4-17-2009

A Semi-Parametric Two-Part Mixed-Effects Heteroscedastic Transformation Model for Correlated Right-Skewed Semi-Continuous Data

Huazhen Lin
Sichuan University

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

A semi-parametric two-part mixed-effects heteroscedastic transformation model for correlated right-skewed semi-continuous data

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

SUMMARY. In longitudinal or hierarchical structure studies, we often encounter a semi-continuous variable that has a certain proportion of a single value and a continuous and skewed distribution among the rest of values. In the paper, we propose a new semi-parametric two-part mixed-effects transformation model to fit correlated skewed semi-continuous data. In our model, we allow the transformation to be non-parametric. Fitting the proposed model faces computational challenges due to intractable numerical integrations. We derive the estimates for the parameter and the transformation function based on an approximate likelihood, which has high order accuracy but less computational burden. We also propose an estimator for the expected value of the semi-continuous outcome on the original-scale. Finally, we apply the proposed methods to a clinical study on effectiveness of a collaborative care treatment on late life depression on health care costs.

KEY WORDS: Semi-continuous; Right-skewed; Mixed-effects; Transformation model; Semi-parametric; Laplace approximation.

1 Introduction

This study is motivated by an analysis to examine the effectiveness of the Improving Mood-Promoting Access to Collaborative Treatment (IMPACT) program for late-life

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

depression (Unutzer et al. 2002). Intervention patients had access for up to 12 months to a depression care manager who offered education, care management, and support of antidepressant management. One primary outcome, the total inpatient cost over a half year period, was collected at month 6, 12, 18 and 24.

We are interested in assessing the cost difference between intervention and control groups and how the difference changes with patient's covariates. This problem can be considered as a special case of inference on a change in the mean cost associated with a change in one or more covariates (e.g. increase in depression; comparison of treatment groups). Statistically, we need to develop accurate regression models for the mean function $\mu(x) = E(Y \mid X = x)$. The main challenge for such an estimation is how to deal with three analytic problems: correlated data, zero inpatient costs for some patients, and a highly skewed distribution of non-zero costs. Unlike estimation on regression coefficients, estimation of $\mu(x)$ may be sensitive to how to treat the correlation and skewness (Manning, 1998; Mullahy, 1998; Blough et al., 1999; Manning et al., 2005).

In the literature, a continuous variable with addition zero values is also called a semi-continuous variable. For cross-sectional data, a two-part model, which has a long history in econometrics, is most appropriate for dealing with semi-continuous data. The two-part model assumes that a semi-continuous response results from two processes: one determining whether the response is zero, and the other determining the actual level if it is non-zero (Duan et al., 1983; Manning et al., 1981; Manning, 1998; Mullahy, 1998). Olsen & Schafer (2001) extended the two-part model to longitudinal data by introducing random effects into the two-part model. Tooze et al. (2002) independently developed a similar extension of the two-part model. Albert & Shen (2005) further extended Olsen & Schafer's model to incorporate serial correlations. All these mixed-effect two-part models use a linear normal model to fit the actual level of non-zero observations, which may not be appropriate for highly skewed data.

Since the transformation of Y can simplify the relationship of Y and $\mathbf{X}$ by inducing a particular type of distribution, e.g. normal, homoscedastic, symmetric distribution, or remove extreme skewness so that more efficient estimators and more appropriate plotting can be obtained (Ruppert, 2001), econometricians and statisticians have historically relied on logarithmic or other specific transformations of Y, followed by regression of the transformed Y on $\mathbf{X}$ using Ordinary Least Square (OLS) estimation, to overcome problems of heteroscedasticity, severe skewness, and kurtosis (Box & Cox, 1964; Duan, 1983; Ruppert, 2001; Manning, 1998; Manning & Mullahy, 2001). Since the parametric transformation in OLS is not based upon any meaningful mechanism and may not be reasonable, Horowitz (1996), Cheng (2002) and Zhou et al. (2008) proposed nonparametric transformation models for non-zero cost data in cross-sectional studies. In the paper, we extend Olsen & Schafer's parametric two-part mixed-effects model to a semi-parametric transformation two-part mixed-effects model.

Fitting our semi-parametric two-part mixed-effects transformation model faces a computational challenge because of intractable numerical integration, which is also encountered in generalized linear random effects models and nonlinear variance component models. In the paper, by transforming the integral in the likelihood function to a "conditional expectation," we obtain an approximation to the likelihood function that has a closed form. The simulation shows that our approximation is even more accurate than the sixth-order Laplace approximation in finite sample sizes. However, the computational requirement on our accurate approximation is minimal; we only need to evaluate first and second derivatives and maximize the two integrands.

This paper is organized as follows. In Section 2, we derive the estimates for the regression parameters and the transformation function based on the approximate log-likelihood and a system of estimating equations. In Section 3, we present a method for calculating the unbiased estimator for the mean of the untransformed cost of a patient given the patient's covariates. In Sections 2 and 3, we also show that under some regularity conditions that our estimators for the unknown transformation function and the mean of the untransformed scale are asymptotically normal, both with the parametric rate of $O(n^{-1/2})$. We report results of simulation studies on the accuracy of our approximation and the robustness and efficiency of our method in Section 4. Finally, we apply our methods to the IMPACT data in Section 5.

2 Model and Estimation

2·1 *Notation and model*

Let Y_{ij} denote a semi-continuous response for subject i at occasion j, where $i = 1, \cdots, n$, and $j = 1, \cdots, n_i$. This response can be recorded as two different responses,

$$U_{ij} = \begin{cases} 1 & \text{if} \quad Y_{ij} \neq 0 \\ 0 & \text{if} \quad Y_{ij} = 0 \end{cases}, \quad \text{and} \quad V_{ij} = \begin{cases} Y_{ij} & \text{if} \quad Y_{ij} \neq 0 \\ \text{irrelevant} & \text{if} \quad Y_{ij} = 0 \end{cases}.$$

We model these two responses by a pair of correlated random-effects models: one for the probability that $U_{ij} = 1$, and one for the continuous response V_{ij}. Let δ_{1i} and δ_{2i} be the random effects due to subject i for the two parts. We allow δ_{1i} and δ_{2i} to be correlated, reflecting possible correlations across the two parts of the model. Denote $\delta_i = (\delta_{1i}, \delta_{2i})'$ and assume that δ_i's are i.i.d. with the density function $f(\delta_i; \psi)$. A common choice of f is a multivariate normal distribution with zero mean vector and covariance matrix, $\psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi'_{12} & \psi_{22} \end{pmatrix}$, where ψ_{11} and ψ_{22} are the covariance matrices of δ_{1i} and δ_{2i}, respectively. We assume that Y_{ij}'s are conditionally independent, given δ_i. Let $\pi_{ij}(\delta_{1i}) = P(U_{ij} = 1|\delta_{1i})$. The first part of the two-part model predicts the probability of having a non-zero cost by the following mixed-effects model:

$$\eta(\pi_{ij}(\delta_{1i})) = X'_{1ij}\alpha + Z'_{1ij}\delta_{1i}, \tag{2.1}$$

where α is a vector of unknown parameters, and η is a known link function. A common choice of η is the logistic function $\eta(x) = \log(x/(1-x))$, but other choices are possible.

The second part of the two-part model predicts the continuous response by the following model:

$$h(V_{ij}) = b_0 + X'_{2ij}\beta + Z'_{2ij}\delta_{2i} + \varepsilon_{ij}, \tag{2.2}$$

where h is a monotone increasing but unknown transformation function, satisfying $h(0) = -\infty$, that makes the distribution of the error term ε_{ij} to be the normal distribution with mean zero and variance σ^2, β is a vector of unknown parameters, ε_{ij} and δ_{2i} are independent. b_0 is a known constant for identifiability. The requirement $h(0) = -\infty$ ensures that $\Phi(a + h(0)) = 0$ for any finite a, where Φ is the distribution function of the standard normal variable.

Let $U_i = (U_{i1}, \ldots, U_{in_i})'$, $X_{1i} = (X_{1i1}, \ldots, X'_{1in_i)})$, and $Z_{1i} = (Z_{1i1}, \ldots, Z_{1in_i})'$. Denote $h(V_i)$, V_i, X_{2i}, and Z_{2i} to be the vectors or matrices of all relevant values of $h(V_{ij})$, V_{ij}, X_{2ij}, and Z_{2ij} for subject i with $U_{ij} = 1$, respectively.

Let $\Theta = (\beta, \alpha, \sigma, \psi)$. Hence Θ and h are the unknown parameters and function to be estimated in our two-part mixed-effects transformation regression model, defined by (2.1) and (2.2). In the rest of the paper, we denote the $(k_1 + k_2 + \cdots)$th order partial derivative of a function $f(x_1, x_2, \cdots)$ by $f^{(k_1, k_2, \cdots)}(x_1, x_2, \cdots)$; that is, $f^{(k_1,k_2,\cdots)}(x_1, x_2, \cdots) = \frac{d^{(k_1+k_2+\cdots)} f(x_1,x_2,\cdots)}{dx_1^{k_1} dx_2^{k_2} \cdots}$.

2·2 *The approximate likelihood for the parameter vector Θ given h*

In this section, we propose a likelihood-based estimation method for Θ given that h is known. Given h, the marginal likelihood for the model defined by (2.1) and (2.2) can be expressed as follows:

$$L = \prod_{i=1}^{n} \int f(U_i|\delta_{1i}) f(h(V_i)|\delta_{1i}, \delta_{2i}) f(\delta_{1i}, \delta_{2i}) d\delta_{1i} d\delta_{2i}.$$

Clearly, $f(h(V_i)|\delta_{1i}, \delta_{2i})f(\delta_{1i}, \delta_{2i})$ can be identified as the joint density of $(h(V_i), \delta_{1i}, \delta_{2i})$ and can be further written as $f(\delta_{2i}|h(V_i), \delta_{1i})f(h(V_i), \delta_{1i})$. Since $\int f(\delta_{2i}|h(V_i), \delta_{1i})d\delta_{2i} = 1$, we can further write the marginal likelihood function as follows:

$$L = \prod_{i=1}^{n} \int f(U_i|\delta_{1i}) f(h(V_i), \delta_{1i}) d\delta_{1i} = \prod_{i=1}^{n} f(h(V_i)) \int f(U_i|\delta_{1i}) f(\delta_{1i}|h(V_i)) d\delta_{1i}, \quad (2.3)$$

where $f(U_i|\delta_{1i}) = \exp\left\{\sum_{j=1}^{n_i}(U_{ij}\eta(\pi_{ij}(\delta_{1i})) + \log(1 - \pi_{ij}(\delta_{1i})))\right\}$ comes from the model that describes the probability of being a zero observation.

When the dimension of δ_{1i} is high, ML estimation of Θ becomes difficult because of the intractable numerical integration in (2.3). One method to avoid this intractable numerical integration is to use a Bayesian simulation method with an MCMC algorithm. However, with an unknown transformation function h, the Bayesian MCMC algorithm may also be time-consuming.

The problem of intractable integration in (2.3) is closely related to that in the marginal likelihood of a generalized linear mixed model. In the literature on maximizing the marginal likelihood of a generalized linear mixed model, several authors have proposed several methods for approximating the integrands in the marginal likelihood functions, including Gauss-Hermite quadrature (Anderson & Aitkin, 1985; Hedeker & Gibbons, 1994) and second-order Laplace approximations (Solomon & Cox, 1992; Liu & Pierce, 1993; Breslow & Clayton, 1993). In general, the Laplace method is easier to implement than the quadrature method, while the quadrature method is more accurate than the Laplace approximation. Recently, Raudenbush et al. (2000) proposed the sixth-order Laplace approximation.

Olsen & Schafer (2001) applied the sixth-order Laplace approximation to the parametric two-part homoscedastic mixed-effects model for the semi-continuous data.

Through the simulation, Raudenbush et al. (2000) found that the sixth-order Laplace approximation is as accurate as the quadrature method but with much less computational time. However, computation of the first to sixth derivatives, required by the sixth-order Laplace approximation, is also a difficult task in our case.

Based on the idea proposed by Tierney and Kadane (1986), we propose a new approximation to the integration in (2.3), which only requires evaluation of first and second derivatives. Through simulations, we find our approximation is more accurate than the sixth-order Laplace approximation in finite sample sizes. We achieve this accurate approximation by writing the integral in (2.3) as the ratio of two integrals,

$$\int f(U_i|\delta_{1i}) f(\delta_{1i}|h(V_i)) d\delta_{1i} \quad = \quad \frac{\int f(U_i|\delta_{1i}) f(\delta_{1i}, h(V_i)) d\delta_{1i}}{\int f(\delta_{1i}, h(V_i)) d\delta_{1i}}. \tag{2.4}$$

We approximate the numerator and denominator in (2.4), respectively, by Laplace's approximation, instead of directly approximating $\int f(U_i|\delta_{1i}) f(\delta_{1i}|h(V_i)) d\delta_{1i}$, as done in a standard Laplace's approximation. In taking the ratio of these two approximations, we can cancel some portion of these residual errors. As a result, we can improve the order of accuracy of the approximation for the ratio.

Next we give a formal statement of the proposed approximation. Denote

$$D_i = \text{diag}(\sigma^2, \cdots, \sigma^2), \quad \Sigma_i = D_i + Z_{2i}\psi_{22}Z'_{2i}, \quad B_i = \psi_{11} - \psi_{12}Z'_{2i}\Sigma_i^{-1}Z_{2i}\psi_{21},$$

$$\pi_i(\delta_{1i}) = (\pi_{i1}(\delta_{1i}), \cdots, \pi_{in_i}(\delta_{1i}))', \quad \Delta_i = \begin{pmatrix} \Sigma_i & Z_{2i}\psi_{21} \\ \psi_{12}Z'_{2i} & \psi_{11} \end{pmatrix},$$

$$\Pi_i(\delta_{1i}) = diag\{\Pi_{ij}(\delta_{1i}), j = 1, \cdots, n_i\}, \quad \text{and} \quad \Pi_{ij}(\delta_{1i}) = \pi_{ij}(\delta_{1i})(1 - \pi_{ij}(\delta_{1i})).$$

Let $\tau_i^*(\delta_{1i}) \equiv \log\{f(U_i|\delta_{1i}) f(\delta_{1i}, h(V_i))\}$ and $\tau_i(\delta_{1i}) \equiv \log\{f(\delta_{1i}, h(V_i))\}$, which correspond to the integrands of the numerator and the denominator in (2.4), respectively. Let $\widehat{\delta}_{1i}^*$ and $\widehat{\delta}_{1i}$ be the modes of $\tau_i^*(\delta_{1i})$ and $\tau_i(\delta_{1i})$, respectively. We obtain $\widehat{\delta}_{1i}$ by solving the equation $\frac{\partial \tau_i(\delta_{1i})}{\partial \delta_{1i}} = 0$, which has an explicit solution,

$$\widehat{\delta}_{1i} = \psi_{12}Z'_{2i}\Sigma_i^{-1}\left(h(V_i) - b_0 1 - X_{2i}\beta\right),$$

where *1* is the vector with all of the component to be 1. By setting $\frac{\partial \tau_i^*(\delta_{1i})}{\partial \delta_{1i}} = 0$, we obtain $\widehat{\delta}_{1i}^*$ by iteratively solving the following equation:

$$\widehat{\delta}_{1i}^* = B_i Z_{1i}' \left(U_i - \pi_i(\widehat{\delta}_{1i}) \right) + \psi_{12} Z_{2i}' \Sigma_i^{-1} \left(h(V_i) - b_0 \mathit{1} - X_{2i}\beta \right).$$

From (2.4), we see that we can further write the integral in (2.3) as the following ratio of the two integrals:

$$\int f(U_i|\delta_{1i}) f(\delta_{1i}|h(V_i)) d\delta_{1i} \quad = \quad \frac{\int \exp(\tau_i^*(\delta_{1i})) d\delta_{1i}}{\int \exp(\tau_i(\delta_{1i})) d\delta_{1i}}. \tag{2.5}$$

Following the same idea as in Tierney and Kadane (1986), we first derive the second-order Laplace's approximations to the numerator and denominator of the ratio in (2.5). Then, taking the ratio of the two approximations, we have a new approximation for the integral in (2.3),

$$\begin{aligned} \int f(U_i|\delta_{1i}) f(\delta_{1i}|h(V_i)) d\delta_{1i} = \left(\frac{| - \tau_i^{(2)}(\widehat{\delta}_{1i})|}{| - \tau_i^{(2)}(\widehat{\delta}_{1i}^*)|} \right)^{1/2} & \exp\{\tau_i(\widehat{\delta}_{1i}^*) - \tau_i(\widehat{\delta}_{1i})\} \\ & \times \left(1 + \frac{a^* - a}{n_i} + O(n_i^{-2}) \right), \end{aligned}$$

where $a = g(\tau_i^{(2)}(\widehat{\delta}_{1i}), \tau_i^{(3)}(\widehat{\delta}_{1i}), \tau_i^{(4)}(\widehat{\delta}_{1i}))$, $a^* = g(\tau_i^{(*2)}(\widehat{\delta}_{1i}^*), \tau_i^{(*3)}(\widehat{\delta}_{1i}^*), \tau_i^{(*4)}(\widehat{\delta}_{1i}^*))$, and g is a known function. For example, when δ_{1i} is one-dimension, denote $\tau_i^{(k)} = \tau_i^{(k)}(\widehat{\delta}_{1i})$, $\sigma^2 = -\left(\tau_i^{(2)}\right)^{-1}$, we have $a = \frac{1}{8}\sigma^4 \tau_i^{(4)} + \frac{5}{24}\sigma^6 \left(\tau_i^{(3)}\right)^2$, a^* is defined in the same way except that τ_i and $\widehat{\delta}_{1i}$ are replaced by τ_i^* and $\widehat{\delta}_{1i}^*$. In Appendix A, we show that our new approximation has the error of order $O(n_i^{-3/2})$,

$$\begin{aligned} \int f(U_i|\delta_{1i}) f(\delta_{1i}|h(V_i)) d\delta_{1i} = \left(\frac{| - \tau_i^{(2)}(\widehat{\delta}_{1i})|}{| - \tau_i^{(2)}(\widehat{\delta}_{1i}^*)|} \right)^{1/2} & \exp\{\tau_i(\widehat{\delta}_{1i}^*) - \tau_i(\widehat{\delta}_{1i})\} \\ & \times \left(1 + O(n_i^{-3/2}) \right). \end{aligned} \tag{2.6}$$

The simulations in Section 4 also demonstrate that our approximation is more accurate than the six-order Laplace's method in finite sample sizes.

Based on (2.6), we obtain the following final approximate likelihood:

$$
\begin{aligned}
l(\Theta; h) = &-\frac{1}{2}\sum_{i=1}^{n} \log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{n} \log|B_i| \\
&-\frac{1}{2}\sum_{i=1}^{n} (h(V_i) - b_0 \mathit{1} - X_{2i}\beta)' \, \Sigma_i^{-1} \, (h(V_i) - b_0 \mathit{1} - X_{2i}\beta) \\
&-\frac{1}{2}\sum_{i=1}^{n} \log| - \tau_i^{(2)}(\widehat{\delta}_{1i})| + \sum_{i=1}^{n}\sum_{j=1}^{n_i} (U_{ij}\eta(\pi_{ij}(\widehat{\delta}_{1i})) + \log(1 - \pi_{ij}(\widehat{\delta}_{1i}))) \\
&-\frac{1}{2}\sum_{i=1}^{n} \left(U_i - \pi_i(\widehat{\delta}_{1i})\right)^T Z_{1i} B_i Z'_{1i} \left(U_i - \pi_i(\widehat{\delta}_{1i})\right). \qquad (2.7)
\end{aligned}
$$

We maximize the function $l(\Theta; h)$ by Newton-Raphson iterative procedure,

$$\Theta^{(t+1)} = \Theta^{(t)} + C^{-1}S,$$

where $C = -\partial^2 l(\Theta; h)/\partial\Theta\partial\Theta'$, and $S = \partial l(\Theta; h)/\partial\Theta$ evaluated at $\Theta = \Theta^{(t)}$. Since the second derivative of the log-likelihood is difficult to calculate, the well-known identity

$$E(\partial^2 l(\Theta; h)/\partial\Theta\partial\Theta') = -E[(\partial l(\Theta; h)/\partial\Theta)(\partial l(\Theta; h)/\partial\Theta)']$$

suggests an approximate scoring procedure with $C \approx \sum_{i=1}^{n} (\partial l_i(\Theta; h)/\partial\Theta)(\partial l_i(\Theta; h)/\partial\Theta)'$, where $l_i(\Theta; h)$ is the contribution of subject i to the approximate log-likelihood. Expressions for the components of the score vector can be obtained from the authors upon a request.

2·3 *Estimation of the transformation function h given Θ*

In this section, we discuss estimation of the transformation function h given all the parameters Θ. Since

$$Pr(V_{ij} \le v) = Pr(h(V_{ij}) \le h(v)) = \Phi\left(\frac{h(v) - b_0 - X'_{2ij}\beta}{\sqrt{Z'_{2ij}\psi_{22}Z_{2ij} + \sigma^2}}\right), \qquad (2.8)$$

where Φ is the cumulative distribution function of the standard normal random variable, we obtain an estimate $\widehat{h}(v)$ for $h(v)$ by solving the following estimating equation:

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(I(V_{ij} \leq v) - \Phi\left(\frac{h(v) - b_0 - X'_{2ij}\beta}{\sqrt{Z'_{2ij}\psi_{22}Z_{2ij} + \sigma^2}}\right)\right) = 0. \qquad (2.9)$$

where $v \in [v_0, v_1]$, the range of the observed V_{ij}.

Using the monotone increasing property of the function Φ, we obtain that the estimator $\widehat{h}(v)$ is a nondecreasing step function in $v \in [v_0, v_1]$ with jumps only at the observed V_{ij}, where $i = 1, \cdots, n, j = 1, \cdots, n_i$. Hence, let $v_1 < \cdots < v_K$ be the set of distinct points of $V_{ij}, i = 1, \cdots, n, j = 1, \cdots, n_i$, then solving the system of estimating equations defined by (2.9) is equivalent to solving the following system of K equations:

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(I(V_{ij} \leq v_k) - \Phi\left(\frac{h(v_k) - b_0 - X'_{2ij}\beta}{\sqrt{Z'_{2ij}\psi_{22}Z_{2ij} + \sigma^2}}\right)\right) = 0, \text{ for } k = 1, \cdots, K. \qquad (2.10)$$

The Newton-Raphson algorithm can be used to solve the system of K estimating equations (2.10). We can see later that the discrete property of $\widehat{h}$ provides us with a large simplification to predict the mean of the original scale. In addition, unlike a traditional nonparametric approach to estimate the transformation function (Horowitz, 1996; Klein & Sherman, 1998), our approach does not involve nonparametric smoothing, and thus does not suffer from smoothing related problems, for example, selection of a smoothing parameter.

We estimate Θ and h iteratively based on the approximation likelihood (2.7) and the system of estimating equations (2.10) until two successive values of Θ do not differ significantly. An initial value of Θ is required to start the iterations, which can be obtained by fitting a generalized linear model for U_i and a transformation model for nonzero with the dependence between models and data being ignored. For simplicity, we set the starting values for ψ_{11} and ψ_{22} to be the identity matrix.

Let $\widehat{h}$ be the estimators of h, $d_{ij}(\Theta) = Z'_{2ij}\psi_{22}Z_{2ij} + \sigma^2$,

$$S(w; v, \Theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(I(V_{ij} \leq v) - \Phi\left(\frac{w - b_0 - X'_{2ij}\beta}{\sqrt{d_{ij}(\Theta)}}\right)\right),$$

$$s_1(v) = \lim_{n\to\infty} S^{(100)}(h_0(v); v, \Theta_0), \quad s_2(v) = \lim_{n\to\infty} S^{(001)}(h_0(v); v, \Theta_0),$$

where h_0 and Θ_0 are the true values of h and Θ, respectively. Then under the conditions given in Appendix B, we have

$$\widehat{h}(v) - h_0(v) \approx s_1^{-1}(v)\left\{S(h_0(v); v, \Theta_0) + s_2(v)'(\widehat{\Theta} - \Theta_0)\right\}, \tag{2.11}$$

where $\widehat{\Theta}$ is the estimate of Θ. Hence, if there exist independent random variablesξ_i with $E(\xi_i) = 0$ and $Var(\xi_i) < \infty$, for $i = 1, 2, \cdots, n$, such that $\widehat{\Theta} - \Theta_0 = \frac{1}{n}\sum_{i=1}^{n}\xi_i + o_p(n^{-1/2})$, we have

$$\widehat{h}(v) - h_0(v) \quad = \quad \frac{1}{ns_1(v)}\sum_{i=1}^{n}\Omega_i(v) + o_p(n^{-1/2}),$$

where $\Omega_i(v) = \sum_{j=1}^{m_i}\left[I(V_{ij} \leq v) - \Phi\left(\frac{h_0(v) - b_0 - X'_{2ij}\beta_0}{\sqrt{d_{ij}(\Theta_0)}}\right)\right] + s_2(v)'\xi_i$. This implies that the distribution of $n^{1/2}(\widehat{h}(v) - h_0(v))$ can be approximated by a normal random variable with mean 0 and variance $\Sigma = \frac{1}{s_1^2(v)}E\Omega_i^2(v)$. Hence, we can estimate the nonparametric function $h(.)$ with a parametric convergent rate if we can estimate the parameters Θ at a rate of $n^{-1/2}$. The similar conclusion, regarding $n^{-1/2}$ convergent rate of the estimated transformation function, can be also found in Horowitz (1996), Chen (2002), Ye and Duan (1997) and Zhou et al. (2008). The conclusion assures that the resulting estimator for the mean of the original scale converges to the true value at a rate of $n^{-1/2}$.

3 Predicting the mean of the original scale

Given the covariates $x = (x'_1, x'_2)'$ and $z = (z'_1, z'_2)'$, we want to estimate $u(x, z) = E(Y|x, z)$, where Y is the response of the outcome for the patient with the covariates

x and z. Unbiased and consistent quantities on the transformed scale may not automatically retransform into unbiased or consistent quantities on the untransformed scale. The smearing estimate, proposed by Duan (1983), is a popular method to consistently estimate an individual's expected response on the untransformed scale. Since the random effects, δ_i's, are unobservable, it is difficult to extend the smearing estimator to the two-part model with the random effects.

In this section, we propose a numerical method to estimate $\mu(x, z)$. Let $\pi(\delta_1) = \eta^{-1}(x_1'\alpha + z_1'\delta_1)$ and $v(\delta_2) = h^{-1}(b_0 + x_2'\beta + z_2'\delta_2 + \sigma\varepsilon)$, where $\delta = (\delta_1', \delta_2')' \sim N(0, \psi)$, $\varepsilon \sim N(0, 1)$, δ and ε are independent. With this notation, we obtain the following expression for $\mu(x, z)$:

$$\begin{aligned} u(x,z) &= E\left(E(Y|x,z,\delta)\right) = E\left(\pi(\delta_1)E\left(v(\delta_2)|\delta\right)\right) \\ &= E\left\{\eta^{-1}(x_1'\alpha + z_1'\delta_1)E\left[h^{-1}(b_0 + x_2'\beta + z_2'\delta_2 + \sigma\varepsilon)|\delta\right]\right\}. \end{aligned}$$

From this expression, we see that one way to estimate $u(x, z)$ is to first estimate $E\left[h^{-1}(b_0 + x_2'\beta + z_2'\delta_2 + \sigma\varepsilon)|\delta\right]$ for any given δ, which can be achieved by the following estimator:

$$\frac{1}{R_1}\sum_{k=1}^{R_1}\widehat{h}^{-1}(b_0 + x_2'\widehat{\beta} + z_2'\delta_2 + \widehat{\sigma}\varepsilon_k),$$

where ε_k is generated from the standard normal distribution. Then, we can obtain the following estimator for $u(x, z)$:

$$\widehat{u}(x,z) = \frac{1}{R_1R_2}\sum_{r=1}^{R_2}\sum_{k=1}^{R_1}\eta^{-1}(x_1'\widehat{\alpha} + z_1'\delta_{1r})\widehat{h}^{-1}(b_0 + x_2'\widehat{\beta} + z_2'\delta_{2r} + \widehat{\sigma}\varepsilon_k), \tag{3.1}$$

where $\delta_r = (\delta_{1r}', \delta_{2r}')'$ is generated from the multivariate normal distribution with mean vector 0 and covariance matrix $\widehat{\psi}$. Next we give an asymptotic result for $\widehat{u}(x, z)$.

Let $\zeta_1 = x_1'\alpha + z_1'\delta_1$ and $\zeta_2 = b_0 + x_2'\beta + z_2'\delta_2 + \sigma\varepsilon$, $\pi(\zeta_1) = \eta^{-1}(\zeta_1)$, $v(\zeta_2) = h^{-1}(\zeta_2)$, $z_{01} = (z', 0')'$, and $z_{02} = (0', z')'$. Suppose that $\delta = (\delta_1', \delta_2')'$ is the normally distributed random vector with mean 0 and covariance matrix ψ, and that ϑ and ε are the standard normal random vector and standard normal random variable, respectively.

Assume that δ, ϑ, and ε are independent. Denote

$$\varrho_1(x,z,\Theta) = E\left\{\pi^{(1)}(\zeta_1)v(\zeta_2)\right\}, \quad \varrho_2(x,z,\Theta) = E\left\{\pi(\zeta_1)v^{(1)}(\zeta_2)\right\},$$

$$\varrho_3(x,z,\Theta) = E\left\{\pi(\zeta_1)v^{(1)}(\zeta_2)\varepsilon\right\}, \quad \varrho_4(x,z,\Theta) = E\left\{\pi(\zeta_1)v^{(1)}(\zeta_2)s_1^{-1}(v(\zeta_2))s_2(v(\zeta_2))\right\},$$

$$\varrho_5(x,z,\Theta) = E\left\{\pi^{(1)}(x_1'\alpha + z_{01}'\psi^{1/2}\vartheta)v(b_0 + x_2'\beta + z_{02}'\psi^{1/2}\vartheta + \sigma\varepsilon)\vartheta\right\},$$

$$\varrho_6(x,z,\Theta) = E\left\{\pi(x_1'\alpha + z_{01}'\psi^{1/2}\vartheta)v^{(1)}(b_0 + x_2'\beta + z_{02}'\psi^{1/2}\vartheta + \sigma\varepsilon)\vartheta\right\},$$

$$\varrho_7'(x,z,\Theta)(\widehat{\Theta}-\Theta) \equiv z_{01}'\left(\widehat{\psi}^{1/2} - \psi^{1/2}\right)\varrho_5(x,z,\Theta) + z_{02}'\left(\widehat{\psi}^{1/2} - \psi^{1/2}\right)\varrho_6(x,z,\Theta),$$

$$\varsigma(x,z,\Theta) = (\varrho_1(x,z,\Theta)x_1', \varrho_2(x,z,\Theta)x_2', \varrho_3(x,z,\Theta), 0)',$$

$$\varrho(x,z,\Theta) = \varsigma(x,z,\Theta) - \varrho_4(x,z,\Theta) + \varrho_7(x,z,\Theta),$$

$$\Upsilon(x,z,x_2^*,z_2^*,v^*,\Theta) = E\left\{\pi(\zeta_1)v^{(1)}(\zeta_2)s_1^{-1}(v(\zeta_2))\left(I(h(v^*)\le\zeta_2) - \Phi\left(\frac{\zeta_2 - b_0 - x_2^{*\prime}\beta}{\sqrt{d(\Theta)}}\right)\right)\right\},$$

where $d(\Theta) = z_2'\psi_{22}z_2 + \sigma^2$. Then under the conditions given in Appendix B, we have

$$\widehat{u}(x,z) - u(x,z) \approx \varrho(x,z,\Theta_0)'(\widehat{\Theta}-\Theta_0) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\Upsilon(x,z,X_{2ij},Z_{2ij},V_{ij},\Theta_0). \quad (3.2)$$

Hence, if there exist independent random variables, $\xi_i, i = 1, 2, \cdots, n$, with $E(\xi_i) = 0$, $Var(\xi_i) < \infty$, such that $\widehat{\Theta} - \Theta_0 = \frac{1}{n}\sum_{i=1}^{n}\xi_i + o_p(n^{-1/2})$, then the distribution of $n^{1/2}(\widehat{u}(x,z) - u(x,z))$ can be approximated by a normal random variable with mean 0 and a finite covariance matrix.

4 Simulation

4·1 *Performance of the approximate log-likelihood*

In this subsection, we investigate the accuracy of our approximate log-likelihood by comparing the estimates based on our approximation with the estimates based on the six-order Laplace approximation, proposed by Olsen & Schafer (2001).

In our simulation study, we use the same setting as in Olsen & Schafer (2001). For each subject, X_i is the matrix of covariates related to fixed effects and is constructed with three columns: a constant equal to 1, a dummy indicator for a non-time-varying covariate drawn from Bernoulli(p=0.5), and a time-varying covariate taking values $0, 1 \cdots, m-1$, where m is the number of occasions. The matrix of covariates Z_i, which are related to random effects, are set to be columns of 1's. The coefficients of the fixed effects are set to $\alpha = (-1, -0.5, 0.4)'$ and $\beta = (-0.3, 0.1, 0.4)'$. The variance parameters are set to be $\psi_{11} = 1$, $\psi_{12} = 0.2$, and $\psi_{22} = 0.5$. The homoscedastic variance of the transformed nonzero response is set to be $\sigma^2 = 0.5$. The transformation function is assumed to be identity. We also vary the number of subjects and the number of occasions in a 2×2 design with $n = 1000$ or 200 and $n_i = m = 10$ or 5.

We summarize the behavior of our new estimators and the Olsen & Schafer's (OS) estimators of α and β in Table 1.

For each scenario, Table 1 lists the average, standard error (SE), and the root of mean square errors (RMSE) of the estimators. Both our estimators and the OS estimators are basically unbiased. However, the standard deviations of our estimators are smaller than those of the OS estimators in all of the settings considered here. As a result, our estimators have smaller RMSE than the OS estimators in all of the settings and hence, are better than the OS estimators.

Although our estimator needs only the first and second derivatives, and the OS estimator needs the first through sixth derivatives, our estimators are still more accurate than the OS estimators.

4·2 *Robustness*

Since our method does not require specification of a parametric form for the transformation function, we expect that the resulting estimates and inferences are more reliable than the parametric method with the misspecified transformation function, for example, the OS estimators. We want to know whether the added robustness is gained at the expense of reduced efficiency. To investigate these two issues, we

Table 1: Simulation results when semi-continuous data are generated from a two-part homoscedastic mixed-effects model

n	m		Average		Empirical SE		RMSE	
			proposed	OS	proposed	OS	proposed	OS
1000	10	α_1	-1.0039	-1.000	0.0668	0.070	0.0669	0.0700
		α_2	-0.4947	-0.503	0.0808	0.083	0.0810	0.0831
		α_3	0.3998	0.400	0.0097	0.012	0.0097	0.0120
		β_1	-0.2993	-0.298	0.0405	0.043	0.0405	0.0430
		β_2	0.0994	0.096	0.0475	0.051	0.0475	0.0512
		β_3	0.3999	0.400	0.0036	0.004	0.0036	0.0040
1000	5	α_1	-0.9948	-1.000	0.0768	0.089	0.0770	0.089
		α_2	-0.4907	-0.501	0.0849	0.105	0.0855	0.105
		α_3	0.3982	0.401	0.0240	0.030	0.0241	0.030
		β_1	-0.3029	-0.300	0.0530	0.059	0.0531	0.059
		β_2	0.1010	0.102	0.0601	0.069	0.0601	0.069
		β_3	0.4004	0.401	0.0123	0.016	0.0123	0.016
200	10	α_1	-1.0144	-0.995	0.1434	0.165	0.1441	0.1651
		α_2	-0.4899	-0.512	0.1784	0.188	0.1787	0.1884
		α_3	0.4019	0.402	0.0231	0.028	0.0232	0.0281
		β_1	-0.2932	-0.297	0.0890	0.097	0.0893	0.0970
		β_2	0.0935	0.097	0.1052	0.117	0.1054	0.1170
		β_3	0.3996	0.400	0.0080	0.010	0.0080	0.0100
200	5	α_1	-0.9876	-0.993	0.1820	0.199	0.1824	0.1991
		α_2	-0.4925	-0.497	0.1972	0.219	0.1973	0.2190
		α_3	0.3972	0.399	0.0565	0.067	0.0566	0.0670
		β_1	-0.2979	-0.303	0.1150	0.133	0.1150	0.1330
		β_2	0.1012	0.100	0.1356	0.145	0.1356	0.1450
		β_3	0.3987	0.400	0.0279	0.034	0.0279	0.0340

examine the performance of the proposed method in comparison with the misspecified transformation (MT) method, where the transformation function is misspecified, and the correctly-specified transformation (CT) method, where the transformation function is correctly specified.

Our proposed model contains key two components, the transformation function, h, and the distribution of the random effects. We would also like to know the relative effect of misspecification of the transformation function and misspecification of the random effect distribution on our inference. To investigate this issue, we want to compare the performance of the proposed method with a misspecified distribution function of the random effects with the performance of the MT method with the correctly specified random effect distribution and a misspecified transformation function.

We conduct two simulation studies to answer the above three issues. In the first simulation study, we simulate data from the setting similar to the above simulation in Section 4.1 except that $\beta = (-0.3, 0.3, 0.4)'$ and the transformation function $h(v) = 3\log(v)$. A total of 200 data sets were generated. For each simulated data set, we obtain estimates for the fixed effect and the mean of original scale $\mu(x)$ at the combination of $x_1 = 1$, $x_2 = 0, 1$ and $x_3 = (0, 1, 2, 3, 4)$ using the proposed approach, the CT method, and the MT method with the misspecified transformation function $h(v) = v^4$.

The MT method fails to converge for 123 of the 200 samples. The results reported in Tables 2 and 3 are based on the remaining samples. Table 2 presents the average, the standard error (SE), the standardized bias (bias as a percent of the SE), and the RMSE for the fixed effect parameters. The MT estimate is severely biased. In contrast, the proposed approach yields an estimate with essentially no bias, once again suggesting that our method is robust.

Table 2: Simulation results when semi-continuous data are generated from a two-part mixed-effects heteroscedastic transformation model

		proposed	CT	MT(OS)		proposed	CT	MT(OS)
bias	α_1	0.0026	0.0022	0.0043	β_1	–	0.0044	-0.9814
SE		0.0588	0.0558	0.0530		–	0.0352	0.0476
RMSE		0.0589	0.0559	0.0532		–	0.0355	0.9825
bias	α_2	0.0104	0.0092	-0.0069	β_2	-0.0117	-0.0010	2.6542
SE		0.0654	0.0617	0.0632		0.0633	0.0386	0.0354
RMSE		0.0662	0.0624	0.0635		0.0644	0.0386	2.6544
bias	α_3	-0.0025	-0.0024	-0.0066	β_3	0.0013	-0.0005	3.1341
SE		0.0160	0.0166	0.0146		0.0078	0.0086	0.0097
RMSE		0.0162	0.0167	0.0160		0.0079	0.0086	3.1341
Bias	σ^2	0.0164	-0.0010	-0.3699	ψ_{11}	-0.1159	-0.1159	-0.1156
SE		0.0534	0.0138	0.0001		0.0783	0.0789	0.0787
RMSE		0.0559	0.0138	0.3699		0.1399	0.1402	0.1399
Bias	ψ_{12}	-0.0048	-0.0014	-0.1958	ψ_{22}	0.0176	0.0004	-0.4002
SE		0.0347	0.0340	0.0014		0.0530	0.0287	0.0001
RMSE		0.0351	0.0340	0.1958		0.0558	0.0287	0.4002

Table 3 below presents the average, SE, and RMSE for the estimated $\mu(x)$ at $x = (1, 4), (1, 5)$ and $(1, 6)$. Since the transformation function is involved only in the second part of the models, the estimates of the regression parameters in the first part of the models are basically unbiased even when we misspecified the transformation function. However, misspecification of the transformation function can lead to severely biased estimates for the parameters related to the second part of the model. In contrast, our method gives estimates close to the truth value of the parameter with the reasonable variances, suggesting that our procedure is robust.

For each simulated data set, we also obtain estimates of the transformation H using the proposed approach. Figure 1 displays the averaged estimated transformation

Table 3: Simulation Results for the original scaled average based on the same simulation of Table 2

x		proposed	CT	MT	x	proposed	CT	MT
(1,1)	Bias	0.0200	-0.0018	0.0646	(0,2)	0.0268	-0.0011	0.1167
	SE	0.0274	0.0125	0.0128		0.0341	0.0147	0.0146
	RMSE	0.0339	0.0126	0.0658		0.0434	0.0147	0.1176
(1,2)	Bias	0.0061	-0.0012	0.0942	(0,3)	0.0074	0.0000	0.1489
	SE	0.0252	0.0156	0.0153		0.0345	0.0188	0.0181
	RMSE	0.0260	0.0156	0.0955		0.0353	0.0188	0.1500
(1,3)	Bias	-0.0133	-0.0001	0.0966	(0,4)	-0.0191	0.0013	0.1374
	SE	0.0334	0.0204	0.0190		0.0494	0.0247	0.0221
	RMSE	0.0359	0.0204	0.0984		0.0529	0.0248	0.1391

function and their 95% empirical pointwise confidence limits, based on 200 simulated data sets; Figure 1 shows that our proposed estimate of the transformation function is very close to the true transformation function.

In the second simulation study, we investigate sensitivity of inferences to the random effects distribution. We generate the data, according to the same setting as in the above simulation study, and then we discretize the generated values of the random effect δ_1 to $-2, -1, 0, 1, 2$ and the generated values of the random effect δ_2 to $-1, 0, 1$. Table 4 presents the bias, SE, and RMSE for the fixed effect parameters, suggesting that our estimator basically is unbiased even when we misspecified the random effect distribution, which implies that our estimator is not sensitive to the random effects distribution. On the other hand, from Table 2 we know that the misspecification of the transformation function leads to biased estimates for the covariate effects. Hence, misspecification of the transformation function has a worse effect on estimation of covariate effects than misspecification of the random effects distribution does.

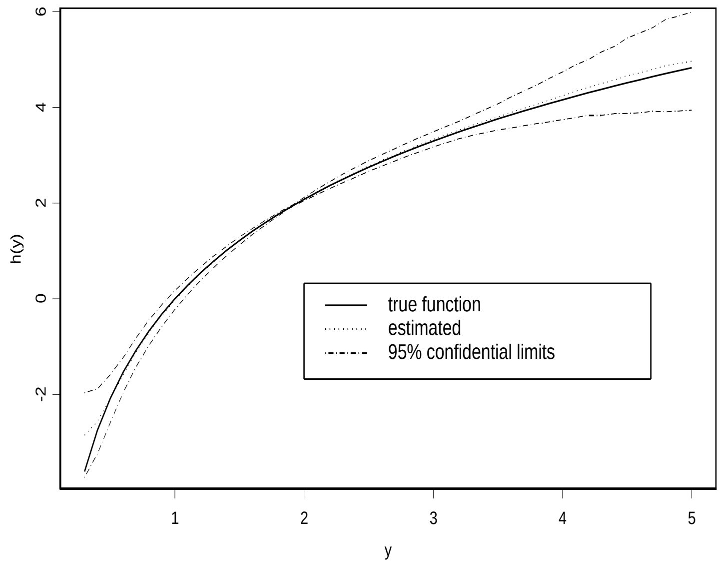

Figure 1: The typical estimates of transformation curve.

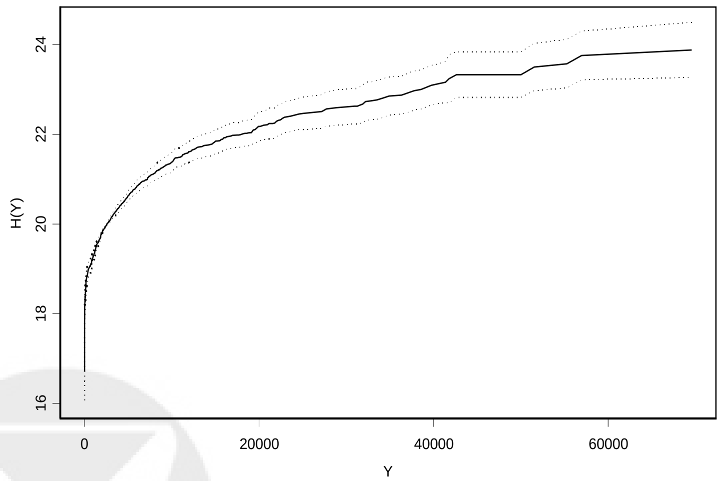

Figure 2: The estimated transformation curve for IMPACT data (Solid— estimated; dashed— 95% confidential limit).

Table 4: Simulation results when the random effects distribution is not normal.

		Proposed		Proposed		Proposed		Proposed
bias	α_1	0.0086	β_1	—	ψ_{11}	-0.0509	σ^2	0.0122
SE		0.0599		—		0.0866		0.0490
RMSE		0.0605		—		0.1005		0.0505
bias	α_2	0.0056	β_2	-0.0080	ψ_{12}	-0.0064		
SE		0.0675		0.0630		0.0388		
RMSE		0.0678		0.0635		0.0393		
bias	α_3	-0.0033	β_3	0.0012	ψ_{22}	0.0973		
SE		0.0171		0.0090		0.0570		
RMSE		0.0174		0.0091		0.1128		

5 Example

The sample used for this study was from a clinical study, examining the effectiveness of the IMPACT collaborative care management program for 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 to usual care ($n = 895$). Intervention patients had access for up to 12 months to a depression care manager who offered education, care management, and support of antidepressant management. The primary outcome, the total inpatient cost over the previous 6 month period, was collected at month 6, 12, 18, and 24. Denote Y_{ij} to be the total inpatient cost over the jth half year for patient i. The two independent variables are $\mathcal{X}_{1ij}$ and $\mathcal{X}_{2ij}$, where $\mathcal{X}_{1ij}$ is the treatment indicator, and $\mathcal{X}_{2ij}$ is the mean score of the 20 depression items from the symptom checklist for the jth observation of patient i. With $n_i = 4$ per subject, we do not have enough information to fit a high dimension random effects model, and hence, we fit the following random intercept only model:

$$\text{logistic}(\pi_{ij}(\delta_{1i})) = \alpha_0 + \mathcal{X}_{1ij}\alpha_1 + \mathcal{X}_{2ij}\alpha_2 + \delta_{1i},$$

and
$$h(V_{ij}) = \beta_0 + \mathcal{X}_{1ij}\beta_1 + \mathcal{X}_{2ij}\beta_2 + \delta_{2i} + \sigma\varepsilon_{ij},$$

where $\delta_i = (\delta_{1i}, \delta_{2i})$ is a bivariate normal vector with mean 0 and covariance matrix $\psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix}$, and ε_{ij} is a standard normal random variable. To reduce the computational time, we first apply a log transformation to the nonzero outcome variable. To compare, we also analyze the cost data by a parametric transformation model with log transformation (termed "LOG-TRAN").

We present parameter estimates in Table 5, which shows that the effects of treatment ($\mathcal{X}_1$) on the mean and variance are not significant and the correlations ($\psi_{12} = -0.0398$) across the two parts of the models are not significant. The results for $\mathcal{X}_2$ show that the patients with higher scores of depression are associated with higher costs and larger variation in cost, although the effect on variance is not significant. Figure 2 presents the estimate and its 95% confidential interval for the transformation function. Using the estimates of the parameters and transformation function, we estimate the average cost of a patient with the given covariate values. Table 6 gives some average costs. For example, for a patient in the intervention group ($\mathcal{X}_1 = 1$) with a depression score of 1.2 (around the mean of $\mathcal{X}_2$), the estimated average cost and its standard deviation are \$1130.028 and \$102.4616, respectively.

Estimating the difference of the means of health medical costs between the intervention and control patients as a function of patients' covariates is also an important target in econometrics, and hence we present some differences in Table 7, which suggests that the differences in cost between intervention and control patients can vary, depending on patients' characteristics.

Our models make some assumptions that should be investigated: normality of δ_i and ε_i, a linear relationships between covariates, and the logit-probability and

Table 5: The estimates of the parameters for IMPACT data

	Proposed	LOG-TRAN
$\widehat{\alpha}_1(SE)$	-2.2363(0.0990)	-2.4595(0.1398)
$\widehat{\alpha}_2(SE)$	-0.0508(0.0806)	0.0390(0.1159)
$\widehat{\alpha}_3(SE)$	0.1792(0.0534)	0.2752(0.0601)
$\widehat{\beta}_1(SE)$	20(0)	7.2558(0.1799)
$\widehat{\beta}_2(SE)$	0.2087(0.0925)	0.2227(0.1566)
$\widehat{\beta}_3(SE)$	0.1355(0.0484)	0.1915(0.0853)
$\widehat{\psi}_{11}(SE)$	1.1556(0.1228)	1.2460(0.1679)
$\widehat{\psi}_{12}(SE)$	-0.0398(0.1022)	0.0835(0.1870)
$\widehat{\psi}_{22}(SE)$	0.4948(0.3153)	1.0850(0.2282)
$\widehat{\sigma}^2(SE)$	1.2281(0.0864)	2.4351(0.2117)

Table 6: The estimates for the mean of original scale for IMPACT data

$\mathcal{X}_1$	$\mathcal{X}_2$	u_0	SE	$\mathcal{X}_1$	$\mathcal{X}_2$	u_0	SE
1	0.6	970.7469	85.9139	0	0.6	836.2906	79.3205
1	1.2	1130.0280	102.4616	0	1.2	974.7967	90.8485
1	1.8	1312.3440	134.2079	0	1.8	1133.4890	115.2832

linear relationships between covariates and the transformation of respondents. For normal mixed-effects models, only a few formal diagnostics have been developed, and practitioners often rely on informal techniques such as normal quantile plots of the estimated random effects. Diagnostics for Generalized Linear Mixed Models (GLMMs) are even more scarce (Olsen & Schafer, 2001).

Here we follow the similar method as used by Olsen & Schafer (2001). We detect the large discrepancies in the model fit by comparing the observed values for $U_{i\cdot} = \sum_{j=1}^{n_i} U_{ij}$ and $V_{i\cdot} = \sum_{j=1}^{n_i} log(V_{ij})$ with their predicted values $\widehat{U}_{i\cdot}$ and

Table 7: The differences of the mean of cost between two groups

group 1	group 2	difference of mean(SE)
$\mathcal{X}_1 = 1, \mathcal{X}_2 = 0.6$	$\mathcal{X}_1 = 0, \mathcal{X}_2 = 0.6$	134.4563 (90.0963)
$\mathcal{X}_1 = 1, \mathcal{X}_2 = 1.2$	$\mathcal{X}_1 = 0, \mathcal{X}_2 = 1.2$	155.2313(105.2384)
$\mathcal{X}_1 = 1, \mathcal{X}_2 = 1.8$	$\mathcal{X}_1 = 0, \mathcal{X}_2 = 1.8$	178.8550 (122.9108)

$\widehat{V}_{i\cdot} = \sum_{j=1}^{n_i} log(\widehat{V}_{ij})$, obtained by substituting the estimates of Θ, H and empirical Bayes estimates of δ_i. Viewing $N(0, \psi)$ as a prior distribution for δ_i, empirical Bayes estimates of δ_i can be obtained by calculating a posterior mean $E(\delta_i|Y_i)$ with the unknown parameters replaced by their estimates. Since the conditional distribution δ_i given Y_i does not have a closed form, we evaluate the integrals required for posterior moments by numerical techniques. In the example δ_i, has two dimensions, and we use numerical techniques methods to evaluate the related integrals.

Table 8 gives the frequency in each cell defined by the observed and the predicted (rounded to the nearest integers) value for $U_{i\cdot}$. The percentage of total agreement between the observed and the predicted values is 83.08%. Figure 3 plots $V_{i\cdot}$ versus $\widehat{V}_{i\cdot}$, showing no significant deviation. Table 8 and Figure 3 suggest our models are reasonable.

6 Discussion

In the paper, we have developed a flexible methodology to estimate the mean of the skewed semi-continuous outcome of a patient and regression parameters in a semi-parametric two-part mixed-effects transformation model with an unknown transformation function. The current existing methods to analyzing correlated right-skewed semi-continuous data require the specification of the transformation, which is a diffi-

Table 8: The frequency in each cell defined by the observed number of zero cost and expected number of zero cost among the four observations

	Observed number				
Expected number	0	1	2	3	4
0	975	2	0	0	0
1	0	424	217	11	0
2	0	0	0	47	8
3	0	0	0	0	0
4	0	0	0	0	0

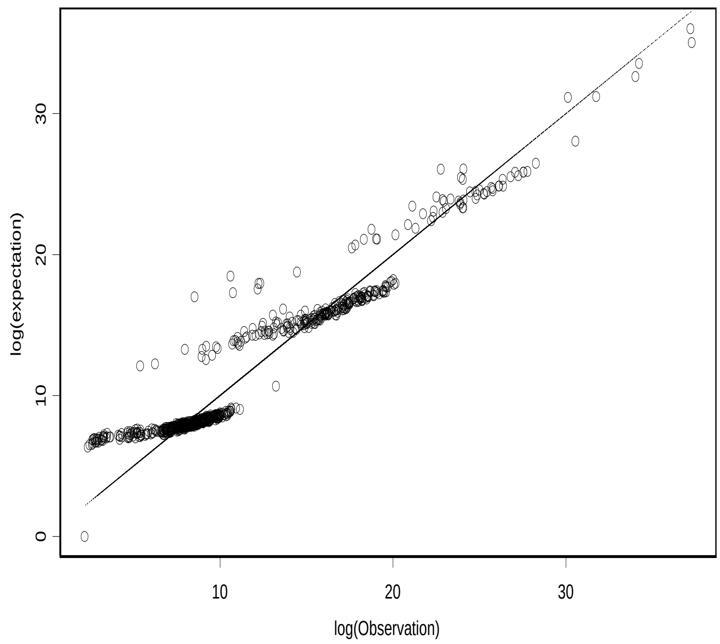

Figure 3: Observed log amount of the cost versus expected log amount of the cost.

cult task in practice. Our paper has several new features over the existing methods. First, our method allows the arbitrary non-parametric transformation function, and thus is more flexible and robust. The asymptotic distribution theory shows that our new estimators for the transformation function converge to their true values at the parametric rate $n^{-1/2}$ if the parameters are estimated at the parametric rate $n^{-1/2}$, suggesting that the extra flexibility is gained at little cost in efficiency. The simulation studies in the paper also show that the efficiency of our new estimators is comparable to the existing parametric method with the correctly specified transformation in finite sample sizes. Finally, we propose a new and more accurate approximate likelihood function to handle intractable numerical integration in the marginal likelihood, and the computational requirement of the new approximate likelihood is rather minimal.

In modeling non-zero data, we need to decide whether to put a parametric assumption on the transformation function or the distribution of the random effects. In our proposed method, we chose to impose a normal distribution assumption on random effects but leave the transformation function unknown. Our simulation study shows that the correctly specified transformation function is more important than the correctly specified distribution function of random effects in our inferences. Hence, our simulation study supports our choose. Future research could explore the possibility of allowing both the transformation function and the distribution functions of random effects unknown.

In our proposed model, we assume that homoscedastic variance for transformed non-zero costs. In some cases, the homoscedastic variance assumption may be not met (Manning, 1998; Mullahy, 1998; Zhou et al., 1997a; Zhou et al., 1997b; and Zhou & Tu, 1999). Mullahy (1998) gave several real situations where two-part regression models with homoscedastic variance after transforming the nonzero responses yield inconsistent inferences on $\mu(x)$. The heteroscedasticity for the non-zero data may be complicated. However, on the other hand, it may be difficult to get a good estimate of the variance if we specify a complicated heteroscedasticity. A possible method for handling heteroscedasticity of the non-zero data is to replace the second model (2.2) with

$$h(V_{ij}) = X'_{2ij}\beta + Z'_{2ij}\delta_{2i} + g(X'_{2ij}\theta)\varepsilon_{ij}, \tag{6.1}$$

where g is a known function, and θ is a vector of unknown parameters. In the model (6.1), heteroscedasticity is modeled by the known function $g(.)$ with a vector of unknown parameters, θ, hence the heteroscedasticity is not linked to the mean level; and the mean and variance may be influenced by covariates in different ways. It is straightforward to extend our proposed method to model (6.1). Another possibility to model the heteroscedasticity is setting $\theta = \beta$, as the literature in the generalized linea model, and leave the variance function unknown.

ACKNOWLEDGEMENT

Lin's research was supported in part by the Fund of National Natural Science (Grant 10771148) of China. 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. Xiao-Hua Zhou is currently a core investigator and Biostatistics Unit Director at the Northwest HSR&D Center of Excellence within the VA Puget Sound Health Care System. The views expressed in this article are those of the author and do not necessarily represent the views of the Department of Veterans Affairs. We also like to thank Dr. Jurgen Unutzer for providing us with the data.

References

Anderson, D. A., & Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society, Ser.B*, **47**, 204-210.

Albert, P. S. & Shen, J. (2005). Modeling longitudinal semi-continuous emesis volume data with serial correlation in an acupuncture clinical trial. *Journal of the Royal Statistical Society, Ser. C*, **54**, 707-720.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9-25.

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

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformation. *Journal of the Royal Statistical Society, Series B*, **26**, 211-243.

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. & Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*, **1**, 115-126.

Hedeker, D. & Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics*, **50**,933-944.

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

Manning, W. G., Basub, A. & Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data, *Journal of Health Economics*, **24**, 465-488.

Manning, W., Morris, C. N., Newhouse, J. P., Orr, L. L., Duan, N., Keeler, E. B., Leibowitz, A., Marquis, K. H., Marquis, M. S. & Phelps, C. E. (1981). A two-part model of the demand for medical care: preliminary results from the health insurance experiment. *Health, Ecomomics and Health Ecomomics eds. J. van der Gaag & M. Perlman, Amsterdam: North-Holland*, 103-104.

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

Lindstrom, M. J. & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673-687.

Liu, Q. & Pierce, D. A. (1993). Heterogeneity in Mantel-Haenszel type models. *Biometrika*, **80**, 543-556.

Olsen, M. K. & Schafer, J. (2001) A two-part random-effects model for semi-continuous longitudinal data. *Journal of the American Statistical Association*, **96**, 730-745.

Raudenbush, S. W., Yang, M., & Yosef, M. (2000). Maximum Likelihood for Generalized Linear Models With Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics*, **9**, 141-157.

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

Soloman, P. J. & Cox, D. R. (1992). Nonlinear component of variance models. *Biometrika*, **79**, 1-11.

Tierney, L. & Kadane, J. B. (1986), Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82-86.

Tooze, J. A., Grunwald, G. K. & Jones, J. H. (2002) Analysis of repeated measures data with clumping zero. *Statistical Methods in Medical Research*, **, 11**, 341-356.

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

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

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

Zhou, X. H., Lin, H. Z and Johnson, E. (2008). Nonparametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *Journal of Royal Statistical Society Series B., in press.*

Zhou, X. H., & 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

In this Appendix, we outline a proof for our approximation (2.6). Denoting $G_i(\delta_{1i}) = \tau_i^*(\delta_{1i}) - \tau_i(\delta_{1i})$, we have

$$G_i^{(2)}(\delta_{1i}) = -\frac{1}{n_i}\left\{\sum_{j=1}^{n_i} \pi_{ij}(\delta_{1i})(1-\pi_{ij}(\delta_{1i}))Z_{1ij}Z'_{1ij}\right\}.$$

Note that

$$\begin{aligned} a^* - a &= g(\tau_i^{(*2)}(\widehat{\delta}_{1i}^*), \tau_i^{(*3)}(\widehat{\delta}_{1i}^*), \tau_i^{(*4)}(\widehat{\delta}_{1i}^*)) - g(\tau_i^{(*2)}(\widehat{\delta}_{1i}), \tau_i^{(*3)}(\widehat{\delta}_{1i}), \tau_i^{(*4)}(\widehat{\delta}_{1i})) \\ &\quad + g(\tau_i^{(*2)}(\widehat{\delta}_{1i}), \tau_i^{(*3)}(\widehat{\delta}_{1i}), \tau_i^{(*4)}(\widehat{\delta}_{1i})) - g(\tau_i^{(2)}(\widehat{\delta}_{1i}), \tau_i^{(3)}(\widehat{\delta}_{1i}), \tau_i^{(4)}(\widehat{\delta}_{1i})) \\ &= O_p(\widehat{\delta}_{1i}^* - \widehat{\delta}_{1i}) + O_p(G^{(2)}(\widehat{\delta}_{1i}) + G^{(3)}(\widehat{\delta}_{1i}) + G^{(4)}(\widehat{\delta}_{1i})). \end{aligned}$$

From the above expression, we see that dependence of $G^{(k)}(\delta_{1i})$, for $k \geq 2$ on δ_{1i} is through $\Pi_{ij}(\delta_{1i}) = \pi_{ij}(\delta_{1i})(1-\pi_{ij}(\delta_{1i}))$, and this dependence is negligble. This negligibility can be justified by using the same argument as in Bates and Watts (1980) and Pinheiro and Bates (2000) for assessing parameter-effects on nonlinearity. There, they showed that the space spanned by the columns of $\Pi_{ij}(\delta_{1i})$ depended only on the intrinsic curvature of the nonlinear model, but not on the parameter-effects curvature in the tangent plane. Therefore, $\Pi_{ij}(\delta_{1i})$ may be assumed to vary slowly with δ_{1i}. This result, coupled with $\widehat{\delta}^*_{1i} - \widehat{\delta}_{1i} = O_p(n_i^{-1/2})$, gives us that $a^* - a = O_p(n_i^{-1/2})$. This completes the proof of the approximation (2.6).

Appendix B

To show our asymptotic results, we need the following conditions.

1. Suppose that $[v_0, v_1]$ is the domain of h. In practice, this would be the range of the observed and fitted V_{ij}'s. Assume that h is strictly increasing and continuous for $v \in [v_0, v_1]$.

2. There exists a sequence $\{\widehat{\Theta}\}$ such that $\widehat{\Theta} - \Theta_0 \to 0$.

3. $(X_{1i}, X_{2i}, Z_{1i}, Z_{2i})$ has bounded support.

4. Denote $\Xi = \{(x, z) : h_0^{-1}(v_0) <= b_0 + x_2'\beta_0 + z_2'\delta_2 + \sigma\varepsilon <= h_0^{-1}(v_1)$ for $\delta_2 \sim N(0, \psi_{220})$ and $\varepsilon \sim N(0, 1)\}$, suppose $Pr(\Xi) > 0$.

5. $n/R_1 = o(1)$ and $n/R_2 = o(1)$.

6. Suppose that $\varrho_j(x, z, \Theta), j = 1, \cdots, 7$ are continuous functions of Θ.

The proof of (2.11).

By the monotonicity and continuity of Φ, for large n, any $\eta > 0$ and $\Theta \in \{\Theta : \|\Theta - \Theta_0\| \leq \eta\}$, uniformly in $v \in [v_0, v_1]$, there exists a unique $\widehat{h}(v; \Theta)$ such that

$$S(\widehat{h}(v; \Theta); v, \Theta) = 0, \tag{A.1}$$

where $S(w; v, \Theta)$ is defined in Section 3. Since $S(h_0(v); v, \Theta_0) \to 0$, we have $\widehat{h}(v; \Theta_0) \to h_0(v)$, so that $\widehat{h}(v; \widehat{\Theta}) \to h_0(v)$ almost surely uniformly in $v \in [v_0, v_1]$.

Now we consider the expansion of $\widehat{h}(v) = \widehat{h}(v; \widehat{\Theta})$. Using a Taylor series expansion of $S(\widehat{h}(v; \widehat{\Theta}); v, \widehat{\Theta})$ with respect to $\widehat{h}(v; \widehat{\Theta})$ around $h_0(v)$, and noting that

$S(\widehat{h}(v;\widehat{\Theta}); v, \widehat{\Theta}) = 0$, we obtain

$$\widehat{h}(v;\widehat{\Theta}) - h_0(v) \approx -\left(S^{(100)}(h_0(v); v, \widehat{\Theta})\right)^{-1} S(h_0(v); v, \widehat{\Theta}).$$

Then using a Taylor series expansion of $S(h_0(v); v, \widehat{\Theta})$ with respect to $\widehat{\Theta}$ around Θ_0, we get

$$\widehat{h}(v;\widehat{\Theta}) - h_0(v) \approx s_1^{-1}(v) \left\{ S(h_0(v); v, \Theta_0) + s_2(v)'(\widehat{\Theta} - \Theta_0) \right\}. \qquad \text{(A.2)}$$

This result, coupled with the condition 2 and the expression of $S(w; v, \Theta)$, leads to (2.11).

The proof of (3.2).

Replace Θ_0, h_0 with Θ and h for notational simplicity. Denote

$$u_n(x, z) = \frac{1}{R_1 R_2} \sum_{r=1}^{R_2} \sum_{k=1}^{R_1} \{\pi(\zeta_{1r}) v(\zeta_{2rk})\},$$

where $\zeta_{1r} = x_1'\alpha + z_1'\delta_{1r}$, $\zeta_{2rk} = b_0 + x_2'\beta + z_2'\delta_{2r} + \sigma\varepsilon_k$, $\pi(\zeta_1) = \eta^{-1}(\zeta_1)$, $v(\zeta_2) = h^{-1}(\zeta_2)$ and $\delta_r = (\delta_{1r}', \delta_{2r}')' \sim N(0, \widehat{\psi})$. Consider (3.1) and use the expansion,

$$\begin{aligned}
&\widehat{u}(x, z) - u_n(x, z) \\
&\approx \frac{1}{R_1 R_2} \sum_{r=1}^{R_2} \sum_{k=1}^{R_1} \pi^{(1)}(\zeta_{1r}) v(\zeta_{2rk}) x_1' (\widehat{\alpha} - \alpha) + \frac{1}{R_1 R_2} \sum_{r=1}^{R_2} \sum_{k=1}^{R_1} \pi(\zeta_{1r}) v^{(1)}(\zeta_{2rk}) x_2' (\widehat{\beta} - \beta) \\
&\quad + \frac{1}{R_1 R_2} \sum_{r=1}^{R_2} \sum_{k=1}^{R_1} \pi(\zeta_{1r}) v^{(1)}(\zeta_{2rk}) (\widehat{\sigma} - \sigma) \varepsilon_k - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \Upsilon(X_{ij}, Z_{ij}, V_{ij}) \\
&\quad - \frac{1}{R_1 R_2} \sum_{r=1}^{R_2} \sum_{k=1}^{R_1} \pi(\zeta_{1r}) v^{(1)}(\zeta_{2rk}) s_1^{-1}(v(\zeta_{2rk})) s_2(v(\zeta_{2rk}))' (\widehat{\Theta} - \Theta) \\
&\approx (\tilde{\varsigma}(x, z, \Theta) - \tilde{\varrho}_4(x, z, \Theta))' (\widehat{\Theta} - \Theta) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \tilde{\Upsilon}(X_{ij}, Z_{ij}, V_{ij}), \qquad \text{(A.3)}
\end{aligned}$$

where $\tilde{\varsigma}$, $\tilde{\varrho}_4$ and $\tilde{\Upsilon}$ are ς, ϱ_4 and Υ defined in Section 4, but with $(\delta_{1r}', \delta_{2r}')' \sim N(0, \widehat{\psi})$.

Denote u by $\widetilde{u}$ if $\delta = (\delta_1', \delta_2')' \sim N(0, \widehat{\psi})$, then

$$\begin{aligned} u_n(x,z) - \widetilde{u}(x,z) &= \frac{1}{R_2}\sum_{r=1}^{R_2}\left\{\frac{1}{R_1}\sum_{k=1}^{R_1}\pi(\zeta_{1r})v(\zeta_{2rk}) - \pi(\zeta_{1r})E\left[v(\zeta_{2rk})|\delta_r\right]\right\} \\ &+\frac{1}{R_2}\sum_{r=1}^{R_2}\left\{\pi(\zeta_{1r})E\left[v(\zeta_{2rk})|\delta_r\right] - E\left\{\pi(x_1'\alpha + z_1'\delta_1)E\left[v(b_0 + x_2'\beta + z_2'\delta_2 + \sigma\varepsilon)|\delta\right]\right\}\right\} \\ &= O(R_1^{-1/2}) + O(R_2^{-1/2}). \qquad \text{(A.4)} \end{aligned}$$

Furthermore, we have

$$\begin{aligned} &\widetilde{u}(x,z) - u(x,z) \\ &\approx z_{01}'\left(\widehat{\psi}^{1/2} - \psi^{1/2}\right)E\left\{\pi^{(1)}(x_1'\alpha + z_{01}'\psi^{1/2}\vartheta)v(b_0 + x_2'\beta + z_{02}'\psi^{1/2}\vartheta + \sigma\varepsilon)\vartheta\right\} \\ &\quad + z_{02}'\left(\widehat{\psi}^{1/2} - \psi^{1/2}\right)E\left\{\pi(x_1'\alpha + z_{01}'\psi^{1/2}\vartheta)v^{(1)}(b_0 + x_2'\beta + z_{02}'\psi^{1/2}\vartheta + \sigma\varepsilon)\vartheta\right\}, \qquad \text{(A.5)} \end{aligned}$$

where ϑ and ε are independent standard normal random vector and variables, respectively, $z_{01} = (z', 0')', z_{02} = (0', z')'$. The result (3.2) follows from (A.3),(A.4) and (A.5).