \vdots \\ J\{\boldsymbol{z}^{(M)}\} - J\{\boldsymbol{z}^{(1)}\} \end{bmatrix} \text{ and } \mathcal{B} = \begin{bmatrix} \boldsymbol{z}^{(2)} - \boldsymbol{z}^{(1)} \\ \vdots \\ \boldsymbol{z}^{(M)} - \boldsymbol{z}^{(1)} \end{bmatrix}.$$

It follows from Horowitz and Härdle (1996) that

$$\gamma = (c_1 - c_0)^{-1} (\mathcal{B}^{\mathrm{T}} \mathcal{B})^{-1} \mathcal{B}^{\mathrm{T}} \Delta J.$$

It suffices to estimate ΔJ. Let

$$J_n = \int_{v_0}^{v_1} \Big[c_0 I\{\widehat{g}_{nz}(v) < c_0\} + c_1 I\{\widehat{g}_{nz}(v) > c_1\} + \widehat{g}_{nz}(v) I\{c_0 \le \widehat{g}_{nz}(v) < c_1\} \Big] dv.$$

As a consequence, we define ΔJ_n by replacing J by J_n in ΔJ, and obtain an estimator of γ

$$\widehat{\gamma}_n = (c_1 - c_0)(\mathcal{B}^{\mathrm{T}} \mathcal{B})^{-1} \mathcal{B}^{\mathrm{T}} \Delta J_n.$$

Under appropriate regularity conditions, $\widehat{\gamma}_n$ is asymptotically normal. See Horowitz and Härdle (1996) for a detailed discussion.

After we have obtained the estimates of β, γ, and $g(.)$, we can then estimate the conditional overall mean $E(Y_i | X_i = x_i, Z_i = z_i, Y_i > 0)$ by the quantity

$$\widehat{\widehat{E}}(Y_i | X_i = x_i, Z_i = z_i, Y_i > 0) = \widehat{g}_{nz}(x_i^{\mathrm{T}} \widehat{\beta}_n + z_i^{\mathrm{T}} \widehat{\gamma}).$$

The mean function $E(Y_i|z) = E(Y_i | X_i = x_i, Z_i = z)$ is estimated by $\hat{E}(Y_i|z) = g_{nz}(x_i^{\mathrm{T}} \widehat{\beta}_n + z_i^{\mathrm{T}} \widehat{\gamma}) \widehat{\pi}_{in}$, where $\widehat{\pi}_{in} = \{1 + \exp(-w_i^{\mathrm{T}} \widehat{\alpha}_n)\}^{-1}$.

Note that $\widehat{g}_{nz}(v)\widehat{\pi}_{in} - g(\bullet)\pi = \widehat{g}_{nz}(v)(\widehat{\pi}_{in} - \pi) - \{\widehat{g}_{nz}(v) - g(\bullet)\}\pi$. Recall (3.1) and (3.2). We can easily show that

$$n^{2/5}\{\widehat{g}_{nz}(v)\widehat{\pi}_{in} - g(\bullet)\pi\} \to N\{M_{(v|z)}\pi, \Sigma_{(v|z)}\pi^2\}. \tag{3.3}$$

For a given value of v, we consistently estimate the bias and variance given in (3.3) by using $\{Y_i, i = 1, \ldots, n\}$, $\widehat{g}$ and $\widehat{f}$ and their derivatives with a standard sandwich method. The resulting estimators are given as follows:

$$\widehat{M}_{n(v|z)} = \{\widehat{g}^{(2)}_{nz}(v)/2 + \widehat{g}'_{nz}(v)\widehat{f}'_{nz}(v)/\widehat{f}_{nz}(v)\} \int s^2 K(s)ds$$

and

$$\widehat{\Sigma}_{n(v|z)} = \widehat{\text{var}}_n(Y|X^{\mathrm{T}}\beta = v, z)/\widehat{f}_{nz}(v) \int K^2(s)ds.$$

Denote $\widehat{V}_i = x_i^{\mathrm{T}}\widehat{\beta}_n + z_i^{\mathrm{T}}\widehat{\gamma}$. Then, for a fixed value of z, we can show that the following statistic is asymptotic pivotal and has the asymptotically standard normal distribution:

$$\left[n^{2/5}\{\widehat{E}(Y_i|z) - E(Y_i|z)\} - \widehat{M}_{n(\widehat{V}_i|z)}\widehat{\pi}_{in}\right] \widehat{\Sigma}^{-1/2}_{n(\widehat{V}_i|z)}\widehat{\pi}^{-1}_{in}.$$

Basing on this statistics, we obtain the following $\xi-$level confidence interval of $E(Y_i|z)$:

$$\left[\widehat{E}(Y_i|z) - \widehat{M}_{n(\widehat{V}_i|z)}\widehat{\pi}_{in}n^{-2/5} - \widehat{\Sigma}^{1/2}_{n(\widehat{V}_i|z)}\widehat{\pi}_{in}n^{-2/5}q_{\xi/2}, \right.$$
$$\left. \widehat{E}(Y_i|z) - \widehat{M}_{n(\widehat{V}_i|z)}\widehat{\pi}_{in}n^{-2/5} + \widehat{\Sigma}^{1/2}_{n(\widehat{V}_i|z)}\widehat{\pi}_{in}n^{-2/5}q_{\xi/2}\right], \qquad (3.4)$$

where $q_{\xi/2}$ is the $(1-\xi/2)$th quantile value of the standard normal distribution.

4 Numerical Results

4.1 Simulation Study

To evaluate the proposed method, we conducted an intensive experiment to explore its performance. We generated a sample size of n observations by a two-stage procedure. In the first stage, we generated zero costs according to a Bernoulli distribution with the probability:

$$P(Y = 0) = \{1 + \exp(-0.3W_1 - 0.4W_2)\}^{-1},$$

where W_1 was a covariate with the uniform distribution, $\text{Uniform}[0.7, 1]$, and W_2 was another covariate with the normal distribution, $\text{Normal}(0, 0.3)$. In the second stage, we generated non-zero costs according to the following non-linear heteroscedastic models:

- case 1: $E(Y|X, Z, Y > 0) = \exp(X_1 + 0.25X_2 + 0.3Z)$,

- case 2: $E(Y|X,Z,Y>0)=\exp(X_1+0.25X_2+0.3Z)\{1+0.1*\exp(X_1+0.25X_2+0.3Z)\}^{-1}$,

respectively. The variance function in the both cases was $\text{var}(Y|X,Z,Y>0)=(X_1+0.25X_2)^2$, and X_1, X_2, and Z were three covariates. Here $X_1 \sim \text{Normal}(0,0.7)$, $X_2 \sim \text{Normal}(0,0.4)$, and $Z \sim \text{Binom}(0.5)$. Therefore, the true regression model for the expected value of Y is

$$E(Y|X,Z,W)=\{1+\exp(-0.3W_1-0.4W_2)\}^{-1}E(Y|X,Z,Y>0).$$

In the simulation experiment we fitted both a nonlinear parametric model and our semi-parametric single index regression model to the simulated data sets. The parametric mode for $E(Y|X,Z,Y>0)$ is assumed to have the form, $\exp(X_1+\beta X_2+\gamma Z)$, which is the same as the case 1 model. Our goal is to investigate the efficiency of our method relative to the parametric approach when the model is correct and to check its robustness when the model is incorrectly specified.

The sample sizes were $n=100, 200$, and 500. Bandwidths were selected as remarked before. We used the kernel function $K(u)=15/16(1-u^2)^2 I_{(|u|\le 1)}$ in nonparametric regression. We generated 1000 data sets in each of six parameter configurations. We computed J_n, defined in Section 4, using the Gauss-Legendre quadrature. To compute c_0 and c_1, we first estimated g_z for each $z \in \Omega_z$ using the standard normal kernel and called the resulting estimate g^*_{nz}. We then computed c_0 and c_1 by the following formula:

$$c_0=\max_{z\in\Omega_z}\max_{X_i\hat\beta_n\le v_{n0}} g^*_{nz}(X_i\hat\beta_n) \quad \text{and} \quad c_1=\min_{z\in\Omega_z}\min_{X_i\hat\beta_n\ge v_{n1}} g^*_{nz}(X_i\hat\beta_n),$$

where

$$v_{n0}=\max_{z\in\Omega_z}\min_{1\le i\le n}\{X_i\hat\beta_n+h_{nz}: Z_i=z\}, \;\; v_{n1}=\min_{z\in\Omega_z}\max_{1\le i\le n}\{X_i\hat\beta_n-h_{nz}: Z_i=z\},$$

$h_{nz}=s_{vz}n_z^{-1/7.5}$, and s_{vz} was the sample standard deviation of $X\hat\beta_n$ conditional on $Z=z\in\Omega_z$.

The computation was implemented in XploRe-an advanced statistical environment developed by Härdle's team, see the website at: `http://www.xplore-stat.de/`.

Table 1 gives the results for the parametric components β and γ. In both the cases, the estimated values of β and γ based on our method are close to the true values, but not as close

as the parametric model based estimates when the parametric model is correctly specified, although the difference may be ignorable. However, when the parametric model is misspecified, the parametric approach leads to biased results, whereas the estimated values based on our method are still close to the true values.

Table 1 goes here

Given points $\{(x_i, w_i, z_i), i = 1, \cdots, m\}$ for some m, we estimated $E(Y|X, Z, W)$ at these given points in each replication. The averages of the estimated values of $E(Y|X = x, Z = z, W = w)$ based on the 1000 replications are our estimates of $E(Y|X, Z, W)$, which are shown in Figure 1.

Figure 1 goes here

In Figure 1, the left-hand panel represents the expectation of Y against $X^{\mathrm{T}}\beta$, and the right-hand panel represents the expectation against $X^{\mathrm{T}}\beta + \gamma$. The solid lines indicate the true curves, the dotted and dashed lines indicate the nonparametric and parametric fitted curves. For example, in the left-hand panel, the solid line corresponds to the function $\{1 + \exp(-0.3w_1 - 0.4w_2)\}^{-1} \exp(x_1 + 0.25x_2 + z)$, and the dotted and dashed lines correspond the estimates, $\{1 + \exp(\hat{\beta}_1 w_1 + \hat{\beta}_2 w_2)\}^{-1} \widehat{g}_{nz}(x_1 + \widehat{\beta}_2 x_2 + \widehat{\gamma} z)$ and $\{1 + \exp(\hat{\beta}_1 w_1 + \hat{\beta}_2 w_2)\}^{-1} \exp(x_1 + \widehat{\beta}_2 x_2 + \widehat{\gamma} z)$, respectively. From Figure 1, we can draw a similar conclusion on estimation of $E(Y|X, Z, W)$ as on estimation of parametric components β and γ; that is, our method is comparable to the parametric one when the parametric model is correctly specified, but beats the parametric one when the model is incorrectly specified.

4.2 Health Care Data Analysis

Effective management of chronic diseases often requires long-term administration of medications. Although many chronic diseases can be treated effectively with medications, there is limited evidence on the effectiveness of proper medications in improving patients' overall functional status and quality of life and in reducing health care charges. In addition, the proliferation of new drugs has increased the potential for adverse drug interactions. These two factors, along with variability in medical training, have led to much variation in treatments for

the same chronic conditions. To determine whether medication-prescribing patterns could be altered to improve patient outcomes in a cost-effective manner, Tierney et al. (1998) conducted a clinical trial of a computer-assisted prospective drug utilization review (DUR) in an urban, hospital-based academic primary care system. The DUR program involved guideline-based, computer-generated treatment recommendations to primary care physicians and hospital-based pharmacists during encounters with their patients. These recommendations were aimed at preventing adverse drug reactions and improving the effectiveness of treatment for three chronic conditions: hypertension, congestive heart failure, and reactive airway disease. In addition to quality of life, outcome variables in this trial included inpatient and outpatient charges. In the current analysis, we will focus on the total inpatient health care charges generated by patients with hypertension during the two year-long trial.

This data set has the two analytic problems: (1) a large number of patients with zero inpatient costs, and (2) a skewed distribution. The formal test for normality gives a p-value of < 0.001 for original non-zero costs and a p-value of 0.003 for log-transformed non-zero costs. Therefore, we know that non-zero costs have a severely skewed distribution, which is not a log-normal distribution. From some preliminary analysis reported elsewhere (Tierney et al, 1998), we found that the following four important covariates which are related to inpatient charges: (a) age, (b) the SF-36 physical function score (from the medical outcomes study 36-item short-form health survey), (c) whether a patient is female, and (d) whether a patient is black. The analytic goal in this paper is to estimate the average cost of a patient with given values of these four covariates.

Let Y_i be the health care cost of the ith patient. Let X_{i1} and X_{i2} denote the continuous-scale covariates, the age and SF-36 physical function score of the ith patient, respectively, and let Z_{i1} and Z_{i2} denote binary covariates, gender and race indicators, for the ith patient, respectively; that is, $Z_{i1} = 1$ if the ith patient is female and 0 otherwise; $Z_{i2} = 1$ if the ith patient is black and 0 otherwise. Denote $X_i = (X_{i1}, X_{i2})'$ and $Z_i = (Z_{i1}, Z_{i2})'$. We model the probability of being the zero cost by the logistic regression model,

$$\text{logit}\{P(Y_i = 0 \mid X_i, Z_i)\} = \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Z_{i1} + \alpha_4 Z_{i2},$$

and we assume that the conditional expectation of the positive costs Y_i given $Y_i > 0$ follows a semi-parametric single-index regression model,

$$Y_i = g(\beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 Z_{i1} + \gamma_2 Z_{i2}),$$

where the function $g(.)$ is unknown. To confirm the logistic assumption for the probability of being a zero cost, we conduct a goodness-of-fit test (le Cessie and van Houwelingen, 1991), and find the assumption is reasonable. Then, the regression model for the overall mean is given as follows:

$$E(Y_i \mid X_i, Z_i) = \frac{1}{1 + \exp(\alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Z_{i1} + \alpha_4 Z_{i4})} g(X_i^T \beta + Z_i^T \gamma).$$

We report the results in Figure 2.

Figure 2 goes here

The solid line on the left panel in Figure 2 displays the estimated values of $E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2, Z_{i1} = z_1, Z_{i2} = z_2)$ versus $x_1\hat{\beta}_1 + x_2\hat{\beta}_2$ for all patients in the sample. Similarly, the solid line on the right panel represents the estimated values of $E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2, Z_{i1} = z_1, Z_{i2} = z_2)$ versus $X_i^{\mathrm{T}}\widehat{\beta} + Z_i^{\mathrm{T}}\widehat{\gamma}$ for all patients in the sample.

For constructing confidence intervals of $E(Y|X, Z)$, we could theoretically use the variance formula given in (3.3) to compute the standardized test statistics for $E(Y|X, Z)$, and then use the normal approximation to construct confidence intervals. Unfortunately, the resulting confidence intervals are not good, giving negative lower bounds. The reason is partly due to the relatively small sample size for non-zero observations, resulting in big bias in the estimation of $\Sigma_{v|z}$. We therefore provided bootstrap confidence intervals of $E(Y|X, Z)$ in Figure 2, in which the dotted lines represent 95% pointwise bootstrap confidence intervals. The pointwise bootstrap confidence intervals were computed at 101 selected points with 200 bootstrap replications by randomly resampling the original cost data with replacement. It is intuitively clear that the bootstrap can be used to construct estimates of standard error, because the estimators of the parameters π, β, γ are regular. The standard bootstrap arguments (Davison and Hinkley, 1997) can justify our statement.

We can also use the proposed model to predict the average cost of a patient with given characteristics of the patient. For example, among 53 years old patients with the SF-36 physical function score of 37.5, a black male patient has an estimated average inpatient cost of $7848

with a 95% confidence interval of ($1390.7, $44326.6), and a black female patient has an estimated average inpatient cost of $6228 with a 95% confidence interval of ($1800.1, $22159.7).

5 Discussion

Effectively analyzing skewed data with excessive zero values is a challenging topic in practice. One additional complication is that non-zero costs may not follow an often assumed log-normal distribution. In this paper, we propose a semi-parametric single-index two-part model that allows us to handle these problems. We have theoretically shown that the proposed estimators are consistent and have asymptotically normal distributions under some regularity conditions. Our theoretical proof is a straightforward extension of theorems in Horowitz and Härdle (1996). The detailed derivation and the discussion of the regularity conditions are referred to Horowitz and Härdle (1996).

It is worthy to mention that we assume the first stage zero versus non-zero model follows a parametric logit model in this paper because our real data follow this distribution. We can easily generalize our method to allow the first stage model also to be a semi-parametric single index model. See Klein and Spady (1993) for a detailed discussion of a single index model with binary response variables. The authors proposed an asymptotically efficient estimator of the index parameter. The convergence rate of the estimator of π is $n^{-1/3}$, which can ensure that the theoretical results of this paper still hold.

References

Aitchison, J. (1955). On the discussion of a positive random variables having a discrete probability mass at the origin. *J. Amer. Statist. Assoc.* 50, 901-908.

Bonneu, M., Delecroix, M., and Malin, E. (1993). Semiparametric versus nonparametric estimation in single index regression model: a computational approach. *Comp. Statist.* 8, 207-222.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, U.K.

Delecroix, M., Härdle, W., and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Mult. Anal.* 86, 213-226.

Duan, N., Manning, W. G., Morris, C. N. and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *J. Business and Econ. Statist.* 1, 115-126.

Friedman, J. H. and Stuetzle,W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* 76, 817-823.

Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the methods of average derivatives. *J. Amer. Statist. Assoc.* 84, 986-995.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* 21, 157-178.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a sample estimation for such models. *Ann. Econom. and Soc. Measurement* 5, 475-592.

Horowitz, J. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* 91, 1632-1640.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* 58, 71-120.

Klein, R.L. and Spady, R.H. (1993). An efficient semiparametric estimator for discrete choice models. *Econometrica* 61, 387-422.

le Cessie, S. and van Houwelingen, J.C. (1991). A goodness-of-test for binary regression models, based on smoothing methods. *Biometrics* 47, 1267-1282.

Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* 86, 316-342.

Olsen, M. K. and Schafer J. L. (2001). A two-part random-effects model for semicontinous longitudinal data. *J. Amer. Statist. Assoc.* 96, 730-745.

Tierney, W. M., Overhage, M., Murray, M., Zhou, X. H., Harris, L., and Wolinsky, F. (1998). "The final report of the computer-based prospective drug utilization review project (1993-1997)". *U. S. Agency for Health Care Policy and Research*, Bethesda, MD.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* 26, 24-36.

Zhou, X.H. and Tu, W.Z. (1999). Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics* 55, 645–651.

Appendix: Assumptions

Let S_v denote the support of the distribution of $V = X^{\mathrm{T}}\beta$. Let $f(v|z)$ be the probability density of V given $Z = z$, let $p(v, \widetilde{x}|z)$ be the joint density of $(V, \widetilde{X})$ conditional on $Z = z$, let $p(z)$ be the probability that $Z = z \in S_z$ and $f(v, z) = f(v|z)p(z)$. Let $r \geq 4$ be an integer and $\|\cdot\|$ denote the Euclidean norm. The following assumptions were given by Horowitz and Härdle (1996) to assure that the asymptotic normality of the estimators of β hold.

Assumption A.1 S_z is a finite set;

$E(\|\widetilde{X}\|^2|Z = z) < \infty$ and $E(|Y|\|\widetilde{X}\|^2|Z = z) < \infty$ for each $z \in S_z$; $\widetilde{X}$ and $\widetilde{\beta}$ denote components 2 through k of X and β if $k > 1$.

$E(|Y|\|\widetilde{X}\|^2|V = v, Z = z)$, $E(|Y|^2|V = v, Z = z)$ and $f(v, z)$ are bounded uniformly over $v \in [v_0 - \zeta, v_0 + \zeta]$ for some $\zeta > 0$ and all $z \in S_z$

For each $z \in S_z$, $p(v, \widetilde{x}|z)$ has continually 3 order derivative with v and uniformly bounded over $(v, \widetilde{x})$

Var$(Y|V = v, Z = z) > 0$ for all $z \in S_z$ and almost every v.

Assumption A.2 $\mathcal{B}^T\mathcal{B}$ is nonsingular.

Assumption A.3 $g(\cdot)$ is r times continually differentiable, and its r derivatives are bounded on all bounded intervals.

Assumption A.4 There are finite numbers v_0, v_1, c_0 and c_1 such that $v_0 < v_1$, $c_0 < c_1$ and $g(v + z\gamma) < c_0$ for each $z \in \Omega_z$ if $v < v_0$, and $g(v + z\gamma) > c_1$ for $z \in \Omega_z$ if $v > v_1$; $f(v|z)$ is bounded away from 0 on an open interval containing $[v_0, v_1]$.

Assumption A.5 If $k > 1$, there are (a) a $n^{1/2}$-consistent estimator of $\widetilde{\beta}$, denoted by $\widehat{b}_n$, and (b) a $(k - 1) \times 1$ vector-valued function $\Psi(y, x, z)$ such that

$$n^{1/2}(\widetilde{b}_n - \widetilde{\beta}) = n^{-1/2}\sum_{i=1}^{n}\Psi(Y_i, X_i, Z_i) + o_p(1)$$

as $n \to \infty$ where $E\Psi(Y, X, Z) = 0$ and $n^{-1/2}\sum_{i=1}^{n}\Psi(Y_i, X_i, Z_i)$ converges to a normal distribution.

Assumption A.6 Kernel function K is bounded, symmetric, differentiable and nonzero on $[-1, 1]$, its derivative is Lipschitz continues. For $0 \leq i \leq r$, K satisfies $\int_{-1}^{1} v^i K(v)dv = 1$ if $i = 0$, 0 if $1 < i < r$ and nonzero if $i = r$.

Assumption A.7 $nh^{r+3} \to \infty$ and $nh^{2r} \to 0$ as $n \to \infty$.

Table 1: Results of the simulation study. 'mean' is the simulation mean, 's.e.' is the Monte Carlo standard error. The methods are 'parametric': parametric fitting; 'SIN': semiparametric approach proposed in this paper.

case	n	parameter	True	SIN		Parametric	
				mean	s.e.	mean	s.e.
1	100	β	0.25	0.25	0.035	0.25	0.006
		γ	0.3	0.296	0.025	0.3	0.003
	200	β	0.25	0.249	0.023	0.25	0.004
		γ	0.3	0.296	0.019	0.3	0.002
	500	β	0.25	0.25	0.019	0.25	0.003
		γ	0.3	0.297	0.014	0.3	0.001
2	100	β	0.25	0.249	0.012	0.174	0.184
		γ	0.3	0.3	0.011	0.003	0.088
	200	β	0.25	0.25	0.008	0.172	0.141
		γ	0.3	0.3	0.006	-0.016	0.084
	500	β	0.25	0.25	0.005	0.156	0.111
		γ	0.3	0.3	0.004	-0.021	0.063

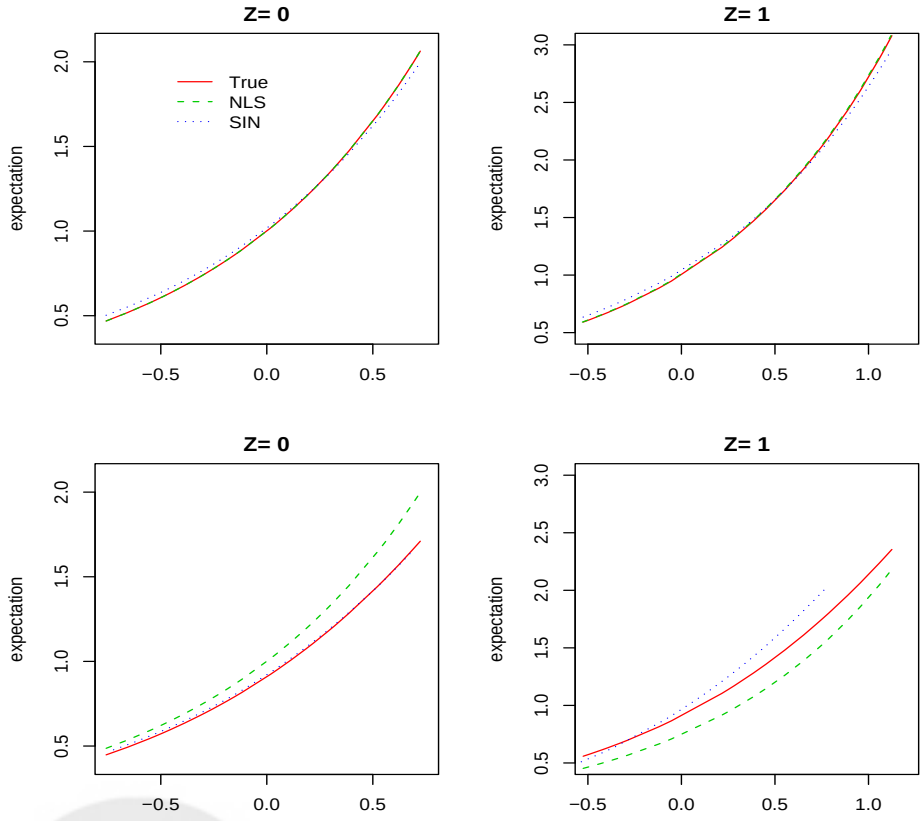

Figure 1: *Pointwise estimated values of the expectation values against* $X^{\mathrm{T}}\beta + Z^{\mathrm{T}}\gamma$ *when* $n = 200$. *The upper panel are for case 1 and the bottom panel for case 2. The solid, dotted, and dashed lines represent the true, nonparametrically fitted, and parametrically fitted curves.*

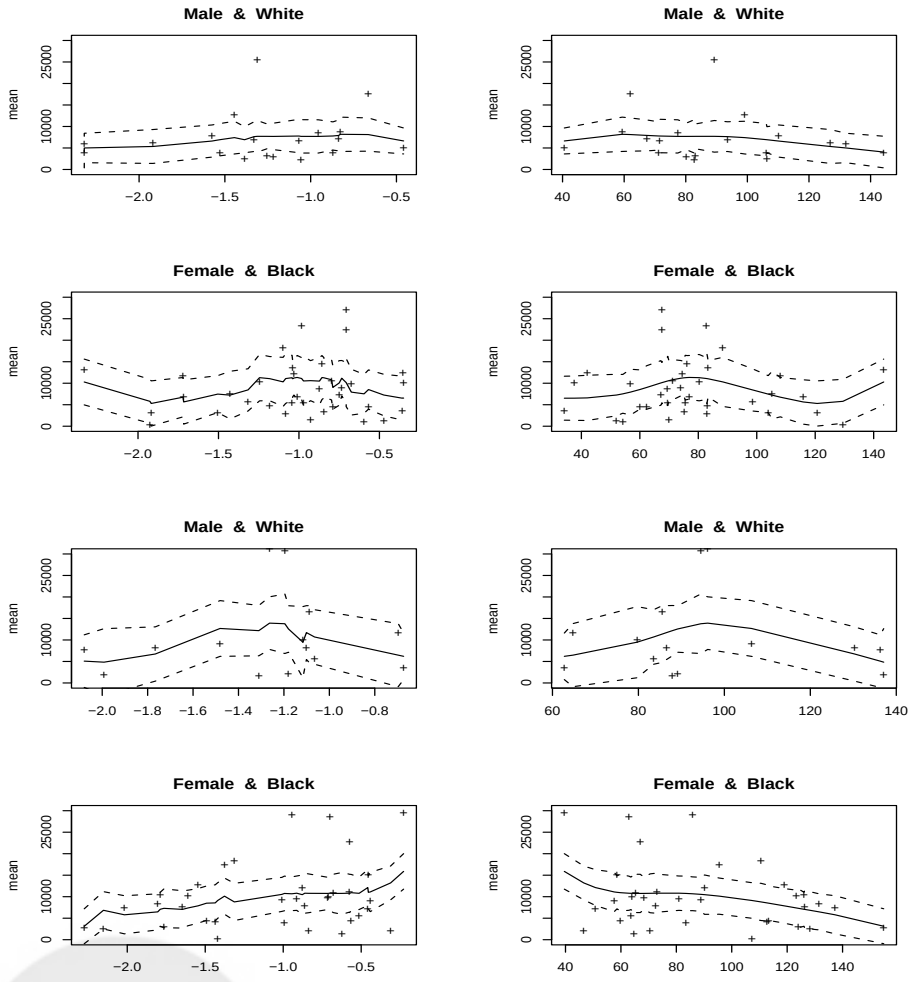

Figure 2: *Pointwise estimates of the expectation values (solid lines) and bootstrap confidence intervals (dotted lines). + represents the observed values. The left panel corresponds to the expectation of Y against $X^{\mathrm{T}}\widehat{\beta}$, and the right panel corresponds to the expectation of Y against $X^{\mathrm{T}}\widehat{\beta} + Z^{\mathrm{T}}\widehat{\gamma}$.*