BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

A probit- log- skew-normal mixture model for repeated measures data with excess zeros, with application to a cohort study of paediatric respiratory symptoms

Sadia Mahmud*[1], WY Wendy Lou[2] and Neil W Johnston[3]

Abstract

Background: A zero-inflated continuous outcome is characterized by occurrence of "excess" zeros that more than a single distribution can explain, with the positive observations forming a skewed distribution. Mixture models are employed for regression analysis of zero-inflated data. Moreover, for repeated measures zero-inflated data the clustering structure should also be modeled for an adequate analysis.

Methods: Diary of Asthma and Viral Infections Study (DAVIS) was a one year (2004) cohort study conducted at McMaster University to monitor viral infection and respiratory symptoms in children aged 5-11 years with and without asthma. Respiratory symptoms were recorded daily using either an Internet or paper-based diary. Changes in symptoms were assessed by study staff and led to collection of nasal fluid specimens for virological testing. The study objectives included investigating the response of respiratory symptoms to respiratory viral infection in children with and without asthma over a one year period. Due to sparse data daily respiratory symptom scores were aggregated into weekly average scores. More than 70% of the weekly average scores were zero, with the positive scores forming a skewed distribution. We propose a random effects probit/log-skew-normal mixture model to analyze the DAVIS data. The model parameters were estimated using a maximum marginal likelihood approach. A simulation study was conducted to assess the performance of the proposed mixture model if the underlying distribution of the positive response is different from log-skew normal.

Results: Viral infection status was highly significant in both probit and log-skew normal model components respectively. The probability of being symptom free was much lower for the week a child was viral positive relative to the week she/he was viral negative. The severity of the symptoms was also greater for the week a child was viral positive. The probability of being symptom free was smaller for asthmatics relative to non-asthmatics throughout the year, whereas there was no difference in the *severity* of the symptoms between the two groups.

Conclusions: A positive association was observed between viral infection status and both the probability of experiencing any respiratory symptoms, and their severity during the year. For DAVIS data the random effects probit -log skew normal model fits significantly better than the random effects probit -log normal model, endorsing our parametric choice for the model. The simulation study indicates that our proposed model seems to be robust to misspecification of the distribution of the positive skewed response.

Background

Zero-inflated data is frequently encountered in health science studies and is characterized by the occurrence of "excess" zeros that more than a single distribution can explain. There is a considerable amount of literature dealing with the problem of zero-inflated count data such as Zero Inflated Poisson (ZIP) or Zero Inflated Binomial (ZIB) mixture models, and their extension to clustered or longitudinal data structures [1-7]. The early research on

* Correspondence: sadia.mahmud@aku.edu
[1] Department of Community Health Sciences, The Aga Khan University, Stadium Road, P O Box 3500, Karachi 74800 Pakistan
Full list of author information is available at the end of the article

modeling zero-inflated continuous data was reported in econometrics literature [8]. Tobin [9] proposed the "Tobit" model assuming an underlying normally distributed variable whose non-positive values were considered as unobserved. In the Tobit model the observed zeros were treated as the unobserved non-positive values of the underlying variable that have been left-censored, and a linear regression model with normally distributed errors was suggested. Thereby the same stochastic process and regression parameters determined whether the response was zero or positive, as well as the magnitude of the positive response. Cragg [10] proposed a "two-part" model for semi-continuous data that separately modeled the dichotomous nature of the response (zero versus positive values), and the magnitude of the positive values respectively. Cragg suggested the probit link for modeling the binary part and a truncated normal density for the positive values, allowing a different set of covariates to be associated with the probability of having a non-zero response and the magnitude of the positive response respectively. Duan et al. [11] suggested a logit/lognormal coupling for the two part model for semi continuous data, and applied this model to demand for medical care. Moulton and Halsey [12] generalized the two-part model with logit/lognormal coupling by incorporating interval censoring, implying that the observed zeros were either a realization of the true zero point distribution or observations from the distribution of the positive outcome observed as zero due to detection limits. Heckman [13,14] extended the Tobit model to a two-part model referred to as the sample selection model that assumed an underlying bivariate normal error. Duan et al [11,15] pointed out that this model has poor numerical and statistical properties. Further discussion and references regarding sample selection model and its comparison with the two-part model are provided by Min and Agresti [8].

Olsen and Schafer [16] and Tooze et al [17] extended the two-part logit- lognormal mixture model, proposed by Duan et al. [11] for cross-sectional data, to repeated measures data by including two subject specific random effects in the logit and log-normal components respectively. These authors assumed that the random effects follow a bivariate normal distribution, and allowed for the two random effects to be correlated. Recently some additional work has been reported in literature extending mixture models for a continuous outcome with a discrete component to clustered data. Li et al [18] presented a zero-inflated log-normal model that takes hierarchical clustering structure of a data into account; they incorporated nested random intercepts in the linear predictors of the logit and lognormal model components respectively, assuming the random effects are independently and normally distributed. Liu et al [19] proposed a multi-level two-part random effects logit-lognormal model; two nested random effects were included in each part to model the nested clustering structure in a data, assuming the respective random effects in the two parts followed a bivariate normal distribution. More recently Su et al [20] showed that bias can be induced for regression coefficients when random effects are truly correlated but misspecified as independent in a 2-part mixed model.

The positive part of a zero-inflated continuous variable is often skewed to the right, logarithmic transformation had been suggested to correct for the skewness. Although Olsen and Schafer [16] allowed a more general transformation (a monotone increasing function) that would make the positive component approximately Gaussian, they only used a log transformation in the illustrative example reported in their paper and did not discuss the choice of the adequate transformation. The customary statistical approach of applying a log transformation in setting of right skewness is ad hoc, and may or may not optimally account for distributional characteristics of the data under study. As a referee noted the log transformation may often over-transform the data making the distribution skewed in the opposite direction. In an attempt to remedy this problem Chai and Bailey [21] extended the cross-sectional two-part model by suggesting a skew-normally distributed error in the regression equation for log-transformed positive values, and proposed a probit/log-skew normal mixture model for cross-sectional data. The skew-normal distribution accommodates asymmetry in a more flexible manner, and can model both positively or negatively skewed data (depending on the sign of the skewness parameter) reducing to the normal distribution when the skewness parameter is zero. Tooze et al [22] and Kipnis et al [23] suggested a different remedy to deal with the problem of skewness of the positive responses in the two-part model for longitudinal semi-continuous data. They introduced the Box-Cox transformation of the positive responses so that on the transformed scale the within subject error and the subject specific random effect in the regression model of the positive part were approximately normally distributed. The normality transformation was done within the modeling step so that the positive responses were transformed to normality conditionally on the covariates in the model, and the Box-Cox transformation parameter was estimated along with the regression parameters in the maximum likelihood procedure. When adopting the Box-Cox transformation approach one has to make an assumption that such a transformation does exist. In the present communication we adopt the approach suggested by Chai and Bailey [21] to model the positive part of semi continuous data using the skew-normal distribution. In the Discussion section we will comment on the comparison of the Box-Cox transformation approach with our present model.

In this paper we present an extension of the cross-sectional two-part Bernoulli-log-skew-normal mixture model, suggested by Chai and Bailey [21], for longitudinal zero-inflated continuous data. We modeled the clustering structure of the data by introducing correlated bivariate normal random effects in both parts, similar to what Olsen and Schafer [16] and Tooze et al [17] did for modeling longitudinal data in the two-part model with normally distributed error for log transformed positive responses. As discussed above Chai and Bailey [21] suggested the Bernoulli-log-skew-normal model for cross-sectional data thereby presenting a more flexible approach for modeling the asymmetry of the positive responses as compared to the ad hoc log transformation. Like Chai and Bailey [21] we used the probit link to model the binary component, however a logit link can also be used. The potential of the proposed model was demonstrated through analysis of a real data from a study titled "Diary of Asthma and Viral Infections Study". In addition, by fitting a random effects probit-lognormal mixture model on the dataset, we conducted a likelihood ratio- test for the skewness parameter in the log skew normal distribution, and demonstrated that the random effects probit/log-skew normal mixture model fits better on the dataset as compared to the random effects Bernoulli-log normal model proposed in references [16-19]. Moreover, in order to assess the aptness of the proposed probit/log-skew normal mixture model we conducted a probit/log-beta regression simulation for repeated measures data.

Methods

Probit log-skew normal mixture model for repeated measures

Let Y_{ij} be an observation from the j_{th} measurement on the i_{th} subject with all $Y_{ij} \geq 0$. The probability density function of Y_{ij} is:

$$\begin{aligned} f(Y_{ij}) &= \Pr(Y_{ij} = 0) = p_{ij} \qquad \text{if } Y_{ij} = 0 \\ &= \Pr(Y_{ij} > 0)\, f(Y_{ij}|Y_{ij} > 0) \text{ if } Y_{ij} > 0 \\ &= (1\text{-}p_{ij})\, f(Y_{ij}|Y_{ij} > 0) \end{aligned} \quad (1)$$

We used the probit link function for p_{ij}:

$$p_{ij} = \Phi[-(X^{T}_{(1)ij}\beta_{(1)} + \tau_{0i})]$$

where Φ is the standard normal cumulative distribution function, $X_{(1)ij}$ is the vector of explanatory variables associated with the probability of the i_{th} subject being symptom free at the j_{th} occasion, and $\beta_{(1)}$ is the vector of corresponding regression parameters. The positive outcome, $Y_{ij} > 0$, was assumed to follow a skew normal (SN) distribution.

$$\log(Y_{ij}|Y_{ij} > 0,\ \tau_{1i}) \sim SN(X^{T}_{(2)ij}\beta_{(2)} + \tau_{1i}, \sigma, \delta)$$

$X_{(2)ij}$ is the vector of explanatory variables associated with the severity of symptoms for the i_{th} subject at the j_{th} occasion, $\beta_{(2)}$ and is the vector of corresponding regression parameters. μ,σ,δ are the parameters of the skew normal distribution (we are setting $\mu = X^{T}_{(2)ij}\beta_{(2)} + \tau_{1i}$), δ is referred to as the skewness parameter (see Additional File 1). The reason for modeling the positive outcome on the logarithmic instead of the original scale is to ensure positive estimation as log of negative numbers does not exist. In order to model the correlation among repeated measurements on the same subject, we included two random effects (τ_{0i}, τ_{1i}) in the linear predictors of the two regression model components respectively. We assumed that these random effects follow a bivariate normal distribution (BVN), that is,

$$\tau = \begin{pmatrix} \tau_{0i} \\ \tau_{1i} \end{pmatrix} \sim BVN(0, \Sigma), \Sigma = \begin{pmatrix} s_{11}s_{12} \\ s_{21}s_{22} \end{pmatrix}$$

S_{11} and S_{22} being the variance of the random effects in the probit and log-skew-normal components respectively and S_{12} being the covariance between the two random effects.

Maximum Marginal Likelihood Estimation

Defining an indicator function, I_{ij} ($I_{ij} = 1$ if $Y_{ij} = 0$, $I_{ij} = 0$ if $Y_{ij} > 0$) likelihood contribution from the i_{th} subject can be expressed as follows:

$$L_i = \int \prod_{j=1}^{m_i} (p_{ij})^{I_{ij}} [(1 - p_{ij}) \times f(Y_{ij}|Y_{ij} > 0)\ ]^{1-I_{ij}} f(\tau_{0i}, \tau_{1i}) d\tau_{0i}\, d\tau_{1i}$$

where m_i is the number of repeated measurements on the i_{th} subject and $f(\tau_{0i}, \tau_{1i})$ is the joint distribution of the two random effects. We assumed that $f(\tau_{0i}, \tau_{1i})$ is bivariate normal.

Assuming the measurements on different subjects are independent, the likelihood to be maximized is:

$$L = \prod_{i=1}^{n} L_i$$

where n is the total number of subjects in the sample. In the Maximum Marginal Likelihood Estimation approach

integration over random effects is approximated by numerical integration. We used Gaussian quadrature to obtain the marginal likelihood and employed Double Dogleg optimization method to maximize the likelihood. This optimization technique combines the concept of the Trust Region and Quasi- Newton methods and works well for medium to moderately large optimization problems [24-26].

Diary of Asthma and Viral Infection Study

Respiratory viral infections (RVI), most commonly of rhinovirus, have been found to coincide with the majority of children's asthma exacerbations throughout the year, including the post summer vacation epidemic periods, in both community and hospital based studies [27-30]. Children admitted to hospital for wheezing have been shown to have a significantly higher rate of RVI [28,29]. Furthermore, asthma exacerbations in children are highly cyclic and follow predictable seasonal patterns [30]. The 'Diary of Asthma and Viral Infection Study (DAVIS)', a 12 month, cohort study was conducted at McMaster university to monitor infection and respiratory symptoms including asthma exacerbations in children aged 5-11 years with and without asthma. The study objectives included investigating the response of respiratory symptoms to RVI in children with and without asthma over a one year period. The study period was the 2004 calendar year. Respiratory symptoms were recorded daily using either an Internet or paper-based diary. Changes in symptoms were assessed by study staff and led to collection of further information, the use of spirometry and collection of nasal fluid specimens for virological testing. Virological testing was conducted using polymerase chain reaction techniques as previously described [31].

The study was designed and executed by academic investigators (with Neil W Johnston as the principal investigator, PI) and was approved by the Research Ethics Board of St. Joseph's Healthcare, Hamilton (R.P. #03-2195). Written informed consent for children to participate was obtained from parents of all subjects and assent from appropriately aged children. The raw data is accessible only to the PI and the research team, as was approved by the Research Ethics Board. Individuals who wish to have access to the data for replicating the study results are advised to contact the PI for necessary Research Ethics Board (or Institutional Review Board) approval.

Six lower respiratory tract (LRT) symptoms (cough during the day, cough during the night, wheeze, difficulty breathing or shortness of breath during the night, difficulty breathing or shortness of breath during the day and breathing problems interfering with child's regular activities during day) were categorised by subjects on a 5 point scale from 0 (none) to 4 (very severe). Overall daily LRT symptom scores were determined by summing the six LRT symptom scores, emulating the approach taken in a previous study [27]. For most of the subjects many daily LRT scores were zero leading to sparse data, hence daily LRT scores were aggregated to give weekly average LRT scores. This was done by writing a SAS macro that executed PROC EXPAND for each subject. PROC EXPAND can change the frequency of a single time series such as conversion from daily measurements to weekly or monthly averages or totals. The weekly interval was defined as Sunday to Saturday, giving a total of 51 weeks for the year 2004. If a daily measurement was missing for a given week for a subject, the missing value was replaced by the weekly average. If more than two measurements were missing for a week, the weekly average was treated as a missing value.

Data for 190 subjects (135 asthmatics and 55 non-asthmatics) were available for the analysis. The majority of the subjects (172) had measurements for all 51 weeks, one subject had measurements for only 41 weeks, whereas 17 subjects had measurements ranging from 44 to 50 weeks. Eleven subjects entered the study later than the 1st week but no later than the 8th week of the year. The subjects started dropping from the study after the 41st week. One hundred and eighty- six subjects had weekly LRT measurements after the 46th week.

A histogram of weekly average LRT scores indicated that about 75% scores were zero and the positive scores seemed to be represented by a continuous skewed distribution (Figure 1). The zeros correspond to the absence of respiratory symptoms during a week, and the positive scores measured the severity of the respiratory symptoms when present. In order to account for the excess zeros and repeated measurements on each subject over the weeks of the year, we propose the mixture model (1) that first considers the response as a dichotomous variable (zero versus greater than zero), and then models the positive response using log skew normal distribution. The clustering structure of the data is modeled by the two correlated random effects in each part. The regression analysis was aimed at investigating the relationship of children's respiratory symptom scores with viral infection status throughout the year 2004, adjusting for their asthmatic status, age and sex.

The mixture model (1) was fitted using PROC NLMIXED on SAS (see SAS codes in Additional File 2). We started with fitting a main effect model with asthmatic/non-asthmatic status, child's sex and age, week of follow-up and viral infection status as independent variables in both model components (-2LL = 14181, BIC = 14270, AIC = 14215). The variable "week" was coded as week = 1 corresponding to the 1st week of January (starting 4th January that was the 1st Sunday of January 2004), week = 2 to the 2nd week and so on. Viral infection status, being a time-dependent variable, was defined as positive

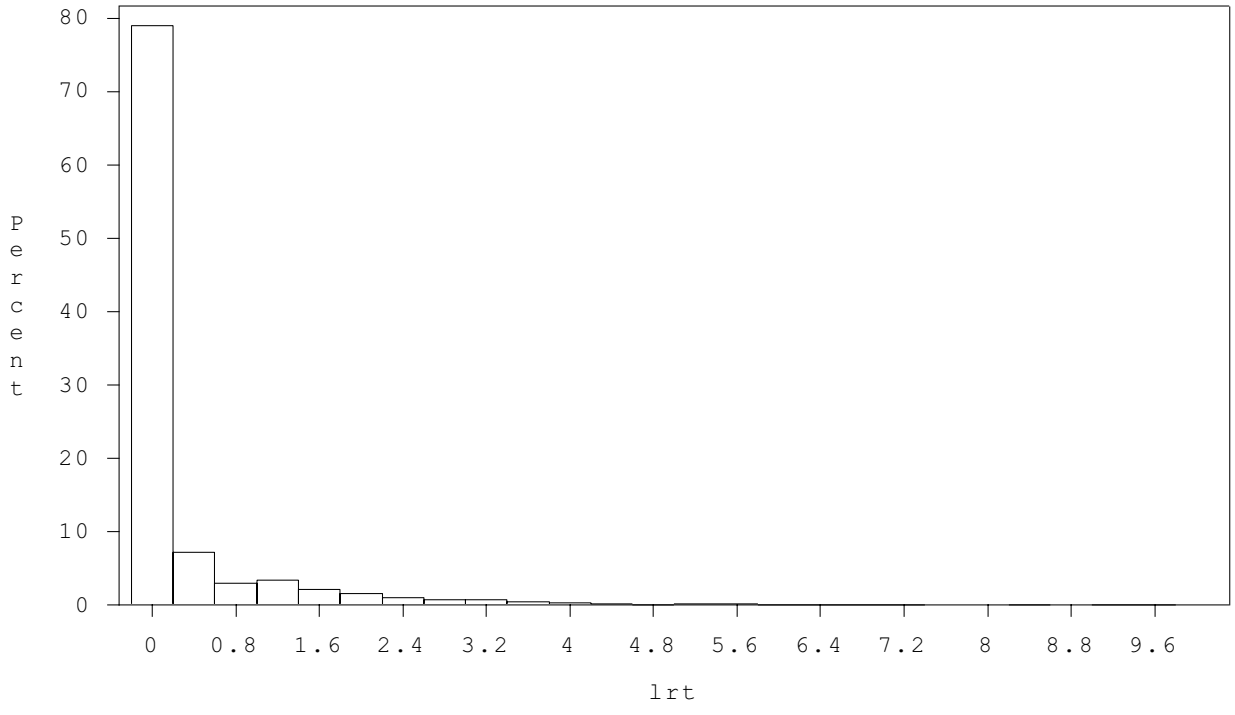

Figure 1 Histogram of weekly average LRT scores.

for a week if the subject had any respiratory virus detected during that week, negative otherwise; it was coded as "1" if the subject was viral positive and "0" if viral negative for a given week. Initially age and week of follow-up were modeled as linear continuous variables. Initial values of parameters were taken from a probit-log skew normal mixture model without random effects (that converged more quickly). In addition the initial parameter values for the variance and covariance of the random effects were set as $s_{11} = 0.5$, $s_{22} = 0.5$, $s_{12} = 0$. Generalized Additive Model (GAM) approach was used to examine the scale of continuous predictor variables, age and week of follow-up, with reference to their regression relationship with weekly LRT scores. GAM is a non-parametric smoothing technique for exploring the scale of an independent continuous variable for a regression model [32]. The exploratory GAM analysis indicated a quadratic scale for both age and week (more pronounced for week). Fitting the mixture model by including square terms for age and week in both model components gave -2LL = 14069, BIC = 14179, AIC = 14111. Wald p-values that were employed for preliminary screening indicated that the square terms for age could be removed from the model. Fitting the mixture model again with week of follow-up modeled as quadratic and age as linear in both model components led to -2LL= 14069, BIC= 14169, AIC= 14107. Hence based on the log-likelihood, AIC and BIC criteria the week of follow-up was modeled as quadratic, whereas age was modeled as linear in both model components respectively. Next all possible two-way interactions were included in both mixture model components (-2LL = 14037, BIC = 14241, AIC = 14115). Using the Wald p-values of the interaction terms for preliminary screening we obtained the final model (- 2LL = 14056, BIC = 14166, AIC = 14098) including two interactions that were both statistically and biologically significant (Table 1).

Simulation Study

We carried out a simulation study to assess the performance of the proposed random effects probit log-skew normal model, if the underlying distribution of the positive response is different from log-skew normal. For the simulation study we generated repeated measures data from a probit log- beta model stated as follows:

$$\begin{aligned} Y_{ij} &= 0 \quad \text{with prob} \quad \Phi[-(\alpha_0 + \beta_0 x_{ij} + \tau_{0i})] \\ &= e^{(\alpha_1 + \beta_1 x_{ij} + \tau_{1i}) + scale(\varepsilon_{ij}))} \\ &\quad \text{with prob} \quad \Phi[(\alpha_0 + \beta_0 x_{ij} + \tau_{0i})] \end{aligned} \tag{2}$$

where x_{ij} is a binary time dependent predictor variable (similar to the viral infection status in the DAVIS data, x_{ij} = 1 if i_{th} subject is viral positive for j_{th} week, x_{ij} = 0 if nega-

Table 1: Probit/log-skew normal model for weekly LRT scores. Parameter estimates (standard errors) n= 190.

	Final Model (LRT)	
	probit	**log-skew-normal**
β_0	-0.635(0.285)**	0.298(0.263)
$\beta_{asthamatic}$	0.598(0.104)***	0.096(0.089)
β_{male}	0.024(0.078)	-0.129(0.076)*
β_{age}	-0.027(0.031)	6.2e-3(0.028)
β_{week}	-0.050(0.005)***	-3.0e-3(0.008)
$\beta_{week*week}$	8.5e-4(0.9e-4) ***	3.5e-4(1.0e-4)***
β_{virus}	2.459(0.126)***	0.744(0.057)***
$\beta_{age*week}$	-	-2.0e-3(0.7e-3)***
$\beta_{asthamatic*week}$	-6.4e-3(2.6e-3)**	-
σ		0.666(0.049)***
δ		-0.952(0.100)***
s11	1.023(0.183)***	
s22		0.214(0.041)***
s21		0.267(0.077)***
-2 Log Likelihood	14056	
AIC	14098	
BIC	14166	
* p-value < 0.10	** p-value < 0.05	*** p-value < 0.01

tive). ε_{ij} follows a beta distribution with probability density:

$$f(\varepsilon) = \frac{(\varepsilon)^{\alpha-1}(1-\varepsilon)^{\beta-1}}{B(\alpha,\beta)} \qquad 0 < \varepsilon < 1$$

where $\alpha > 0, \beta > 0$ are the parameters of the beta distribution and $B(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du$ is the beta function. A random variable Y follows a four parameter beta distribution if:

$$Y = \Theta + scale(\varepsilon)$$

where $\varepsilon \sim BETA(\alpha, \beta)$ and parameters Θ and (Θ + *scale*) define the minimum and maximum values of Y respectively. Hence in model (2) log $(Y_{ij} \mid Y_{ij} > 0)$ follow the four parameter beta distribution with $\Theta = (\alpha_1 + \beta_1 x_{ij} + \tau_{1i})$. τ_{0i} and τ_{1i} are the subject specific random effects such that

$$\tau = \begin{pmatrix} \tau_{0i} \\ \tau_{1i} \end{pmatrix} \sim BVN(0,\Sigma), \Sigma = \begin{pmatrix} s_{11}s_{12} \\ s_{21}s_{22} \end{pmatrix}$$

The rationale for selecting the beta distribution to model the positive response in our simulation study was that, similar to the skew normal distribution, it can be positively ($\beta > \alpha$) or negatively ($\beta < \alpha$) skewed, as well as symmetric ($\beta = \alpha$). The skewness of the beta distribution is:

$$Skewness = 2\frac{(\beta-\alpha)}{(\alpha+\beta+2)}\sqrt{\frac{(\alpha+\beta+1)}{\alpha\beta}}$$

We generated 50 repeated measurements corresponding to each week for each of 200 independent clusters (subjects) and assigned the following values for the model parameters, $\alpha_0 = -1, \beta_0 = 2.5, \alpha_1 = -20, \beta_1 = 0.75, s_{11} = 1, s_{22} = 0.2, s_{21} = 0.2$, *scale* = 30. For each

subject at every week the time dependent binary covariate x_{ij} was generated as a Bernoulli variable with probability = 0.03 (of all weekly average LRT scores in DAVIS data for about 3% viral infection variable was positive). The true values of α_1 parameters and *scale* were specified so that the minimum and the maximum values of $(Y_{ij} \mid Y_{ij} > 0)$ are approximately equal to $e^{-20} \approx 0$ and $e^{(-20+30)} = e^{10}$ (that is a very large number) respectively, thereby simulating a situation where the positive outcome can be considered as a continuous variable bounded below at zero.

We considered two scenarios with respect to specifying the parameters of the beta distribution, (i) a negatively skewed beta distribution (β = 70, α = 130, *skewness* = -0.0883) and (ii) a symmetric beta distribution ($\beta = \alpha$ = 100, *skewness* = 0). For each of the two scenarios 200 datasets were generated from model (2) and probit log-skew normal model (1) was fitted on each dataset. We also did some simulation runs generating data from a positively skewed beta distribution (β = 130, α = 70, *skewness* = -0.0883) and fitted random effects probit-log-skew normal model.

Results

Analysis of DAVIS

The final fit of the mixture model (1) to DAVIS data is reported in Table 1. The covariates significantly associated with the probability of having no LRT symptoms were asthmatic/non-asthmatic status, week of follow-up and viral infection status (probit component). The highly significant positive estimate of β_{virus} indicates that for the week a child was viral positive, the probability of being LRT symptom free was much less than that for the week the child was viral negative (p-value < 0.0001, beta = 2.459). The variable week was modeled as quadratic in the linear predictor of the regression model of the probit component. Moreover, there was a significant interaction between the asthmatic/non-asthmatic status and (the linear term of) week of follow-up (Wald p-value = 0.0150). We also examined the interaction of asthmatic/non-asth-

matic status with the quadratic term in week (that is asthmatic*week*week) but that was insignificant based on the Wald, likelihood, AIC and BIC criteria. The probability of being LRT symptom free was lower in the beginning of the year, increased from January to August, and after that decreased until December (Figure 2). The association between the probability of being symptom free and the asthmatic/non-asthmatic status can be clearly seen from Figure 2; the probability of being symptom free is smaller for asthmatics relative to non-asthmatics throughout the year, the difference being more pronounced in the beginning of the year.

Covariates associated with the *severity* of LRT symptoms were subject's age and sex, viral infection status and week of follow-up (log skew normal component). There was no significant difference in the severity of the LRT symptoms between asthmatic and non-asthmatic children (p-value = 0.2758). For the week a child was viral positive the severity of LRT symptoms was significantly greater than for the week he/she was viral negative (p-value < 0.0001, beta = 0.744). There was some marginal evidence that the severity of LRT symptoms was lesser for the male relative to the female children (p-value = 0.0938, beta = -0.1286). As for the probit component, the quadratic term for week was significant in the log-skew normal component. Moreover, there was a significant interaction between the age of a child and (the linear term of) the week of follow (p-value = 0.0069). We also examined the interaction of age with the quadratic term in week (that is age*week*week), however that was insignificant based on the Wald, likelihood, AIC and BIC criteria. In Figure 3 we plot predicted values of log(LRT > 0) from the fitted model versus week of the year for two age groups, < 8 years (mean age) and ≥8 years, along with mean log(LRT > 0) values at each week computed from the data. (In Figure 4 mean LRT > 0 scores on the original scale, computed from the data, are plotted versus week of follow-up). The severity of LRT symptoms was higher in the beginning of the year, decreased in summer and increased again by the end of the year. The severity of LRT symptoms was higher for children younger than 8 years relative to older children, and seemed to exhibit a more pronounced seasonal pattern.

From the model fit in Table 1 we note that the Wald p-value for the skewness parameter (δ) was highly significant (p-value < 0.0001, estimate of δ = -0.952). This suggests the importance of using the log-skew-normal distribution to model the positive response for this data, using the log-normal distribution would be inadequate. We also conducted the likelihood ratio test for testing the hypothesis H_0: δ = 0 by fitting a probit log-normal model; chi-square test statistic, x^2 (df = 1) = 10, p-value < 0.005 indicating significance of the skewness parameter (δ). In

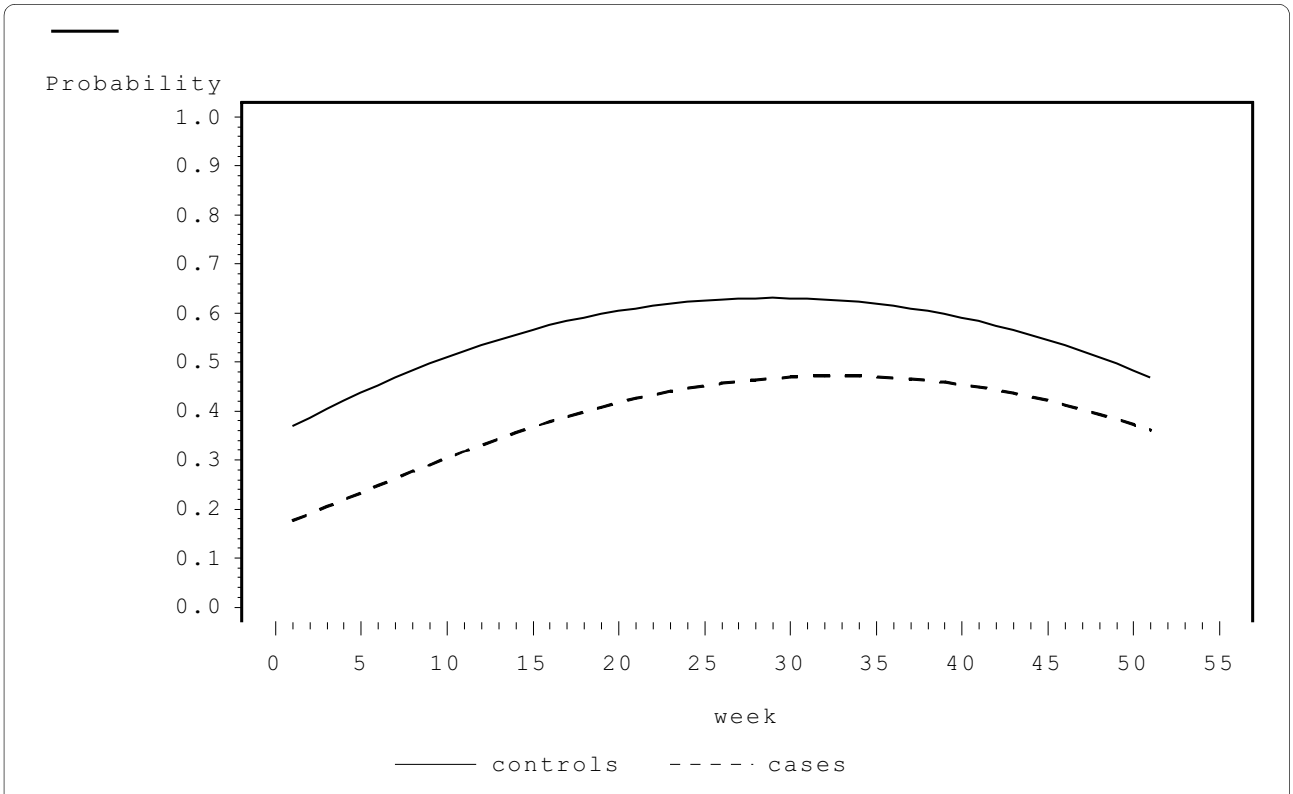

Figure 2 Probability of being LRT symptom free versus the week of the year.

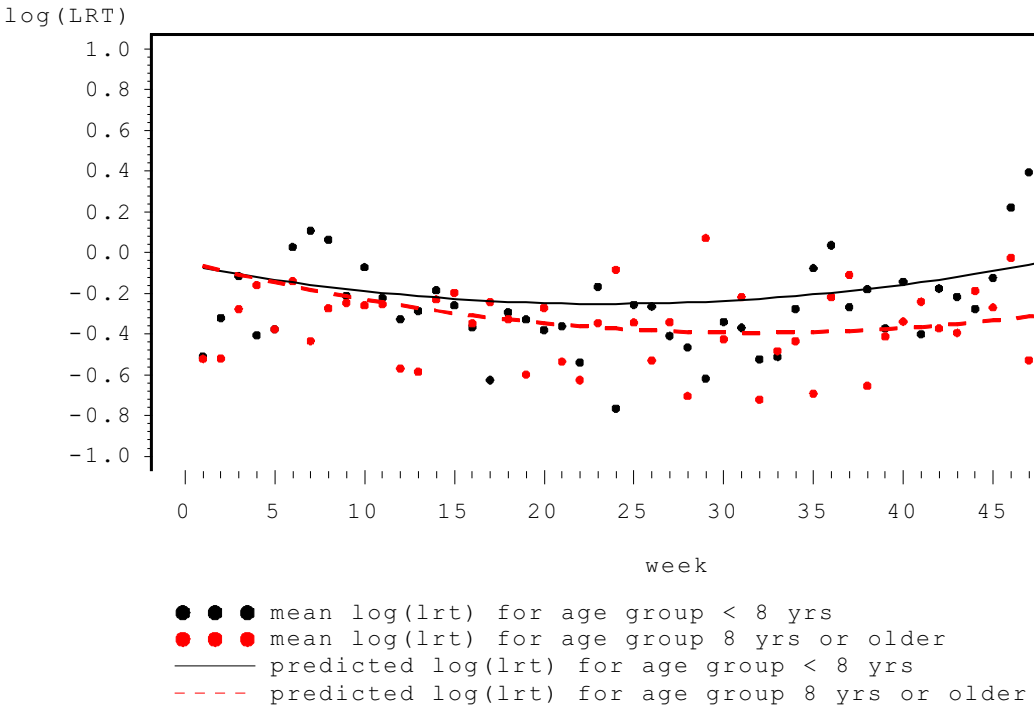

Figure 3 Log (LRT > 0) versus the week of the year

addition we note that for the probit-log-normal mixture model AIC = 14106 and BIC = 14171, these criteria also suggest that the probit-log- skew normal mixture model reported in Table 1 is superior to probit log-normal model. We note the sign of the estimate of the skewness parameter indicating a negatively skewed distribution for the log of positive LRT scores. We would like to point out here that for certain parameterization of skew normal distribution the Fisher information matrix is singular when asymmetry parameter is equal to zero. This leads to difficulties regarding asymptotic distributions for maximum likelihood estimators and likelihood ratio statistic [33]. However, for alternative parameterization of skew normal distribution these difficulties are resolved, and the likelihood function exhibits a more regular behaviour without a stationary point when the asymmetry parameter is equal to zero [33]. In our model we used the skew normal parameterization suggested by Sahu et al [34] that was also employed by Chai and Bailey [21] in their probit -log- skew normal mixture model for zero-inflated continuous cross-sectional data. Chai and Bailey [21] conducted and reported the likelihood ratio test for testing the null hypothesis: asymmetry parameter (for the skew-normal distribution) = 0.

Moreover, the Wald p-values for the variance of the random effects (s_{11}, s_{22}) in the two model components respectively were highly significant (p-value < 0.0001) indicating the random effects were needed in the model to account for the correlation among measurements on the same subject. The significant *positive* covariance (s_{21}) between the two random effects has an intuitively appealing interpretation; the higher the probability of a subject of being positive for LRT symptoms, the greater the severity of LRT symptoms.

Results of the simulation study

The results of the simulation study are presented in Table 2. In the simulation study we generated data from random effects probit log-beta model (2) and fitted random effects probit log-skew normal model (1). In Table 2 we report the bias and mean square error for the estimated values of the intercept (in the probit component), regression coefficients corresponding to the time dependent binary variable, x_{ij} and the variance and covariance of the random effects in the two model components respectively from the simulation runs. We note that the regression coefficient corresponding to x_{ij} in the continuous part, for both log-skew-normal and log-beta model, is the difference in the expected value of $\log(Y_{ij} \mid Y_{ij} > 0)$ when $x_{ij} = 1$ versus $x_{ij} = 0$ and hence is comparable. However, the intercepts in the log-skew-normal and log-beta models are not comparable. As discussed in the Methods sec-

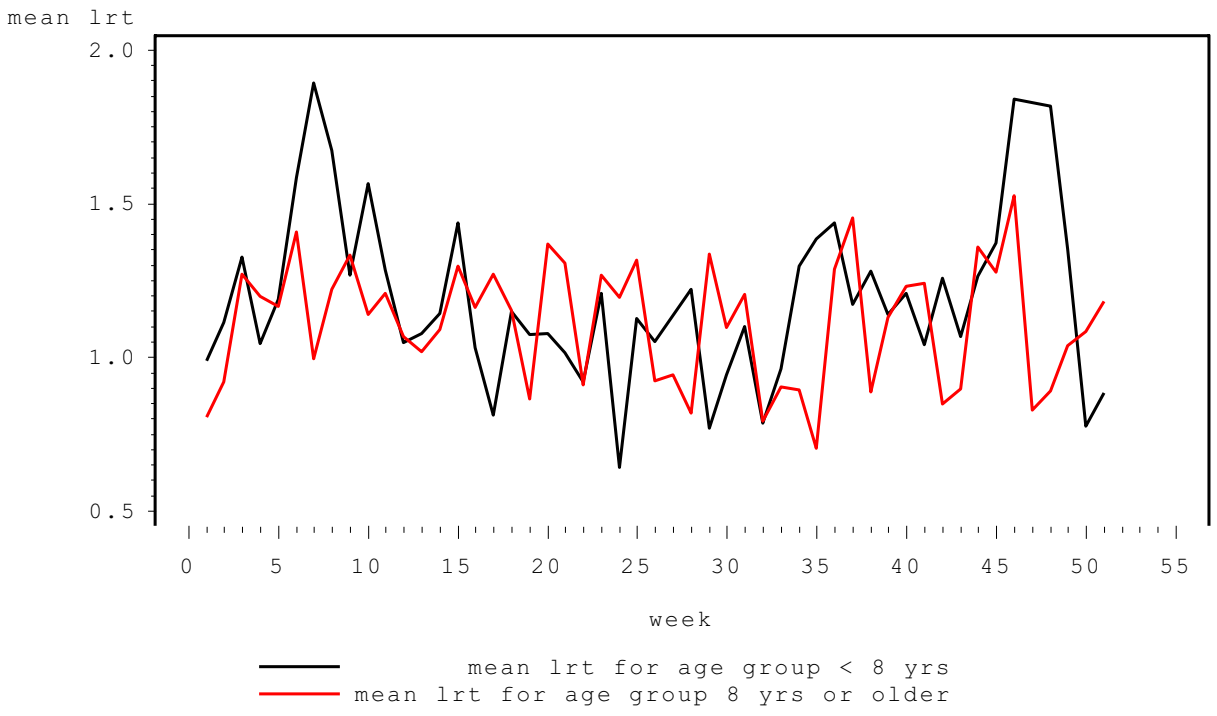

Figure 4 Mean LRT > 0 versus the week of the year

tion, we considered two scenarios with respect to specifying the parameters of the beta distribution (i) a negatively skewed beta distribution, and a (ii) symmetric beta distribution. For both these scenarios the simulation results indicate that the estimates seem to be unbiased, particularly for the regression coefficients corresponding to the time dependent predictor variable in both model components. This suggests that the probit-log skew normal model performs reasonably well, as the primary goal of the simulation study was to assess the ability of the model to estimate the affect of an explanatory variable on the response.

Table 2: Simulation Results from 200 simulation runs. Fitted model is probit-log-skew-normal.

PROBIT component	**$\beta < \alpha$ [1] (negatively skewed)**		**$\beta = \alpha$ [2] (symmetric)**	
	Bias	**MSE [4]**	**Bias**	**MSE**
Intercept (-1)[3]	0.0303	0.0128	0.0054	0.0116
$\beta_{\text{Viral Infection}}$ (2.5)	-0.0186	0.0130	-0.0064	0.0112
s11 (1)	-0.0927	0.0243	-0.0914	0.0235
Log-beta component				
Intercept (-20)				
$\beta_{\text{Viral Infection}}$ (0.75)	0.0004	0.0050	-0.0090	0.0055
s22 (0.2)	-0.0059	0.0010	-0.0085	0.0012
s12 (0.2)	-0.0167	0.0028	-0.0199	0.0033

1. $\beta = 70$, $\alpha = 130$
2. $\beta = \alpha = 100$
3. Numbers in parentheses refer to true parameter values
4. Mean Square Error (MSE)

For scenario (i) where we generated the datasets from a negatively skewed beta distribution, the mean (standard deviation) of estimates of δ (the skewness parameter in the probit-log-skew normal model) was -0.7031 (0.1951) and for scenario (ii) corresponding to a symmetric beta distribution, the mean (standard deviation) of estimates of δ was -0.3171 (0.2760) from the 200 simulation runs. This suggests that the log-skew normal model correctly identified the *negatively* skewed as well as the symmetric distribution; for the former the mean estimated value of δ was negative, and for the latter though the mean estimated value was somewhat negative but does not appear to be significantly different from zero due to the large standard deviation. We also did some simulation runs generating data from a positively skewed beta distribution ($\beta = 130$, $\alpha = 70$, *skewness* = 0.0883) and fitted probit-log-skew normal model that gave similar results.

Discussion

In this communication we have presented a probit/log-skew-normal mixture model for continuous repeated measures data with a discrete component at zero. We modeled the clustering

structure of the data by including two random effects in the probit and log-skew-normal model components respectively, assuming the random effects follow a bivariate normal distribution. In case of longitudinal data structure, in addition to random intercepts it may be of interest to include a random slope for time in the continuous component as suggested by Su et al [20]. This can be done through a straightforward extension of the model we proposed by including a random slope for time in the log-skew normal component, assuming the three random effects follow a trivariate normal distribution. Recent research has focused on the impact of misspecification of random effects distribution on the maximum likelihood estimates for generalized linear mixed models (GLMM). For the case of linear mixed models (that correspond to the identity link function for GLMM) the parameter estimates are rather robust with respect to deviation from normality of random effects. However, for random-intercept logistic models the estimates of the mean structure parameters can have substantial bias upon misspecification of random effects distribution in case of large random effects variance [35]. In our present analysis the maximum likelihood estimates of the variances of random intercepts in the two parts are rather small (of the order of magnitude of 1 in the binary and 0.2 in the continuous component) thereby suggesting that the potential bias in the estimates of fixed effects in case of misspecification of random effects distribution could be small. However, there can be considerable bias in the estimate of variance components in case of misspecification of the random effects distribution, thereby making it difficult to distinguish between the small or large variance scenarios [35]. This suggests the need for further research to investigate the impact of misspecification of random effects distribution on the estimates of fixed and random effects in a two-part mixture model for semi-continuous data. This research will be particularly relevant as the continuous model component constitutes a non-linear mixed model. Investigating the impact of the distribution of random effects on parameter estimation in mixture models for clustered semi continuous data will be taken up as future research.

For longitudinal data, subject specific random effect models account for the correlation among measurements on the same subject through the concept of heterogeneity among subjects; some subjects are intrinsically high responders, others low-responders. However, serial correlation models time varying stochastic process within a subject [36]. In our proposed random effects mixture model, serial correlation can be incorporated by including a lagged response variable as a predictor variable in the model [37].

The potential of the proposed model was demonstrated through analysis of a real dataset from DAVIS. The response variable of interest was the weekly average LRT symptom score; about 75% of these scores were zero and the positive scores formed a skewed distribution. We assumed that the zeros correspond to 'true zeros' indicating absence of any LRT symptoms. This assumption seems reasonable in the context of LRT symptoms reported by subjects in DAVIS, where zero scores correspond to "No symptoms". Incorporating interval censoring in a zero-inflated mixture model implies that the observed zeros are either a realization of a 'true zero' point distribution, or an observation from the distribution of the positive outcome observed as zero due to detection limitation [12]. The latter does not seem to be relevant for the self-reported LRT symptoms in DAVIS, detection limits are usually relevant for measurements involving laboratory markers [12,21].

As discussed above, in addition to random intercepts we also included a random time slope in the log-skew normal component for the main effect mixture model for DAVIS; the estimate of the variance of the random slope, though significant, was quite small (0.0002) implying that the random effect of time did not vary substantially from subject to subject.

For the positive outcome the log-skew-normal distribution fits significantly better, for the dataset we used as an example, as compared to log-normal distribution suggested by authors in references [17-20]. For the binary component of the mixture model either a probit or a logit link can be used, authors in references [17-20] presented a logit-lognormal coupling for their mixture models. We also fitted a logit-log-skew-normal random effects mixture model on DAVIS data, however with the probit link the likelihood of the fitted model was higher than that with the logit link, though a likelihood ratio test could not be conducted as the models were not nested.

We conducted a simulation study, to assess the performance of log skew normal distribution in modeling the positive component of the response, by generating repeated measures data from a probit log-beta model, and fitting the proposed probit-log-skew normal mixture model. The results of the simulation study indicate that the probit-log-skew normal mixture model performs reasonably well in estimating the true regression parameters, when the underlying distribution of the log transformed positive response was a beta distribution. Like the skew-normal distribution, the beta distribution can be positively or negatively skewed or symmetric. The skew-normal distribution did demonstrate an ability to recognize a negatively skewed and a symmetric beta distribution cor-

rectly, though the skewness parameters for the skew-normal and the beta distribution were not directly comparable.

Finally the Box-Cox transformation approach to normally transform both the skewed positive observations of a semi continuous longitudinal outcome, and the subject specific random effects in the regression model of the positive part [22,23] seems to be a potentially interesting alternative to the random effects regression model with log-skew-normally distributed errors we have suggested in this paper. A formal comparison between these two different approaches will be taken up as future research.

Conclusions

Mixture models offer an informative and elegant regression approach, allowing assessment of association of a potential risk factor with *both* the probability of being symptom free and the severity of symptoms for a response with clustering at zero. We proposed a probit-log skew normal mixture model for zero-inflated repeated measures data, and demonstrated its potential by analyzing real data from DAVIS. We showed that for this data probit-log skew normal mixture model fits significantly better than the Bernoulli -log normal model proposed in previous references. The probability of a child being free of lower respiratory track symptoms was lower for a week he/she was positive for viral infection relative to a week viral infection was negative. Moreover, the severity of the respiratory symptoms was greater for the week the child was viral positive. The results of our simulation study indicate that our proposed model performs reasonably well even if the underlying distribution of the positive outcome is misspecified.

List of abbreviations

AIC: (Akaike's Information Criterion); BIC: (Bayesian Information Criterion); BVN: (Bivariate Normal Distribution); DAVIS: (Diary of Asthma and Viral Infections Study); GAM: (Generalized Additive Model); GLMM: (Generalized Linear Mixed Models); LRT: (lower respiratory tract); LL: (log-likelihood); PI: (principal investigator); RVI: (respiratory viral infections); ZIB: (Zero Inflated Binomial); ZIP: (Zero Inflated Poisson).

Additional material

Additional file 1 Skew-Normal Distribution. Probability Density Function of Skew-Normal Distribution.

Additional file 2 SAS codes. SAS codes for implementing Maximum Marginal Likelihood Estimation of the Random Effects Probit- Log Skew Normal model and Random Effects Logit- Log Skew Normal model respectively.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SM conducted the analysis and simulation reported in the paper, and wrote the statistical component of the manuscript. WL supervised post doctoral research (reported in the paper) conducted by SM. NWJ provided original data and conducted the study (DAVIS) from which it was derived. He provided input for writing the introduction and methods for conducting DAVIS.
All authors read and approved the final manuscript.

Acknowledgements

Sadia Mahmud conducted the modeling and analysis in course of a post doctoral fellowship at University of Toronto, sponsored by Higher Education Commission of Pakistan.
DAVIS was conducted with support by an unrestricted grant from GlaxoSmith-Kline Canada.

Author Details

[1]Department of Community Health Sciences, The Aga Khan University, Stadium Road, P O Box 3500, Karachi 74800 Pakistan, [2]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada and [3]Firestone Institute for Respiratory Health, St Joseph's Healthcare and McMaster University Department of Medicine, Hamilton, Ontario, Canada

Received: 11 January 2010 Accepted: 14 June 2010
Published: 14 June 2010

References

1. Lambert D: **Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing.** *Technometrics* 1992, **34:**1-14.
2. Hall DB: **Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study.** *Biometrics* 2000, **56:**1030-1039.
3. Yau KKW, Lee AH: **Zero-Inflated Poisson regression with random effects to evaluate an occupational injury prevention programme.** *Statistics in Medicine* 2001, **20:**2907-2920.
4. Hall DB, Berenhaut KS: **Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models.** *The Canadian Journal of Statistics* 2002, **30:**415-430.
5. Min Y, Agresti A: **Random effect models for repeated measures of zero-inflated count data.** *Statistical Modelling* 2005, **5:**1-19.
6. Lee AH, Wang K, Scott JA, Yau KKW, McLachlan GJ: **Multi-level zero-Inflated Poisson regression modelling of correlated count data with excess zeros.** *Statistical Methods in medical Research* 2006, **15:**47-61.
7. Ma R, Hasan MT, Sneddon G: **Modelling heterogeneity in clustered count data with extra zeros using compound Poisson random effect.** *Statistics in Medicine* 2009, **28:**2356-69.
8. Min Y, Agresti A: **Modeling nonnegative data with clumping at zero: A survey.** *JIRSS* 2002, **1:**7-33.
9. Tobin J: **Estimation of relationships for limited dependent variables.** *Econometrica* 1958, **26:**24-36.
10. Cragg JG: **Some statistical models for limited dependent variables with application to the demand for durable goods.** *Econometrica* 1971, **39:**829-844.
11. Duan N, Manning WG Jr, Morris CN, Newhourse JP: **A comparison of alternative models for the demand of medical care.** *Journal of Business and Economic Statistics* 1983, **1:**115-126.
12. Moulton LH, Halsey NA: **A Mixture Model with Detection Limits for Regression Analyses of Antibody Response to Vaccine.** *Biometrics* 1995, **51:**1570-1578.
13. Heckman J: **Shadow prices, market wages, and labor supply.** *Econometrica* 1974, **42:**679-694.
14. Heckman J: **Sample selection bias as a specification error.** *Econometrica* 1979, **47:**153-161.
15. Duan N, Manning WG Jr, Morris CN, Newhourse JP: **Choosing between the sample selection model and the multi-part model.** *Journal of Business and Economic Statistics* 1984, **2:**283-289.
16. Olsen , Schafer : **A two-part random-effects model for semicontinuous longitudinal data.** *Journal of American Statistical Association* 2001, **96:**730-745.
17. Tooze JA, Grunwald GK, Jones RH: **Analysis of repeated measures data with clumping at zero.** *Statistical Methods in medical Research* 2002, **11:**341-355.

18. Li N, Elashoff DA, Robbins WA, Xun L: **A hierarchical zero-inflated log-normal model for skewed responses.** *Statistical Methods in medical Research* 2008, **00:**1-15.
19. Liu L, Ma JZ, Johnson BA: **A multi-level two-part random effects model, with application to an alcohol-dependence study.** *Statistics in Medicine* 2008, **27:**3528-3539.
20. Su L, Tom BDM, Farewell VT: **Bias in 2-part mixed models for longitudinal semicontinuous data.** *Biostatistics* 2009, **10:**374-389.
21. Chai HS, Bailey KR: **Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero.** *Statistics in Medicine* 2008, **27:**3643-55.
22. Tooze JA, Midthune D, Dodd KW, Freedman LS, Krebs-Smith SM, Subar AF, Guenther PM: *Journal of American Dietetic Association* 2006, **106:**1575-87.
23. Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, Subar AF, Tooze JA, Carroll RJ, Freedman LS: **Modeling data with excess zeros and measurement error: Application to Evaluating relationships between episodically consumed foods and health outcomes.** *Biometrics* 2009, **65:**1003-10.
24. Denise JE, Mei HHW: **Two new unconstrained optimization algorithms which use function and gradient values.** *Journal of Optimization Theory and Applications* 1979, **28:**453-482.
25. Gay DM: **Subroutines for unconstrained minimization using a model/trust region approach.** *ACM Transactions on Mathematical Software* 1983, **9:**503-524. ALGORITHM 611
26. Dean EJ: **A model trust-region modification of Newton's method for non-linear two-point boundary value problems.** *Journal of Optimization Theory and Applications* 1992, **75:**297-312.
27. Johnston SL, Pattemore PK, Sanderson G, Smith S, Lampe F, Josephs L, Symington P, O'Toole S, Myint SH, Tyrrell DAJ, Holgate ST: **Community study of role of viral infections in exacerbations of asthma in 9-11 year old children.** *BMJ* 1995, **310:**1225-1229.
28. Rakes GP, Arruda E, Ingram JM, Hoover GE, Zambrano JC, Hayden FG, Platts-Mills TA, Heymann PW: **Rhinovirus and respiratory syncytial virus in wheezing children requiring emergency hospital care.** *Am J Respir Crit Care Med* 1999, **159:**785-790.
29. Heymann PW, Carper HT, Murphy DD, Platts-Mills TA, Patrie J, McLaughlin AP, Erwin EA, Shaker MS, Hellems M, Peerzada J, Hayden FG, Hatley TK, Chamberlain R: **Viral infections in relation to age, atopy and season of admission among children hospitalized for wheezing.** *J Allergy Clin Immunol* 2004, **114:**239-247.
30. Johnston NW, Johnston SL, Dai J, Norman GR, Sears MR: **The September epidemic of asthma exacerbations: School children as disease vectors.** *J Allergy Clin Immunol* 2006, **117:**557-62.
31. Chauhan AJ, Inskip HM, Linaker CH, Smith S, Schreiber J, Johnston SL, Holgate ST: **Personal exposure to nitrogen dioxide (NO_2) and the severity of virus-induced asthma in children.** *Lancet* 2003, **361:**1939-44.
32. Hastie H, Tibshirani R: **Generalized Additive Models.** *Statistical Science* 1986, **1:**297-310.
33. Arrellano-Valle RB, Azzalini A: **The centered parametrization for the multivariate skew-normal distribution.** *Journal of Multivariate Analysis* 2008, **99:**1362-1382.
34. Sahu SK, Dey DK, Branco MD: **A new class of multivariate skew distributions with applications to Bayesian regression models.** *The Canadian Journal of Statistics* 2003, **31:**129-150.
35. Alonso A, Litière S, Molenberghs G: **A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models.** *Computational Statistics and Data Analysis* 2008, **52:**4474-86.
36. Diggle PJ, Liang KY, Zeger SL: *Analysis of Longitudinal Data* New York: Oxford University Press; 1994.
37. Grunwald GK, Jones RH: **Markov models for time series with mixed distribution.** *Environmetrics* 2000, **11:**327-339.

Pre-publication history
The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1471-2288/10/55/prepub

doi: 10.1186/1471-2288-10-55
Cite this article as: Mahmud *et al.*, 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 Medical Research Methodology* 2010, **10**:55