Analysis of repeated measures data with clumping at zero

Janet A Tooze, Gary K Grunwald and Richard H Jones

Stat Methods Med Res 2002; 11; 341
DOI: 10.1191/0962280202sm291ra

Statistical Methods in Medical Research 2002; **11**: 341–355

Analysis of repeated measures data with clumping at zero

Janet A Tooze National Cancer Institute, Bethesda, Maryland, USA, **Gary K Grunwald** Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, Colorado, USA and **Richard H Jones** Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, Colorado, USA

Longitudinal or repeated measures data with clumping at zero occur in many applications in biometrics, including health policy research, epidemiology, nutrition, and meteorology. These data exhibit correlation because they are measured on the same subject over time or because subjects may be considered repeated measures within a larger unit such as a family. They present special challenges because of the extreme non-normality of the distributions involved. A model for repeated measures data with clumping at zero, using a mixed-effects mixed-distribution model with correlated random effects, is presented. The model contains components to model the probability of a nonzero value and the mean of nonzero values, allowing for repeated measurements using random effects and allowing for correlation between the two components. Methods for describing the effect of predictor variables on the probability of nonzero values, on the mean of nonzero values, and on the overall mean amount are given. This interpretation also applies to the mixed-distribution model for cross-sectional data. The proposed methods are illustrated with analyses of effects of several covariates on medical expenditures in 1996 for subjects clustered within households using data from the Medical Expenditure Panel Survey.

1 Introduction

Data with clumping at zero commonly occur in biometrics. Typically the outcome variable measures an amount that must be non-negative and may in some cases be zero. The positive values are generally skewed, often extremely so. Examples include concentrations of compounds, amounts of health or insurance expenditures, or amounts of rainfall or pollutants. Distributions of data of this type follow a common form: there is a spike or discrete probability mass at zero, followed by a bump or ramp describing positive values. Since the variable of interest describes an amount there is often interest in estimating the mean amount, including zeros, perhaps in order to estimate total amounts. For example, in estimating mean per person medical expenditures, it must be taken into account that some subjects will have no expenditures during the period of interest. From these means, group totals could be estimated.

Various approaches to the problem of data clumped at zero have been proposed, but most of them have drawbacks.[1] If the data are treated as if they come from a normal distribution, the clumping at zero is ignored as well as the tendency of the positive data

Address for correspondence: J Austin Tooze, Cancer Prevention Fellow, National Cancer Institute, Executive Plaza North, Suite 3131, 6130 Executive Blvd, MSC7354, Bethesda, MD 20892-7354, USA. E-mail: toozej@mail.nih.gov

10.1191/0962280202sm291ra

to be skewed. If a nonparametric approach utilizing the distribution of the ranks is employed, a large number of ties will exist corresponding to the zero observations, and the distribution will not be symmetric. In addition, it is not possible to obtain predictions of the response variable or to estimate totals using a nonparametric approach. Another approach to analyzing data of this type is to divide the data into two parts—those data with a value equal to zero and those greater than zero. If only the data greater than zero are used in the analysis, important information about subjects with zero response is lost, and estimates of totals will not include zero values. When one is relying on estimates from such analyses to make policy decisions, inaccurate conclusions may be made, which may lead to policies that are inadequate or inappropriate for the population of interest. In addition, this method does not account for the relationship that may exist between the probability of a nonzero response and the level of the nonzero response.

The majority of the literature in the area of data that are clumped at zero addresses the cross-sectional case where the unit of observation is measured once.[1–4] Clumping at zero may also occur with repeated measures or longitudinal data. In addition to sharing the problems of cross-sectional data with clumping at zero, the correlation among measurements on the same unit of observation must be accounted for.

We propose a mixed-distribution model based on the work of Lachenbruch[1,2] for cross-sectional data and Grunwald and Jones[5] for time series data. The model is also similar to the 'two-part model' used for cross-sectional data in econometrics.[3,6] All of these approaches combine models for the probability of occurrence of a nonzero value (a probit or logit model) and for the probability distribution of the nonzero values (a lognormal or exponential family distribution). The term 'mixed-distribution model' refers to a mixture-of-distributions model that takes the general form

$$f(y) = \begin{cases} \Pr(Y = 0), & \text{if } y = 0 \\ [1 - \Pr(Y = 0)]h(y) & \text{if } y > 0 \\ 0 & \text{if } y < 0 \end{cases} \qquad (1)$$

where $h(y)$ is a probability density defined when $y > 0$.[1,2] We draw on methods for modeling non-normal responses with random effects[7] to incorporate random unit (subject) effects into the two parts of the model to account for the correlation due to multiple observations made on the same subject or unit. We also allow the random unit effects for the probability of a nonzero value and for the distribution of nonzero values to be correlated with each other. This allows units with higher rates of occurrence to also have higher (or lower) mean nonzero responses. Correlation between the random effects in the two model components is similar to the cross-sectional correlation between the random normal errors in the two model components of the Heckman, or Type II Tobit model.[8]

Section 2 outlines the proposed mixed-distribution model for longitudinal data with correlated random effects, shows how the methods of generalized linear mixed models (GLMM) and nonlinear mixed models may be used to fit the model, and addresses the interpretation of the model parameters in terms of the total amount, including zeros. In particular, a covariate may affect the mean amount by affecting both the probability of

occurrence of a nonzero value and also the mean of the nonzero values, and we give an approach to quantifying and separating these two effects. In Section 3, results from simulation studies are presented. Section 4 illustrates application of the mixed-distribution model for repeated measures data using data from the Medical Expenditure Panel Survey, and Section 5 provides a summary and discusses areas for further research.

2 Mixed-distribution model with correlated random effects

In this section a mixed-distribution model for repeated measures data with clumping at zero and correlated random effects is introduced. This model will be referred to as the correlated mixed-distribution model. An extension of the mixed-distribution model was chosen to model repeated measures data because it provides a general statistical modeling approach using existing methodologies (generalized linear and nonlinear mixed-effects models). The model gives information about the separate occurrence and nonzero amount components of the model as well as the overall mean. The correlated mixed-distribution model relates the two components of the model by assuming a bivariate normal distribution for the random effects.

2.1 Model

For a random variable Y_{ij}, which represents the *amount* of a quantity with observed value y_{ij} for a unit of observation i at time j, let R_{ij} represent the *occurrence variable* where

$$R_{ij} = \begin{cases} 0, & \text{if } Y_{ij} = 0 \\ 1, & \text{if } Y_{ij} > 0 \end{cases}$$

R_{ij} has conditional probabilities

$$\Pr(R_{ij} = r_{ij} \mid \boldsymbol{\theta}_1) = \begin{cases} 1 - p_{ij}(\boldsymbol{\theta}_1), & \text{if } r_{ij} = 0 \\ p_{ij}(\boldsymbol{\theta}_1), & \text{if } r_{ij} = 1 \end{cases}$$

where $\boldsymbol{\theta}_1 = [\boldsymbol{\beta}_1', u_{1i}]'$ is a vector of fixed occurrence effects $\boldsymbol{\beta}_1$, and random unit occurrence effect u_{1i}. We assume a logistic model for occurrence so that

$$\text{logit}(p_{ij}(\boldsymbol{\theta}_1)) = \mathbf{X}_{1ij}'\boldsymbol{\beta}_1 + u_{1i} \tag{2}$$

where $\mathbf{X}_{1ij}$ is a vector of covariates for occurrence.

Define $S_{ij} \equiv [Y_{ij} \mid R_{ij} = 1]$ to be the *intensity variable* with p.d.f. $f(s_{ij}|\boldsymbol{\theta}_2)$ for $s_{ij} > 0$ and mean $E(S_{ij}|\boldsymbol{\theta}_2) = \mu_{s_{ij}}(\boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2 = [\boldsymbol{\beta}_2', u_{2i}]'$ is a vector of fixed intensity effects $\boldsymbol{\beta}_2$ and random unit intensity effect u_{2i}. We assume a lognormal model for intensity so that

$$\log(S_{ij} \mid \boldsymbol{\theta}_2) \sim N(\mathbf{X}_{2ij}'\boldsymbol{\beta}_2 + u_{2i}, \sigma_e^2) \tag{3}$$

where $\mathbf{X}_{2ij}$ is a vector of covariates for intensity. We allow the random effects for occurrence and intensity to be correlated by assuming that

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim BVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right). \tag{4}$$

Under this assumption the subject-specific mean intensity is

$$\mathrm{E}(S_{ij} \mid \boldsymbol{\theta}_2) = \exp\left(\mathbf{X}'_{2ij}\boldsymbol{\beta}_2 + u_{2i} + \frac{\sigma_e^2}{2}\right) \tag{5}$$

and the marginal mean intensity is

$$\mathrm{E}(S_{ij} \mid \boldsymbol{\beta}_2) = \exp\left(\mathbf{X}'_{2ij}\boldsymbol{\beta}_2 + \frac{\sigma_2^2}{2} + \frac{\sigma_e^2}{2}\right) \tag{6}$$

Note in particular that the values and interpretations of the fixed effects parameters $\boldsymbol{\beta}_2$ are identical in (5) and (6) except for the intercept.[7]

The p.d.f. of Y_{ij} is

$$\begin{aligned} f(y_{ij} \mid \boldsymbol{\theta}) &= \Pr(R_{ij} = 0 \mid \boldsymbol{\theta}_1)\delta_0(y_{ij}) + \Pr(R_{ij} = 1 \mid \boldsymbol{\theta}_1)f(s_{ij} \mid \boldsymbol{\theta}_2) \\ &= [1 - p_{ij}(\boldsymbol{\theta}_1)]\delta_0(y_{ij}) + p_{ij}(\boldsymbol{\theta}_1)f(s_{ij} \mid \boldsymbol{\theta}_2) \end{aligned}$$

where $\boldsymbol{\theta} = [\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2]$ and $\delta_0(y)$ is a Dirac delta function[9] such that

$$\begin{cases} \int_{-\infty}^{\infty} \delta_0(y)dy_{ij} = 1 \\ \delta_0(y) = 0 \text{ when } y_{ij} \neq 0 \end{cases}$$

The conditional expectation of Y_{ij} is:

$$E(Y_{ij} \mid \boldsymbol{\theta}) = p_{ij}(\boldsymbol{\theta}_1)\mu_{S_{ij}}(\boldsymbol{\theta}_2), \tag{7}$$

and the conditional variance is:[10]

$$\mathrm{var}(Y_{ij} \mid \boldsymbol{\theta}) = p_{ij}(\boldsymbol{\theta}_1)\mathrm{var}(S_{ij} \mid \boldsymbol{\theta}_2) + (p_{ij}(\boldsymbol{\theta}_1))[1 - (p_{ij}(\boldsymbol{\theta}_1))]\mu_{S_{ij}}(\boldsymbol{\theta}_2)^2.$$

The contribution to the likelihood for the ith subject ($i = 1, \ldots, m$) is

$$\begin{aligned} &L_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2, \sigma_e, \rho; y_{i1}, \ldots, y_{in_i}) \\ &\quad = \int_{u_{1i}} \int_{u_{2i}} \prod_{j=1}^{n_1} f(y_{ij} \mid \beta_1, \beta_2, u_{1i}, u_{2i}) f(u_{1i}, u_{2i} \mid \sigma_1, \sigma_2, \sigma_e, \rho) du_{1i} du_{2i}. \end{aligned}$$

The likelihood is then

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2, \sigma_e, \rho) = \prod_{i=1}^{m} \int_{u_{1i}} \int_{u_{2i}} \prod_{j=1}^{n_i} f(y_{ij} \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, u_{1i}, u_{2i}) f(u_{1i}, u_{2i} \mid \sigma_1, \sigma_2, \sigma_e, \rho) du_{1i} du_{2i}$$

$$= \prod_{i=1}^{m} \int_{u_{1i}} \int_{u_{2i}} \prod_{j=1}^{n_i} [1 - p_{ij}(\boldsymbol{\beta}_1, u_{1i})]^{1-r_{ij}} [p_{ij}(\boldsymbol{\beta}_1, u_{1i})]^{r_{ij}}$$

$$\times f(s_{ij} \mid \boldsymbol{\beta}_2, u_{2i}) f(u_{1i}, u_{2i} \mid \sigma_1, \sigma_2, \sigma_e, \rho) du_{1i} du_{2i} \qquad (8)$$

In the correlated mixed-distribution model, it is not assumed that the random effects are independent and, as a result, the components of (8) for occurrence and intensity contain a common parameter, ρ. Therefore, the two components of the likelihood cannot be maximized separately as in Lachenbruch[2] or Grunwald and Jones.[5] In model (8) it is also possible for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to share common parameters. However, because $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are on different scales, doing so may lead to parameter estimates that are difficult to interpret.

With the assumptions that u_{1i} and u_{2i} are independent, i.e. that $\rho = 0$, the likelihood may be factored into two parts that correspond to the occurrence process and the intensity process:

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2, \sigma_e) = \prod_{i=1}^{m} \int_{u_{1i}} \prod_{j=1}^{n_i} [1 - p_{ij}(\boldsymbol{\theta}_1)]^{1-r_{ij}} [p_{ij}(\boldsymbol{\theta}_1)]^{r_{ij}} f(u_{1i} \mid \sigma_1) du_{1i}$$

$$\times \prod_{i=1}^{m} \int_{u_{2i}} \prod_{j=1}^{n_i} f(s_{ij} \mid \boldsymbol{\theta}_2) f(u_{2i} \mid \sigma_2, \sigma_e) du_{2i}$$

The first component is the likelihood for the occurrence process, $L_R(\boldsymbol{\beta}_1, \sigma_1)$, and the second component is the likelihood for the intensity process, $L_S(\boldsymbol{\beta}_2, \sigma_2, \sigma_e)$. With the further assumption that $\boldsymbol{\theta}_1$ has no parameters in common with $\boldsymbol{\theta}_2$, $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2, \sigma_e)$ is maximized when each component is maximized separately. When $\rho = 0$, the model is referred to as the uncorrelated mixed-distribution model.

2.2 Model fitting

If u_{1i} and u_{2i} are assumed to be independent, then maximum likelihood methods may be used to maximize both components of the likelihood separately. Wolfinger and O'Connell's pseudo-likelihood approach,[11] Breslow and Clayton's penalized quasi-likelihood approach,[12] or optimization of the likelihood approximated by adaptive Gaussian quadrature,[13] may be used to maximize $L_R(\boldsymbol{\beta}_1, \sigma_1)$ and $L_S(\boldsymbol{\beta}_2, \sigma_2, \sigma_e)$ separately. The overall likelihood $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2, \sigma_e)$ is the product of $L_R(\boldsymbol{\beta}_1, \sigma_1)$ and $L_S(\boldsymbol{\beta}_2, \sigma_2, \sigma_e)$, and the maximum of $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2, \sigma_e)$ occurs when $L_R(\boldsymbol{\beta}_1, \sigma_1)$ and $L_S(\boldsymbol{\beta}_2, \sigma_2, \sigma_e)$ are maximized separately. If the models contain no random effects, this allows the mixed-distribution model to be estimated using standard software of generalized linear models[14]. When correlated random effects are present these special cases are useful for obtaining initial estimates when optimizing the correlated model likelihood (8).

The full likelihood (8) for the correlated mixed-distribution model can be maximized using quasi-Newton optimization of a likelihood approximated by adaptive Gaussian quadrature.[13] This method is implemented in the SAS PROC NLMIXED procedure (SAS Institute, Cary, NC, Version 8). This procedure allows the user to specify a general likelihood, in particular one of the form (8), and also allows great flexibility for specification of the distribution of S_{ij}. We assume a logistic-lognormal-normal model, where 'logistic' refers to the modeling of the occurrence part of the model (2), 'lognormal' to the modeling of the intensity part of the model (3), and 'normal' to the assumption that the random effects are assumed to have a bivariate normal distribution (4).

To fit this model we developed a SAS macro (MIXCORR, available from the authors) that calls PROC GENMOD and PROC NLMIXED. The user must specify the dataset, the outcome variable, covariates for the binomial component of the model and for the lognormal component of the model, and the variable that identifies the random unit. The macro estimates a binomial model for the occurrence and a lognormal model for the intensity (both without random effects) using PROC GENMOD. These parameter estimates are used as starting values in estimating the separate occurrence and intensity models with uncorrelated random effects using SAS PROC NLMIXED. Finally, the parameter estimates from the two uncorrelated random effects models are used as starting values for the mixed-distribution model with correlated random effects in a final PROC NLMIXED run. The starting value for the covariance of the random effects is calculated using the estimates of σ_1^2 and σ_2^2 and $\rho = 0.5$.

2.3 Model checking

The model assumes normality and constant variance of random effects, μ_{1i} and μ_{2i}, and the residuals of the intensity distribution. Standard regression diagnostics may be used to assess the goodness of fit of the model. Quantile–quantile plots can be constructed for $\hat{u}_{1i}$ and $\hat{u}_{2i}$, and for the residuals for the intensity variable, given by $\ln(s_{ij}) - (\mathbf{X}'_{2ij}\boldsymbol{\beta}_2 + u_{2i})$. If the normality assumption is not violated, the data will fall in a straight line. A plot of the residuals for the intensity distribution versus fitted values will indicate if the assumption of constant variance is violated. A nonrandom pattern indicates departure from this assumption.

2.4 Interpretation of Fixed-Effects Parameters

The separate effects of the fixed-effect occurrence and intensity parameters, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, have the same interpretations for occurrence and intensity as they would have if the two components of the model were fit separately (e.g., logistic and lognormal regression). If a variable is used in both the occurrence and intensity models, however, there may be interest in quantifying the overall effect of the variable on the total amount Y. This can be carried out as follows.

Assume that Z is a covariate in both the occurrence and intensity models (2) and (3), and that $\mathbf{X}_1$ and $\mathbf{X}_2$ are vectors of the other occurrence and intensity covariates, respectively. For simplicity we suppress the subscripts i and j. Then from (2) and (3),

$$\Pr(R = 1 \mid \boldsymbol{\theta}_1) = \frac{\exp(\mathbf{X}'_1\boldsymbol{\beta}_1 + \alpha_1 z + u_1)}{1 + \exp(\mathbf{X}'_1\boldsymbol{\beta}_1 + \alpha_1 z + u_1)} \tag{9}$$

and

$$E(S \mid \boldsymbol{\theta}_2) = \exp\left(\mathbf{X}_2'\boldsymbol{\beta}_2 + \alpha_2 z + u_2 + \frac{\sigma_e^2}{2}\right). \tag{10}$$

Then the ratio of mean amount of Y when $Z = z + 1$ to that when $Z = z$ is

$$\frac{\mathrm{E}(Y \mid Z = z+1, \boldsymbol{\theta})}{\mathrm{E}(Y \mid Z = z, \boldsymbol{\theta})} = \left[\frac{\Pr(R = 1 \mid Z = z+1, \boldsymbol{\theta}_1)}{\Pr(R = 1 \mid Z = z, \boldsymbol{\theta}_1)}\right]\left[\frac{\mathrm{E}(S \mid Z = z+1, \boldsymbol{\theta}_2)}{\mathrm{E}(S \mid Z = z, \boldsymbol{\theta}_2)}\right] \tag{11}$$

From (10) the second term in (11) is $\exp(\alpha_2)$. In general the first term in (11) depends on $\mathbf{X}_1'\boldsymbol{\beta}_1 + u_1$ as well as on α_1 and z. However, some insight can be gained by substituting (9) into (11) and noting that the function

$$\left(\frac{\exp(k + \alpha_1)}{1 + \exp(k + \alpha_1)}\right) \Big/ \left(\frac{\exp(k)}{1 + \exp(k)}\right) \to 1 \text{ as } k \to \infty$$

and

$$\left(\frac{\exp(k + \alpha_1)}{1 + \exp(k + \alpha_1)}\right) \Big/ \left(\frac{\exp(k)}{1 + \exp(k)}\right) \to \exp(\alpha_1) \text{ as } k \to -\infty.$$

Thus in (11),

$$\frac{\mathrm{E}(Y \mid Z = z+1, \boldsymbol{\theta})}{\mathrm{E}(Y \mid Z = z, \boldsymbol{\theta})} \approx \begin{cases} \exp(\alpha_2) & \text{when } \mathbf{X}_1'\boldsymbol{\beta}_1 + u_1 \text{ is large and positive} \\ \exp(\alpha_1)\exp(\alpha_2) & \text{when } \mathbf{X}_1'\boldsymbol{\beta}_1 + u_1 \text{ is large and negative} \end{cases}$$

When $\mathbf{X}_1'\boldsymbol{\beta}_1 + u_1$ is large and positive, $\Pr(R = 1) \approx 1$ so there are few zeros, $\mathrm{E}(Y) \approx \mathrm{E}(S)$, and the effect of Z on Y is mainly via the mean of the nonzero values. When $\mathbf{X}_1'\boldsymbol{\beta}_1 + u_1$ is large and negative, the ratio of means in (11) is a combination of occurrence and intensity effects. The term $\Pr(R = 1 \mid Z = z + 1, \boldsymbol{\theta}_1)/\Pr(R = 1 \mid Z = z, \boldsymbol{\theta}_1)$ is the risk ratio for occurrence per one unit change in Z. When $\Pr(R = 1) \approx 0$, as when $\mathbf{X}_1'\boldsymbol{\beta}_1 + u_1$ is large and negative, this term is close to the odds ratio for occurrence per one unit change in Z, which is $\exp(\alpha_1)$ as in the usual logistic regression interpretation. Special cases of all of these results hold if Z enters into only the occurrence model ($\alpha_2 = 0$) or only into the intensity model ($\alpha_1 = 0$).

In practice, neither of the limiting cases of (11) may apply. In order to determine the range of the effect of a common covariate Z on Y, the ratio of the means in (11) can be computed for the minimum, maximum, and median values of Z, and for the minimum and maximum values of the other covariates. The limit $\exp(\alpha_1)\exp(\alpha_2)$ in (11) provides an upper (lower) limit for the combined effect of Z on Y when α_1 is positive (negative). Note that these results also hold when no random effects are present and thus provide an interpretation of the combined effect of a variable in a mixed-distribution regression model.

2.5 Interpretation of random effects

The random effects in the correlated mixed-distribution model, u_{1i} and u_{2i}, account for unobserved heterogeneity among units. In the occurrence part of the model, the random intercept on the link (e.g., logit) scale, $\beta_{10i} = \beta_{10} + u_{1i}$, allows some units to have a consistently low or high probability of a nonzero response. The variance of the random effect, σ_1^2, indicates the variability of the probability of a nonzero response among units with similar covariate patterns. The random intercept, $\beta_{20i} = \beta_{20} + u_{2i}$, in the intensity part of the model, allows some units to have consistently low or high mean of nonzero values. If σ_2^2 is large it indicates that there is a great deal of heterogeneity of mean nonzero responses among units with similar covariate patterns.

Allowing correlation of the random effects u_{1i} and u_{2i} allows units with consistently high occurrence probability to have consistently high (low) mean of nonzero values when the correlation between u_{1i} and u_{2i}, ρ, is positive (negative).

3 Simulation results

A simulation study was performed to study the performance of the parameter estimates from Section 2.2. Using a method adapted from Zeger and Karim,[15] data were simulated from the logistic-lognormal-normal mixed-distribution model with

$$p_{ij}(\boldsymbol{\theta}_1) = \beta_{10} + \beta_{11}t_j + \beta_{12}x_i + \beta_{13}x_i t_j + u_{1i}$$
$$\log(S_{ij} \mid u_{2i}) \sim N(\beta_{20} + \beta_{21}t_j + \beta_{22}x_i + u_{2i}, \sigma_e^2),$$

and correlated random effects u_{1i} and u_{2i} as in (4).

One hundred datasets with $m = 100$ units (clusters or subjects) of size $n_i = 7$ were generated using each of the two sets of parameter values shown in Table 1. The number of quadrature points specified in NLMIXED was held to the maximum number

Table 1 Simulation results for the correlated mixed-distribution model using $m = 100$ simulated datasets from the model given in Section 3 with each of the two sets of true parameter values

True value	Mean of 100 estimates		True value	Mean of 100 estimates	
β_{10}	2.50	2.51	β_{10}	2.50	2.58
β_{11}	0.10	0.10	β_{11}	0.10	0.09
β_{12}	−1.00	−1.04	β_{12}	−1.00	−1.13
β_{13}	0.05	0.05	β_{13}	0.05	0.06
σ_1^2	1.00	0.97	σ_1^2	**10.00**	9.98
β_{20}	4.00	4.00	β_{20}	4.00	3.98
β_{21}	0.50	0.50	β_{21}	0.50	0.50
β_{22}	1.50	1.49	β_{22}	1.50	1.39
σ_2^2	1.44	1.41	σ_2^2	**14.40**	14.33
σ_e^2	1.00	1.00	σ_e^2	1.00	1.01
$\rho\sigma_1\sigma_2$	0.60	0.58	$\rho\sigma_1\sigma_2$	6.00	6.15

determined adaptively, seven. The estimates from NLMIXED appear to be unbiased (Table 1).

4 Application

The Medical Expenditure Panel Survey (MEPS) is a longitudinal survey conducted by the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS). MEPS data may be used to obtain estimates of health care use, medical expenditures, and insurance coverage in the United States. In the Household Component of the MEPS, data were collected on health care use and expenditures, demographic characteristics, medical conditions, health status, and insurance coverage on 22 601 persons in 10 596 households. Although the expenditure and use data are collected longitudinally, they are aggregated by year; only data for 1996 were analyzed. However, due to the multiple subjects within households, these data exhibit clustering, and the techniques described in this paper are applicable with household as the unit of repeated measurement. Although the MEPS is a representative sample and weighted and unweighted frequencies are provided in order to provide data analysts the ability to make population-level estimates, the analysis presented in this paper was not weighted.

For this analysis, the impact of age, sex, health rating, the presence of a medical condition, census region (Northeast, Midwest, South, West), the presence of physical limitations, and insurance status on total medical expenditures in 1996 were modeled. The health rating was assessed on a scale of 1 to 5 with 1 corresponding to 'Excellent' and 5 corresponding to 'Poor'. Whether or not a subject had a medical condition was based on household-reported medical conditions collected in 1996. A subject was considered to have a limitation if they were found to have any type of limitation with activities of daily living (ADLs: including bathing, dressing, and getting around the house), instrumental activities of daily living (IADLs: including using the telephone, paying bills, taking medications, preparing light meals, doing laundry, and going shopping), physical limitations (such as walking, climbing stairs, grasping objects, reaching overhead, lifting, bending or stooping, and standing for long periods of time), any limitation that impeded their work, housework, or school activities, or vision or hearing limitations. The presence or absence of any insurance (including coverage under CHAMPUS/CHAMPVA, Medicare, Medicaid or other public hospital/physician or private hospital/physician insurance) was reported for each month in 1996. The portion of the year that the respondent was insured was used as a covariate in the analysis. There were from one to fourteen persons in a family; the median number of family members was three. Owing to missing data on the limitation, health rating variable, insurance status, region, age, or sex, 746 respondents were excluded from the analysis.

Both models with and without correlated random effects were fit using the MIXCORR macro and a backwards selection procedure. In all cases the model with correlated random effects was found to be better than the model with uncorrelated random effects (based on a likelihood ratio test and AIC). A model with all covariates was the best of the models considered. Parameter estimates from the models with uncorrelated and correlated random effects are given in Table 2.

Table 2 Parameter estimates and model comparisons for final model fit to MEPS data

Parameter	Uncorrelated		Correlated	
	Estimate (S.E.)	$p>\|t\|$	Estimate (S.E)	$p>\|t\|$
Occurrence (Logistic)				
Intercept	−2.8292(0.1129)	<0.0001	−2.8131(0.1129)	<0.0001
Medical condition (N = 0/Y = 1)	3.0342(0.0724)	<0.0001	2.9792(0.0717)	<0.0001
Limitations (N = 0/Y = 1)	0.5574(0.0894)	<0.0001	0.5498(0.0897)	<0.0001
Portion of year insured (0–1)	1.7152(0.0680)	<0.0001	1.7262(0.0679)	<0.0001
Age (years)	0.0051(0.0014)	0.0003	0.0040(0.0014)	0.0043
Health rating (1–5)	0.1782(0.0292)	<0.0001	0.2181(0.0296)	<0.0001
Sex (M = 0/F = 1)	0.6122(0.0509)	<0.0001	0.6318(0.0510)	<0.0001
Region 1 (Northeast)	0.5173(0.0881)	<0.0001	0.5184(0.0884)	<0.0001
Region 2 (Midwest)	0.5547(0.0867)	<0.0001	0.5465(0.0869)	<0.0001
Region 3 (South)	0.1359(0.0724)	0.0606	0.1236(0.0726)	0.0886
σ_1^2	1.1502(0.1140)	<0.0001	1.1852(0.1149)	<0.0001
Intensity (Lognormal)				
Intercept	3.0459(0.0619)	<0.0001	2.8653(0.0641)	<0.0001
Medical condition (N = 0/Y = 1)	1.0485(0.0473)	<0.0001	1.1503(0.0482)	<0.0001
Limitations (N = 0/Y = 1)	0.5681(0.0299)	<0.0001	0.5743(0.0299)	<0.0001
Portion of year insured (0–1)	0.8702(0.0347)	<0.0001	0.9047(0.0348)	<0.0001
Age (years)	0.0189(0.0005)	<0.0001	0.0187(0.0005)	<0.0001
Health rating (1–5)	0.2609(0.0111)	<0.0001	0.2697(0.0112)	<0.0001
Sex (M = 0/F = 1)	0.2235(0.0206)	<0.0001	0.2366(0.0206)	<0.0001
Region 1 (Northeast)	0.1188(0.0355)	0.0008	0.1237(0.0356)	0.0005
Region 2 (Midwest)	0.1314(0.0341)	0.0001	0.1383(0.0342)	<0.0001
Region 3 (South)	0.0126(0.0313)	0.6878	0.0145(0.0313)	0.6435
σ_2^2	1.6959(0.0239)	<0.0001	1.6960(0.0238)	<0.0001
σ_2^2	0.2368(0.0190)	<0.0001	0.2468(0.0192)	<0.0001
$\rho\sigma_1\sigma_2$	—	—	0.3523(0.0347)	<0.0001 ($\rho = 0.6514$)
Name	Value		Value	Difference in −2 log likelihood
AIC	293 107.6		293 002.0	
−2 ll	293 061.6		292 954.0	107.59 ($p<0.0001$)

Checks of the goodness of fit of the model, as described in Section 2.3, were performed. The quantile–quantile plots for the random effects showed no indication of departure from a straight line. Plots of residuals versus fitted values for the lognormal intensity model did not show any indications of heteroscedasticity of variance.

The separate and combined effects of the variables included in the model are presented in Table 3. In this table each column is referenced by a lower case letter. Recall that from (11) the ratio of the overall mean for a one unit change in a common covariate Z may be represented as follows:

$$\underset{(k)}{\left[\frac{\mathrm{E}(Y \mid Z=z+1)}{\mathrm{E}(Y \mid Z=z)}\right]} = \underset{(j)}{\exp(\alpha_2)} \underbrace{\underset{(h)}{\exp(\alpha_1)} \underset{(g)}{\left[\frac{\Pr(R=0 \mid Z=z+1)}{\Pr(R=0 \mid Z=z)}\right]}}_{(i)} \tag{12}$$

Table 3 Effects on medical expenditure in 1996 MEPS data for covariates (a–f) on probability of occurrence (*i*), on intensity (*j*), and on mean amount (*k*)

Variable	(*a*) Medical condition	(*b*) Limitation	(*c*) Health rating	(*d*) Sex	(*e*) Insurance	(*f*) Region	(*g*) Ratio of Prob.*	(*h*) e^{α_1}	(*i*) $e^{\alpha_1} *$ ratio (*f*)	(*j*) e^{α_2}	(*k*) Ratio of means
Any medical condition	N/Y	N	Ex	M	0	4	0.404	19.672	7.946	3.159	25.101
(N = 0/Y = 1)	N/Y	Y	Poor	F	1	2	0.058	19.672	1.145	3.159	3.617
Any limitation	N	N/Y	Ex	M	0	4	0.945	1.733	1.638	1.776	2.909
(N = 0/Y = 1)	Y	N/Y	Poor	F	1	2	0.580	1.733	1.006	1.776	1.786
Health rating (1 = Ex, 2 = Very	N	N	Ex/VG	M	0	4	0.981	1.244	1.220	1.310	1.598
Good, 3 = Good, 4 = Fair, 5 = Poor)	Y	Y	Ex/VG	F	1	2	0.807	1.244	1.004	1.310	1.314
Sex (M = 0/F = 1)	N	N	Ex	M/F	0	4	0.935	1.881	1.759	1.267	2.228
	Y	Y	Poor	M/F	1	2	0.535	1.881	1.007	1.267	1.276
Portion of year insured (0–1)	N	N	Ex	M	0/1	4	0.733	5.619	4.116	2.471	10.172
	Y	Y	Poor	F	0/1	2	0.184	5.619	1.036	2.471	2.560
Region (1 = Northeast,	N	N	Ex	M	0	4/2	0.946	1.727	1.633	1.148	1.876
2 = Midwest, 3 = South, 4 = West)	Y	Y	Poor	F	1	4/2	0.582	1.727	1.006	1.148	1.155

*Age set equal to the mean, 34.8 years.
Ex = excellent, VG = very good.
The terms are as given in equation (12).

The variable listed in the first column of Table 3 is z in the equation. Because the values of the other variables in the model impact the ratio of probabilities (g), various scenarios for values of the other variables are given in columns (a)–(f). In general, the 'low' condition, in which the other covariates in the model are at their lowest value, is given on the first row for the variable, and the 'high' condition, in which the other covariates in the model are at their highest value, is given on the following row.

Presence of a medical condition was associated with increased mean medical expenditure in 1996. The increase ranged from 3.6 times (for subjects with otherwise 'high risk' covariate patterns) to 25.1 times (for subjects with otherwise 'low risk' covariate patterns). Differences in this effect were due to differences in the effect of a medical condition on the probability of some medical expenditure. The mean medical expenditure for respondents with a physical limitation was from 1.8 to almost 3 times the mean of respondents without physical limitations. Having insurance for the entire year was associated with increased mean medical expenditures from 2.5 to 10.2 times that of persons who did not have insurance for the entire year, with the larger increase for patients with an otherwise low risk covariate pattern. A one unit increase in the health rating scale, which actually corresponded to a decline in health, increased the mean amount of health expenditures by 1.3 to 1.6 times. The difference between a male subject and a similar female increased the mean amount of expenditure from 1.3 to 2.2 times. Lastly, living in the Midwest increased the mean amount of expenditure from 1.2 to 1.9 times that of those living in the West. In none of these cases was there a uniform dominance of the occurrence effect over the intensity effect (or vice versa) on total expenditure.

The significant random effects variance for the occurrence shows that after accounting for covariate differences among subjects, some families have a greater probability of seeking medical care than others. Similarly, the highly significant random effect variance for intensity indicates that after accounting for covariate differences, some families have consistently higher (or lower) expenditures when they do seek medical care than the norm. The positive correlation between the occurrence and intensity random effects indicates that after accounting for covariate differences, families with a greater tendency to seek medical care tended also to report a higher mean amount of positive expenditures.

5 Discussion

We have proposed a model for longitudinal or repeated measures data with clumping at zero, using a mixed-effects, mixed-distribution model. The model includes features of the cross-sectional statistical models of Lachenbruch,[1,2] the cross-sectional econometric models of Heckman,[8] Duan et al.,[3] and Manning et al.,[6] and the time series model of Grunwald and Jones.[5] In addition, by including correlated random errors, the occurrence and intensity parts of the model are linked. An interpretation of fixed-effects parameters was given, which also applies to mixed-distribution models for cross-sectional data.

We have shown how the proposed model may be estimated using standard software for non-linear and generalized linear mixed models such as SAS PROC NLMIXED.

Simulations indicate that this method of estimation gives unbiased results for both fixed and random effects. We chose this method due to its good performance on simulation studies, and because it can be easily implemented in SAS. However, other methods of model fitting appropriate for GLMMs and nonlinear mixed-effects models[13] potentially could be used to fit our model, including penalized quasi-likelihood[12] or a Monte Carlo method within a Bayesian framework.[15]

We used the approach to model the association between several covariates including demographic characteristics, insurance coverage, and health status on health care expenditures of subjects, using random effects to account for clustering of subjects into families. We noted strong fixed effects of most covariates on total amount of expenditure, through both the probability of nonzero expenditure and the mean of nonzero expenditures. We also noted strong random effects due to clustering of subjects within families. Further, adjusting for covariates, there was a tendency for subjects in families that had a higher probability of some health care expenditure to also have higher mean nonzero expenditure.

The model proposed in this paper is appropriate for data with true zeros. Although this method may appear to be applicable to the case where data are left censored or missing, a zero in these cases is not a real zero and should not be treated as such when calculating the mean amount.

One byproduct of our work is a method for interpreting effects of covariates. Estimation of the mean amount, including the probability of zeros, is in our view one of the main reasons for developing models for the combined response when zeros are included. Totals, such as total expenditure for a group over a period of time including the fact that some subjects will have no expenditures, can be estimated from these means. The method we propose gives information about the effect of a covariate on this mean amount and how that effect arises as a combination of the covariate's effect on occurrence probability and on mean nonzero amount. The methods we propose are also applicable in the cross-sectional case.

Many modifications and extensions of our methods are possible. Some types of data with clumping at zero may exhibit serial correlation, particularly if repeated measurements are made longitudinally. One possible extension of the model described in this paper is to a transition model or an autoregressive error structure to account for the type of autoregressive pattern that longitudinal data might exhibit. Another direction for extension would be toward the Heckman[8] econometric model, which uses correlated random errors to allow the probability of occurrence and the mean intensity to be related in a cross-sectional model. We have adapted that approach to include correlated random unit effects, our main interest. Our model could be modified to include correlated within-subject random components as well. Such a model could again be estimated using standard methods for GLMMs and SAS PROC NLMIXED. Further extensions might include both a transition component and a random effect. Other extensions of the correlation structure, such as a stochastic parameter model including random slopes as well as random intercepts, would be possible as well. However, as the correlation structures become more complex and additional parameters are added to the model, the model becomes less parsimonious and more difficult to fit.

In this paper we have assumed that the nonzero amounts follow a lognormal distribution, as in the two-part models of Duan *et al*.[3] This distribution is appropriate

for skewed, positive, continuous data and is frequently used for analysis of cost data. The gamma distribution would be an alternative choice for the intensity distribution, as in Grunwald and Jones[5] and Hyndman and Grunwald.[16] The Weibull distribution could also be chosen. All of these are distributions on $(0, \infty)$ and can be accommodated by the model. Because all of these distributions are capable of modeling a variety of positively skewed shapes, the exact form assumed for the errors would not be expected to have a substantial effect on the estimated model parameters or inferences. However, if quantiles of the nonzero amounts are to be estimated (as in Grunwald and Jones[5]), more care is needed to specify and check the form of the error distribution. A nonparametric density estimate[17] could also be considered for estimating the shape of the error distribution. This approach potentially could provide better estimation of quantiles, although sparse data in the tails of the highly skewed distributions may cause difficulties. We are not aware of any applications of nonparametric density estimation to data with clumping. Some care would be needed so that the estimates were applied only to the nonzero data rather than smoothing across the zeros as well. It is unclear how multiple covariates and random effects could be included.

In our model, an intensity model appropriate for $y_i > 0$ was chosen so that it may be assumed that zeros only arise when $r_i = 0$. Otherwise, it is unknown whether the zeros arise from the distribution for the occurrence component of the model, or from the intensity component of the model. An example of a mixture of distributions that contains both type of zeros is a Binomial–Poisson mixture. Lambert[18] has proposed zero-inflated Poisson (ZIP) regression for handling data that arise from this mixture of distributions. Dunson and Haseman[19] extended ZIP regression to a transition model for longitudinal data with an application to carcinogenicity in animal studies. Hall[20] adapted Lambert's methodology to an upper-bounded count situation by using a zero-inflated binomial model. He also incorporated random effects into the ZIP regression model to accommodate repeated measures data. Our model was developed for the case where the nonzero data arise from a continuous distribution. The Poisson would not be an appropriate distribution for the intensity variable for the medical expenditure data described in this paper, as these data are not independent counts.

In the econometric literature there has been an increased interest in semiparametric approaches to fitting data with clumping at zero.[21,22] In addition, Hyndman and Grunwald[16] have developed a generalized additive mixed-distribution model with a first-order Markov structure for time series data. Another extension to the model described in this paper could involve a semiparametric modeling approach.

Because the correlated mixed-distribution model is a nonlinear model that incorporates the models and methods of GLMMs, as the methodology advances in the area of nonlinear models and GLMMs, especially with regard to model fitting and diagnostics, the methodology of the correlated mixed-distribution model will be advanced as well.

Acknowledgments

This research was partially supported by the National Institute of General Medical Studies, Grant GM38519 (RHJ). The authors would also like to acknowledge Dr. Gary Zerbe, Dr. David Young and Dr. Becki Bucher Bartelson for their guidance with this

research. Dr. Tooze is a fellow in the National Cancer Institute's Cancer Prevention Fellowship Program in the Division of Cancer Prevention.

References

1 Lachenbruch P. Analysis of data with clumping at zero. *Biometrische Zeitschrift* 1976; **18**: 351–56.

2 Lachenbruch P. Utility of regression analysis in epidemiologic studies of the elderly. In: Wallace R, Woolson R, eds. *The epidemiologic study of the elderly.* Oxford University Press, 1992.

3 Duan N, Manning WG, Morris CN, Newhouse JP. *A comparison of alternative models for the demand for medical care.* Santa Monica, California: The RAND Corporation, 1982. R-2754-HHS.

4 Amemiya T. *Advanced econometrics.* Cambridge, Massachusetts: Harvard University Press, 1985.

5 Grunwald GK, Jones RH. Markov models for time series with mixed distribution. *Environmetrics* 2000; **11**: 327–39.

6 Manning W, Duan N, Rogers W. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* 1987; **35**: 59–82.

7 Diggle PJ, Liang K-Y, Zeger SL. *Analysis of longitudinal data.* Oxford: Oxford University Press, 1994.

8 Heckman JJ. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator of such models. *Annals of Economic and Social Measurement* 1976; **5**: 475–92.

9 Robertson JS, Bolinger K, Glasser LM, Sloane NJ, Gross R. Chapter 1. In: Zwillinger D, ed. *CRC standard mathematical tables and formulae.* Boca Raton: CRC Press, 1996: 71.

10 Aitchinson J. On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* 1955; **50**: 901–8.

11 Wolfinger R, O'Connell M. Generalized linear models. *Journal of Statistical Computer Simulation* 1993; **48**: 233–43.

12 Breslow N, Clayton D. Approximate inference in generalized linear models. *Journal of the American Statistical Association* 1993; **88**: 9–25.

13 Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* 1995; **4**: 12–35.

14 McCullagh P, Nelder J. *Generalized linear models* (2nd ed). London: Chapman & Hall, 1989.

15 Zeger SL, Karim MR. Generalized linear models with random effects. *Journal of the American Statistical Association* 1991; **86**: 79–86.

16 Hyndman RJ, Grunwald GK. Generalized additive modeling of mixed distribution markov models with application to Melbourne's rainfall. *Australian and New Zealand Journal of Statistics* 2000; **42**: 145–58.

17 Silverman BW. *Density estimation for statistics and data analysis*. London: Chapman and Hall, 1986.

18 Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.

19 Dunson D and Haseman J. Modeling tumor onset and multiplicity using transition models with latent variables. *Biometrics* 1999; **55**: 965–70.

20 Hall D. Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* 2000; **56**: 1030–39.

21 Maddala G. Limited dependent variable models using panel data. *Journal of Human Resources* 1987; **22**: 307–38.

22 Vella F. Estimating models with sample selection bias. *Journal of Human Resources* 1998; **33**: 127–69.