Statistical Science
2019, Vol. 34, No. 2, 253–279
https://doi.org/10.1214/18-STS681

Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review

Lei Liu, Ya-Chen Tina Shih, Robert L. Strawderman, Daowen Zhang, Bankole A. Johnson and Haitao Chai

Abstract. Zero-inflated nonnegative continuous (or semicontinuous) data arise frequently in biomedical, economical, and ecological studies. Examples include substance abuse, medical costs, medical care utilization, biomarkers (e.g., CD4 cell counts, coronary artery calcium scores), single cell gene expression rates, and (relative) abundance of microbiome. Such data are often characterized by the presence of a large portion of zero values and positive continuous values that are skewed to the right and heteroscedastic. Both of these features suggest that no simple parametric distribution may be suitable for modeling such type of outcomes. In this paper, we review statistical methods for analyzing zero-inflated nonnegative outcome data. We will start with the cross-sectional setting, discussing ways to separate zero and positive values and introducing flexible models to characterize right skewness and heteroscedasticity in the positive values. We will then present models of correlated zero-inflated nonnegative continuous data, using random effects to tackle the correlation on repeated measures from the same subject and that across different parts of the model. We will also discuss expansion to related topics, for example, zero-inflated count and survival data, nonlinear covariate effects, and joint models of longitudinal zero-inflated nonnegative continuous data and survival. Finally, we will present applications to three real datasets (i.e., microbiome, medical costs, and alcohol drinking) to illustrate these methods. Example code will be provided to facilitate applications of these methods.

Key words and phrases: Two-part model, Tobit model, health econometrics, semiparametric regression, joint model, cure rate, frailty model, splines.

1. INTRODUCTION

Zero-inflated nonnegative continuous (or semicontinuous) data are frequently encountered in biomedical, economic, and ecological studies. Such data often have two distinct features: (i) the presence of a large portion of zero values, and (ii) right skewness and heteroscedasticity for positive continuous values. Examples include medical costs (Manning et al., 1981, Duan et al., 1983, Zhou and Tu, 1999, Liu, 2009, Liu

Lei Liu is Professor of Biostatistics, Division of Biostatistics, Washington University in St. Louis, St. Louis, Missouri 63110, USA (e-mail: lei.liu@wustl.edu). Ya-Chen Tina Shih is Professor, Department of Health Services Research, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA (e-mail: yashih@mdanderson.org). Robert L. Strawderman is the Donald M. Foster M.D. Distinguished Professor of Biostatistics and Chair, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York 14642, USA (e-mail: robert_strawderman@urmc.rochester.edu). Daowen Zhang is Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, USA (e-mail: daowen_zhang@ncsu.edu). Bankole A. Johnson is Professor and Chair, Department of Psychiatry, University of Maryland, Baltimore, Maryland 21201, USA (e-mail: bjohnson@psych.umaryland.edu). Haitao Chai is Ph.D. student, Institute for Financial Studies, Shandong University, Jinan, Shandong 250100, China (e-mail: cht0816@163.com).

et al., 2008, 2010), medical care utilization (Xie et al., 2004), daily precipitation levels (Hyndman and Grunwald, 2000), alcohol consumption (Liu, Ma and Johnson, 2008, Liu et al., 2016b, Han et al., 2018), health assessment score (Boag, 1949), single cell gene expression rates (McDavid et al., 2013, Finak et al., 2015), and relative abundance of microbiome data (Chen and Li, 2016, Chai et al., 2018). Of note, there also exist zero-inflated data when the continuous portion of the data can be positive and negative but with excessive zeroes, for example, changes of alveolar bone height of the tooth sites as considered in Lu, Lin and Shih (2004). In this paper, we will focus on zero-inflated nonnegative continuous outcomes, and hereafter may omit the word "nonnegative" without causing further confusion.

Three motivating applications are considered in this paper to illustrate the use of these models. The first example is microbiome composition data. In such a study, the raw measures of microbial abundances (in read counts) are not comparable across samples. They are normalized to relative abundances, a proportion in $[0, 1)$, for each species. However, not all species are present, resulting in zero values in relative abundances. For example, Figure 1(A) shows the distribution of the relative abundance of *Haemophilus*, which has excessive zeros (39%). The continuous nonzero values are right skewed, ranging from 0.0000106 to 0.4440, having a mean of 0.0285. The second example is medical cost data. Figure 1(B) shows monthly medical costs of heart failure patients in the University of Virginia Health System (Liu, 2009). In any given month, some patients had no costs while other patients incurred tremendous medical costs which may increase with disease severity. Specifically, in this dataset, (i) a large portion of monthly medical costs (49%) are zero; and (ii) positive monthly costs have a mean of $2982 and median of $365. Thus, the data exhibit a pattern of a point mass at 0, and the positive values are right skewed with possible heteroscedasticity. In the third example, we are interested in the daily drinking level (converted to number of standard drinks) in a topiramate trial to treat alcohol dependence (Johnson et al., 2007). A "standard drink" is 0.5 oz of absolute alcohol, equivalent to 10 oz of beer, 4 oz of wine, or 1 oz of 100-proof liquor. After conversion to the unit of standard drinks, alcohol consumption data are present as fractional drinking levels, resulting in continuous data instead of count data. On a given day, some alcohol dependent individuals were abstinent, that is, had zero drinking values, while others drank heavily. As shown in Figure 1(C), there is a presence of a large portion (32%) of zero values. The continuous nonzero (positive) values have a mean of 7.6 standard drinks with a range of 0.072–58.9, demonstrating the right skewness.

A common analytical approach for such zero-inflated nonnegative outcome data is to do a linear regression on the log transformed value $\log(Y + c)$, where c is a small positive number, for example, $1 for medical costs. This is the so-called "one part" model. However, the transformed outcome still has a point mass at $\log c$. Furthermore, it is of clinical significance to distinguish zero from positive values in the original zero-inflated continuous data. For example, for alcohol consumption, a "0" value means abstinence, which is an important treatment target to achieve. In medical cost data, individuals with $0 in medical costs may be healthy and have not sought any medical treatment. However, it is also possible that the absence of any medical costs reflects an inability to seek medical care due to financial or other access barriers.

In this paper, we will explore more advanced and rigorous methods to handle the features of zero-inflated nonnegative continuous outcomes. The first challenge is to distinguish between zero and positive values, which will be discussed in Section 2. Then, we present more flexible models to address the right skewed and often heteroscedastic positive values in Section 3. Correlated zero-inflated nonnegative data, for example, clustered medical costs or longitudinal drinking outcomes, will be tackled in Section 4. In Section 5, we will discuss other forms of zero-inflated data, for example, zero-inflated count and zero-inflated survival models. Next, we will consider other issues commonly encountered in zero-inflated continuous data, for example, nonlinear covariate effects in Section 6 and joint models of longitudinal zero-inflated outcomes and a terminal event in Section 7. In Section 8, we illustrate the applications of some of these methods to three datasets. Concluding remarks and future methodological considerations are given in Section 9. Sample programming codes are provided in the web materials.

Of note, a recent tutorial on modeling zero-inflated count and semicontinuous data was published in *Statistics in Medicine* by Neelon, O'Malley and Smith (2016). While our paper has some overlap in the general model set-up, we placed considerably more attention on analytical challenges associated with the positive values of the semicontinuous data, such as the right skewness and heteroscedasticity (Sections 3 and 4) and nonlinear covariate effects (Section 6). In addition, this

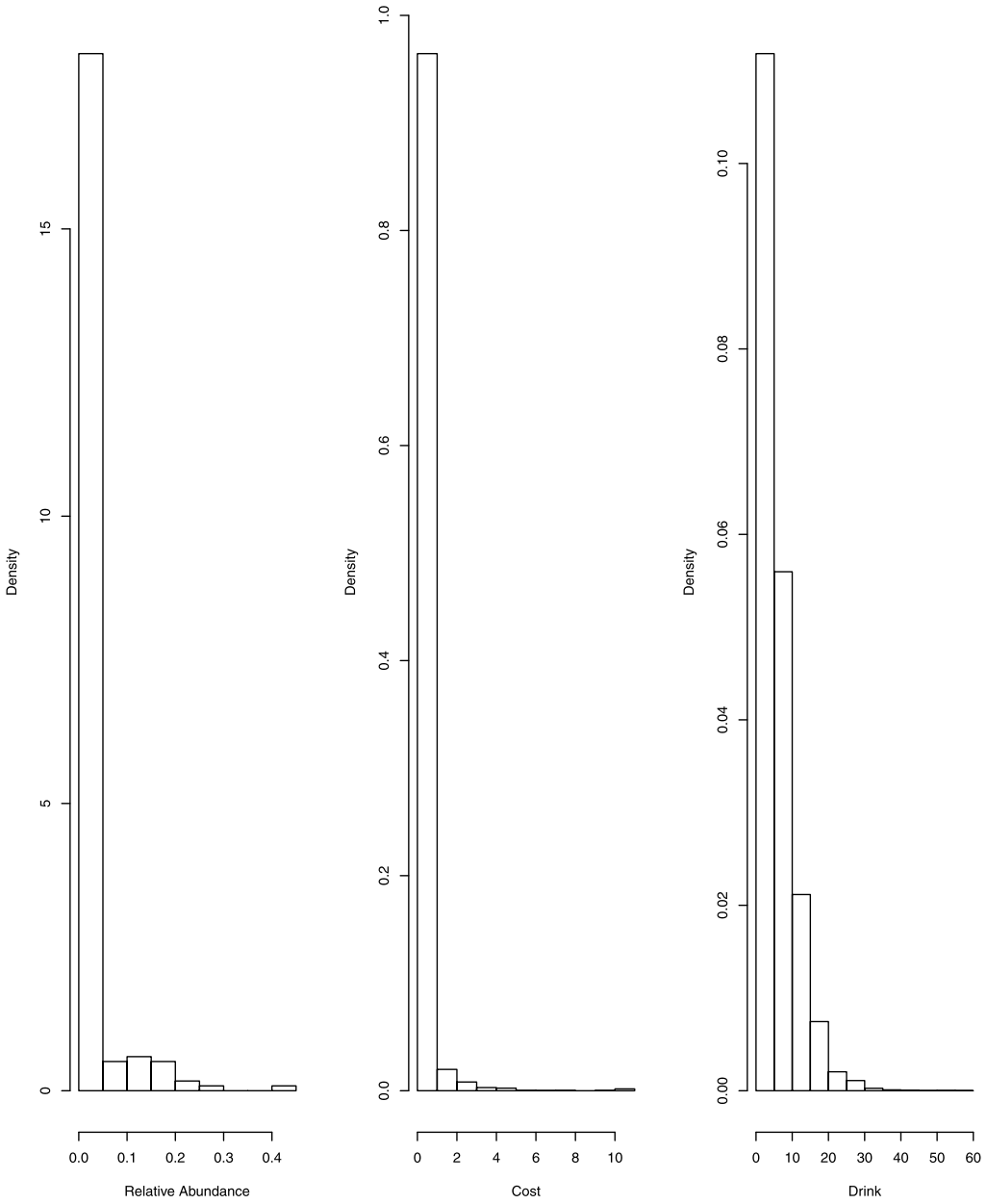

FIG. 1. (A) *Histogram of relative abundance of Haemophilus*. (B) *Histogram of monthly medical costs* (*in* 10,000 *U.S. dollars*) *for heart failure patients*. (C) *Histogram of daily drinking levels in the topiramate trial*.

review covers several more advanced topics in zero-inflated models, including zero-inflated survival models (Section 5) and joint models of semicontinuous data with survival or other clinical outcomes (Section 7).

2. ZERO VS. POSITIVE VALUES

Two approaches are generally adopted to describe a semicontinuous outcome: a Tobit model (Tobin, 1958) where zero values are considered as "censored" observations, or a two-part model (2PM or TPM, Manning et al., 1981) which separately describes the probability of the outcome being positive and the magnitude of positive values. These two approaches will be discussed in the first two subsections. We will then describe a zero-inflated Tobit model (Moulton and Halsey, 1995) which accommodates the characteristics of both the Tobit and the two-part models.

2.1 Tobit Model

In this approach, the actual (underlying true or latent) outcome Y^* is continuous and positive. However, for example, due to the detection limit of a measuring instrument, we may not be able to observe the true outcome below such a detection limit. Denoting the observed value by Y, we have

$$Y = \begin{cases} Y^* & Y^* > y_{\min}, \\ 0 & Y^* \le y_{\min}, \end{cases} \tag{1}$$

where $y_{\min}$ is the smallest observed positive value (e.g., the detection limit). The actual outcome Y^* is left censored at $y_{\min}$ if it drops below the detection limit (compared to right censoring in survival models), resulting in a point mass at 0 for observed Y. Importantly, the choice of value for Y is arbitrary when $Y^* \le y_{\min}$; with zero, $y_{\min}$, or $y_{\min}/2$ being 3 such example; however, zero is mostly used for simplicity and ease in interpretation.

Covariates can be included, such as in the following Tobit model (Tobin, 1958): given a covariate vector X,

$$\log Y^* | X \sim X^T \beta + e, \tag{2}$$

where β is a coefficient vector and e is an error term that is independent of X. If we assume $e \sim N(0, \sigma^2)$, the underlying Y^* has a lognormal distribution.

There are many variations of the Tobit model. As an example, the Heckman sample selection model (Heckman, 1979), also known as the Type II Tobit model, introduces a second latent variable Y_2^* in addition to the first latent variable Y_1^*: let

$$Y_2 = \begin{cases} Y_2^* & \text{if } Y_1^* > y_{\min}, \\ 0 & \text{o.w.}, \end{cases} \tag{3}$$

where

$$\begin{aligned} \log Y_1^* | X &\sim X^T \beta_1 + e_1, \\ \log Y_2^* | X &\sim X^T \beta_2 + e_2. \end{aligned} \tag{4}$$

That is, Y_2^* is observed if and only if $Y_1^* > y_{\min}$. Other variations of Tobit models (Types III–V) can be found in Amemiya (1994).

2.2 Two-Part Model (2PM)

In contrast to the Tobit model which considers the true outcome to be left censored, the two-part model does not distinguish between Y and Y^* and treats zero values as true observations. More accurately, it models the observed data as a mixture, separating the zero and positive values explicitly by two submodels (parts). Let Y_0 be a Bernoulli random variable such that

$$\operatorname{logit} P(Y_0 = 1 | X) = X^T \alpha. \tag{5}$$

Let $Y_+ > 0$ be a continuous random variable such that

$$\log Y_+ | X \sim X^T \beta + e, \tag{6}$$

where X and e are independent. Then, defining the nonnegative random variable $Y = Y_0 Y_+$, the above specifications induce a conditional cumulative distribution function of the form

$$\begin{aligned} P(Y \le y | X) = 1 - p(X) \\ + I(y > 0) p(X) F_e\big(\log y - X^T \beta\big), \end{aligned}$$

where $F_e(u) = P(e \le u)$ and $p(X) = \operatorname{expit}(X^T \alpha)$. This last expression generates the so-called two-part model (2PM: Aitchison, 1955; Manning et al., 1981); that is, if only Y is observed, the probability model specification involves two parts:

- Part I: $P(Y > 0 | X) = p(X)$ and $P(Y = 0 | X) = 1 - p(X)$;
- Part II: the probability distribution $[Y | Y > 0, X]$ is given by $[Y_+ | X]$, with $E(\log Y | Y > 0, X) = E(\log Y_+ | X) = X^T \beta$.

Variations on this model that use a probit link in place of the logit link, or some alternative monotone transformation linking Y_+ to X, are clearly possible. Throughout the remainder of the paper, and with a slight abuse of notation, we will not make a distinction between the conditional distributions $[Y_+ | X]$ and $[Y | Y > 0, X]$, understanding that the former is being used to represent the latter.

2.3 Tobit vs. 2PM

The Tobit model and 2PM are not directly comparable through their respective likelihoods due to the inclusion of $y_{\min}$ as part of the Tobit model, and in particular, the requirement that $y_{\min} > 0$ for there to be a point mass at zero. In principle, the Tobit model is more plausible when there exists a detection limit. However, for many data types including alcohol intake, the Tobit model is not appropriate because there is no clinically meaningful definition of the detection limit. More importantly, a zero daily drinking level is considered to be a "true" zero, indicating abstinence which is an important goal to achieve in alcohol dependence studies. The same applies to medical cost data: a zero cost is a "true" zero, indicating a condition free of medical service usage.

The major difference between these two classes of models is that the error term in Equation (6) is only meaningful when $Y > 0$, while in the Type II Tobit model (4) there is no such condition. Manning, Duan and Rogers (1987) conducted a simulation study to evaluate the merits of both models and concluded in favor of the two-part model. Leung and Yu (1996) carried out an independent study and concluded that, in general, the two-part and sample selection model classes are designed to answer distinct inferential questions and both are useful in their respective contexts.

2.4 Zero-Inflated Tobit Model

When there are a large portion of zeros (below the detection limit), the Tobit model needs a large spread (variance) to cover the probability of zeros, which often results in poor fit. Moulton and Halsey (1995) accommodated the features of both the two-part and Tobit models, and proposed a zero-inflated Tobit (ZIT) model. In the ZIT model, the *observed* zero values are assumed to come from two sources: either a "true" zero as in a two-part model, or a left-censored "actual but unobserved" outcome. Formally, given X, the ZIT model is defined as

$$Y = \begin{cases} Y^* I(Y^* > y_{\min}) & \text{with probability } p(X), \\ 0 & \text{with probability } 1 - p(X), \end{cases} \tag{7}$$

where Y^* is as given in (2) and

$$\operatorname{logit}\big(p(X)\big) = X^T\gamma. \tag{8}$$

The case where $p(X)$ is independent of X corresponds to $\gamma = 0$ (i.e., except possibly the intercept); other link functions are also possible.

The ZIT model is another example of a finite mixture model with two components: a point mass at zero and a Tobit model. It contains the Tobit model (i.e., if $p(X) = 1$ for every X) and two part model (i.e., if $p(X) \neq 1$ for every X and $y_{\min} \leq 0$) as special cases. However, an important challenge with the ZIT model is identifiability, similarly to other finite mixture models.

2.5 Marginalized 2PM

The notion of "marginalization" in the context of a 2PM really refers to the overall effect of X on the mean of the observed Y, or $E(Y|X)$. In a 2PM, one has $E(Y|X) = P(Y > 0|X)E(Y_+|X)$; since both $P(Y > 0|X)$ and $E(Y_+|X)$ involve a set of regression parameters intended to capture the effect of X, the product form creates challenges in easily determining the effect of a given covariate on $E(Y|X)$.

Suppose X contains a continuous component, say X_s, and let x_s denote realizations of this component. Then, using models (5) and (6), Liu et al. (2010) considered the overall impact of a covariate on $E(Y|X)$ through

$$\begin{aligned} \frac{\partial \log E(Y|X)}{\partial x_s} &= \frac{\partial \log P(Y > 0|X)}{\partial x_s} + \frac{\partial \log E(Y_+|X)}{\partial x_s} \\ &= \alpha_s P(Y = 0|X) + \beta_s, \end{aligned}$$

where α_s and β_s are the regression coefficients corresponding to X_s. Multiplying $\partial \log E(Y|X)/\partial x_s$ by x_s, one can write

$$x_s \frac{\partial \log E(Y|X)}{\partial x_s} = \frac{\partial E(Y|X)/E(Y|X)}{\partial x_s/x_s},$$

which is the "partial elasticity" of the mean cost with respect to x_s (e.g., Wooldridge, 2002, page 16). Liu et al. (2010) also derive a related formula for binary X.

Smith et al. (2014) proposed an alternative form of the marginalized 2PM for $E(Y|X)$, parameterizing instead the point mass probability $p(X) = P(Y > 0|X)$ as $p(X) = \operatorname{expit}(X^T\alpha)$ and the "marginalized" mean $E(Y|X) = \exp(X^T\gamma)$. In certain cases, it is possible to fit the marginalized 2PM model using software capable of fitting a general 2PM. For example, suppose $[Y_+|X]$ in (6) follows a lognormal distribution with parameters $\mu(X)$ and σ^2; that is, $\log Y_+|X \sim \mu(X) + e$, where $e \sim N(0, \sigma^2)$. Then, under the 2PM,

$$E(Y|X) = p(X)\exp\big(\mu(X) + \sigma^2/2\big).$$

Letting $\mu(X) = X^T\gamma - \log p(X) - \sigma^2/2$, it follows that

$$E(Y|X) = p(X)\exp\big(\mu(X) + \sigma^2/2\big) = \exp\big(X^T\gamma\big),$$

showing that the conventional 2PM can be parameterized in a way that yields the desired marginalized 2PM (c.f. Smith et al., 2014). The corresponding likelihood can be maximized using SAS PROC NLMIXED. Other types of distributions introduced in Section 3, for example, log skew normal (Smith et al., 2014), generalized gamma (Smith et al., 2017a) and Beta (Chai et al., 2018), can also be considered in the marginalized 2PM.

2.6 Remarks

We will mostly focus on the two-part model framework throughout the rest of the review article since, as indicated in our motivating examples, neither medical costs or drinking data have the detection limit feature. Statistically, the Tobit model may be appropriate

when there exists an inflated but small percentage of zero values, for example, to allow for a somewhat better fit compared to a continuous model. SAS code for fitting a Tobit model can be found at http://www.ats.ucla.edu/stat/sas/faq/cens_nlmixed.htm, and SAS code designed to fit the ZIT model may be found in Berk and Lachenbruch (2002).

3. MODELING POSITIVE VALUES IN PART II OF THE 2PM

In this section, we will review methods to model positive values in Part II of the 2PM. Recall that we use $[Y_+|X]$ to represent the conditional distribution $[Y|Y>0, X]$, and as a continuation of that slight abuse in notation, we will use Y_+ to represent the positive values of Y.

Historically, a logarithmic transformation of the response Y_+ is most commonly used to tackle positive values with skewness and heteroscedasticity, followed by regressing the transformed Y_+ on covariates X, as in Model (6). However, using a transformed Y_+ (e.g., $\log Y_+$) for regression no longer models the mean response on the original scale of interest. This presents a problem, especially in the area of predicting medical costs. Below, we discuss the retransformation issue and contrast that approach with one based on generalized linear models (GLMs), where the need for retransformation is avoided. In addition, although commonly used in practice, the use of a log normal distribution for $[Y_+|X]$ may not be adequate for describing right skewness and potential heteroscedasticity; hence, also reviewed below are three useful parametric extensions of the log normal distribution and some related approaches for tackling heteroscedasticity. Finally, we will discuss several other possible adaptations of the 2PM.

3.1 GLM vs. Log Transformation

In Model (6), a transformed Y_+ (e.g., $\log Y_+$) is taken as the dependent variable. However, we are often more interested in $E(Y_+|X)$ (e.g., dollars) than $E(\log Y_+|X)$ (log dollars), especially for prediction purposes. Therefore, a retransformation mapping the results of the analysis done on a transformed scale back to the scale of interest is needed to draw inference about $E(Y_+|X)$ (Manning, 1998, Manning and Mullahy, 2001).

In Model (6), $E(Y_+|X) = \exp(X^T\beta)E(\exp(e))$. If the error $e \sim N(0,\sigma^2)$, then $E(Y_+|X) = \exp(X^T\beta) \times \exp(\sigma^2/2)$. If the error distribution is unknown but homoscedastic (having the same error variance), retransformation is possible using a "smearing estimate" (Duan, 1983) that uses the average of $\exp(\hat{e})$ to estimate $E(\exp(e))$, where $\hat{e}$ is the residual in Model (6). However, in the presence of heteroscedasticity, the smearing estimate becomes very complicated and can sometimes fail (Manning and Mullahy, 2001).

To avoid the need for retransformation, generalized linear models (GLMs) have been proposed, for example, see Blough, Madden and Hornbrook (1999) and Mullahy (1998). For example, instead of estimating $E(\log Y_+|X)$, a GLM can be set up to model $\log E(Y_+|X)$ to obtain the mean on the desired scale. These models simultaneously describe the link and variance structure by pre-specified functions, for example, log link with a gamma error distribution. However, an issue in GLMs is to determine the proper distributional form. Manning and Mullahy (2001) proposed a modified Park's test (Park, 1966) by regressing the log squared residual $\hat{e}^2$ on log of the prediction $\hat{Y}_+$, that is, $\log(\hat{e}^2) = \alpha_0 + \alpha_1 \log \hat{Y}_+ + \epsilon$. The significance in testing $\alpha_1 = 0$ indicates heteroscedasticity, and the magnitude of α_1 can be used to determine the distribution, for example, $\alpha_1 = 0$: Gaussian; $\alpha_1 = 1$: Poisson; $\alpha_1 = 2$: Gamma; and $\alpha_1 = 3$: Inverse Gaussian or Wald. Further methods for characterizing and modeling heteroscedasticy are given in Section 3.3.

3.2 Three Parametric Extensions to Log Normal Distribution

Liu et al. (2016b) considered 3 extensions of the log normal distribution: generalized gamma distribution, log skew normal distribution, and normal distribution after Box–Cox transformation. All 3 extensions have 3 parameters, and all contain the 2-parameter log normal distribution as a special case.

3.2.1 *Generalized gamma distribution*. For a generalized gamma distribution with 3 parameters (κ: shape, μ: location, σ: scale), the density function is (Manning, Basu and Mullahy, 2005, Liu et al., 2010, 2016b)

$$f(y_+;\kappa,\mu,\sigma) = \frac{\eta^\eta}{\sigma y_+ \Gamma(\eta)\sqrt{\eta}} \times \exp\left[u\sqrt{\eta} - \eta\exp(|\kappa|u)\right], \tag{9}$$

where $\eta = |\kappa|^{-2} > 0$ and $u = \text{sign}(\kappa)(\log y_+ - \mu)/\sigma$ depends on the sign of κ. If Y_+ is a random variable with density (9),

$$E(Y_+) = \exp\left\{\mu + \frac{\sigma\log(\kappa^2)}{\kappa} + \log\left[\Gamma\left(1/\kappa^2 + \sigma/\kappa\right)\right] - \log\left[\Gamma\left(1/\kappa^2\right)\right]\right\}$$

and

$$\operatorname{Var}(Y_+) = \{\exp(\mu)\kappa^{2\sigma/\kappa}\}^2\left\{\frac{\Gamma(1/\kappa^2+2\sigma/\kappa)}{\Gamma(1/\kappa^2)} - \left[\frac{\Gamma(1/\kappa^2+2\sigma/\kappa)}{\Gamma(1/\kappa^2)}\right]^{-2}\right\}.$$

The generalized gamma distribution is very flexible, including the standard gamma, inverse gamma, Weibull and lognormal distributions as special or limiting cases. For example, if $\sigma = \kappa$, the generalized gamma distribution (9) reduces to a standard gamma with shape parameter $\eta = \kappa^{-2}$ and scale parameter $\nu = \kappa^2\exp(\mu)$, that is,

$$f(y_+;\nu,\eta) = \frac{1}{\nu^\eta\Gamma(\eta)}y_+^{\eta-1}\exp(-y_+/\nu),$$

with mean $\exp(\mu)$ and variance $\kappa^2\exp(2\mu)$. If we set $\kappa = -\sigma$ with $\sigma > 0$, we will obtain the inverse gamma distribution with density

$$f(y_+;\epsilon,\eta) = \frac{\epsilon^\eta}{\Gamma(\eta)}\left(\frac{1}{y_+}\right)^{\eta+1}\exp(-\epsilon/y_+),$$

with shape parameter η and scale parameter $\epsilon = \eta e^\mu$ (e.g., Robert, 2007, page 520). The Weibull and lognormal distributions are obtained by fixing κ. Specifically, setting $\kappa = 1$ reduces the generalized gamma to a Weibull distribution with shape parameter $1/\sigma$ and scale parameter e^μ. Alternatively, as $\kappa \to 0$, we obtain a log normal probability density function with parameters μ and σ:

$$f(y_+;\mu,\sigma) = \frac{1}{\sigma y_+\sqrt{2\pi}}\exp\left\{-\frac{(\log y_+ - \mu)^2}{2\sigma^2}\right\}.$$

3.2.2 *Log skew normal distribution.* The density of the skew normal distribution is given by (e.g., see Chai and Bailey, 2008)

$$f(z;\lambda,\mu,\sigma) = \frac{2}{\sqrt{\sigma^2+\lambda^2}}\phi\left(\frac{z-\mu}{\sqrt{\sigma^2+\lambda^2}}\right) \times \Phi\left(\frac{\lambda}{\sigma}\frac{z-\mu}{\sqrt{\sigma^2+\lambda^2}}\right),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions, respectively. The three parameters are: λ: skewness; μ: location; σ: scale. If $\lambda = 0$, the skew normal distribution reduces to the normal distribution.

The density of the log skew normal distribution for $y_+ = \exp(z)$ is

$$f(y_+;\lambda,\mu,\sigma) = \frac{1}{y_+}\frac{2}{\sqrt{\sigma^2+\lambda^2}}\phi\left(\frac{\log y_+ - \mu}{\sqrt{\sigma^2+\lambda^2}}\right) \times \Phi\left(\frac{\lambda}{\sigma}\frac{\log y_+ - \mu}{\sqrt{\sigma^2+\lambda^2}}\right). \tag{10}$$

3.2.3 *Normal after Box–Cox transformation.* The Box–Cox transformation (Box and Cox, 1964) is given by

$$v(y_+) = \begin{cases}\gamma^{-1}(y_+^\gamma - 1) & \gamma \neq 0,\\ \log y_+ & \gamma = 0,\end{cases}$$

where γ is a parameter to be estimated. Assuming that Y_+ has some probability distribution, the "Normal after Box–Cox transformation" assumes that $v(Y_+) \sim N(\mu,\sigma^2)$, that is, the transformed value follows a normal distribution. The log normal distribution for Y_+ is a special case at $\gamma = 0$.

3.3 Modeling Heteroscedasticity

In both the ordinary least squares (OLS) and GLM frameworks, a misspecified variance function can lead to a substantial loss of efficiency for parameter estimates. In this section, we will review several possible methods for modeling heteroscedasticity.

A simple solution is to model the scale parameter as a function of covariates. For example, in all three distributions introduced in Section 3.2, we can have the following model for heterogeneity (scale parameter):

$$\sigma^2(X) = \exp\bigl(X^T\delta\bigr), \tag{11}$$

where δ is a parameter to be estimated.

Alternatively, we can assume the scale parameter to be a function of the mean. For example, Basu and Rathous (2005) assumed

$$\operatorname{Var}(Y_+|X) = h\bigl(\mu(X)\bigr), \tag{12}$$

where $h(\cdot)$ has a known functional form, for example, $h(\mu) = \theta_1\mu^{\theta_2}$ or $h(\mu) = \theta_1\mu + \theta_2\mu^2$ for parameters θ_1 and θ_2, and μ is the mean. It is also possible to take $\log(h(\mu(X)))$ as an unknown but smooth function whose form could be estimated by nonparametric regression, for example, splines, as in Chen et al. (2013a). Since there is no need to assume any specific form of the variance structure, this approach is more robust in fitting data with heteroscedasticity.

Finally, along the lines of Rigby and Stasinopoulos (2005), we can also consider

$$\sigma^2(X,z_1,\ldots,z_m) = \exp\bigl(X^T\alpha + h_1(z_1) + \cdots + h_m(z_m)\bigr), \tag{13}$$

where $(z_1,\ldots,z_m)^T$ is a $m \times 1$ vector of continuous variables. This model is more appealing when one is interested in the associations between covariates and the variance structure. However, when there exist multiple smooth functions for nonlinear covariate associations, it becomes more complicated computationally than assuming a single smooth function (of mean μ) as in Chen et al. (2013a).

3.4 Other Models for Positive Values

3.4.1 *Cox proportional hazards model*. The Cox proportional hazards model has been used for positive medical costs (Dudley et al., 1993, Lipscomb et al., 1998). Of note, we assume there is no censoring in such data—we observe the complete cost for each subject, which is different from censored medical costs where cumulative medical costs could be censored by events such as end of study or drop out (e.g., Lin et al., 1997; Jain and Strawderman, 2002).

The Cox model is a semiparametric alternative to the OLS and GLM methods, where $[Y_+|X]$ is modeled through the hazard function specification

$$\lambda(y_+|X) = \lambda_0(y_+)\exp\bigl(X^T\alpha\bigr), \tag{14}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function (at $X = 0$). This model can accommodate an arbitrary shape for the distribution of Y_+. The baseline cumulative hazard function $\Lambda_0(y_+) = \int_0^{y_+} \lambda_0(u)\,du$ can be estimated nonparametrically by the Nelson–Alalen or Kaplan–Meier estimator (Kalbfleisch and Prentice, 2002, Section 4.3); alternatively, splines or other forms of smoothing can be used to obtain a smooth estimate of $\lambda_0(y_+)$ (Lipscomb et al., 1998). It should be noted that the interpretation of the regression coefficient α (or hazard ratio) is different from that in the OLS or GLM models. For example, if Y_+ represents medical cost and X is a binary treatment indicator, then $\exp(\alpha)$ indicates the ratio of conditional probabilities between treatment and placebo of no additional cost given y_+ dollars have been spent. Thus, it is not appropriate to directly compare hazard ratios with regression coefficients in the OLS or GLM models.

The corresponding conditional density function of Y_+ can be written

$$p(y_+|X) = \lambda_0(y_+)\exp\bigl(X^T\alpha\bigr) \times \exp\bigl(-\exp\bigl(X^T\alpha\bigr)\Lambda_0(y_+)\bigr),$$

where

$$E(Y_+|X) = \int_0^\infty sp(s|X)\,ds = \int_0^\infty \exp\bigl(-\exp\bigl(X^T\alpha\bigr)\Lambda_0(s)\bigr)\,ds.$$

The key to the Cox model is the proportionality assumption, which requires the effects of the predictor variables upon the hazard function to be additive on the indicated scale and constant over "time" (e.g., cost). The appropriateness of the proportional hazards assumption can be informally tested by examining the interaction of the fixed effects with certain functions of time. Schoenfeld residuals are often used to give graphical evidence of the nonproportionality (Schoenfeld, 1982). Therneau and Grambsch (2000) (Chapter 6) gave a more general discussion on testing proportional hazards. Basu and Manning (2006) proposed a novel test to detect the nonproportionality of hazards within the class of exponential conditional mean models, based on the coefficients of the generalized gamma regression model.

In simulation studies, Basu, Manning and Mullahy (2004) demonstrated the poor performance of the Cox model in estimating the expected outcome, $E(Y_+|X)$ when the proportionality assumption fails. To tackle this issue, survival analysis techniques without the proportional hazards assumption can be used in modeling medical cost data, for example, Martinussen and Scheike (2006), Tian, Zucker and Wei (2005), Jain and Strawderman (2002), and Wooldridge (2002).

3.4.2 *Four-part model*. In medical cost data, there is often a marked difference between outpatient and inpatient costs when the medical costs are positive. In the example of heart failure patients in the clinical data repository (CDR) database, the monthly costs for individuals who had only outpatient services have a mean of \$813 and standard deviation of \$2058, while the monthly costs among individuals with at least one inpatient service have a mean of \$15,457 and standard deviation of \$30,004. Duan et al. (1983) considered a four-part model for medical cost data, further separating outpatient and inpatient costs. Part I is a logistic model for the indicator of any positive costs

$$\operatorname{logit} P(U = 1|X) = X^T\alpha. \tag{15}$$

Part II is another logistic model for the indicator of any inpatient costs conditional on any medical services

$$\operatorname{logit} P(V = 1|U = 1, X) = X^T\beta. \tag{16}$$

Part III describes the magnitude of outpatient (i.e., $U = 1$, $V = 0$) costs

$$\log(Y)|\{U = 1, V = 0, X\} \sim X^T\gamma + e^c, \tag{17}$$

and Part IV describes the inpatient ($V = 1$) costs

$$\log(Y)|\{U = 1, V = 1, X\} \sim X^T\delta + e^d. \tag{18}$$

3.4.3 *Comparison of one-part vs. multipart model*. Duan et al. (1983) compared the one-part, two-part, and four-part models with applications to a large portion of the Rand Health Insurance Experiment (RHIE) data. They found that both the two-part model and the four-part model performed better than the one-part

model in terms of consistency and accuracy for making predictions. They expected the four-part model would outperform the two-part model if more data were available for analysis.

Smith et al. (2017a) advised that one-part GLMs be avoided since they may yield biased and unreliable results for datasets with >10% zeros. They also showed that parametric two-part marginalized models outperform one-part GLMs in many settings, particularly when the focus is on estimating treatment effects.

3.4.4 *Beta distribution for proportions*. Zero-inflated proportions are commonly present in compositional data. For example, microbiome compositional data (denoted by relative abundance) can be represented as proportions, though may be skewed and contain many zeros. Such data can be modeled by a two-part model with a Beta distribution for modeling proportions, that is,

$$\text{logit}\big(p(X)\big) = X^T\alpha, \tag{19}$$

$$\text{logit}\big(\mu(X)\big) = X^T\beta, \tag{20}$$

where $[Y|Y>0]$ follows a Beta distribution with mean $\mu(X) = E(Y|Y>0, X) \in (0,1)$; see, for example, Chen and Li (2016). A logistic transformation is used to address the presence of skewness in the proportion outcome.

4. CORRELATED ZERO-INFLATED CONTINUOUS DATA

4.1 Random Effects 2PM

More recently, there has been an increasing interest in analyzing correlated zero-inflated continuous data. The correlation may stem from the structure of clustered data, where the outcomes (e.g., medical costs) of subjects from the same cluster are correlated due to similarities in their health status, socioeconomic characteristics, and shared genetic traits. Another source of correlation arises in longitudinal data where repeated measures (e.g., daily drinking levels) are correlated for the same subject. Examples of correlated zero-inflated continuous data include pharmacy costs for patients clustered within physicians (Zhang et al., 2006, Liu et al., 2010), longitudinal (e.g., monthly) medical costs (Liu et al., 2008), longitudinal drinking levels (Liu, Ma and Johnson, 2008, Liu et al., 2016b), and repeated measures of microbiome data (Chen and Li, 2016).

Olsen and Schafer (2001) and Tooze, Grunwald and Jones (2002) proposed a random effects two-part model to describe repeated measures of semicontinuous data. We will use daily drinking outcomes as an example to illustrate this model. Denote by Y_{ij} a semicontinuous repeated measure for the jth observation of subject i, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$. The random effects 2PM can be written as

$$\text{logit}\, P(Y_{ij} > 0|X_{ij}, Z_{ij}, a_i, b_i) = X_{ij}^T\alpha + Z_{ij}^T a_i, \tag{21}$$

$$\log Y_{+ij}|\{X_{ij}, Z_{ij}, a_i, b_i\} \sim X_{ij}^T\beta + Z_{ij}^T b_i + e_{ij}, \tag{22}$$

where $(a_i^T, b_i^T)^T \sim N(0, \Sigma)$ are correlated random effects, Σ is a positive definite matrix, X_{ij} and Z_{ij} are respectively covariate vectors for fixed and random effects, and extending an earlier notational convention that will be used going forward, $[Y_{+ij}|X_{ij}, Z_{ij}, a_i, b_i]$ represents a model for $[Y_{ij}|Y_{ij} > 0, X_{ij}, Z_{ij}, a_i, b_i]$. We therefore have two correlated models: a generalized linear mixed model (21) (e.g., for the longitudinal binary outcome of drinks being positive) and a linear mixed model (22) (e.g., for the level of positive drinks). Zhang et al. (2006) consider a Bayesian version of this model.

The random effects are used to describe two types of correlation: within-part and cross-part correlations. The within-part correlation for longitudinal drinking data arises among repeated measures of abstinence, which is characterized by random effect a_i in Part I; while in Part II, the repeated measures of the amount of alcohol on a drinking day are correlated for the same subject, indicated by random effect b_i. The cross-part correlation, specified by the correlation between a_i and b_i (or the covariance in the off-diagonal elements of matrix Σ), is used to describe the relation between frequency ("how often") and amount ("how much") of drinking; for example, subjects more likely to be abstinent may tend to drink less on a drinking day. Cross-part correlation may also exist for medical cost data: patients in poorer health may be more likely to seek medical treatments, and in the event that they do, one may expect their costs to be higher.

Statistically, ignoring the within-part correlation may lead to bias in the estimation of standard errors of parameter estimates, while ignoring the cross-part correlation could result in biased estimates in coefficients of Model (22). For example, Albert (2005) and Su, Tom and Farewell (2009) showed that a positive cross-part correlation between random intercepts in Models (21) and (22) results in a positively biased estimate of intercept β_0 in (22) if one estimates the model parameters assuming cross-part independence, as was done in Lu, Lin and Shih (2004). Conceptually, the source

of this bias may be viewed as arising from the presence of an informative cluster size; when the random effects a_i and b_i are correlated, the propensity for having a positive response (i.e., Part I) will influence the observed number of positive responses available for estimating the parameters in Part II. A GEE-type approach that ignores this dependence will estimate the parameters of Part II (i.e., (22)) separately from those in Part I (i.e., (21)), and leads to a biased estimate of the (subject-specific) regression effects in (22) because larger clusters of observations will exhibit greater influence on the regression parameter estimates in comparison to smaller clusters (Williamson, Datta and Satten, 2003). A related view of this same problem can be adapted from the discussion provided by Su, Tom and Farewell (2009). Consider a randomly selected subject i and suppose that a_i and b_i are each scalar (i.e., a random intercept model). When a_i is large and positive, Model (21) shows that there is a higher propensity of seeing positive responses across j. Suppose also that $\rho = \operatorname{cor}(a_i, b_i) > 0$; then, b_i is more likely to be large and positive, the propensity being dependent on the magnitude of ρ. In combination, this results in an increased number of positive, larger Y_{ij}s. In a GEE-type approach that essentially treats the estimation of the two parts separately, those subjects with larger values of b_i can be expected to contribute more observations to estimating β, and these observations will also tend to be larger than average. Consequently, an estimate of β in Model (22) that is derived from the reduced sample of subjects having positive observations will typically be biased, and the magnitude of this bias will depend on the magnitude of ρ. Although the situation is more complicated with random effects models that are more complex, similar conclusions apply; see, for example, Su, Tom and Farewell (2009). Thus, GEE-type marginal models are not recommended when there exists the possibility of a strong cross-part correlation.

4.2 Extensions in Part II

In Part II of the random effect 2PM (22) we assume that the positive values follow a log normal distribution. As in Section 3.2 for cross-sectional data, we can use extensions of the log normal distribution for correlated semicontinuous data. For example, Liu et al. (2010) considered a generalized gamma distribution in Part II, with the location parameter

$$\mu_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i, \tag{23}$$

and scale parameter (heteroscedasticity term)

$$\sigma_{ij}^2 = \exp(X_{ij}^T\delta). \tag{24}$$

A random effects 2PM with a log skew normal or normal after Box–Cox transformation in Part II can be similarly defined (Liu et al., 2016b). Of note, Mahmud, Lou and Johnston (2010) proposed a log skew normal distribution, while Tooze et al. (2006) proposed a Box–Cox transformation in Part II of the random effects 2PM. However, neither addressed heteroscedasticity as in Model (24).

4.3 Random Effects 4PM

Liu et al. (2008) extended the 4 part model in Section 3.4.2 to correlated zero-inflated data with random effects. Part I is a logistic model for the indicator of any positive costs

$$\operatorname{logit} P(U_{ij} = 1|X_{ij}) = X_{ij}^T\alpha + a_i. \tag{25}$$

Another logistic model is used in Part II for the indicator of any inpatient costs conditional on any medical services

$$\operatorname{logit} P(V_{ij} = 1|U_{ij} = 1, X_{ij}) = X_{ij}^T\beta + b_i. \tag{26}$$

The level of outpatient and inpatient costs are described in Parts III and IV:

$$\log(Y_{ij})|\{X_{ij}, U_{ij}, V_{ij}\} \sim \begin{cases} X_{ij}^T\gamma + c_i + e_{ij}^c & \text{for monthly outpatient costs } (U_{ij} = 1, V_{ij} = 0), \\ X_{ij}^T\delta + d_i + e_{ij}^d & \text{for monthly inpatient costs } (U_{ij} = 1, V_{ij} = 1). \end{cases} \tag{27}$$

Although not explicit in the notation, each of these parts is also specified conditionally upon the full set of random effects, that is (a_i, b_i, c_i, d_i). The within- and cross-part correlations are denoted by the correlation among the four random effects $(a_i, b_i, c_i, d_i)^T \sim N(0, \Sigma)$, where Σ is a positive definite matrix. In an example of monthly medical costs of heart failure patients, Liu et al. (2008) found several highly significant cross-part associations: (i) patients with higher odds of positive monthly medical costs were more likely to incur higher outpatient costs (Parts I and III); (ii) higher odds of hospitalization were correlated to higher outpatient costs (Parts II and III); (iii) higher outpatient costs were associated with higher inpatient costs (Parts III and IV).

4.4 Random Effects 2PM for Proportions

The two-part model with a Beta distribution in Section 3.4.3 can be extended to correlated proportion data by adding random effects to capture the "within-part" and "cross-part" correlation:

$$\text{logit}(p_{ij}) = X_{ij}^T\alpha + Z_{ij}^T a_i, \tag{28}$$

$$\text{logit}(\mu_{ij}) = X_{ij}^T\beta + Z_{ij}^T b_i, \tag{29}$$

where $(a_i, b_i)^T \sim N(0, \Sigma)$ are correlated random effects and Σ is a positive definite matrix. The model specifications for p_{ij} and μ_{ij} are both conditional on X_{ij}, Z_{ij}, a_i and b_i. Of note, Chen and Li (2016) developed such a two-part logistic-Beta regression model with random effects to account for the "within-part correlation," but did not consider "cross-part correlation" (i.e., Σ is diagonal).

4.5 Multilevel 2PM

Semicontinuous data may have a hierarchical structure. For example, longitudinal zero-inflated continuous outcomes are recorded for subjects, which are further clustered within families, hospitals, or communities. Consider a three-level 2PM, where level 1 consists of longitudinal daily drinking records for each subject, level 2 is the subject level, and level 3 is the cluster (e.g., family) level. Denote by Y_{ijk} the kth daily drinking record of subject j in cluster i. Liu, Ma and Johnson (2008) proposed a three-level random effects 2PM:

$$\begin{aligned}&\text{logit}\, P(Y_{ijk} > 0 | X_{ijk}, u_i, v_i, a_{ij}, b_{ij})\\ &\quad = X_{ijk}^T\alpha + u_i + a_{ij},\end{aligned} \tag{30}$$

$$\begin{aligned}&\log Y_{+ijk} | \{X_{ijk}, u_i, v_i, a_{ij}, b_{ij}\}\\ &\quad \sim X_{ijk}^T\beta + v_i + b_{ij} + e_{ijk},\end{aligned} \tag{31}$$

where u_i and a_{ij} are random effects in Part I, v_i and b_{ij} are random effects in Part II, and Models (30) and (31) are each specified conditionally on $X_{ijk}, u_i, v_i, a_{ij}, b_{ij}$. Of note, u_i and v_i are random effects at the cluster level (level 3), and a_{ij} and b_{ij} are random effects at the subject level (level 2). We assume $(u_i, v_i)^T \sim N(0, \Sigma_1)$ is independent of $(a_{ij}, b_{ij})^T \sim N(0, \Sigma_2)$ for all i and j. The error term $e_{ijk} \sim N(0, \sigma_e^2)$ is independent of random effects $(u_i, v_i, a_{ij}, b_{ij})$.

4.6 Estimation in Multipart Models with Random Effects

Parameter estimation in the afore-mentioned random effects models is often challenging due to the appearance of an integral of a complicated nonlinear function with respect to the multivariate normal density for the random effects in the loglikelihood function. Several estimation options are available. Olsen and Schafer (2001) adopted the high (sixth) order Laplace approximation (Raudenbush, Yang and Yosef, 2000), while Tooze, Grunwald and Jones (2002) used the adaptive Gaussian quadrature (AGQ). Based on our experiences (Liu et al., 2008, 2010), both methods yield accurate estimates. The high order Laplace approximation is much faster in computational time and it can also handle more random effects then AGQ. However, its implementation is very challenging as substantial mathematical derivation involving higher order matrix and vector differentiation is needed. As a result, no ready-to-use software is available for this method. On the contrary, the AGQ is much easier to implement, for example, in SAS Proc NLMIXED (Littell et al., 2006).

However, for the multilevel zero-inflated continuous model (30) and (31), SAS Proc NLMIXED cannot handle the multilevel random effects well. Nonadaptive Gaussian quadrature, such as the aML Multiprocess Multilevel Modeling software available at http://www.applied-ml.com, can be used for multilevel data.

In addition to these frequentist methods, Bayesian approaches have been adopted as an alternative solution to this often intractable problem (e.g., Zhang et al., 2006, Cooper et al., 2007, Neelon, O'Malley and Normand, 2011, Smith et al., 2017b).

4.7 Random Effects Tobit/ZIT Model

The random effects Tobit model is an alternative to the random effects multipart models for correlated semicontinuous data. Random effects Tobit models have been used to assess the effects of the conversion of eligible hospitals to critical access hospitals on hospital patient safety (Li, Schneider and Ward, 2007) and to investigate the association between arsenic exposure and oxidative stress (Breton et al., 2007), among others. Twisk and Rijmen (2009) compared the performance of the random effects Tobit model with the linear mixed model using data from a longitudinal rehabilitation study among stroke patients.

Denote by Y_{ij}^* the actual outcome measured for the jth observation of subject i. The corresponding random effects Tobit model is

$$Y_{ij}^* | \{X_{ij}, Z_{ij}, r_i\} \sim X_{ij}^T\beta + Z_{ij}^T r_i + e_{ij}, \tag{32}$$

$$Y_{ij} = Y_{ij}^* I\big(Y_{ij}^* > y_{\min}\big), \tag{33}$$

where X_{ij} and Z_{ij} are the covariate vectors for fixed effect β and random effect $r_i \sim N(0, \Sigma_r)$, with Σ_r being a positive definite covariance matrix.

Berk and Lachenbruch (2002) extended the ZIT model with the addition of random effects and applied the model to a study of the private speech of children. Bjerre et al. (2007) used the model to study the hospital care utilization and sick leave in an intervention program to decrease alcohol intake in driving-while-impaired offenders. The random effects ZIT model is

$$Y_{ij}^*|\{X_{ij}, Z_{ij}, r_i\} \sim X_{ij}^T\beta + Z_{ij}^T r_i + e_{ij}, \tag{34}$$

$$Y_{ij} = \begin{cases} Y_{ij}^* I(Y_{ij}^* > y_{\min}) \\ \quad \text{with probability } p_{ij}, \\ 0 \quad \text{with probability } 1 - p_{ij}. \end{cases} \tag{35}$$

The probability p_{ij} may depend on covariates and random effects, such as

$$\text{logit}(p_{ij}) = X_{ij}^T\delta + Z_{ij}^T a_i, \tag{36}$$

in which case the full model specification is given conditionally on $\{X_{ij}, Z_{ij}, r_i, a_i\}$ with $(r_i^T, a_i^T)^T \sim N(0, \Sigma_d)$.

The random effects Tobit model can be fit in several software packages. The Tobit model with random intercept can be fit by function *survreg* in R. SAS Proc NLMIXED can be used to fit Tobit models with more complicated random effects and the random effects ZIT model (Berk and Lachenbruch, 2002).

4.8 Marginalized 2PM with Random Effects

Smith et al. (2017b) extended the marginalized two-part model in Section 2.5 to longitudinal semicontinuous data, proposing a marginalized random effects 2PM:

$$\text{logit}\, P(Y_{ij} > 0|X_{ij}, Z_{ij}, a_i, b_i) = X_{ij}^T\alpha + Z_{ij}^T a_i, \tag{37}$$

$$\log E(Y_{ij}|X_{ij}, Z_{ij}, a_i, b_i) = X_{ij}^T\gamma + Z_{ij}^T b_i, \tag{38}$$

where both models are specified conditionally on all covariates and random effects. A fully Bayesian approach that can accommodate correlated multidimensional random effects was used for estimation to improve computational tractability relative to maximum likelihood. It is of interest to adapt such marginalized random effects 2PM to other distributions, for example, generalized gamma in (38).

5. OTHER TYPES OF ZERO-INFLATED DATA

Zero-inflation is commonly encountered with other types of data, particularly counts. Examples of zero-inflated count data include health care utilization (e.g., number of days of hospitalization) and school attendance (e.g., number of days absent), among others. Suppose that a subject has a probability p of being $Y = 0$, and a probability of $1 - p$ to have Poisson or negative binomial distribution (Lambert, 1992, Hall, 2000). Unlike the continuous setting, a Poisson or negative binomial distribution can have zero as a possible realization; hence, observed zeros can either represent a "true zero" or a realization of 0 from the Poisson or negative binomial distribution ("random zero"). This differs from the 2PM for zero-inflated continuous data, where true zeros are easily distinguished from positive values.

Although not technically "zero inflated" in the sense considered thus far, "cure models" for single event data share some important similarities to zero-inflated count data. In such models, there is the presumed presence of long term survivors, which may be reflected by heavy censoring at the end of the follow-up period. Viewed from a counting process perspective, where $N_i(t)$ is a monotone binary process that starts and remains at zero until an event occurs for subject i, the resulting event counts at the end of follow-up may be characterized by a substantial proportion of zeros. Similarly to the case of zero-inflated count data, these zeros may arise because a subject is a long-term survivor (i.e., not susceptible to the event) or because a subject is susceptible but has merely been right-censored.

As will be seen in the following two subsections, both types of data can be viewed as special cases of finite mixture (or latent class) models involving two distributions, similar in spirit to the zero-inflated Tobit model.

5.1 Zero-Inflated Count Data

Lambert (1992) first proposed a zero-inflated Poisson (ZIP) model for count data with excess zeros. In this model, zero values can come from either a true (i.e., structural) zero or the realization of a random zero from the Poisson distribution. Let $A = 1$ denote a "true zero" and $A = 2$ otherwise. Then, the ZIP model is

$$\text{logit}\, P(A = 2|X) = X^T\alpha, \tag{39}$$

$$Y|\{X, A = 2\} \sim \text{Poisson}\big(\theta(X)\big), \tag{40}$$

where $\theta(X) = \exp(X^T\beta)$. The corresponding loglikelihood is

$$\begin{aligned} &I(Y = 0)\log\big(1 - p(X) + p(X)e^{-\theta(X)}\big) \\ &\quad + I(Y > 0)\big[\log p(X) - \theta(X) + Y\log\big(\theta(X)\big) \\ &\quad - \log(Y!)\big], \end{aligned}$$

where $p(X) = P(A = 2|X)$. The ZIP model can be fitted in SAS Proc GENMOD, for example, http://

www.ats.ucla.edu/stat/sas/dae/zipreg.htm. A zero inflated negative binomial (ZINB) model can be similarly constructed.

Hall (2000) extended the ZIP and ZINB models to correlated data with random effects. Denote by Y_{ij} the count outcome for subject j in cluster i, and X_{ij}, A_{ij} defined similarly as above, the model is

$$\text{(41)}\quad \text{logit}\, P(A_{ij}=2|X_{ij},b_i)=X_{ij}^T\alpha,$$

$$\text{(42)}\quad Y_{ij}|\{X_{ij},b_i,A_{ij}=2\}\sim \text{Poisson}(\theta_{ij}),$$

where $\theta_{ij}=\exp(X_{ij}^T\beta+b_i)$, and b_i is a random effect to capture the association among the correlated outcomes from the Poisson distribution. A random effect can also be added to (41) (e.g., Min and Agresti, 2005).

5.2 Cure Models

Cure models for a single event type (Boag, 1949, Farewell, 1982, Kuk and Chen, 1992, Peng, 2000, 2003, Sy and Taylor, 2000 and Peng, Taylor and Yu, 2007) have been a useful tool to analyze survival data, including in cancer studies (Sposto, 2002, Othus et al., 2012). Let $A=1$ denote "cured" (i.e., not susceptible to the event) and $A=2$ otherwise. Then, a Cox proportional hazards cure model for a single event is

$$\text{(43)}\quad \text{logit}\, P(A=2|X)=X^T\alpha,$$

$$\text{(44)}\quad \lambda(t|X,A=2)=\lambda_0(t)\exp\bigl(X^T\beta\bigr).$$

This model assumes that a subject's cure status is a baseline quantity that is inherent to the subject, and that both the propensity of cure and cure status does not change with time. A subject that experiences the event during follow-up clearly has $A=2$. However, subjects that do not experience an event before being censored represent a potential mixture of subjects with $A=1$ and $A=2$. Similarly to zero-inflated count data, a challenge with this model is that the ascertainment of cure is difficult to identify from the observed data without further information. For example, patients who do not experience a cancer-related event over a long period of time (e.g., 5 years) might be considered "cured" from a clinical perspective. If this is a priori taken to be the definition of cure, the ability to identify subjects having $A=1$ clearly improves, despite the reality that these patients may simply be at greatly reduced risk for a future event.

The model given by (43) and (44) can also be viewed as a special case of a latent class model with two classes, where

$$\text{(45)}\quad \lambda(t|X)=\begin{cases}\lambda_0(t)\exp\bigl(X^T\beta\bigr) & \text{with probability } p(X)=P(A=2|X),\\ 0 & \text{with probability } 1-p(X)=P(A=1|X).\end{cases}$$

As an alternative, we can also consider other survival models, such as the accelerated failure time (AFT) model, to describe the survival time of noncured subjects. For example, the model (44) can be replaced by its log-linear model equivalent

$$\text{(46)}\quad \log T|\{X,A=2\}\sim -X^T\beta+e.$$

The baseline distribution, equivalently error distribution e, in (46) can either be parametric, as in (Yamaguchi, 1992), or nonparametric, as in (Li and Taylor, 2002).

In related fashion, recurrent event data may also exhibit the characteristic inflation of zero counts associated with the standard cure model. We have previously analyzed data on four different diseases: recurrent inpatient services for end stage renal disease patients (Liu, Wolfe and Huang, 2004), recurrent tumor occurrences in a soft tissue sarcoma study (Liu and Huang, 2008), recurrent hospital visits for heart failure patients (Liu et al., 2008, Liu, Ma and Johnson, 2008), and recurrent opportunistic diseases for AIDS patients (Liu, 2009). In each of these studies, a large portion (often 1/2 to 2/3) of subjects had no recurrent events by the end of follow-up. When this occurs, there may be nonsusceptible subjects whose event intensity is zero, that is, "cured" in the same sense described above.

Let $A_i=1$ denote "nonsusceptible" and $A_i=2$ "susceptible" for subject i. As with a single event, subjects experiencing zero recurrent events may fall into either class. Rondeau et al. (2013) proposed a zero-inflated frailty model to analyze such data:

$$\text{(47)}\quad \text{logit}\, P(A_i=2|X_i)=X_i^T\alpha,$$

$$\text{(48)}\quad \lambda(t|X_i,\nu_i,A_i=2)=\lambda_0(t)\exp\bigl(X_i^T\beta+\nu_i\bigr),$$

where ν_i is the frailty (random effect) shared by recurrent events experienced by the ith subject and (48) describes the recurrent event intensity among susceptible subjects that remain at risk.

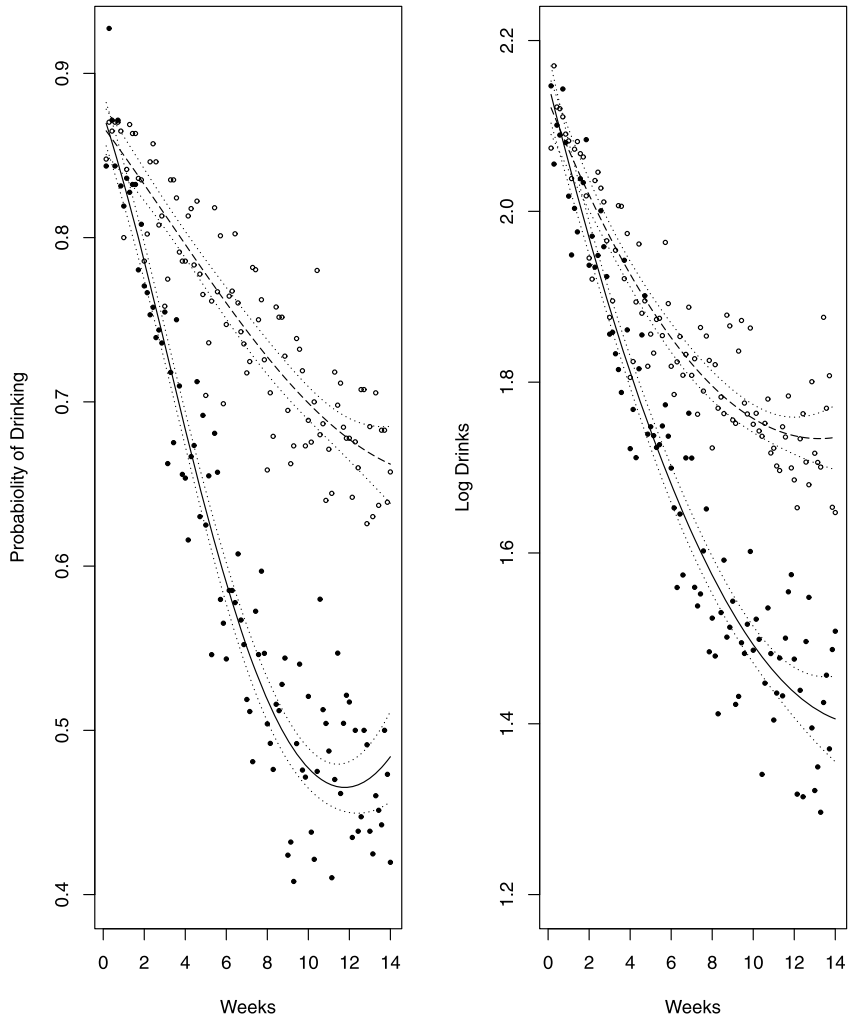

FIG. 2. *Trajectory analysis of the topiramate study* (*Chen et al*., 2012). (A) *Probability of drinking*. (B) *Log number of drinks on drinking days*. *The solid* (*dashed*) *lines are the estimated curves for topiramate* (*placebo*). *The dotted lines are* 95% *pointwise confidence intervals*.

6. NONLINEAR COVARIATE EFFECTS

6.1 Nonlinear Trajectories in 2PM

In Figure 2, we present the trajectories of the probability of drinking alcohol and number of daily drinks in a topiramate study (Chen et al., 2012). The estimated trajectory plots clearly show nonlinear temporal trends in drinking outcomes. The decrease in the drinking measure tends to stabilize during the end of the trial. This trajectory analysis can be taken as an exploratory tool to evaluate the efficacy of pharmacotherapy trials. For example, we can use the plots to determine the "grace period" wherein the data are only analyzed for efficacy after sufficient time has elapsed for the treatment to achieve its full effect (Food and Drug Administration, 2006, Falk et al., 2010).

However, Figure 2 does not consider the cross-part correlation between the frequency and amount of drinking in the longitudinal drinking records. To tackle this issue, Chen et al. (2013a) proposed a random effects 2PM with nonlinear trajectories; using notational conventions introduced earlier,

$$\text{logit}\, P(Y_{ij} > 0 | u_i, v_i, t_{ij}) = f_1(t_{ij}) + u_i, \tag{49}$$

$$\log Y_{+ij} | \{u_i, v_i, t_{ij}\} \sim f_2(t_{ij}) + v_i + e_{ij}, \tag{50}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are unknown functions of time (denoted by t_{ij} for the jth observation of subject i). The functions $f_1(\cdot)$ and $f_2(\cdot)$ can be estimated by penalized splines; correlated subject-level random effects $(u_i, v_i)^T \sim N(0, \Sigma)$ are used to account for the "cross-part" correlation.

6.2 Semiparametric Model for Positive Values

Chen et al. (2013b) considered a generalized semiparametric model with an unknown variance function

$$g(\mu) = X^T\boldsymbol{\beta} + f_1(z_1) + \cdots + f_m(z_m), \tag{51}$$

$$\log\bigl(\mathrm{Var}(Y_+)\bigr) = \mathcal{V}(\mu), \tag{52}$$

where $\mu = E(Y_+|X, z_1, \ldots, z_m)$ is the mean of the positive values, $g(\cdot)$ is a known link function (e.g., $g(\cdot) = \log(\cdot)$ for medical costs), $f_j(\cdot)$, $j = 1, \ldots, m$ are unknown smooth functions, and $\mathcal{V}(\cdot)$ is an unknown but smooth function of the mean. In this model, there are two types of covariates: a $p \times 1$ vector X that has a linear effect on the scale of the link function; and a $m \times 1$ vector of continuous variables $\mathbf{z} = (z_1, \ldots, z_m)^T$ that may have distinct nonlinear effects. This model is very flexible, and does not assume any specific form of the variance structure, increasing robustness when fitting data with heteroscedasticity. Assuming that $f_i(\cdot)$'s are each splines, this model can also be fitted using penalized quasi-likelihood, avoiding parametric assumptions on the distribution of Y_+, hence Y.

Chen et al. (2016) extended Models (51) and (52) to correlated medical costs, for example, for subjects within the same family. They used a working correlation matrix for $\mathrm{cov}(Y_+)$, for example, Compound Symmetry or AR(1). The estimation is carried out using extended generalized estimating equations (EGEE) (Hall and Severini, 1998). The method is applied to a subset of Medical Expenditure Panel Survey (MEPS) data: a nonlinear age effect in medical costs, together with substantial heteroscedasticity as shown in Figure 3.

7. JOINT MODEL OF ZERO-INFLATED OUTCOMES WITH SURVIVAL

7.1 Joint Model for Longitudinal Semicontinuous Data and Survival

In many longitudinal studies, subjects may drop-out or experience a terminal event before the end of study. These events may be correlated with the longitudinal outcome of interest. For example, in alcohol trials, there is concern with whether drop-out is associated with drinking outcomes, for example, subjects drinking heavily are more likely to drop-out (informative drop-out). While in longitudinal medical cost studies, "frailer" patients tend to have a higher mortality rate and accumulate medical cost faster, resulting in potential correlation between longitudinal medical costs and

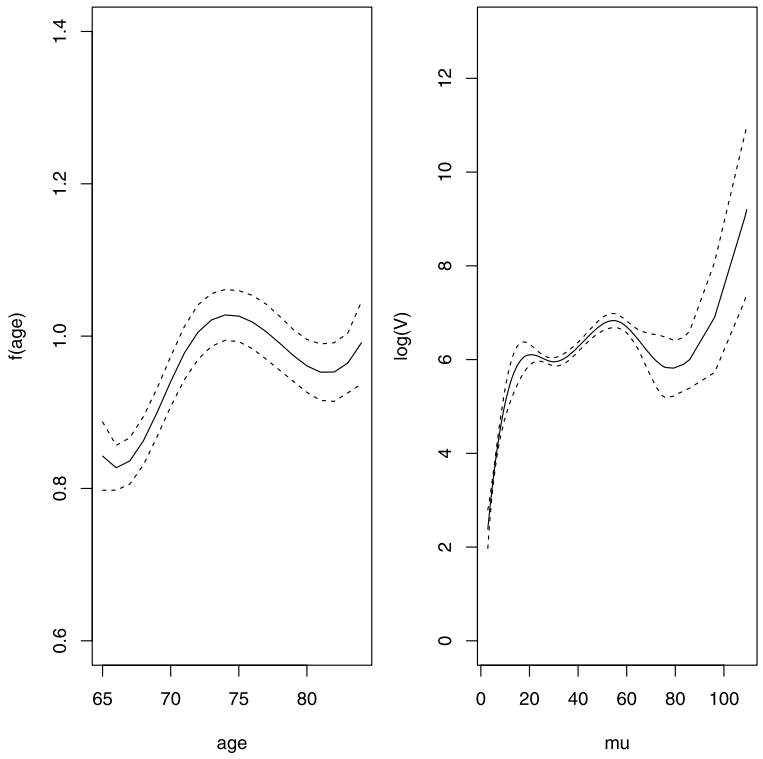

FIG. 3. *Curve estimation for MEPS data. Left*: *estimated curve for age with* 95% *point-wise confidence interval*; *right*: *estimated variance function* (*"mu" represents the mean of positive medical costs in the unit of US*$1000.)

survival. Failure to account for such correlation would result in biases in parameter estimates.

Joint models of longitudinal and survival data have been developed to take into account the potential correlation between longitudinal and survival data using shared random effects, for example, Wulfsohn and Tsiatis (1997), Henderson, Diggle and Dobson (2000), Vonesh, Greene and Schluchter (2006) and Liu and Liu (2015). Clinically, a joint model of longitudinal medical costs and survival can present a clear picture of the cost accrual process until death, clarifying the misleading relationship between total medical costs and health status. It is particularly attractive in cost effectiveness studies, when both costs and survival are of interest simultaneously (Pullenayegum and Willan, 2007). As another example, a joint model of drinking outcomes and the time to drop-out can provide a sensitivity analysis on the mechanism of missing data (Johnson et al., 2013).

Liu (2009) proposed a joint model of longitudinal semicontinuous medical costs and survival. One version of this model appears below: for independent subjects $i = 1, \ldots, n$,

$$\begin{aligned}(53)\quad &\operatorname{logit} P(Y_{ij} > 0 | X_{ij}, a_i, b_i) \\ &\quad = X_{ij}^T \alpha_1 + Z_i^T \alpha_2 + a_i,\end{aligned}$$

$$\begin{aligned}(54)\quad &\log Y_{+ij} | \{X_{ij}, a_i, b_i\} \\ &\quad \sim X_{ij}^T \beta_1 + Z_i^T \beta_2 + \delta_1 a_i + b_i + e_{ij},\end{aligned}$$

$$\begin{aligned}(55)\quad &\lambda(t | X_i, a_i, b_i) \\ &\quad = \lambda_0(t) \exp\bigl(Z_i^T \gamma + \delta_2 a_i + \delta_3 b_i\bigr),\end{aligned}$$

where $\{X_{ij}, j \geq 0\}$ represents a longitudinally measured vector of covariates and Z_i represents a vector baseline covariates. Part I is a logistic model for monthly costs being positive, and Part II models the amount of positive monthly costs. All models are conditional on covariates and random effects and implicitly condition on a subject remaining at risk. Model (55) links the semicontinuous medical costs with the terminal event by the shared random effects a_i and b_i, where $a_i \sim N(0, \sigma_a^2)$ and $b_i \sim N(0, \sigma_b^2)$ are independent. Model association is denoted by δ_1, δ_2, δ_3. For example, δ_2 and δ_3 in Model (55) indicate the relation between frequency and costs of medical services and death. As shown in Section 8.2, if they are both significantly greater than 0, patients who are more likely to seek medical treatment and/or incur higher monthly medical costs tend to have higher mortality rates.

The specification of Model (55) can be easily extended to the setting where X_i is replaced by an external time dependent covariate, say $X_i(t)$, $t \geq 0$. In principle, it should also be possible to extend this model to the case of an internal time-dependent covariate, such as $X_i(t) = X_{i,N_i(t-)}$, where $N_i(t)$ is a counting process that describes the times at which longitudinal measurements have been taken through time t.

7.2 Joint Model for Zero-Inflated Recurrent and Terminal Events

Recurrent and terminal events are often correlated; for example, in AIDS studies, patients with poor health tend to experience more recurrences of opportunistic diseases and are also at a higher risk of death. Building on earlier literature that jointly modeled recurrent and terminal events (e.g., Wang, Qin and Chiang, 2001; Liu, Wolfe and Huang, 2004), Liu et al. (2016a) proposed a joint model of zero-inflated recurrent and terminal events that adds a Cox regression model for the terminal event to models (47) and (48) from Section 5.2:

$$(56)\quad \operatorname{logit} P(A_i = 2 | X_i, \nu_i) = X_i^T \alpha,$$

$$(57)\quad \lambda(t | X_i, A_i = 2, \nu_i) = \lambda_0(t) \exp\bigl(X_i^T \beta + \nu_i\bigr),$$

$$(58)\quad h(t | X_i, \nu_i) = h_0(t) \exp\bigl(X_i^T \eta + \nu_i\bigr).$$

The frailty term ν_i in the recurrent events model also appears in Model (58) for the terminal event, introducing the correlation between the recurrent and terminal events. A higher intensity of recurrent events is associated with a higher mortality rate. If a subject is "unsusceptible", s/he cannot experience any recurrent events.

There are two situations for the relation between zero inflation ("susceptibility") and the terminal event. For AIDS patients, there is no "cure", so even patients who are unsusceptible to recurrent opportunistic diseases would die from AIDS. Therefore, there is no condition $A_i = 2$ in Model (58).

On the other hand, some cancers, for example, sarcomas, can be "cured": the cured subjects can experience neither recurrent tumors nor death from sarcoma. We can change model (58) to

$$(59)\quad h(t | X_i, \nu_i, A_i = 2) = h_0(t) \exp\bigl(X_i^T \eta + \nu_i\bigr).$$

By adding the condition $A_i = 2$, cured subjects cannot experience death due to sarcoma, that is, $h(t | \nu_i, A_i = 1) = 0$. For detailed likelihood and estimation methods, please see Liu et al. (2016a).

8. APPLICATIONS

In this section, we consider the application of these methods to three datasets. In the first application, the methods of Section 4.4 are applied to a microbiome study. In the second application, the methods of Section 4.2 are applied to an oral topiramate trial for the treatment of alcohol dependence. In the third application, the methods of Section 7.1 are used to analyze monthly medical costs for heart failure patients.

8.1 Two-Part Random Effects Model for Microbiome Data

The first example is a microbiome study which compares different therapies for pediatric inflammatory bowel disease (IBD) patients (Lewis et al., 2015, Chen and Li, 2016). The goal of this study is to identify the bacterial taxa that showed overall different abundances in patients who were given different treatments. The longitudinal data contains 236 samples for 69 children with IBD. Among them, 47 received anti-tumour necrosis factor (TNFα) antibodies, and 22 received exclusive enteral nutrition (EEN). Gut microbiome samples were collected for each subject at four time points: baseline, 1 week, 4 weeks and 8 weeks into the therapy. In microbiome studies, microbial abundances measured in read counts are not comparable across samples. The read counts are often normalized to relative abundances, which are bounded in $[0, 1)$. Consequently, the relative abundances of all microbes in one sample sum to one (Tyler, Smith and Silverberg, 2014).

The data contain the relative abundances at the genus level of the 18 most common bacterial genera. Figure 1(A) shows the distribution of the relative abundance of a bacterial genus named *Haemophilus*; it has a large portion of zero values (39%). The continuous positive values have a mean of 0.0285 with a range of 0.0000106 to 0.4440, demonstrating the right skewness. In addition, the repeated measures of the relative abundance of the bacteria from the same sample across time points are expected to be correlated. Correlated random effects are thus used to capture the association within and between the two parts.

Let Y_{ij} be the relative abundance of subject i at the jth time for a given bacterial genus. Models (28) and (29) can be used to model the relative abundance:

$$\operatorname{logit}(p_{ij}) = X_{ij}^T\alpha + a_i,$$

$$\operatorname{logit}(\mu_{ij}) = X_{ij}^T\beta + b_i,$$

where $p_{ij} = P(Y_{ij} > 0|X_{ij}, a_i, b_i)$ and $\mu_{ij} = E(Y_{+ij}|X_{ij}, a_i, b_i)$; the random effects a_i and b_i are used to account for the correlation among the repeated measurements on the same sample. Chen and Li (2016) assumed that a_i and b_i are independent and respectively follow normal distributions:

$$a_i \sim N(0, \sigma_1^2), \quad b_i \sim N(0, \sigma_2^2). \tag{60}$$

However, their model does not allow the "cross-part correlation" between the two parts, that is, as the relative abundance is more likely to be positive, its magnitude tends to be larger. Hence, we consider a two-part random effects model for the microbiome data with a cross-part correlation. Covariates of interest in both parts of the model include time (in weeks) and treatment. The results for *Haemophilus* are shown in Table 1. For comparison, we also show the results from a model without the cross-part correlation in the right panel of Table 1.

The results suggest that the model with the cross-part correlation provide a better fit, with a covariance estimate $\hat{\sigma}_{12} = 0.88$ and $p = 0.01$. In this model, treatment is significant in both parts: it reduces the odds of the relative abundance being positive and the magnitude if the abundance is positive. There is a decreasing temporal trend in Part I of the model. For comparison, we find that the parameter estimate for treatment in Part II is not significant ($p = 0.08$) in the model without the cross-part correlation. Thus, ignoring the cross-part correlation could lead to an erroneous conclusion on the treatment effect. Note that here and below we use the adjustment by Stram and Lee (1994) to address the boundary testing issue for variance components.

8.2 Zero-Inflated Models for Alcohol Drinking Data

The second application comes from a double-blind randomized controlled 14-week trial on the safety and efficacy of oral topiramate (Topa) for the treatment of alcohol dependence (Johnson et al., 2007). Three hundred seventy one subjects were enrolled in this study. Among them, 183 alcoholics received Topa and 188 were given a placebo. The outcome of interest was daily drinking records, which were assessed every week by the timeline follow-back method (Sobell and Sobell, 1992). A dose escalation mechanism was designed for the study: the dose of Topa (or the placebo) started at 25 mg for week 1 and was increased to 300 mg for weeks 6–14. Since this is a proof-of-concept trial, we are not interested in the dose response. Rather, the objective is to assess the *overall* Topa treatment effect at improving drinking outcomes. Thus, treatment is taken as a binary variable: Topa vs. placebo.

TABLE 1
Random effects two-part model for Haemophilus

Covariates	With cross-part correlation			Without cross-part correlation		
	Est.	SE	P-value	Est.	SE	P-value
			Part I			
Intercept	1.30	0.32	0.0001	1.27	0.32	0.0002
Time	−0.10	0.05	0.04	−0.10	0.05	0.04
Treatment	−1.86	0.57	0.002	−1.81	0.56	0.002
			Part II			
Intercept	−3.84	0.21	<0.0001	−3.72	0.21	<0.0001
Time	−0.041	0.028	0.14	−0.034	0.029	0.24
Treatment	−0.96	0.40	0.02	−0.70	0.39	0.08
			Variance components			
σ_1^2	1.39	0.72	0.03	1.24	0.67	0.03
σ_2^2	0.59	0.22	0.004	0.48	0.18	0.004
σ_{12}	0.88	0.35	0.01			
			Dispersion parameter			
ϕ	18.0	3.9	<0.0001	17.4	3.9	<0.0001
			Model comparison			
-2Loglik		−736.6			−726.4	

As shown in Figure 1(B), there is a large portion (32%) of zero values, that is, on average, participants were abstinent in about 1/3 of the follow up period. The continuous nonzero (positive) values have a mean of 7.6 (range: 0.072–58.9) standard drinks. In the original paper, Johnson et al. (2007) utilized a data reduction technique to calculate the weekly average of drinks per drinking day (DDD) and percentage of days abstinent (PDA) for the follow-up period. However, the weekly averaged drinking outcome is not as efficient as the original daily outcome.

We fit two-part random effects models for daily drinking records using three flexible distributions in Part II (Liu et al., 2016a): generalized gamma (Model A); log skew normal (Model B); and Box–Cox transformation (Model C). Random intercepts and slopes are included in both parts, resulting in a total of 4 random effects for each model. We also consider heteroscedasticity as in Model (24). Covariates of interest in each part of Models A–C include gender, age, baseline drinking level, treatment (Trt), time since onset, and the interaction between time and treatment. The results are shown in Table 2.

As expected, the parameter estimates are almost identical in Part I across the three models. However, the covariate effects in Part II are quite different. For example, age is not significant in Model B, but highly significant in Models A and C. Also, strong heteroscedasticity exists in all three models. Each of the shape/skewness parameters is highly significant, indicating that each model fits better than the log normal distribution.

For interpretation of covariate effects, the treatment by week interaction is highly significant in both parts. Topa is significantly better than placebo in reducing the odds of drinking over time, and the amount of drinking on any drinking day was reduced more quickly in the Topa group. Males tended to drink less often but consumed more alcohol on drinking days. Age is also in the opposite direction in the two parts: being older is correlated with increased odds of drinking, but also significantly associated with consuming less alcohol on a drinking day.

The estimate of variance components show that all 4 random effects (random intercept and slope) are highly significant, so heterogeneity exists in both parts. There is a significant positive cross-part correlation $\hat{\sigma}_{24}$ (for random slope in Parts I and II): patients more likely to decrease their odds of drinking were also more likely to decrease their amount of drinking.

TABLE 2
Random effects two-part model for alcohol drinking data

	Generalized gamma distribution			Log skew normal distribution			Box–Cox transformation		
Covariates	Est.	SE	P-value	Est.	SE	P-value	Est.	SE	P-value
				Part I					
Intercept	−0.50	0.83	0.55	−0.50	0.84	0.55	−0.50	0.89	0.57
Male	−0.27	0.32	0.39	−0.27	0.32	0.39	−0.27	0.34	0.42
Age	0.047	0.015	0.002	0.048	0.015	0.002	0.048	0.016	0.003
Base drinking	0.17	0.033	<0.0001	0.17	0.033	<0.0001	0.17	0.036	<0.0001
Treatment	−0.22	0.28	0.43	−0.22	0.29	0.44	−0.22	0.31	0.47
Time	−0.10	0.023	<0.0001	−0.11	0.023	<0.0001	−0.12	0.023	<0.0001
Trt × Time	−0.17	0.033	<0.0001	−0.18	0.033	<0.0001	−0.17	0.033	<0.0001
				Part II					
Intercept	1.80	0.11	<0.0001	1.81	0.11	<0.0001	2.69	0.23	<0.0001
Male	0.17	0.040	<0.0001	0.21	0.041	<0.0001	0.31	0.083	0.0002
Age	−0.0066	0.0019	**0.0007**	−0.0022	0.0020	**0.27**	−0.019	0.0041	<**0.0001**
Base drinking	0.064	0.004	<0.0001	0.069	0.004	<0.0001	0.13	0.009	<0.0001
Treatment	−0.015	0.036	0.68	0.054	0.037	0.14	−0.13	0.080	0.11
Time	−0.033	0.0035	<0.0001	−0.032	0.0035	<0.0001	−0.067	0.0069	<0.0001
Trt × Time	−0.033	0.0055	<0.0001	−0.037	0.0056	<0.0001	−0.053	0.011	<0.0001
				Heteroscedasticity					
Intercept	−1.20	0.07	<0.0001	−1.13	0.15	<0.0001	0.18	0.073	0.02
Male	−0.031	0.024	0.20	−0.13	0.052	0.02	0.10	0.023	<0.0001
Age	−0.013	0.0012	<0.0001	−0.029	0.0025	<0.0001	−0.018	0.0012	<0.0001
Base drinking	−0.010	0.002	<0.0001	−0.031	0.005	<0.0001	0.033	0.0023	<0.0001
Treatment	0.13	0.041	0.002	0.21	0.088	0.02	0.043	0.039	0.27
Time	−0.0022	0.0033	0.51	−0.013	0.0076	0.10	−0.026	0.0031	<0.0001
Trt × Time	0.0087	0.0052	0.09	0.032	0.011	0.003	−0.0047	0.0048	0.33
				Variance components					
σ_1^2	5.70	0.66	<0.0001	5.70	0.66	<0.0001	6.79	0.88	<0.0001
σ_2^2	0.066	0.0088	<0.0001	0.064	0.0083	<0.0001	0.066	0.0087	<0.0001
σ_3^2	0.10	0.0091	<0.0001	0.11	0.0096	<0.0001	0.51	0.049	<0.0001
σ_4^2	0.0020	0.00021	<0.0001	0.0020	0.00022	<0.0001	0.007	0.00082	<0.0001
σ_{12}	−0.092	0.049	0.06	−0.095	0.048	0.05	−0.14	0.058	0.02
σ_{13}	−0.059	0.049	0.23	−0.055	0.050	0.27	0.16	0.13	0.20
σ_{14}	−0.0064	0.007	0.38	−0.0070	0.007	0.33	−0.035	0.017	0.03
σ_{23}	−0.0037	0.0054	0.50	−0.0020	0.0053	0.71	−0.018	0.012	0.15
σ_{24}	0.0061	0.0010	<0.0001	0.0062	0.001	<0.0001	0.014	0.0020	<0.0001
σ_{34}	−0.0025	0.0010	0.01	−0.0020	0.0010	0.04	−0.021	0.0045	<0.0001
				Shape/skewness					
κ	0.70	0.017	<0.0001						
λ				−0.60	0.006	<0.0001			
γ							0.36	0.008	<0.0001
				Model comparison					
Loglik		−64,114			−64,228			−64,170	

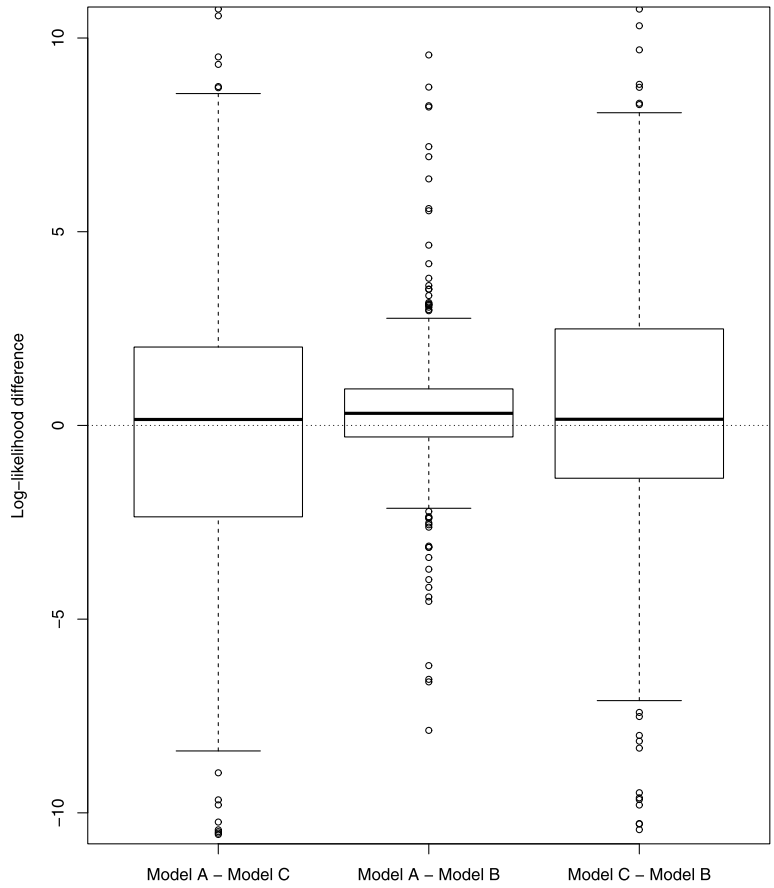

FIG. 4. *Boxplots of the difference in marginal loglikelihoods for each subject. Solid bold lines in the boxplot denote the corresponding mean (rather than median) difference in marginal loglikelihoods. Y-axis limits are truncated at* $\{-10, 10\}$. *From left to right, minimum and maximum differences are respectively* $\{-50.21, 45.83\}$, $\{-38.02, 22.38\}$, $\{-36.27, 38.20\}$.

Next we compare the performance of the three models. Generalized gamma distribution (Model A) has the largest likelihood. However, the three models are not nested within each other, so the likelihood ratio test cannot be applied. Since the three models have the same number of parameters, comparison of AIC/BIC results yields the same conclusion as that of loglikelihood. Instead, we considered the likelihood-based test statistic in Vuong (1989) for nonnested hypotheses, which uses the Kullback–Leibler Information Criterion to measure the closeness of a model to the truth. Since SAS Proc NLMIXED cannot provide the contribution of each subject to the marginal loglikelihood directly, we used the R package *gaussquad* to compute the marginal likelihood for each subject using adaptive Gaussian quadrature.

The results of model comparisons are shown in Figure 4. For pairwise model comparisons, the p-values are: (i) Model A (generalized gamma) vs. Model B (log skew normal): 0.099; (ii) Model A vs. Model C (Box–Cox transformation): 0.70; (iii) Model B vs. Model C: 0.64. It is reasonable to conclude that Model A provides the best overall description of the data because of the more pronounced difference between Models A and B, as well as the more straightforward interpretation of fixed effects afforded by Model A in comparison to Model C (i.e., using $y^{0.36}$ as the outcome).

8.3 Joint Model of Longitudinal Semicontinuous Medical Costs and Survival

The third application concerns a dataset of medical costs for heart failure patients in the clinical data repository (CDR) from the University of Virginia (UVa) Health System. Heart failure is the only cardiac disease growing in prevalence, with 670,000 new patients diagnosed each year. A total of 5.7 million heart failure patients reside in USA. It is one of the most expensive health care problems in the U.S., costing $39.2 billion in 2010. Also, heart failure is the leading cause of hospitalization among people 65 and older in the United

TABLE 3
Summary of the heart failure data

	Mean (Percent)	SD
Age	72.4	7.8
Male	54.7%	
White	73.2%	
Follow up months per subject	18.8	8.7
Months with nonzero cost per subject	49%	
Positive monthly costs	$2982	$14,383

States. Thus, it is important to identify risk factors and explore strategies to reduce medical costs for this common and expensive disease.

A total of 1475 patients over 60 years old were first diagnosed with and treated for heart failure in 2004. The follow up ended with each patient's last hospital admission, or July 31, 2006, or death date extracted from the Death Certificate Database. The outcome of interest is the UVa health system costs (actual monetary expenses of the hospital). Table 3 shows a summary of the data.

Medical costs are collected longitudinally. Here medical costs are grouped by month, forming monthly medical costs. Grouping by other time units, for example, quarter or year, could also be considered. Analysis of longitudinal medical costs is more efficient than that of cross sectional medical costs. It is essential to understand the dynamics of medical costs, illustrating the processes of cost accumulation and the reasons for cost differences between arms (Heitjan, Kim and Li, 2004). Moreover, it is useful in estimating incidence costs (medical costs after diagnosis) and predicting future medical costs, offering useful information for policy makers when making decisions about resource allocation and coverage of specific treatments (Yabroff et al., 2009). Finally, it can help to assess the effect of health care policy change on medical costs (Cotter et al., 2006).

There is often a U-shape (or bathtub shape) pattern observed for longitudinal medical costs. For example, Yabroff et al. (2009) showed three phases of longitudinal medical costs for colorectal cancer: high costs in the initial period following diagnosis, high costs (up to 30% of lifetime medical costs) during the end of life period (e.g., final 12 months before death), and low and relatively stable costs in between these two periods. A similar U-shape pattern is presented in Figure 5(B) when using the Lowess estimate on medical costs of heart failure patients.

Liu (2009) applied the joint model of semicontinuous medical costs and survival to this data; that is, models (52)–(54). In this model, Part I is a logistic model for monthly costs being positive, Part II models the amount of positive monthly costs, and a Cox model

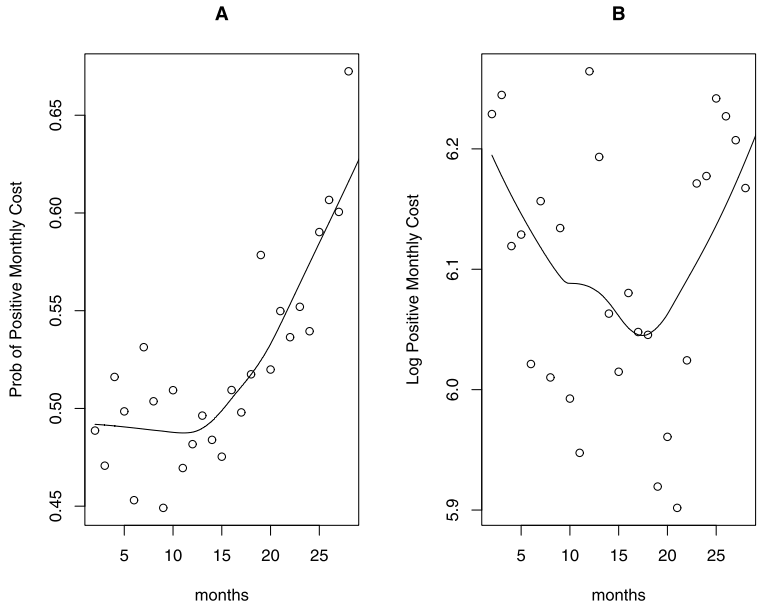

FIG. 5. *Time pattern of monthly medical costs for heart failure patients in the CDR database.* (A) *Probability of cost being positive*; (B) *amount of positive monthly costs* (*in log scale*). *The cross-sectional mean for each month is denoted by a circle. Lowess estimate is provided to show the temporal pattern.*

TABLE 4
Results for heart failure data

Par	Est	SE	P-value
	Part I: Monthly cost being positive		
Intercept	0.151	0.077	0.05
Male	−0.051	0.070	0.47
White	−0.159	0.079	0.04
Age	0.201	0.045	<0.0001
Linear Rise	0.013	0.009	0.16
Quadratic Rise	0.015	0.007	0.03
	Part II: Log positive monthly cost		
Intercept	6.304	0.090	<0.0001
Male	0.120	0.069	0.08
White	−0.247	0.078	0.002
Age	−0.080	0.046	0.09
Age^2	−0.185	0.052	0.0004
Time	−0.158	0.084	0.06
Time^2	0.126	0.035	0.0004
	Survival		
Male	0.194	0.130	0.14
White	−0.178	0.143	0.21
Age	0.505	0.084	<0.0001
	Model association		
δ_1	0.325	0.035	<0.0001
δ_2	0.347	0.065	<0.0001
δ_3	0.689	0.074	<0.0001
	Variance components		
σ_a^2	1.320	0.071	<0.0001
σ_b^2	1.127	0.060	<0.0001

with shared random effects is used to model survival. Covariates of interest are shown in Table 4. Age is centered at 72 years old. "Linear rise" (= max(month − 12, 0)) and "quadratic rise" (square of linear rise) are used to describe the pattern shown in Figure 5(A), where the probability of monthly costs being positive has a quadratically increasing temporal trend from 12 months after entry into study. A quadratic time effect is included to describe the U-shape pattern for longitudinal medical costs shown in Figure 5(B). A quadratic age effect is incorporated in Part II based on preliminary analysis.

There is a significant increasing quadratic time effect happening 12 months after entry on the logit of probability of positive monthly costs. There also exists a significant quadratic time effect on the amount of positive monthly costs; that is, there is a high initial cost (due to diagnosis and treatment) and a high cost in the end of follow up (due to intensive care during the end-of-life stage). This is consistent with the "bathtub" shape of monthly outpatient Erythropoietin (EPO) medical cost data for dialysis patients (Liu, Wolfe and Kalbfleisch, 2007), or the U-shape pattern of incidence cancer costs (Yabroff et al., 2009).

No significant gender difference is found. However, a racial difference is identified, as white patients tended to have a lower odds ($OR = 0.85$, $p = 0.04$) of seeking medical treatments. They had lower positive monthly costs (−0.247 in log scale, $p = 0.002$) in Part II; they were also at a lower risk of mortality, although this is not statistically significant. These results reflect the fact that white patients had better outcomes than non-white patients, indicating possible racial disparities in the care of heart failure patients. This is consistent with a study by Jha et al. (2003) which showed that black women less often received appropriate preventive therapy and adequate risk factor control despite having a greater risk of coronary heart disease events; consequently, they often began treatment at more advanced disease states and incurred higher costs when treated.

The model also reveals differences in care based on age. Older patients were more likely to seek medical care ($OR = 1.22$ for every 10-year increase in age, $p < 0.0001$). Older age is associated with a higher mortality rate ($OR = 1.66$ for every 10-year increase in age, $p < 0.0001$). Finally, age has a quadratic effect on the amount of positive monthly medical costs: the trend first increases, then decreases after age 70. The decrease after age 70 can be explained by the fact that older patients were usually treated less aggressively, thus incurring less medical costs. For example, Gatsonis et al. (1995) showed less frequent utilization of coronary angiography for elderly patients. Furthermore, Stukel, Lucas and Wennberg (2005) and Stukel et al. (2007) showed that younger patients with heart diseases were more likely to receive invasive treatments and medical therapies.

Association across the three outcomes is captured in the estimates of δ_1, δ_2, δ_3; all three parameters are positive and highly significant. Patients seeking medical treatments more often had higher medical costs and higher mortality rates. Furthermore, patients with higher costs were at higher risk of mortality. Finally, from the estimates of random effects variance σ_a^2 and σ_b^2, heterogeneity appears to exist in both use and costs of health services.

9. DISCUSSION

In this paper, we reviewed contemporary statistical methods to analyze zero-inflated nonnegative continuous data. Taking microbiome, alcohol consumption, and medical cost data as examples, we showed various approaches to separating zero and positive values and modeling the right skewed and heteroscedastic positive values. We considered both cross-sectional and correlated data. We also discussed related issues of zero-inflated count/survival data, nonparametric regression, and joint models of zero-inflated data and time to event.

There are several issues worthy of further study. First, methodological research on Part II of the 2PM to date has largely focused on statistical inferences of mean, while little attention has been paid to other segments of the entire distribution. From both policy and clinical perspectives, the group of "high spenders" or "heavy drinkers" is of utmost interest as policies or interventions affecting this group would likely achieve the largest impact, either in terms of reducing healthcare costs or improving health outcomes. Importantly, behaviors of high spenders or heavy drinkers cannot be understood by studying the mean. Refinement of robust approaches on the higher tail of the distribution, for example, 90th or 95th percentile, deserves more consideration. Quantile regression methods (e.g., Bang and Tsiatis, 2002; Dominici and Zeger, 2005; Dominici et al., 2005) could be employed for the positive values of semicontinuous data. This analytical approach can fully utilize information from the entire distribution, thus providing important insights to more comprehensively understand the characteristics of individuals with excessively high medical costs or drinking records.

Second, more complicated correlation structures could be imposed in the zero-inflated continuous data. For example, Neelon, Zhu and Neelon (2015, 2016) considered spatial and spatial-temporal models for zero-inflated count and semicontinuous data. For microbiome composition data, the correlation among the semicontinuous relative abundance of different species is dependent on the structure of the phylogenetic tree, with a constraint that the relative abundance of all species sums up to 1. Modeling the semicontinuous micriobiome composition data is a field worthy of further investigation.

Third, variable selection is necessary when there exists a large number of covariates for the semicontinuous outcome. Sparse estimation via the regularized method is attractive for enhanced prediction accuracy and model interpretability. Recently, Han et al. (2018) proposed a feasible way of conducting variable selection for the random effects two-part model (21) and (22) on the basis of the "minimum information criterion" (MIC) method (Su et al., 2016). Adaptation of this method to other complicated models for semicontinuous data forms another topic for future research.

Finally, most of the methods reviewed in this paper are in the frequentist framework. Bayesian methods have been adopted in the analysis of zero-inflated continuous data. For example, adopting a hierarchical Bayesian approach, Zhang et al. (2006) modeled provider effects on pharmacy cost data in a random effects 2PM model in which the probability of positive costs and observed positive costs are each modeled using dependent mixed models. Neelon, O'Malley and Normand (2011) proposed a related two-part latent class model for longitudinal medical expenditure data, fitting a random effects 2PM within each latent class to separately describe the probability of medical service utilization and the mean spending trajectories among those having used services. The deviance information criterion (DIC) was used to determine the number of classes. Bayesian versions of the two-part random effects model with nonlinear covariate effects (Models (49) and (50)) was considered by Ghosh and Albert (2009). Bayesian approaches for complicated semicontinuous data merit future interest due to their computational and often inferential advantages.

ACKNOWLEDGEMENTS

This research is partly supported by AHRQ R01 HS 020263, NIH/NCI R01 CA 85848, and NSF DMS-1308009. Dr. Liu is a consultant to Celladon, Zensun, and Outcome Research Solutions, Inc. We are grateful to Drs. Jinsong Chen, Xuelin Huang, Mingyao Li for helpful discussions and comments. Copyright permission has been granted by Wiley and Sage in reusing tables and figures in Section 8. Part of the content has been given in a short course in 2016 Joint Statistical Meetings. The authors thank Gary Deyter, technical writer from the Department of Health Services Research at The University of Texas MD Anderson Cancer Center, for his editorial assistance.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review" (https://github.com/joyfulstones/zero-inflated-

continuous). Data and programming codes are available at https://github.com/joyfulstones/zero-inflated-continuous.

REFERENCES

AITCHISON, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *J. Amer. Statist. Assoc.* **50** 901–908. MR0071685

ALBERT, P. S. (2005). Letter to the editor. *Biometrics* **61** 879–881. MR2196179

AMEMIYA, T. (1994). *Introduction to Statistics and Econometrics*. Harvard Univ. Press, Boston, MA.

BANG, H. and TSIATIS, A. A. (2002). Median regression with censored cost data. *Biometrics* **58** 643–649. MR1926117

BASU, A. and MANNING, W. G. (2006). A test for proportional hazards assumption within the exponential conditional mean framework. *Health Serv. Outcomes Res. Methodol.* **6** 81–100.

BASU, A., MANNING, W. G. and MULLAHY, J. (2004). Comparing alternative models: Log vs Cox proportional hazard? *Health Econ.* **13** 749–765.

BASU, A. and RATHOUS, P. J. (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6** 93–109.

BERK, K. N. and LACHENBRUCH, P. A. (2002). Repeated measures with zeros. *Stat. Methods Med. Res.* **11** 303–316.

BJERRE, B., MARQUES, P., SELEN, J. and THORSSON, U. (2007). Swedish alcohol ignition interlock programme for drink-drivers: Effects on hospital care utilization and sick leave. *Addiction* **102** 560–570.

BLOUGH, D. K., MADDEN, C. W. and HORNBROOK, M. C. (1999). Modeling risk using generalized linear models. *J. Health Econ.* **18** 153–171.

BOAG, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc.* **11** 15–53.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. (With discussion.) *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **26** 211–252. MR0192611

BRETON, C. V., KILE, M. L., CATALANO, P. J., HOFFMAN, E., QUAMRUZZAMAN, Q., RAHMAN, M., MAHIUDDIN, G. and CHRISTIANI, D. C. (2007). GSTM1 and APE1 genotypes affect arsenic-induced oxidative stress: A repeated measures study. *Environ. Health* **6** 39.

CHAI, H. S. and BAILEY, K. R. (2008). Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. *Stat. Med.* **27** 3643–3655. MR2523977

CHAI, H., JIANG, H., LIN, L. and LIU, L. (2018). A marginalized two-part Beta regression model for microbiome compositional data. *PLoS Comput. Biol.* **14** e1006329.

CHEN, E. Z. and LI, H. (2016). A two-part mixed-effect model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32** 2611–2617.

CHEN, J., JOHNSON, B. A., WANG, X. Q., O'QUIGLEY, J., ISAAC, M., ZHANG, D. and LIU, L. (2012). Trajectory analyses in alcohol treatment research. *Alcohol. Clin. Exp. Res.* **36** 1442–1448.

CHEN, J., LIU, L., JOHNSON, B. A. and O'QUIGLEY, J. (2013a). Penalized likelihood estimation for semiparametric mixed models, with application to alcohol treatment research. *Stat. Med.* **32** 335–346. MR3041871

CHEN, J., LIU, L., ZHANG, D. and SHIH, Y.-C. T. (2013b). A flexible model for the mean and variance functions, with application to medical cost data. *Stat. Med.* **32** 4306–4318. MR3118356

CHEN, J., LIU, L., SHIH, Y.-C. T., ZHANG, D. and SEVERINI, T. A. (2016). A flexible model for correlated medical costs, with application to medical expenditure panel survey data. *Stat. Med.* **35** 883–894. MR3457613

COOPER, N. J., LAMBERT, P. C., ABRAMS, K. R. and SUTTON, A. J. (2007). Predicting costs over time using Bayesian Markov chain Monte Carlo methods: An application to early inflammatory polyarthritis. *Health Econ.* **16** 37–56.

COTTER, D., THAMER, M., NARASIMHAN, K., ZHANG, Y. and BULLOCK, K. (2006). Translating epoetin research into practice: The role of government and the use of scientific evidence. *Health Aff.* **25** 1249–1259.

DOMINICI, F. and ZEGER, S. L. (2005). Smooth quantile ratio estimation with regression: Estimating medical expenditures for smoking-attributable diseases. *Biostatistics* **6** 505–519.

DOMINICI, F., COPE, L., NAIMAN, D. Q. and ZEGER, S. L. (2005). Smooth quantile ratio estimation. *Biometrika* **92** 543–557. MR2202645

DOW, W. H. and NORTON, E. C. (2003). Choosing between and interpreting the heckit and two-part models for corner solutions. *Health Serv. Outcomes Res. Methodol.* **4** 5–18.

DUAN, N. (1983). Smearing estimate: A nonparametric retransformation method. *J. Amer. Statist. Assoc.* **78** 605–610. MR0721207

DUAN, N., MANNING, W. G., MORRIS, C. and NEWHOUSE, J. P. (1983). A comparison of alternative models for the demand for medical care. *J. Bus. Econom. Statist.* **1** 115–126.

DUDLEY, R. A., HARRELL, F. E. JR, SMITH, L. R., MARK, D. B., CALIFF, R. M., PRYOR, D. B., GLOWER, D., LIPSCOMB, J. and HLATKY, M. (1993). Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *J. Clin. Epidemiol.* **46** 261–271.

FALK, D., WANG, X. Q., LIU, L., FERTIG, J., MATTSON, M., RYAN, M., JOHNSON, B., STOUT, R. and LITTEN, R. Z. (2010). Percentage of subjects with no heavy drinking days: Evaluation as an efficacy endpoint for alcohol clinical trials. *Alcohol. Clin. Exp. Res.* **34** 2022–2034.

FAREWELL, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38** 1041–1046.

FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCELRATH, M. J., PRLIC, M., LINSLEY, P. S. and GOTTARDO, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278.

FOOD AND DRUG ADMINISTRATION (2006). *Medical Review of Vivitrol* 21-897. U.S. Government, Rockville, MD.

GATSONIS, C., EPSTEIN, A. M., NEWHOUSE, J. P., NORMAND, S. L. and MCNEIL, B. J. (1995). Variations in the utilization of coronary angiography for elderly patients with an acute myocaridal infaction: An analysis using hierarchical logistic regression. *Med. Care* **33** 625–642.

GHOSH, P. and ALBERT, P. S. (2009). A Bayesian analysis for longitudinal semicontinuous data with an application to an

acupuncture clinical trial. *Comput. Statist. Data Anal.* **53** 699–706. MR2654581

HALL, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **56** 1030–1039. MR1815581

HALL, D. B. and SEVERINI, T. A. (1998). Extended generalized estimating equations for clustered data. *J. Amer. Statist. Assoc.* **93** 1365–1375. MR1666633

HAN, D., LIU, L., SU, X., JOHNSON, B. and SUN, L. (2018). Variable selection for random effects two-part model. *Stat. Methods Med. Res.* DOI:10.1177/0962280218784712.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161. MR0518832

HEITJAN, D. F., KIM, C. Y. and LI, H. (2004). Bayesian estimation of cost-effectiveness from censored data. *Stat. Med.* **23** 1297–1309.

HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480.

HYNDMAN, R. and GRUNWALD, G. (2000). Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall. *Aust. N. Z. J. Stat.* **42** 145–158.

JAIN, A. K. and STRAWDERMAN, R. L. (2002). Flexible hazard regression modeling for medical cost data. *Biostatistics* **3** 101–118.

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning*: *With Applications in R*. *Springer Texts in Statistics* **103**. Springer, New York. MR3100153

JHA, A. K., VAROSY, P. D., KANAYA, A. K., HUNNINGHAKE, D. B., HLATKY, M. A., WATERS, D. D., FURBERG, C. D. and SHLIPAK, M. G. (2003). Differences in medical care and disease outcomes among black and white women with heart disease. *Circulation* **108** 1089–1094.

JOHNSON, B. A., ROSENTHAL, N., CAPECE, J. A., WIEGAND, F., MAO, L., BAYERS, K., MCKAY, A., AIT-DAOUD, N., ANTON, R. F., CIRAULO, D. A., KRANZLER, H. R., MANN, K., O'MALLEY, S. S. and SWIFT, R. M. (2007). Topiramate for treating alcohol dependence—a randomized controlled trial. *J. Am. Med. Assoc.* **298** 1641–1651.

JOHNSON, B. A., AIT-DAOUD, N., WANG, X.-Q., PENBERTHY, J. K., JAVORS, M. A., SENEVIRATNE, C. and LIU, L. (2013). Topiramate for the treatment of cocaine addiction: A randomized clinical trial. *J. Am. Med. Dir. Assoc. Psychiatr.* **70** 1338–1346.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1924807

KUK, A. Y. C. and CHEN, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79** 531–541.

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

LEUNG, S. F. and YU, S. (1996). On the choice between sample selection and two-part models. *J. Econometrics* **72** 197–229.

LEWIS, J. D., CHEN, E. Z., BALDASSANO, R. N., OTLEY, A. R., GRIFFITHS, A. M., LEE, D., BITTINGER, K., BAILEY, A., FRIEDMAN, E. S., HOFFMANN, C., ALBENBERG, L., SINHA, R., COMPHER, C., GILROY, E., NESSEL, L., GRANT, A., CHEHOUD, C., LI, H., WU, G. D. and BUSHMAN, F. D. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell Host Microbe* **18** 489–500.

LI, P., SCHNEIDER, J. E. and WARD, M. M. (2007). Effect of critical access hospital conversion on patient safety. *Health Serv. Res.* **42** 2089–2108; discussion 2294–2323.

LI, C.-S. and TAYLOR, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Stat. Med.* **21** 3235–3247.

LIN, D. Y., ETZIONI, R., FEUER, E. J. and WAX, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53** 419–434.

LIPSCOMB, J., ANCUKIEWICZ, M., PARMIGIANI, G., HASSELBLAD, V., SAMSA, G. and MATCHAR, D. B. (1998). Predicting the cost of illness: A comparison of alternative models applied to stroke. *Med. Decis. Mak.* **18** S39–S56.

LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W., WOLFINGER, R. D. and SCHABERNBERGER, O. (2006). *SAS for Mixed Model*, 2nd ed. SAS Institute Inc., Cary, NC.

LIU, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Stat. Med.* **28** 972–986. MR2518360

LIU, L. and HUANG, X. (2008). The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Stat. Med.* **27** 2665–2683. MR2440058

LIU, Y. and LIU, L. (2015). Joint models for longitudinal data and time-to-event occurrence. In *Routledge International Handbook of Advanced Quantitative Methods in Nursing Research* (S. J. Henly, ed.) 253–263. Taylor and Francis, London.

LIU, L., MA, J. Z. and JOHNSON, B. A. (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Stat. Med.* **27** 3528–3539. MR2523969

LIU, L., WOLFE, R. A. and HUANG, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60** 747–756. MR2089451

LIU, L., WOLFE, R. A. and KALBFLEISCH, J. D. (2007). A shared random effects model for censored medical costs and mortality. *Stat. Med.* **26** 139–155. MR2312704

LIU, L., CONAWAY, M. R., KNAUS, W. A. and BERGIN, J. D. (2008). A random effects four-part model, with application to correlated medical costs. *Comput. Statist. Data Anal.* **52** 4458–4473. MR2432473

LIU, L., STRAWDERMAN, R. L., COWEN, M. E. and SHIH, Y. C. T. (2010). A flexible two-part random effects model for correlated medical costs. *J. Health Econ.* **29** 110–123.

LIU, L., HUANG, X., YAROSHINSKY, A. and CORMIER, J. N. (2016a). Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics* **72** 204–214. MR3500589

LIU, L., STRAWDERMAN, R. L., JOHNSON, B. A. and O'QUIGLEY, J. M. (2016b). Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. *Stat. Methods Med. Res.* **25** 133–152. MR3460432

LU, S.-E., LIN, Y. and SHIH, W.-C. J. (2004). Analyzing excessive no changes in clinical trials with clustered data. *Biometrics* **60** 257–267. MR2044122

MAHMUD, S., LOU, W. W. and JOHNSTON, N. W. (2010). A probit- log- skew-normal mixture model for repeated measures data with excess zeros, with application to a cohort study

of paediatric respiratory symptoms. *BMC Med. Res. Methodol.* **10** 55.

MANNING, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *J. Health Econ.* **17** 283–295.

MANNING, W. G., BASU, A. and MULLAHY, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *J. Health Econ.* **20** 465–488.

MANNING, W. G., DUAN, N. and ROGERS, W. H. (1987). Monte-Carlo evidence on the choice between sample selection and 2-part models. *J. Econometrics* **35** 59–82.

MANNING, W. G. and MULLAHY, J. (2001). Estimating log models: To transform or not to transform? *J. Health Econ.* **20** 461–494.

MANNING, W., MORRIS, C., NEWHOUSE, J. et al. (1981). A two-part model of the demand for medical care: Preliminary results from the health insurance study. In *Health, Economics, and Health Economics* (J. van der Gaag and M. Perlman, eds.) 103–123. North-Holland, Amsterdam.

MARTINUSSEN, T. and SCHEIKE, T. H. (2006). *Dynamic Regression Models for Survival Data. Statistics for Biology and Health*. Springer, New York. MR2214443

MCDAVID, A., FINAK, G., CHATTOPADYAY, P. K., DOMINGUEZ, M., LAMOREAUX, L., MA, S. S., ROEDERER, M. and GOTTARDO, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29** 461–467.

MIN, Y. and AGRESTI, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Stat. Model.* **5** 1–19. MR2133525

MOULTON, L. and HALSEY, N. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* **51** 1570–1578.

MULLAHY, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *J. Health Econ.* **17** 247–281.

NEELON, B., O'MALLEY, A. J. and NORMAND, S.-L. T. (2011). A Bayesian two-part latent class model for longitudinal medical expenditure data: Assessing the impact of mental health and substance abuse parity. *Biometrics* **67** 280–289. MR2898840

NEELON, B., O'MALLEY, A. J. and SMITH, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research part 1: Background and overview. *Stat. Med.* **35** 5070–5093. MR3569914

NEELON, B., ZHU, L. and NEELON, S. E. B. (2015). Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures. *Biostatistics* **16** 465–479. MR3365440

NEELON, B., CHANG, H. H., LING, Q. and HASTINGS, N. S. (2016). Spatiotemporal hurdle models for zero-inflated count data: Exploring trends in emergency department visits. *Stat. Methods Med. Res.* **25** 2558–2576. MR3572870

OLSEN, M. K. and SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.* **96** 730–745. MR1946438

OTHUS, M., BARLOGIE, B., LEBLANC, M. L. and CROWLEY, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clin. Cancer Res.* **18** 3731–3736.

PARK, R. E. (1966). Estimation with heteroscedastic error terms. *Econometrica* **34** 888.

PENG, Y. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56** 237–243.

PENG, Y. (2003). Fitting semiparametric cure models. *Comput. Statist. Data Anal.* **41** 481–490. MR1973725

PENG, Y., TAYLOR, J. M. G. and YU, B. (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Anal.* **13** 351–369. MR2409955

PULLENAYEGUM, E. M. and WILLAN, A. R. (2007). Semiparametric regression models for cost-effectiveness analysis: Improving the efficiency of estimation from censored data. *Stat. Med.* **26** 3274–3299. MR2380581

RAUDENBUSH, S. W., YANG, M.-L. and YOSEF, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J. Comput. Graph. Statist.* **9** 141–157. MR1826278

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. MR2137253

ROBERT, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2723361

RONDEAU, V., SCHAFFNER, E., CORBIÈRE, F., GONZALEZ, J. R. and MATHOULIN-PÉLISSIER, S. (2013). Cure frailty models for survival data: Application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Stat. Methods Med. Res.* **22** 243–260. MR3190656

SCHOENFELD, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69** 239–241.

SMITH, V. A., PREISSER, J. S., NEELON, B. and MACIEJEWSKI, M. L. (2014). A marginalized two-part model for semicontinuous data. *Stat. Med.* **33** 4891–4903. MR3276507

SMITH, V. A., NEELON, B., MACIEJEWSKI, M. L. and PREISSER, J. S. (2017a). Two parts are better than one. *Health Serv. Outcomes Res. Methodol.* **17** 198–218.

SMITH, V. A., NEELON, B., PREISSER, J. S. and MACIEJEWSKI, M. L. (2017b). A marginalized two-part model for longitudinal semicontinuous data. *Stat. Methods Med. Res.* **26** 1949–1968. MR3687189

SOBELL, L. C. and SOBELL, M. B. (1992). Timeline follow-back: A technique for assessing self-reported alcohol consumption. In *Measuring Alcohol Consumption: Psychosocial and Biochemical Methods* (R. Z. Litten and J. P. Allen, eds.) 41–72. Humana Press Inc., Totowa, NJ.

SPOSTO, R. (2002). Cure model analysis in cancer: An application to data from the children's cancer group. *Stat. Med.* **21** 293–312.

STRAM, D. O. and LEE, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50** 1171–1177.

STUKEL, T. A., LUCAS, F. L. and WENNBERG, D. E. (2005). Long-term outcomes of regional variations in intensity of invasive vs medical management of medicare patients with acute myocardial infarction. *J. Am. Med. Assoc.* **293** 1329–1337.

STUKEL, T. A., FISHER, E. S., WENNBERG, D. E., ALTER, D. A., GOTTLIEB, D. J. and VERMEULEN, M. J. (2007). Analysis of observational studies in the presence of treatment selection bias effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *J. Am. Med. Assoc.* **297** 278–285.

SU, L., TOM, B. D. M. and FAREWELL, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10** 374–389.

SU, X., WIJAYASINGHE, C. S., FAN, J. and ZHANG, Y. (2016). Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics* **72** 751–759. MR3545668

SY, J. P. and TAYLOR, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56** 227–236. MR1767631

THERNEAU, T. M. and GRAMBSCH, P. M. (2000). *Modeling Survival Data*: *Extending the Cox Model*. *Statistics for Biology and Health*. Springer, New York. MR1774977

TIAN, L., ZUCKER, D. and WEI, L. J. (2005). On the Cox model with time-varying regression coefficients. *J. Amer. Statist. Assoc*. **100** 172–183. MR2156827

TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462

TOOZE, J. A., GRUNWALD, G. K. and JONES, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Stat. Methods Med. Res*. **11** 341–355.

TOOZE, J. A., MIDTHUNE, D., DODD, K. W., FREEDMAN, L. S., KREBS-SMITH, S. M., SUBAR, A. F., GUENTHER, P. M., CARROLL, R. J. and KIPNIS, V. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J. Am. Diet. Assoc*. **106** 1575–1587.

TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. MR2087974

TWISK, J. and RIJMEN, F. (2009). Longitudinal tobit regression: A new approach to analyze outcome variables with floor or ceiling effects. *J. Clin. Epidemiol*. **62** 953–958.

TYLER, A. D., SMITH, M. I. and SILVERBERG, M. S. (2014). Analyzing the human microbiome: A how to guide for physicians. *Am. J. Gastroenterol*. **109** 983–993.

VONESH, E. F., GREENE, T. and SCHLUCHTER, M. D. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Stat. Med*. **25** 143–163. MR2222079

VUONG, Q. H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* **57** 307–333. MR0996939

WANG, M.-C., QIN, J. and CHIANG, C.-T. (2001). Analyzing recurrent event data with informative censoring. *J. Amer. Statist. Assoc*. **96** 1057–1065. MR1947253

WILLIAMSON, J. M., DATTA, S. and SATTEN, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59** 36–42. MR1978471

WOOLDRIDGE, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186

XIE, H., MCHUGO, G., SENGUPTA, A., CLARK, R. and DRAKE, R. (2004). A method for analyzing long longitudinal outcomes with many zeros. *Ment. Health Serv. Res*. **6** 239–246.

YABROFF, K. R., WARREN, J. L., SCHRAG, D., MARIOTTO, A., MEEKINS, A., TOPOR, M. and BROWN, M. L. (2009). Comparison of approaches for estimating incidence costs of care for colorectal cancer patients. *Med. Care* **47** S56–S63.

YAMAGUCHI, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "Permanent Employment" in Japan. *J. Amer. Statist. Assoc*. **87** 284–292.

YU, Z., LIU, L., BRAVATA, D. M., WILLIAMS, L. S. and TEPPER, R. S. (2013). A semiparametric recurrent events model with time-varying coefficients. *Stat. Med*. **32** 1016–1026. MR3042854

ZHANG, M., STRAWDERMAN, R. L., COWEN, M. E. and WELLS, M. T. (2006). Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care. *J. Amer. Statist. Assoc*. **101** 934–945. MR2324094

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