UW Biostatistics Working Paper Series

9-29-2004

Estimating the Retransformed Mean in a Heteroscedastic Two-Part Model

Alan H. Welsh
University of Southampton, A.H.Welsh@maths.soton.ac.uk

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

Suggested Citation
Welsh, Alan H. and Zhou, Xiao-Hua, "Estimating the Retransformed Mean in a Heteroscedastic Two-Part Model" (September 2004). *UW Biostatistics Working Paper Series.* Working Paper 215.
http://biostats.bepress.com/uwbiostat/paper215

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

1. INTRODUCTION

The analysis of right-skewed, heteroscedastic data can often be simplified by applying a monotone transformation and then analyzing the data on the transformed scale. This approach is particularly attractive when a transformation which achieves linearity with additive, normal and homoscedastic errors can be found. The main complication in this case is the retransformation bias which arises when we try to transform back to the original scale for prediction and forecasting. In practice, there is often no single transformation which simultaneously achieves additivity, normality and homoscedasticity so that, if we achieve linearity with additive errors, we may still have to deal with non-normality and heteroscedasticity. A further complication in the transformation approach arises when, in addition, the data contain a certain proportion of zero measurements. This is a common occurrence when measuring diverse phenomena such as rainfall, tumor size, reaction time, resource usage, etc where the non-occurrence or absence of the phenomenon leads to a zero observation. See Panel on Nonstandard Mixtures of Distributions (1989) for discussion and further examples. Our purpose in this paper is to develop a flexible methodology which enables us to handle retransformation bias when the transformation achieves linearity and additivity but not necessarily normality and homoscedasticity, and the data contain zero measurements.

The complications of using transformation based models described above are well documented in the literature on analyzing health care cost data. Duan et al (1983) proposed fitting standard linear regression models to transformed cost data. However, for cost data, the assumption of homoscedastic variance after transformation is often not met; see Manning (1998), Mullahy (1998), Zhou et al (1997a), Zhou et al (1997b), and Zhou and Tu, (1999). Mullahy (1998) gave several real situations where two-part regression models which assume homoscedasticity after transforming the nonzero responses yield inconsistent inferences about important policy parameters and has warned against their automatic application in health econometrics when interest is focused on the mean of the original responses.

Retransformation bias has been treated previously under different assumptions about what the transformation achieves in the data. For problems without zeros, Duan (1983) assumed that a known transformation achieved linearity with additive, homoscedastic but not necessarily normal errors. He proposed a nonparametric estimator of the mean on the original scale which he called the smearing estimator. He showed that the smearing estimator is consistent under mild conditions but did not give its asymptotic distribution, leaving open the problem of setting approximate confidence and prediction intervals on the original scale. Carroll and Ruppert (1984) suggested using the smearing estimator when the transformation has been estimated from the data and Taylor (1986) explored the properties of this estimator by

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

simulation. Taylor (1986) also proposed a parametric estimator of the mean on the original scale which is appropriate when the transformation additionally achieves normality. In his simulation, he showed that the performance of the two methods is very similar.

The presence of zero observations can be handled by fitting a nonstandard mixture model with a degenerate component at zero. That zero/nonzero response can be modeled by binary regression and the magnitude of the nonzero responses can be modelled conditionally by a continuous distribution. To allow for the fact that transformation of the nonzero responses may not achieve normality and homoscedasticity, we fit a heteroscedastic regression model proposed by Welsh et al (1994) to the transformed nonzero responses. Although the zero and nonzero responses are effectively modelled separately, the two models need to be combined to produce estimates of the mean response on the original scale. We propose extensions of Duan's smearing estimator which combine the two parts of the model to produce estimates of the mean response on the original scale.

We describe our semi-parametric two-part heteroscedastic regression model for a skewed population with additional zero observations in Section 2. In Section 3, we specify estimators of the regression parameters on the transformed scale. Then, in Section 4, we propose two nonparametric estimators for the overall mean on the original scale; these non-parametric estimators are extensions of Duan's smearing estimator to the semi-parametric two-part heteroscedastic regression model. We show in Section 5 that both the estimators are consistent and asymptotically normally distributed and show how to construct approximate confidence intervals for the mean response on the original scale. In Section 6 we illustrate the application of the estimators in a real clinical study and in Section 7 we report a simulation study of the finite-sample performance of these two estimators.

2. A HETEROSCEDASTIC TWO-PART REGRESSION MODEL

We treat the observations as realizations of independently distributed random variables $Y_1, \ldots, Y_n$ which have a density function

$$f^*(y_i, \pi_i, \phi_i) = \begin{cases} \pi_i & \text{if } y_i = 0 \\ (1-\pi_i) f(y_i, \phi_i) & \text{if } y_i > 0 \end{cases}, \tag{1}$$

where $f(y_i, \phi_i)$ is a proper density function. Clearly, $\pi_i = Pr(Y_i = 0)$ and $f(y_i, \phi_i)$ is the conditional density of Y_i given that $Y_i > 0$.

As in a standard generalized linear model (see for example McCullagh and Nelder, 1989), π_i can be related to known vectors of covariates z_i through a known link function l so that

$$l(\pi_i) = z_i^T \alpha_0, \tag{2}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

where α_0 is a vector of unknown parameters. Here the symbol T denotes the transpose of a matrix or vector. A common choice of l is the logistic function $l(x) = \log(x/(1-x))$ but other choices are possible.

As $f(y_i, \phi_i)$ is often asymmetric, we adopt a conditional transformation model to relate Y_i to vectors of covariates x_i. Specifically, given that $Y_i > 0$, we assume that for a monotone transformation h,

$$h^{-1}(Y_i) = x_i^T \beta_0 + g_i(\beta_0, \theta_0)\epsilon_i, \tag{3}$$

where β_0 and θ_0 are vectors of unknown parameters with dimensions q_1 and q_2, respectively, g_i is a known function allowing scaling and heteroscedasticity on the transformed scale, and $\{\epsilon_i\}$ are independent and identically distributed random variables with common density function f_ϵ with mean zero and variance one. It is traditional to assume that the transformation h^{-1} makes the mean linear and the residual variation both homoscedastic and normally distributed. In model (3), we assume only that the transformation h^{-1} makes the mean linear, leaving us to model any heteroscedasticity through the g_i and to account for possible non-normality of the $\{\epsilon_i\}$. The transformation h can be known or unknown, depending on the application: for simplicity, in this paper, we treat h as known, reserving comment on the case of estimated h to the final discussion.

Note that the covariates in z_i and x_i may be but are not necessarily different and that the function g_i can depend on x_i and/or other covariates. Put $\xi = (\beta^T, \theta^T)^T$ and define

$$e_i(\xi) = \frac{h^{-1}(Y_i) - x_i^T \beta}{g_i(\xi)}. \tag{4}$$

Here the dimension of ξ is $q = q_1 + q_2$. Then the conditional transformation model implies that $f(y_i, \phi_i) = \frac{1}{h'(h^{-1}(y_i))g_i(\xi_0)} f_\epsilon(e_i(\xi_0))$, with $\phi_i = (x_i^T \beta_0, g_i(\beta_0, \theta_0))^T$.

3. PARAMETER ESTIMATION

The log-likelihood for the model (1) is

$$\begin{aligned} \ell(\alpha, \xi) &= \sum_{i=1}^n \left\{ I\left(y_i = 0\right) \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right\} + \sum_{i=1}^n I(y_i > 0) \log f(y_i, \phi_i) \\ &= \ell_1(\alpha) + \ell_2(\xi). \end{aligned}$$

This factorization shows that the parameters α_0 and ξ_0 are orthogonal so, without any loss of efficiency, can be estimated separately.

Estimation of α_0 by maximizing $\ell_1(\alpha)$ is a standard binary regression problem. Under mild conditions, it follows from standard estimating equation theory (see for example Diggle et al, 2002) that

$$\hat{\alpha} - \alpha_0 = \frac{1}{n} A^{-1} \sum_{i=1}^n \rho_i \{ I(y_i = 0) - \pi_i \} + o_p(\frac{1}{\sqrt{n}}), \tag{5}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where $\rho_i = z_i/(l'(\pi_i)\pi_i(1-\pi_i))$ and

$$A = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \rho_i z_i^T / l'(\pi_i).$$

Similarly, we can maximize $\ell_2(\xi)$ to estimate ξ_0 but it is useful to consider a wider class of estimators. We therefore consider estimators which satisfy estimating equations of the form

$$\sum_{i=1}^{n} \Psi_i(Y_i, \xi) = 0, \tag{6}$$

where $\Psi_i(Y_i, \xi)$ and 0 are q dimensional vectors. Writing the derivatives of g as

$$g_i^{(1)}(\xi) = \frac{\partial g_i(\xi)}{\partial \beta}, \text{ and } g_i^{(2)}(\xi) = \frac{\partial g_i(\xi)}{\partial \theta},$$

the maximum likelihood estimator satisfies (6) with

$$\Psi_i(Y_i, \xi) = I(Y_i > 0)(\psi(e_i(\xi))x_i^T/g_i(\xi) + \chi(e_i(\xi))g_i^{(1)}(\xi)^T/g_i(\xi), \chi(e_i(\xi))g_i^{(2)}(\xi)^T/g_i(\xi))^T,$$

where $\psi(x) = -f_\epsilon'(x)/f_\epsilon(x)$ and $\chi(x) = x\psi(x) - 1$. The pseudo likelihood estimator (Carroll and Ruppert, 1982) satisfies (6) with

$$\Psi_i(Y_i, \xi) = I(Y_i > 0)(\psi(e_i(\xi))x_i^T, \chi(e_i(\xi))g_i^{(2)}(\xi)^T)^T.$$

When the $\{\epsilon_i\}$ are normally distributed, $\psi(x) = x$ and $\chi(x) = x^2 - 1$.

From the standard theory of estimating equations (see for example Diggle et al, 2002), we can show that under mild conditions

$$\hat{\xi} - \xi_0 = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} B_\beta \Psi_i(Y_i, \xi_0) \\ B_\theta \Psi_i(Y_i, \xi_0) \end{pmatrix} + o_p(\frac{1}{\sqrt{n}}), \tag{7}$$

where B_β and B_θ are defined by

$$\begin{pmatrix} B_\beta \\ B_\theta \end{pmatrix} = \left\{ -\frac{1}{n}\sum_{i=1}^{n} E\frac{\partial}{\partial \xi}\Psi_i(Y_i, \xi)|_{\xi=\xi_0} \right\}^{-1}.$$

Here B_β is a $q_1 \times q$ matrix, and B_θ is a $q_2 \times q$ matrix.

4. ESTIMATING THE MEAN ON THE ORIGINAL SCALE

When a linear regression model is fitted on the transformed scale, it is often of interest to use the estimated coefficients to estimate the (unconditional) mean of the response on the original-scale. That is,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

given the covariates x and z, we want to estimate $u = E(Y \mid x, z)$, where Y is the response of the outcome on the patient with the covariates x and z. Define

$$\eta_i(\xi) = x^T\beta + g(\xi)e_i(\xi),$$

where $e_i(\xi)$ is defined by (4), and $g(\xi)$ is the value of $g_i(\xi)$ when $x_i = x$ and $z_i = z$. For simplicity, we write $\eta_i(\xi_0) = \eta_i$. Since $e_i(\xi_0) = \epsilon_i$ and the $\{\epsilon_i\}$ are assumed to be independent and identically distributed, for fixed x and z, the random variables $\{\eta_i\}$ are independent and identically distributed. In this notation, we have $Eh(\eta_i) = Eh(\eta_1)$ and that

$$u = (1 - \pi)Eh(\eta_1),$$

where $\pi = l^{-1}(z^T\alpha_0)$.

We consider two different estimators of u. Both estimators are generalizations of the smearing estimator of Duan (1983). Put $\hat{\xi} = (\hat{\beta}^T, \hat{\theta}^T)^T$ and $\hat{\pi}_i = l^{-1}(z_i^T\hat{\alpha})$. Then we have the "externally" weighted estimator

$$\hat{u}^* = \frac{1 - \hat{\pi}}{1 - \bar{\hat{\pi}}}\hat{m}^*, \tag{8}$$

where $\bar{\hat{\pi}} = \frac{1}{n}\sum_{i=1}^n \hat{\pi}_i$ and $\hat{m}^* = \frac{1}{n}\sum_{i=1}^n I(y_i > 0)h(\eta_i(\hat{\xi}))$, and the "internally" weighted estimator

$$\hat{u} = (1 - \hat{\pi})\hat{m}, \tag{9}$$

where $\hat{m} = \frac{1}{n}\sum_{i=1}^n \frac{I(y_i>0)}{1-\hat{\pi}_i}h(\eta_i(\hat{\xi}))$.

5. ASYMPTOTIC PROPERTIES

To analyze the asymptotic properties of $\hat{u}^*$ and $\hat{u}$, we require conditions on the $\{\epsilon_i\}$, the covariates z_i and x_i, and smoothness conditions on g_i and h. These conditions (C) and (D) are given in Appendix A. The conditions for $\hat{u}$ are clearly stronger than those for $\hat{u}^*$. In both cases, we avoid assuming that either the covariates z_i and x_i, or the functions g_i and h are bounded; the conditions can be simplified considerably if boundedness assumptions are appropriate and if h and its derivatives are monotone. This point is also made in Duan (1983). In either case, our conditions are stronger than those of Duan (1983) because he proved only consistency of the estimator using the linear least squares estimator under a homoscedastic regression model while we establish central limit theorems using nonlinear estimators for both the regression and heteroscedasticity parameters, both of which require expansions for their treatment.

We introduce the notation

$$\mu_i(\xi) = x - \frac{g(\xi)}{g_i(\xi)}x_i, \quad \nu_i(\xi) = g^{(1)}(\xi) - \frac{g(\xi)}{g_i(\xi)}g_i^{(1)}(\xi), \text{ and } \tau_i(\xi) = g^{(2)}(\xi) - \frac{g(\xi)}{g_i(\xi)}g_i^{(2)}(\xi),$$

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

and write $\mu_i(\xi_0) = \mu_i$, $\nu_i(\xi_0) = \nu_i$ and $\tau_i(\xi_0) = \tau_i$ for simplicity.

5.1 The externally weighted estimator $\hat{u}^*$

We first consider the externally weighted estimator $\hat{u}^*$. Define

$$w^* = \begin{pmatrix} 1 \\ B_\beta^T\{Eh'(\eta_1)\bar{\mu}^* + E\epsilon_1 h'(\eta_1)\bar{\nu}^*\} \\ B_\theta^T\{E\epsilon_1 h'(\eta_1)\bar{\tau}^*\} \end{pmatrix}$$

and

$$\Omega_i^* = \begin{pmatrix} I(Y_i > 0)h(\eta_i) - (1-\pi_i)Eh(\eta_1) \\ \Psi_i(Y_i, \xi_0) \\ \Psi_i(Y_i, \xi_0) \end{pmatrix},$$

where $\bar{\mu}^* = \frac{1}{n}\sum_{i=1}^n (1-\pi_i)\mu_i$, $\bar{\nu}^* = \frac{1}{n}\sum_{i=1}^n (1-\pi_i)\nu_i$, $\bar{\tau}^* = \frac{1}{n}\sum_{i=1}^n (1-\pi_i)\tau_i$, and $\bar{\pi} = \frac{1}{n}\sum_{i=1}^n \pi_i$. Here ω^* and Ω_i^* are $(2q+1)$ dimensional vectors.

From now on, we use $\|\cdot\|$ to denote Euclidean norm for a vector and $|\cdot|$ to denote the absolute value for a scalar. The basis of our analysis is the following asymptotic linearity result.

Theorem 1

Suppose that conditions (C) hold. Then, as $n \to \infty$,

$$\hat{m}_0^* - (1-\bar{\pi})Eh(\eta_1) = \frac{1}{n}\sum_{i=1}^n w^{*T}\Omega_i^* + o_p(\frac{1}{\sqrt{n}}).$$

The proof of the theorem is given in Appendix B.

Next, put

$$\Sigma^* = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n Var(\Omega_i^*)$$

and

$$d^* = Eh(\eta_1)\left\{\frac{z}{l'(\pi)} - \frac{1}{n}\frac{1-\pi}{1-\bar{\pi}}\sum_{i=1}^n \frac{z_i}{l'(\pi_i)}\right\}.$$

Then we have the following expansion for $\hat{u}^*$.

Theorem 2

Suppose that (5), (7) and the conditions (C) hold. Then, as $n \to \infty$

$$\hat{u}^* - u = \frac{1}{n}\sum_{i=1}^n \left\{\frac{1-\pi}{1-\bar{\pi}} w^{*T}\Omega_i^* - d^{*T}A^{-1}\rho_i(I(Y_i = 0) - \pi_i)\right\} + o_p(\frac{1}{\sqrt{n}}).$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

Moreover, if for some $\lambda > 0$,

$$\lim_{n\to\infty} \frac{1}{n^{1+\lambda/2}} \sum_{i=1}^{n} E \left| \frac{1-\pi}{1-\bar{\pi}} w^{*T}\Omega_i^* + d^{*T} A^{-1} \rho_i (I(Y_i = 0) - \pi_i) \right|^{2+\lambda} = 0,$$

then

$$n^{1/2}(\hat{u}^* - u) \xrightarrow{\mathcal{D}} N\left(0, \left(\frac{1-\pi}{1-\bar{\pi}}\right)^2 w^{*T}\Sigma^* w^* + d^{*T} A^{-1} d^*\right).$$

The theorem follows from the fact that we can write

$$\begin{aligned}
&|\hat{u}^* - u - \frac{1}{n}\frac{1-\pi}{1-\bar{\pi}} \sum_{i=1}^{n} w^{*T}\Omega_i^* + \frac{1}{n} d^{*T} \sum_{i=1}^{n} A^{-1}\rho_i(I(Y_i = 0) - \pi_i)| \\
&\le |\frac{1-\hat{\pi}}{1-\bar{\hat{\pi}}}\hat{m}^* - u - \frac{1-\pi}{1-\bar{\pi}}(\hat{m}^* - (1-\bar{\pi})Eh(\eta_1)) + d^{*T}(\hat{\alpha} - \alpha_0)| \\
&+ \|d^*\| \, \|\hat{\alpha} - \alpha_0 - \frac{1}{n}\sum_{i=1}^{n} A^{-1}\rho_i(I(Y_i = 0) - \pi_i)\| \\
&+ \frac{|1-\pi|}{|1-\bar{\pi}|} \left| \hat{m}^* - (1-\bar{\pi})Eh(\eta_1) - \frac{1}{n}\sum_{i=1}^{n} w^{*T}\Omega_i^* \right|
\end{aligned}$$

and each term on the right hand side is $o_p(1)$.

Define $\hat{B}_\beta$ and $\hat{B}_\theta$ by

$$\begin{pmatrix} \hat{B}_\beta \\ \hat{B}_\theta \end{pmatrix} = \left\{ -\frac{1}{n}\sum_{i=1}^{n} E \frac{\partial}{\partial \xi} \Psi_i(Y_i, \xi)|_{\xi=\hat{\xi}} \right\}^{-1}$$

and set

$$\hat{\Omega}_i^* = \begin{pmatrix} I(Y_i > 0)h(\eta_i(\hat{\xi})) - (1-\hat{\pi}_i)\frac{1}{n}\sum_{j=1}^{n} \frac{I(Y_j>0)}{1-\hat{\pi}_j} h(\eta_j(\hat{\xi})) \\ \Psi_i(Y_i, \hat{\xi}) \\ \Psi_i(Y_i, \hat{\xi}) \end{pmatrix}.$$

Also, put $\bar{\hat{\mu}}^* = \frac{1}{n}\sum_{i=1}^{n}(1-\hat{\pi}_i)\mu_i(\hat{\xi})$, $\bar{\hat{\nu}}^* = \frac{1}{n}\sum_{i=1}^{n}(1-\hat{\pi}_i)\nu_i(\hat{\xi})$ and $\bar{\hat{\tau}}^* = \frac{1}{n}\sum_{i=1}^{n}(1-\hat{\pi}_i)\tau_i(\hat{\xi})$. Then we can estimate the asymptotic variance of $\hat{u}^*$ by

$$\hat{v}^* = \frac{1}{n}\left(\frac{1-\hat{\pi}}{1-\bar{\hat{\pi}}}\right)^2 \hat{w}^{*T}\hat{\Sigma}^*\hat{w}^* + \frac{1}{n}\hat{d}^{*T}\hat{A}^{-1}\hat{d}^*,$$

where

$$\hat{w}^* = \begin{pmatrix} 1 \\ \hat{B}_\beta^T \left\{ \left(\frac{1}{n}\sum_{i=1}^{n} \frac{I(Y_i>0)}{1-\hat{\pi}_i} h'(\eta_i(\hat{\xi}))\right) \bar{\hat{\mu}}^* + \left(\frac{1}{n}\sum_{i=1}^{n} \frac{I(Y_i>0)}{1-\hat{\pi}_i} e_i(\hat{\xi}) h'(\eta_i(\hat{\xi}))\right) \bar{\hat{\nu}}^* \right\} \\ \hat{B}_\theta^T \left\{ \left(\frac{1}{n}\sum_{i=1}^{n} \frac{I(Y_i>0)}{1-\hat{\pi}_i} e_i(\hat{\xi}) h'(\eta_i(\hat{\xi}))\right) \bar{\hat{\tau}}^* \right\} \end{pmatrix},$$

$$\hat{\Sigma}^* = \frac{1}{n}\sum_{i=1}^{n} \hat{\Omega}_i^* \hat{\Omega}_i^{*T},$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$$\hat{d}^* = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i > 0)}{1 - \hat{\pi}_i} h(\eta_i(\hat{\xi})) \right\} \left\{ \frac{z}{l'(\hat{\pi})} - \frac{1}{n} \frac{1 - \hat{\pi}}{1 - \bar{\hat{\pi}}} \sum_{i=1}^{n} \frac{z_i}{l'(\hat{\pi}_i)} \right\}$$

and

$$\hat{A} = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \frac{1}{l'(\hat{\pi}_i)^2 \hat{\pi}_i (1 - \hat{\pi}_i)}.$$

An approximate $100(1-\gamma)\%$ confidence interval for u is then given by

$$\left[\hat{u}^* - \Phi^{-1}(1-\gamma/2)\sqrt{\hat{v}^*}, \quad \hat{u}^* + \Phi^{-1}(1-\gamma/2)\sqrt{\hat{v}^*} \right],$$

where Φ is the cumulative distribution function of the standard normal distribution.

5.2 Duan's (1983) problem

Duan (1983) considered the case in which there are no zeros ($\pi_i = 0$), no heteroscedasticity ($g_i = 1$), the ϵ_i are normally distributed (at least for variance calculations), and $\hat{\beta}$ is the least squares estimator. Theorem 2 generalizes the results given in Duan (1983) for this case. Even without assuming that ϵ_i are normally distributed, we have $B_\beta^{-1} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$, $\Omega_i^* = \left(h(\eta_i) - Eh(\eta_1), x_i^T \epsilon_i \right)^T$ so that

$$\Sigma^* = \begin{pmatrix} E\{h(\eta_1) - Eh(\eta_1)\}^2 & \bar{x}^T E\{\epsilon_1 h(\eta_1)\} \\ \bar{x} E\{\epsilon_1 h(\eta_1)\} & B_\beta^{-1} E\epsilon_1^2 \end{pmatrix},$$

where $\bar{x} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i$, and

$$w^* = \begin{pmatrix} 1 \\ Eh'(\eta_1) B_\beta^T (x - \bar{x}) \end{pmatrix},$$

from which the asymptotic variance is readily obtained as $w^{*T} \Sigma^* w^*$.

Note that for Duan's problem, conditions (C) can be weakened considerably: we can replace conditions (ii)-(iv) by

(ii) The following moment conditions hold: $Eh'(\eta_1)^2 < \infty$, $E\epsilon_1^2 h'(\eta_1)^2 < \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} E \left\{ \sup_{\|b\| \le M} \left| h''(\eta_i + \frac{1}{\sqrt{n}} (x - x_i)^T b) \right| \right\}^2 = O(1)$$

and

$$\frac{1}{n^2} \sum_{i=1}^{n} E \left\{ \sup_{\|b\| \le M} \left| h''(\eta_i + \frac{1}{\sqrt{n}} (x - x_i)^T b) \right| \right\}^4 = o(1).$$

(iii) the limits $B_\beta^{-1} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$ and $\bar{x} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i$ exist.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

We can estimate the asymptotic variance of Duan's smearing estimator by $\frac{1}{n}\hat{w}^{*T}\hat{\Sigma}^*\hat{w}^*$, where

$$\hat{w}^* = \begin{pmatrix} 1 \\ \left\{\frac{1}{n}\sum_{i=1}^n h'(\eta_i(\hat{\xi}))\right\}\left\{\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right\}^{-1}(x-\bar{x}) \end{pmatrix},$$

and

$$\hat{\Sigma}^* = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} h(\eta_i(\hat{\xi})) - \frac{1}{n}\sum_{j=1}^n h(\eta_j(\hat{\xi})) \\ x_i e_i(\hat{\xi}) \end{pmatrix} \left(h(\eta_i(\hat{\xi})) - \frac{1}{n}\sum_{j=1}^n h(\eta_j(\hat{\xi})), x_i^T e_i(\hat{\xi}) \right).$$

An approximate $100(1-\gamma)\%$ confidence interval for $Eh(\eta_1)$ is then given by

$$\left[\hat{u} - \Phi^{-1}(1-\gamma/2)\sqrt{\frac{1}{n}\hat{w}^{*T}\hat{\Sigma}^*\hat{w}^*}, \quad \hat{u} + \Phi^{-1}(1-\gamma/2)\sqrt{\frac{1}{n}\hat{w}^{*T}\hat{\Sigma}^*\hat{w}^*}\right],$$

where Φ is the cumulative distribution function of the standard normal distribution.

5.3 The internally weighted estimator $\hat{u}$

Now consider the internally weighted estimator $\hat{u}$. Define

$$w = \begin{pmatrix} 1 \\ B_\beta^T\{Eh'(\eta_1)\bar{\mu} + E\epsilon_1 h'(\eta_1)\bar{\nu}\} \\ B_\theta^T\{E\epsilon_1 h'(\eta_1)\bar{\tau}\} \end{pmatrix}$$

and

$$\Omega_i = \begin{pmatrix} \frac{I(Y_i>0)}{1-\pi_i}h(\eta_i) - Eh(\eta_1) \\ \Psi_i(Y_i,\xi_0) \\ \Psi_i(Y_i,\xi_0) \end{pmatrix},$$

where $\bar{\mu} = \frac{1}{n}\sum_{i=1}^n \mu_i$, $\bar{\nu} = \frac{1}{n}\sum_{i=1}^n \nu_i$, $\bar{\tau} = \frac{1}{n}\sum_{i=1}^n \tau_i$, and $\bar{\rho}^* = \frac{1}{n}\sum_{i=1}^n \pi_i\rho_i$.

We have the following asymptotic linearity result.

THEOREM 3

Suppose that conditions (D) hold. Then, as $n \to \infty$

$$\hat{m} - Eh(\eta_1) = \frac{1}{n}\sum_{i=1}^n w^T\Omega_i + o_p(\frac{1}{\sqrt{n}}).$$

The proof of Theorem 3 is similar to the proof of Theorem 1 so is omitted.

Next, put

$$\Sigma = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n Var(\Omega_i)$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

and

$$d = Eh(\eta_1)\left\{\frac{z}{l'(\pi)} - \frac{1}{n}(1-\pi)\sum_{i=1}^{n}\frac{z_i}{l'(\pi_i)(1-\pi_i)}\right\}.$$

Then we have the following result for $\hat{u}$.

THEOREM 4

Suppose that (5), (7) and the conditions (D) hold. Then, as $n \to \infty$

$$\hat{u} - u = \frac{1}{n}\sum_{i=1}^{n}\{(1-\pi)w^T\Omega_i + d^T A^{-1}\rho_i(I(Y_i=0)-\pi_i)\} + o_p(\frac{1}{\sqrt{n}}).$$

Moreover, if for some $\lambda > 0$,

$$\lim_{n\to\infty}\frac{1}{n^{1+\lambda/2}}\sum_{i=1}^{n}E\left|(1-\pi)w^T\Omega_i + d^T A^{-1}\rho_i(I(Y_i=0)-\pi_i)\right|^{2+\lambda} = 0,$$

then

$$n^{1/2}(\hat{u}-u) \xrightarrow{\mathcal{D}} N(0,(1-\pi)^2 w^T\Sigma w + d^T A^{-1} d).$$

The result follows from the fact that we can write

$$\begin{aligned}
&|\hat{u} - u + \frac{1}{n}Eh(\eta_1)\frac{z^T}{l'(\pi)}\sum_{i=1}^{n}A^{-1}\rho_i(I(Y_i=0)-\pi_i) \\
&\quad -\frac{1}{n}(1-\pi)\sum_{i=1}^{n}\{w^T\Omega_i + Eh(\eta_1)\bar{\rho}^{*T}A^{-1}\rho_i(I(Y_i=0)-\pi_i)\}| \\
\leq\ & |(1-\hat{\pi})\hat{m} - u + (\hat{\pi}-\pi)Eh(\eta_1) - (1-\pi)(\hat{m} - Eh(\eta_0))| \\
&\quad +|Eh(\eta_1)||\hat{\pi} - \pi - \frac{1}{n}\frac{z^T}{l'(\pi)}\sum_{i=1}^{n}A^{-1}\rho_i(I(Y_i=0)-\pi_i)| \\
&\quad +|1-\pi||\hat{m} - Eh(\eta_0) - \frac{1}{n}\sum_{i=1}^{n}\{w^T\Omega_i + Eh(\eta_1)\bar{\rho}^{*T}A^{-1}\rho_i(I(Y_i=0)-\pi_i)\}|
\end{aligned}$$

and each term on the right hand side is $o_p(1)$.

We can estimate the asymptotic variance of $\hat{u}$ by

$$\hat{v} = \frac{1}{n}(1-\hat{\pi})^2\hat{w}^T\hat{\Sigma}\hat{w} + \frac{1}{n}\hat{d}^T\hat{A}^{-1}\hat{d},$$

where

$$\hat{w} = \begin{pmatrix} 1 \\ \hat{B}_\beta^T\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{I(Y_i>0)}{1-\hat{\pi}_i}h'(\eta_i(\hat{\xi}))\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i(\hat{\xi})\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{I(Y_i>0)}{1-\hat{\pi}_i}e_i(\hat{\xi})h'(\eta_i(\hat{\xi}))\right)\left(\frac{1}{n}\sum_{i=1}^{n}\nu_i(\hat{\xi})\right)\right\} \\ \hat{B}_\theta^T\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{I(Y_i>0)}{1-\hat{\pi}_i}e_i(\hat{\xi})h'(\eta_i(\hat{\xi}))\right)\left(\frac{1}{n}\sum_{i=1}\tau_i(\hat{\xi})\right)\right\} \end{pmatrix},$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\hat{\Omega}_i\hat{\Omega}_i^T,$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

with

$$\hat{\Omega}_i = \begin{pmatrix} \frac{I(Y_i>0)}{1-\hat{\pi}_i} h(\eta_i(\hat{\xi})) - \frac{1}{n}\sum_{j=1}^{n} \frac{I(Y_j>0)}{1-\hat{\pi}_j} h(\eta_j(\hat{\xi})) \\ \Psi_i(Y_i, \hat{\xi}) \\ \Psi_i(Y_i, \hat{\xi}) \end{pmatrix},$$

and

$$\hat{d} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i > 0)}{1 - \hat{\pi}_i} h(\eta_i(\hat{\xi})) \right\} \left\{ \frac{z}{l'(\hat{\pi})} - \frac{1}{n}(1 - \hat{\pi}) \sum_{i=1}^{n} \frac{z_i}{(1 - \hat{\pi}_i) l'(\hat{\pi}_i)} \right\}.$$

An approximate $100(1-\gamma)\%$ confidence interval for u is then given by

$$\left[\hat{u} - \Phi^{-1}(1-\gamma/2)\sqrt{\hat{v}}, \quad \hat{u} + \Phi^{-1}(1-\gamma/2)\sqrt{\hat{v}} \right],$$

where Φ is the cumulative distribution function of the standard normal distribution.

6. AN REAL EXAMPLE

We illustrate the application of our method in a subset from real ongoing clinical study (Fortney et al, 2003) on the impact of establishing veterans' health administration (VHA) Community Based Outpatient Clinics in underserved areas on utilization and costs. Our data set consists of 1,785 female veterans, and the outcome variable in this analysis is the year 1998 total cost for a veteran. There are 483 veterans in our sample who do not incur any cost during the year 1998, and hence they have zero cost outcomes.

In the data set we have the following important explanatory variables that have been shown to be associated with VA costs in the literature. We have information on demographics of veterans, including age, sex, race, and marital status. We also have information on the degree to which a veteran's condition was related to their military experience, as well as means test category (Category A - not service connected, Category A - service connect, Category C and Category not applicable). In addition, travel distance to the closest VHA hospital was included to control for access differences. Euclidean distance to VHA facilities for every zip code was determined using the longitude and latitude of each VHA facility and the longitude and latitude of zip code centroids, based on the ArcInfo/ArcView Geographic Information System (GIS). Finally, we have information on Diagnostic Cost Group (DCG) risk category. The risk DCG score is a very widely used diagnosis-based case-mix instrument.

Let Y_i be the total health care cost of the ith patient, and her corresponding covariates are defined as follows. X_{i1} is her travel distance to the closest VHA hospital; X_{i2} represents her 1997 DCG score; X_{i3} represents her age; X_{i4} represents her marital status; X_{i5} represents the percentage of her service connection; X_{i6} and X_{i7} represent her mean test category A NSC and category A SC, respectively. Then,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

for $i = 1, \ldots, n$, we model the probability of non-zero cost by the logistic regression model,

$$\log \frac{P(Y_i = 0 \mid X_{i1}, \ldots, X_{i7})}{P(Y_i > 0 \mid X_{i1}, \ldots, X_{i7})} = \alpha_0 + \alpha_1 X_{i1} + \ldots \alpha_7 X_{i7}, \tag{10}$$

and we model the conditional magnitude of the positive costs Y_i given $Y_i > 0$ by the log-transformed, heteroscedastic linear regression model

$$\log Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_7 X_{i7} + \exp\{(\theta_0 + \theta_1 X_{i1} + \ldots + \theta_7 X_{i7})/2\}\epsilon_i. \tag{11}$$

The parameters in the model were estimated using (6) with $\psi(x) = x$ and $\chi(x) = x^2 - 1$. The estimates for these parameters and their standard errors are given in Table 1. Using these parameter estimates, we can estimate the average cost of a patient with the given covariate values and an associated standard deviation. For example, for a 49-years old unmarried female veteran with the travel distance of 31 miles to her closest VHA hospital, not service connected, with 1997 DCG score 0.56, the estimated average cost is $\hat{u}_0^*$ = \$855.3 using the externally weighted estimator with the estimated standard deviation of \$150.9 and $\hat{u}_0$ = \$823.8 using the internally weighted estimator with the estimated standard deviation of \$134.5. According to the simulation results presented in the next section, we would use the externally weighted estimate for the average cost of such patients during the study period. Hence, we would estimate the average cost for this patient to be \$855.3 with the corresponding 95% confidence interval of (\$559.6,\$1151.1).

7. A SIMULATION STUDY

We conducted a simulation study to assess the finite-sample performance of the proposed estimators $\hat{u}^*$ and $\hat{u}$.

We adopted the two-part regression model with a continuous covariate $X_1 \sim N(2, 1)$ and a binary covariate $X_2 \sim$ Bernoulli(0.3). We used a two-stage procedure to generate the response variable. We first generated a sample of size n from a Bernoulli distribution with the probability of zero defined by (10) with $\alpha_0 = -0.5$, $\alpha_1 = -0.1$, and $\alpha_2 = -0.9$. Let n_1 denote the number of nonzero observations in the Bernoulli sample. Then we generated a random sample of n_1 errors with $\epsilon_1, \ldots, \epsilon_{n_1}$from the standard normal distribution. The logarithms of the non-zero observations were then given by (11) with $\beta_0 = 3.4$, $\beta_1 = 0.01$, $\beta_2 = -0.3$, $\theta_0 = 0.1$, $\theta_1 = 0.5$, and $\theta_2 = 0.15$. We explored the effect of sample size by letting $n = 130, 150, 200, 500, 1000$.

For each simulated dataset, we estimated the parameters using (6) with $\psi(x) = x$ and $\chi(x) = x^2 - 1$, as in the previous section, and then computed the externally weighted estimate $\hat{u}_0^*$ and the internally weighted

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

estimate $\hat{u}_0$ of $E(Y|X = x_0)$. We compare the relative performance of these two estimators in terms of bias and mean squared error (MSE) in Table 2. The results are based on $10,000$ simulated data sets.

The results suggest that, both the internally weighted estimator and the externally weighted estimator have very similar bias and MSE, and the externally weighted estimator has slightly smaller MSE than the internally weighted estimator.

8. DISCUSSION

For modelling skewed, heteroscedastic data with zeros, we used a two-part regression model which enabled us to treat the zeros and the positive observations separately. We proposed applying a transformation to the positive responses to achieve linearity, leaving us to model heteroscedasticity and handle possible non-normality explicitly . We then considered the problem of estimating the mean on the original scale. This entails bias-adjusted back transformation for the positive part of the model and adjustment for the zeros. We proposed two nonparametric estimators of the mean on the original scale. These estimators are extensions of Duan's smearing estimator to our more general context. In particular, our estimators of the mean on the original scale accommodate the zeros, the heteroscedasticity and the possible nonnormality of the positive part of our model. We showed the consistency and asymptotic normality of the two estimators and derived closed-forms for their asymptotic variances. We applied the estimators to a real data set and explored their properties in a simulation study.

A useful extension of the methodology would be to allow the transformation h to be estimated from the data. This is not conceptually difficult but would make the theory more complicated and, in particular, result in much more complicated expressions for the asymptotic variance of the estimators. Specifically, the estimation of h changes the expansion of the estimator $\hat{\xi}$ and then of the conditional mean estimators $\hat{m}_0$ and $\hat{m}_0^*$. These changes tend to increase the asymptotic variance of the estimators.

ACKNOWLEDGEMENTS

We would Like to thank two reviewers and one associate editor for their helpful comments that has results in an much improved version of the manuscript. We would also like to thank Dr. Maciejewski for providing us the example data set that was from a study funded by Department of Veterans Affairs Health Services Research and Development Service, ACC 97068-2, to Dr. Fortney. This work is supported in part by the grant AHRQ R01HS013105-01.

References

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Carroll, R.J. and Ruppert, D. (1982), "Robust estimation in heteroscedastic linear models," *The Annals of Statistics* **10** 429–441.

Carroll, R.J. and Ruppert, D. (1984), "Comment on "The analysis of transformed data" by D.V. Hinkley and G. Runger," *Journal of the American Statistical Association* **79** 312–313.

Diggle, P., Heagerty, P., Liang, K. Y. and Zeger, S. (2002), *The analysis of longitudinal data (Second Edition)*, Oxford University Press, Oxford, UK.

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

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983), "A comparison of alternative models for the demand for medical care," *Journal of Business and Economic Statistics*, **1**, 115–126.

Fortney JC, Maciejewski ML, Warren J and Burgess JF. (2003), "Does Establishing VHA Community Based Outpatient Clinics in Underserved Areas Impact Patterns of Utilization and Costs?" submitted.

McCullagh, P. and Nelder, J.A. (1989), *Generalized linear models (Second Edition)*, Chapman and Hall, London.

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

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

Panel on Nonstandard Mixtures of Distributions. (1989), "Statistical models and analysis in auditing," *Statistical Science,* 4, 2–33.

Taylor, J. M. G. (1986), "The retransformed mean after a fitted power transformation," *Journal of the American Statistical Association,* 81, 114–118.

Welsh, A.H., Carroll, R.J. and Ruppert, D. (1994), "Fitting heteroscedastic regression models," *Journal of the American Statistical Association* **89** 100–116.

Zhou, X. H., Gao, S., and Hui, S. L. (1997a), "Methods for comparing the means of two independent log-normal samples," *Biometrics,* 53, 1129–1135.

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

Zhou, X. H., Melfi, A., and Hui, S. L. (1997b), "Methods for comparison of cost data," *Annals of Internal Medicine,* 127, 752–756

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

Appendix A: Conditions

Let $k = (b^T, t^T)^T$ be a q-dimensional vector. It is tedious to keep writing $\xi_0 + \frac{1}{\sqrt{n}}k$ in the conditions and the proofs. We therefore write $g_i(k) = g_i(\xi_0 + \frac{1}{\sqrt{n}}k)$ and $g_i = g_i(\xi_0)$ and then $\eta_i(k) = \eta_i(\xi_0 + \frac{1}{\sqrt{n}}k)$, $e_i(k) = e_i(\xi_0 + \frac{1}{\sqrt{n}}k)$, $\mu_i(k) = \mu_i(\xi_0 + \frac{1}{\sqrt{n}}k)$, $\nu_i(k) = \nu_i(\xi_0 + \frac{1}{\sqrt{n}}k)$ and $\tau_i(k) = \tau_i(\xi_0 + \frac{1}{\sqrt{n}}k)$. Similarly, it is convenient to write $\pi_i(a) = l^{-1}(z_i^T\alpha_0 + \frac{1}{\sqrt{n}}z_i^T a)$. Here $g_i(k)$, g_i, $\eta_i(k)$, $e_i(k)$ are scalar, $\mu_i(k)$ and $\nu_i(k)$ are $q_1 \times 1$ vectors, and $\tau_i(k)$ is a $q_2 \times 1$ vector. To simplify our notation, we introduce the following notations:

$$\begin{aligned} H_{in}^{(r)}(\epsilon_i) &= \sup_{\|k\|\le M} \Bigg| h^{(r)} \Bigg\{ \eta_i + \frac{1}{\sqrt{n}}\mu_i(k)^T b + \frac{1}{\sqrt{n}}\nu_i(k)^T b \frac{g_i}{g_i(k)}\epsilon_i \\ &\quad - \frac{1}{n}\nu_i(k)^T b \frac{x_i^T b}{g_i(k)} + \frac{1}{\sqrt{n}}\tau_i(k)^T t \frac{g_i}{g_i(k)}\epsilon_i - \frac{1}{n}\tau_i(k)^T t \frac{x_i^T b}{g_i(k)} \Bigg\} \Bigg| \end{aligned}$$

for $r = 0, 1, 2$;

$$K_{n1}^{(c)} = \frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \frac{\|\mu_i'(k)\|^2}{(1-\pi_i)^{2-c}},$$

$$K_{n2}^{(c)} = \frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \frac{\|\nu_i'(k)\|^2}{(1-\pi_i)^{2-c}} \left(\frac{g_i}{g_i(k)}\right)^2,$$

$$K_{n3}^{(c)} = \frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \frac{\|\nu_i(k)\|^2}{(1-\pi_i)^{2-c}} \left(\frac{g_i^2}{g_i(k)^4}\right) \left\|(g_i^{(1)}(k)^T, g_i^{(2)}(k))^T\right\|^2,$$

$$K_{n4}^{(c)} = \frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \frac{\|\tau_i'(k)\|^2}{(1-\pi_i)^{2-c}} \left(\frac{g_i^2}{g_i(k)^2}\right),$$

and

$$K_{n4}^{(c)} = \frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \frac{\|\tau_i(k)\|^2}{(1-\pi_i)^{2-c}} \left(\frac{g_i^2}{g_i(k)^4}\right) \left\|(g_i^{(1)}(k)^T, g_i^{(2)}(k))^T\right\|^2,$$

where $c = 0$,2.

Conditions (C) refers to the following:

(i) The estimator $\hat{\xi}$ satisfies $\hat{\xi} - \xi_0 = O_p(\frac{1}{\sqrt{n}})$.

(ii) The following moment conditions hold:

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

(a) $E\epsilon_1^2 h'(\eta_1)^2 < \infty$ and $Eh'(\eta_1)^2 < \infty$.

(b)

$$\frac{1}{n}\sum_{i=1}^{n} E\epsilon_i^2 H_{in}^{(2)}(\epsilon_i)^2 = O(1), \quad \frac{1}{n^2}\sum_{i=1}^{n} E\epsilon_i^4 H_{in}^{(2)}(\epsilon_i)^4 = o(1)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} EH_{in}^{(2)}(\epsilon_i)^2 = O(1), \quad \frac{1}{n^2}\sum_{i=1}^{n} EH_{in}^{(2)}(\epsilon_i)^4 = o(1).$$

(iii) The following conditions hold on g_i:

$$\frac{1}{n^2}\sum_{i=1}^{n} \|\nu_i\|^2 = o(1), \quad \frac{1}{n^2}\sum_{i=1}^{n} \|\tau_i\|^2 = o(1), \quad \frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \|\mu_i(k)\|^4 = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \|\nu_i(k)\|^4 \left(\frac{g_i}{g_i(k)}\right)^4 = o(1), \quad \frac{1}{n^3}\sum_{i=1}^{n} \sup_{\|k\|\le M} \|\nu_i(k)\|^4 \left\|\frac{x_i}{g_i(k)}\right\|^4 = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \|\tau_i(k)\|^4 \left(\frac{g_i}{g_i(k)}\right)^4 = o(1) \text{ and } \frac{1}{n^3}\sum_{i=1}^{n} \sup_{\|k\|\le M} \|\tau_i(k)\|^4 \left\|\frac{x_i}{g_i(k)}\right\|^4 = o(1).$$

(iv) The following conditions hold on g_i'

$$K_{n1}^{(2)} = o(1), K_{n2}^{(2)} = o(1), K_{n3}^{(2)} = o(1), K_{n4}^{(2)} = o(1), \text{ and } K_{n5}^{(2)} = o(1).$$

Note that Conditions (iib) are implied by $\frac{1}{n}\sum_{i=1}^{n} E\epsilon_i^4 H_{in}^{(2)}(\epsilon_i)^4 = O(1)$ and $\frac{1}{n}\sum_{i=1}^{n} EH_{in}^{(2)}(\epsilon_i)^4 = O(1)$ because then

$$\frac{1}{n}\sum_{i=1}^{n} E\epsilon_i^2 H_{in}^{(2)}(\epsilon_i)^2 \le \left(\frac{1}{n}\sum_{i=1}^{n} E\epsilon_i^4 H_{in}^{(2)}(\epsilon_i)^4\right)^{1/2} = O(1)$$

and

$$\frac{1}{n^2}\sum_{i=1}^{n} E\epsilon_i^4 H_{in}^{(2)}(\epsilon_i)^4 = O\left(\frac{1}{n}\right).$$

Similarly, since

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n} \|\mu_i\|^2 &\le \frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|\le M} \|\mu_i(k)\|^2 \\ &\le \left(\frac{1}{n^3}\sum_{i=1}^{n} \sup_{\|k\|\le M} \|\mu_i(k)\|^4\right)^{1/2} = o\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

the conditions (iii) imply the convergence to zero of similar terms with squared instead of fourth power summands.

The conditions for the internally weighted estimator are stronger than conditions (C). We require the following conditions (D).

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

(i) The estimator $\hat{\xi}$ satisfies $\hat{\xi} - \xi_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\hat{\alpha}$ satisfies $\hat{\alpha} - \alpha_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$.

(ii) The following moment conditions hold:

(a) $Eh(\eta_1)^2 < \infty$, $E\epsilon_1^2 h'(\eta_1)^2 < \infty$ and $Eh'(\eta_1)^2 < \infty$.

(b)

$$\frac{1}{n}\sum_{i=1}^{n} EH_{in}^{(r)}(\epsilon_i)^2 = O(1), \quad \frac{1}{n^2}\sum_{i=1}^{n} EH_{in}^{(r)}(\epsilon_i)^4 = o(1),$$

for $r = 0, 1, 2$ and

$$\frac{1}{n}\sum_{i=1}^{n} E\epsilon_i^{2r} H_{in}^{(r)}(\epsilon_i)^2 = O(1), \quad \frac{1}{n^2}\sum_{i=1}^{n} E\epsilon_i^{4r} H_{in}^{(r)}(\epsilon_i)^4 = o(1)$$

for $r = 1, 2$.

(iii)

$$\frac{1}{n^2}\sum_{i=1}^{n} \frac{\|z_i\|^2}{(1-\pi_i)^3 l'(\pi_i)^2} = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|a\|\le M} \frac{\|z_i\|^4}{(1-\pi_i(a))^6 l'(\pi_i(a))^4} = o(1),$$

and

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|a\|\le M} \frac{\|z_i\|^4 l''(\pi_i(a))^2}{(1-\pi_i(a))^4 l'(\pi_i(a))^6} = o(1),$$

(iv) The following conditions hold on g_i:

$$\frac{1}{n^2}\sum_{i=1}^{n} \frac{\|\nu_i\|^2}{1-\pi_i} = o(1), \quad \frac{1}{n^2}\sum_{i=1}^{n} \frac{\|\tau_i(k)\|^2}{1-\pi_i} = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\mu_i(k)\|^4}{(1-\pi_i(a))^2} = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\nu_i(k)\|^4}{(1-\pi_i(a))^2}\left(\frac{g_i}{g_i(k)}\right)^4 = o(1),$$

$$\frac{1}{n^4}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\nu_i(k)\|^4}{(1-\pi_i(a))^2}\left\|\frac{x_i}{g_i(k)}\right\|^4 = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\tau_i(k)\|^4}{(1-\pi_i(a))^2}\left(\frac{g_i}{g_i(k)}\right)^4 = o(1)$$

and

$$\frac{1}{n^4}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\tau_i(k)\|^4}{(1-\pi_i(a))^2}\left\|\frac{x_i}{g_i(k)}\right\|^4 = o(1).$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

(v) The following conditions hold on g_i'

$$K_{n1}^{(0)} = o(1), K_{n2}^{(0)} = o(1), K_{n3}^{(0)} = o(1), K_{n4}^{(0)} = o(1), \text{ and } K_{n5}^{(0)} = o(1).$$

(vi) The following joint conditions hold:

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\mu_i(k)\|^2\|z_i\|^2}{(1-\pi_i(a))^4 l'(\pi_i(a))^2} = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\nu_i(k)\|^2\|z_i\|^2}{(1-\pi_i(a))^4 l'(\pi_i(a))^2} \left(\frac{g_i}{g_i(k)}\right)^2 = o(1),$$

$$\frac{1}{n^4}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\nu_i(k)\|^2\|z_i\|^2}{(1-\pi_i(a))^4 l'(\pi_i(a))^2} \left\|\frac{x_i}{g_i(k)}\right\|^2 = o(1),$$

$$\frac{1}{n^2}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\tau_i(k)\|^2\|z_i\|^2}{(1-\pi_i(a))^4 l'(\pi_i(a))^2} \left(\frac{g_i}{g_i(k)}\right)^2 = o(1)$$

and

$$\frac{1}{n^4}\sum_{i=1}^{n} \sup_{\|k\|,\|a\|\le M} \frac{\|\tau_i(k)\|^2\|z_i\|^2}{(1-\pi_i(a))^4 l'(\pi_i(a))^2} \left\|\frac{x_i}{g_i(k)}\right\|^2 = o(1).$$

Appendix B: Proof of Theorem 1

Proof. Our proof is based on a Taylor expansion of $h(\eta_i(\hat{\xi}))$ in $\hat{m}_0^* = \frac{1}{n}\sum_{i=1}^{n} I(Y_i > 0)h(\eta_i(\hat{\xi}))$. To write down this Taylor expansion we need to compute the derivative of $\eta_i(\xi)$ w.r.t ξ. Note that $\eta_i(\xi)$ can be written as

$$\eta_i(\xi) = x_0^T\beta + \frac{g_0(\xi)}{g_i(\xi)}(h^{-1}(Y_i) - x_i^T\beta).$$

We can show that

$$\frac{\partial \eta_i(\xi)}{\partial \beta} = (x_0 - \frac{g_0(\xi)}{g_i(\xi)}x_i) + \frac{g_0^{(1)}(\xi)g_i(\xi) - g_0(\xi)g_i^{(1)}(\xi)}{g_i^2(\xi)}(h^{-1}(Y_i) - x_i^T\beta) = \mu_i(\xi) + \nu_i(\xi)e_i(\xi)$$

and that

$$\frac{\partial \eta_i(\xi)}{\partial \theta} = \frac{g_0^{(2)}(\xi)g_i(\xi) - g_0(\xi)g_i^{(2)}(\xi)}{g_i^2(\xi)}(h^{-1}(Y_i) - x_i^T\beta) = \tau_i(\xi)e_i(\xi).$$

For $k = (b^T, t^T)^T$, put

$$\begin{aligned} T(k) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(y_i > 0)h(\eta_i(k)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(y_i > 0)h(\eta_i) \\ &- \{Eh'(\eta_1)\bar{\mu}^* + E\epsilon_1 h'(\eta_1)\bar{\nu}^*\}^T b - E\epsilon_1 h'(\eta_1)\bar{\tau}^{*T} t \end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

and $\delta_i = \eta_i(k) - \eta_i$. Then

$$\begin{aligned}
|T(k)| &\leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(y_i > 0)\{h(\eta_i(k)) - h(\eta_i) - \delta_i h'(\eta_i)\} \right| \\
&\quad + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(y_i > 0) h'(\eta_i)[\delta_i - \frac{1}{\sqrt{n}}\{\mu_i + \nu_i \epsilon_i\}^T b - \frac{1}{\sqrt{n}} \tau_i \epsilon_i^T t] \right| \\
&\quad + M \left\| \frac{1}{n} \sum_{i=1}^{n} \mu_i \{I(y_i > 0) h'(\eta_i) - (1 - \pi_i) E h'(\eta_1)\} \right\| \\
&\quad + M \left\| \frac{1}{n} \sum_{i=1}^{n} \nu_i \{I(y_i > 0) \epsilon_i^T h'(\eta_i) - (1 - \pi_i) E \epsilon_1 h'(\eta_1)\} \right\| \\
&\quad + M \left\| \frac{1}{n} \sum_{i=1}^{n} \tau_i \{I(y_i > 0) \epsilon_i^T h'(\eta_i) - (1 - \pi_i) E \epsilon_1 h'(\eta_1)\} \right\| \\
&= T_1(k) + T_2(k) + T_3 + T_4 + T_5,
\end{aligned}$$

say, and the result will follow if we can show that

$$\sup_{\|k\| \leq M} T_j(k) = o_p(1), \quad j = 1, \ldots, 5.$$

The terms T_3-T_5 involve weighted averages of independent random variables with mean zero and variances which converge to zero by conditions (iii) and (iv) so they converge in probability to zero.

Next, note that for $\|\tilde{k}\| \leq \|k\| \leq M$, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}} \delta_i &= \mu_i(\tilde{k})^T b + \nu_i(\tilde{k})^T b \frac{g_i}{g_i(\tilde{k})} \epsilon_i - \frac{1}{\sqrt{n}} \nu_i(\tilde{k})^T b \frac{x_i^T b}{g_i(\tilde{k})} \\
&\quad + \tau_i(\tilde{k})^T t \frac{g_i}{g_i(\tilde{k})} \epsilon_i - \frac{1}{\sqrt{n}} \tau_i(\tilde{k})^T t \frac{x_i^T b}{g_i(\tilde{k})}.
\end{aligned}$$

For notational simplicity, we drop the tilde and simply write $\tilde{k}$ as k. Then

$$\begin{aligned}
T_2(k) &\leq M \left\| \frac{1}{n} \sum_{i=1}^{n} (\mu_i(k) - \mu_i) I(y_i > 0) h'(\eta_i) \right\| \\
&\quad + M \left\| \frac{1}{n} \sum_{i=1}^{n} (\nu_i(k) \frac{g_i}{g_i(k)} - \nu_i) I(y_i > 0) \epsilon_i h'(\eta_i) \right\| \\
&\quad + M^2 \frac{1}{n^{3/2}} \sum_{i=1}^{n} \|\nu_i(k)\| \left\| \frac{x_i}{g_i(k)} \right\| |I(y_i > 0) h'(\eta_i)| \\
&\quad + M \left\| \frac{1}{n} \sum_{i=1}^{n} (\tau_i(k)^T \frac{g_i}{g_i(k)} - \tau_i) I(y_i > 0) \epsilon_i h'(\eta_i) \right\| \\
&\quad + M^2 \frac{1}{n^{3/2}} \sum_{i=1}^{n} \|\tau_i(k)\| \left\| \frac{x_i}{g_i(k)} \right\| |I(y_i > 0) h'(\eta_i)|. \qquad (12)
\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Now, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\{\mu_i(k)-\mu_i\}I(y_i>0)h'(\eta_i)\right\| &\leq \left(\frac{1}{n}\sum_{i=1}^{n}\|\mu_i(k)-\mu_i\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}h'(\eta_i)^2\right)^{1/2}\\ &\leq M\left(\frac{1}{n^2}\sum_{i=1}^{n}\sup_{\|k\|\leq M}\|\mu_i'(k)\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}h'(\eta_i)^2\right)^{1/2}\end{aligned}$$

so the first term in (12) converges to zero by conditions (iii) and (v). Similarly, the remaining terms in (12) converge to zero.

Finally,

$$\begin{aligned}T_1(k) &\leq 2M^2\frac{1}{n^{3/2}}\sum_{i=1}^{n}\|\mu_i(k)\|^2|h''(\eta_i+s_i\delta_i)|\\ &+ 2M^2\frac{1}{n^{3/2}}\sum_{i=1}^{n}\left\|\nu_i(k)\frac{g_i}{g_i(k)}\right\|^2\epsilon_i^2|h''(\eta_i+s_i\delta_i)|\\ &+ 2M^4\frac{1}{n^{5/2}}\sum_{i=1}^{n}\|\nu_i(k)\|^2\left\|\frac{x_i}{g_i(k)}\right\|^2|h''(\eta_i+s_i\delta_i)|\\ &+ M^2\frac{1}{n^{3/2}}\sum_{i=1}^{n}\left\|\tau_i(k)\frac{g_i}{g_i(k)}\right\|^2\epsilon_i^2|h''(\eta_i+s_i\delta_i)|\\ &+ M^4\frac{1}{n^{5/2}}\sum_{i=1}^{n}\|\tau_i(k)\|^2\left\|\frac{x_i}{g_i(k)}\right\||h''(\eta_i+s_i\delta_i)| \qquad (13)\end{aligned}$$

Now argue as before to bound h'' by $H_{in}^{(2)}(\epsilon_i)$ so

$$\begin{aligned}\frac{1}{n^{3/2}}\sum_{i=1}^{n}\|\mu_i(k)\|^2|h''(\eta_i+s_i\delta_i)| &\leq \frac{1}{n^{3/2}}\sum_{i=1}^{n}\|\mu_i(k)\|^2H_{in}^{(2)}(\epsilon_i)\\ &\leq \left(\frac{1}{n^2}\sum_{i=1}^{n}\|\mu_i(k)\|^4\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}H_{in}^{(2)}(\epsilon_i)^2\right)^{1/2} \qquad (14)\end{aligned}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}H_{in}^{(2)}(\epsilon_i)=\frac{1}{n}\sum_{i=1}^{n}EH_{in}^{(2)}(\epsilon_i)+\frac{1}{n}\sum_{i=1}^{n}\{H_{in}^{(2)}(\epsilon_i)-EH_{in}^{(2)}(\epsilon_i)\}=O(1)$$

by condition (iiib) so (14) is $o_p(1)$ by condition (iv). Similar arguments establish that the remaining terms in (13) are also $o_p(1)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215

Table 1. Parameter Estimates and Associated Standard Deviations for VA Health Care Costs (n = 1785)

	Parameter estimate (standard deviation)
Covariates in the logistic regression model for non-zero versus zero costs	
Constant	-0.4308 (0.2069)
Travel Distance (miles)	0.0033 (0.0025)
1997 DCG score	-0.4393 (0.1481)
Age	0.0115 (0.0037)
Married	-0.8636 (0.1376)
Service connection (%)	-0.0018 (0.0055)
Means test	
- Cat A NSC (%)	-1.4901 (0.1632)
- Cat A SC (%)	-1.9371 (0.2558)
Covariates in the regression model for positive costs	
Constant	5.8311 (0.1306)
Travel Distance (miles)	0.0041 (0.0015)
1997 DCG score	0.9486 (0.0698)
Age	-0.0017 (0.0023)
Married	0.2082 (0.0700)
Service connection (%)	0.0078 (0.0019)
Mean test	
- Cat A NSC (%)	0.6745 (0.0884)
- Cat A SC (%)	0.7026 (0.1058)
Covariates in the variance model for positive costs	
Constant	0.5989 (0.1492)
Travel Distance (miles)	-0.0024 (0.0017)
1997 DCG score	0.1737 (0.0720)
Age	-0.0007 (0.0027)
Married	-0.1342 (0.0827)
Service connection	0.0013 (0.0022)
Mean test	
- Cat A NSC (%)	-0.3238 (0.1033)
- Cat A SC (%)	-0.2571 (0.1229)

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 2. Simulation results for $\hat{u}_0^*$ and $\hat{u}_0$ estimating the average cost for patients with covariates $x_0 = (x_{01}, x_{02})$

$x_{01} = 1.00$ and $x_{02} = 0$										
	n=130		n=150		n=200		n=500		n=1000	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{u}_0^*$	0.025	328.25	-0.425	287.24	-0.103	223.32	-0.314	72.86	-0.201	38.08
$\hat{u}_0$	0.038	329.98	-0.412	288.95	-0.092	224.76	-0.309	72.92	-0.199	38.14
$x_{01} = 2.00$ and $x_{02} = 0$										
	n=130		n=150		n=200		n=500		n=1000	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{u}_0^*$	-5.055	1624.73	-4.684	1436.66	-2.872	1132.14	-1.562	501.94	-1.367	235.35
$\hat{u}_0$	-4.996	1651.20	-4.622	1478.44	-2.801	1159.98	-1.534	502.23	-1.358	237.27

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/uwbiostat/paper215