Maternity Length of Stay Modelling by Gamma Mixture Regression with Random Effects

Andy H. Lee[*, 1], **Kui Wang**[1], **Kelvin K. W. Yau**[2], **Geoffrey J. McLachlan**[3], and **Shu Kay Ng**[3]

[1] Department of Epidemiology and Biostatistics, School of Public Health,
Curtin University of Technology, GPO Box U 1987, Perth, WA 6845, Australia
[2] Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon,
Hong Kong
[3] Department of Mathematics and Institute for Molecular Bioscience, University of Queensland,
St. Lucia, QLD 4072, Australia

Received 30 June 2006, revised 22 February 2007, accepted 16 March 2007

Summary

Maternity length of stay (LOS) is an important measure of hospital activity, but its empirical distribution is often positively skewed. A two-component gamma mixture regression model has been proposed to analyze the heterogeneous maternity LOS. The problem is that observations collected from the same hospital are often correlated, which can lead to spurious associations and misleading inferences. To account for the inherent correlation, random effects are incorporated within the linear predictors of the two-component gamma mixture regression model. An EM algorithm is developed for the residual maximum quasi-likelihood estimation of the regression coefficients and variance component parameters. The approach enables the correct identification and assessment of risk factors affecting the short-stay and long-stay patient subgroups. In addition, the predicted random effects can provide information on the inter-hospital variations after adjustment for patient characteristics and health provision factors. A simulation study shows that the estimators obtained via the EM algorithm perform well in all the settings considered. Application to a set of maternity LOS data for women having obstetrical delivery with multiple complicating diagnoses is illustrated.

Key words: Clustered data; EM algorithm; Gamma mixture regression; Length of stay; Random effects.

1 Introduction

Obstetrical delivery is the most frequent cause of hospitalization in many countries. Therefore, comprehensive and accurate information about maternity length of stay (LOS) have important implications in the strategic planning and deployment of financial, human, and physical resources for health care providers and agencies (Solomon, 1996; Lee, Ng and Yau, 2001; Yau, Lee and Ng, 2003). In addition, the determination of patient-related characteristics affecting maternity LOS can help obstetricians optimize care and rationalize their medical practice, as well as potentially improve patient satisfaction and quality of maternal care (Xiao et al., 1997; Xiao, Lee and Vemuri, 1999). But the skewness of the LOS variable poses a problem for modeling and analysis (Marrazi et al., 1998). Instead of assuming a homogeneous distribution, a two-component gamma mixture regression model appears to be a suitable alternative to describe the variations in maternity LOS; the first component corresponding to women with short stays while the second component corresponding to the long-stay sub-population (Lee et al.,

[*] Corresponding author: e-mail: Andy.Lee@curtin.edu.au, Phone: +61 (0) 8 9266 4180, Fax: +61 (0) 8 9266 2958

2001). The significant factors identified may then be compared between the two sub-groups. Outcomes of the mixture analysis may also enable hospitals to justify a disproportionately high burden of long-stay patients (Lee et al., 2002).

A limitation of the gamma mixture regression is that observations collected from the same hospital are often correlated. The dependence of clustered data (patients nested within hospitals) may result in spurious associations and misleading inferences (Leung et al., 1998). There has been little research on hospital discrepancies as a relevant attribute of LOS variations. The aim of this paper is to develop a two-component gamma mixture regression model with random effects that accommodates simultaneously the inherent correlation structure and heterogeneity of LOS outcomes. The approach is based on the generalized linear mixed model formulation (McGilchrist, 1994), whereby random effects are incorporated in the linear predictor of each mixture component. Residual maximum quasi-likelihood estimation of the regression coefficients and the random component parameters is achieved via an iterative EM algorithm (Lee et al., 2005). The predicted random effects from fitting the mixture model also provide information on inter-hospital variations after adjustment for patient characteristics and relevant risk factors (Yau, Wang and Lee, 2003; Ng et al., 2004).

This paper is organized as follows. In Section 2, a two-component gamma mixture regression model with random effects is presented, followed by details of the numerical estimation procedure in Section 3. A simulation study is conducted in Section 4 to assess the performance of the residual maximum quasi-likelihood estimators. In Section 5, an empirical data set on maternity LOS for women admitted for child birth in Western Australia is used to illustrate the practical application of the methodology. Some concluding remarks are then given in Section 6.

2 Two-Component Gamma Mixture Regression Model

Let Y_{ij} $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n_i)$ represents the maternity LOS for the j-th individual in the i-th hospital, where m is the number of hospitals, n_i is the number of patients within hospital i, the total number of observations being $n = \sum_{i=1}^{m} n_i$. The probability density function of Y is assumed to be a two-component gamma mixture:

$$f(y_{ij}) = pf_1(y_{ij}) + (1-p) f_2(y_{ij})$$

where $0 < p < 1$ gives the proportion of patients belonging to the first component or sub-population, and $f_k(y_{ij})$ is the k-th component gamma distribution with mean μ_k and shape parameter υ_k:

$$f_k(y_{ij}) = \frac{1}{y_{ij}\Gamma(\upsilon_k)} \left(\frac{\upsilon_k y_{ij}}{\mu_{k,ij}}\right)^{\upsilon_k} \mathrm{e}^{-\frac{\upsilon_k y_{ij}}{\mu_{k,ij}}} k = 1, 2; \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n_i.$$

A plausible interpretation is thus in terms of its unobserved heterogeneity: a sub-population of usual patients with relatively short LOS, and another sub-population whose members tend to stay longer due to unexpected complications after delivery.

In the manner of Lee et al. (2001), the underlying relationships between maternity LOS and its risk factors are modeled by a gamma mixture regression, where both $\log(\mu_{1,ij})$ and $\log(\mu_{2,ij})$ are assumed to be linear functions of covariates. The covariate vectors x_k appearing in these two parts are not necessarily the same:

$$\log(\mu_{k,ij}) = x_{k,ij}^T \beta_k \quad k = 1, 2; \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n_i.$$

Maximum likelihood estimates of the parameters p, υ_{k} and regression coefficients β_k can be obtained as a solution of the score equation, i.e. setting the derivative of the log-likelihood function for the mixture regression model to zero, which is solved via an EM algorithm alternating expectation and maximization (Lee et al., 2001).

The gamma mixture regression model described above is not suitable for dependent data because of the study design or data collection procedure. To accommodate the inherent dependency between observations clustered within the same hospital, random effects u_k are incorporated into the linear predictors η_k as:

$$\log(\mu_{k,ij}) = \eta_{k,ij} = x_{k,ij}^T\beta_k + u_{ki} \quad k = 1, 2; \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n_i .$$

For simplicity, $u_1 = (u_{11}, \ldots, u_{1m})^T$ and $u_2 = (u_{21}, \ldots, u_{2m})^T$ are assumed to be independent and distributed as $N(0, \sigma_{u_1}^2 I_m)$ and $N(0, \sigma_{u_2}^2 I_m)$ respectively, where I_m denotes an $m \times m$ identity matrix. Here, u_{ki} represents the unobservable random effect of the i-th hospital on the k-th component.

Following the generalized linear mixed model formulation (McGilchrist, 1994), the estimation procedure commences with the penalized log-likelihood and extends to obtain residual maximum quasi-likelihood (REMQL) estimators. The penalized log-likelihood is given by $l = l_1 + l_2$,

$$l_1 = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log(pf_1(y_{ij}) + (1-p)f_2(y_{ij})),$$

$$l_2 = -\frac{1}{2}\left[m\log(2\pi\sigma_{u_1}^2) + \sigma_{u_1}^{-2}u_1^Tu_1 + m\log(2\pi\sigma_{u_2}^2) + \sigma_{u_2}^{-2}u_2^Tu_2\right],$$

with l_1 being the log-likelihood function when the random effects are conditionally fixed, while l_2 can be regarded as the penalty function for the conditional log-likelihood. Estimation may be performed iteratively. In the initial step, coefficients in the linear predictors are estimated, for fixed variance components, by maximizing the above penalized log-likelihood. Estimation of variance component parameters is then achieved using REMQL estimating equations (Lee et al., 2005). Details of the estimation procedure via an EM algorithm are given in the next section.

3 An EM Algorithm for Estimation

To ensure convergence, parameter estimates are obtained via an EM algorithm alternating expectation and maximization. The penalized log-likelihood l is first expressed as a complete-data log-likelihood $l_C = l_\xi + l_{\eta_1} + l_{\eta_2}$, where

$$l_\xi = \sum_{i=1}^{m}\sum_{j=1}^{n_i} (z_{ij}\xi_{ij} - \log[1 + \exp(\xi_{ij})]), \qquad \xi_{ij} = \log(p/(1-p)),$$

$$l_{\eta_1} = \sum_{i=1}^{m}\sum_{j=1}^{n_j} z_{ij}\log f_1(y_{ij}) - \frac{1}{2}\left(m\log(2\pi\sigma_{u_1}^2) + \frac{1}{\sigma_{u_1}^2}\, u_1^Tu_1\right),$$

$$l_{\eta_2} = \sum_{i=1}^{m}\sum_{j=1}^{n_j} (1 - z_{ij})\log f_2(y_{ij}) - \frac{1}{2}\left(m\log(2\pi\sigma_{u_2}^2) + \frac{1}{\sigma_{u_2}^2}\, u_2^Tu_2\right).$$

Here, z_{ij} is an *unobserved* variable indicating whether y_{ij} comes from the first component ($z_{ij} = 1$) or the second component ($z_{ij} = 0$). At each iteration, the EM algorithm proceeds with the E-step, in which z_{ij} is replaced by its conditional expectation $z_{ij}^{(r)}$ under the current parameter estimates $\{\hat{p}^{(r)}, \hat{\beta}_1^{(r)}, \hat{\beta}_2^{(r)}\}$ and $\{\hat{u}_1^{(r)}, \hat{u}_2^{(r)}\}$:

$$z_{ij}^{(r)} = \frac{\hat{p}^{(r)}\hat{f}_1(y_{ij})}{\hat{p}^{(r)}\hat{f}_1(y_{ij}) + (1 - \hat{p}^{(r)})\hat{f}_2(y_{ij})}.$$

At the M-step, updated estimates $\hat{p}^{(r+1)}$, $\{\hat{\beta}_1^{(r+1)}, \hat{u}_1^{(r+1)}\}$ and $\{\hat{\beta}_2^{(r+1)}, \hat{u}_2^{(r+1)}\}$ can be obtained by maximizing l_ξ, l_{η_1}, l_{η_2}separately, in view of the orthogonal partition of l_C into three components. The M-step can be implemented by solving the following three sets of recursive equations based on the derivatives of l_{η_1} and l_{η_2},

$$\hat{p}^{(r+1)} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}^{(r)}/n,$$

$$\begin{bmatrix} \hat{\beta}_k \\ \hat{u}_k \end{bmatrix} = \begin{bmatrix} \beta_{k,0} \\ u_{k,0} \end{bmatrix} + \Im^{-1}_{\beta_k, u_k} \begin{bmatrix} \dfrac{\partial l_{\eta_k}}{\partial \beta_k} \\ \dfrac{\partial l_{\eta_k}}{\partial u_k} \end{bmatrix}, \quad k = 1, 2.$$

Note that $\beta_{k,0}$ and $u_{k,0}$ are initial values of the corresponding quantities, which are then replaced by their updated estimates at each iteration, for given values of $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$. The matrix $\Im_{\beta_k, u_k}$represents the negative second derivative of $l_{\eta_k}$with respect to $\{\beta_k,\ u_k\}$, k = 1, 2, formula of which is given in Appendix A.1. The shape parameter v_k is estimated by maximizing $l_{\eta_k} = \sum_{i=1}^{m}\sum_{j=1}^{n_j} z_{ij} \log f_k(y_{ij})$ using updated estimates at the end of the M-step.

A step-by-step initialization procedure is adopted. Firstly, the two-component gamma mixture model without any covariate is fitted to the data, using a wide range of starting values. The estimates corresponding to the largest likelihood value are subsequently used as initial estimates for fitting the two-component gamma mixture regression model without any random effects. The results so obtained are then set as the initial estimates for the final mixture model with random effects.

At each cycle of the EM algorithm, it is assumed that the variance components $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ are given. In practice, they are unknown and needed to be estimated. When convergence is attained, $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ are estimated by the most updated values $\hat{u}_1$, $\hat{u}_2$ and the corresponding elements of the information matrix. The process continues until all parameter estimates converge. Convergence is determined when the absolute changes of estimates between two iterations are all less than 0.001. For the estimation of variance components, the REMQL approach is adopted to correct the bias due to maximum likelihood estimation (McGilchrist, 1994; Lee et al., 2005). Suppose the observed information matrix $\Im_{p,\beta_1,\beta_2,u_1,u_2}$ is partitioned conformally to $p \mid \beta_1 \mid \beta_2 \mid u_1 \mid u_2$ as

$$\Im^{-1}_{p,\beta_1,\beta_2,u_1,u_2} = [A_{ij}], \quad i,j = 1, \ldots, 5,$$

then estimators of variance components are given by:

$$\hat{\sigma}^2_{u_1} = m^{-1}[\text{trace}\,(A_{44}) + \hat{u}_1^T \hat{u}_1],$$

$$\hat{\sigma}^2_{u_2} = m^{-1}[\text{trace}\,(A_{55}) + \hat{u}_2^T \hat{u}_2].$$

Finally, asymptotic standard errors of the regression coefficients are obtained from the corresponding partition of $\Im^{-1}_{p,\beta_1,\beta_2,u_1,u_2}$, details of the derivatives involved are provided in Appendix A.2.

4 Simulation Study

A simulation study is conducted to assess the performance of the REMQL estimators obtained via the EM algorithm. Data are simulated under a two-component gamma mixture regression model with random effects (Section 2). In each simulated data set, $m = 20$ clusters and $n_i = 30$ observations are fixed within each cluster, which gives 600 observations per simulated data set. The covariate vector x_{ij} consists of the constant 1 (representing the intercept term) and values randomly generated from the uniform (0, 1) distribution, with associated regression coefficients (1.5, -1, 0.5) for the first mixture component and (-2, 0.5, 1) for the second mixture component. The mixing probability p is chosen to be 0.3, 0.5, and 0.8. For each p, values (2, 2), (1, 3) and (2, 3) are considered for the shape parameters (v_1, v_2), whereas moderate variations of 0.3 and 0.5 are assumed for $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$, respectively. According to this simulation design, 20 realizations of the random cluster effects from a normal distribution are generated for each mixture component. These random effects, together with the corresponding fixed effects, are incorporated into the linear predictor of the model, and subsequently $30 \times 20 = 600$ gamma data records are generated for each mixture component. The mixed responses are then randomly selected from the first component or the second component distribution with mixing probability p. The simulation study is thus designed to evaluate the performance of the estimators with respect to a plausible range of shape and mixing probability values. The number of replications is 500 for each setting considered.

Results of the simulation study are presented in Tables 1, 2 and 3, which report the average bias and mean square error (MSE) of the parameter estimates over the 500 replications. It is evident that the REMQL estimators of the regression coefficients and the mixture proportion have negligible biases and relatively small MSE. For variance component parameters, the REMQL estimators of $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ also perform reasonably well in each of the settings considered. The simulation results thus

Table 1 Bias and mean square error (MSE) of REMQL estimators based on 500 replications of the two-component gamma mixture regression model with $m = 20$, $n_i = 30$, variance component parameters $\sigma^2_{u_1} = 0.3$, $\sigma^2_{u_2} = 0.5$, shape parameters v_1 = 2, v_2 = 2, and varying mixing proportions.

	$p = 0.3$		$p = 0.5$		$p = 0.8$	
	Bias	MSE	Bias	MSE	Bias	MSE
p	−0.006	0.001	−0.005	0.007	−0.003	0.001
β_{10}	−0.022	0.035	−0.015	0.024	−0.015	0.021
β_{11}	0.013	0.005	0.007	0.003	0.003	0.001
β_{12}	0.002	0.049	0.000	0.025	0.002	0.015
β_{20}	−0.005	0.032	−0.010	0.034	−0.066	0.081
β_{21}	−0.002	0.002	−0.002	0.003	−0.015	0.013
β_{22}	−0.008	0.018	−0.012	0.031	−0.014	0.139
$\sigma^2_{u_1}$	0.015	0.017	0.002	0.013	−0.004	0.011
$\sigma^2_{u_2}$	−0.010	0.028	−0.010	0.031	0.010	0.053

Table 2 Bias and mean square error (MSE) of REMQL estimators based on 500 replications of the two-component gamma mixture regression model with $m = 20$, $n_i = 30$, variance component parameters $\sigma^2_{u_1} = 0.3$, $\sigma^2_{u_2} = 0.5$, shape parameters $v_1 = 1$, $v_2 = 3$, and varying mixing proportions.

	$p = 0.3$		$p = 0.5$		$p = 0.8$	
	Bias	MSE	Bias	MSE	Bias	MSE
p	−0.005	0.001	−0.004	0.001	−0.007	0.001
β_{10}	−0.057	0.057	−0.040	0.032	−0.005	0.027
β_{11}	0.007	0.010	−0.001	0.005	0.000	0.003
β_{12}	−0.001	0.098	0.018	0.052	−0.013	0.031
β_{20}	−0.021	0.029	−0.018	0.033	−0.071	0.079
β_{21}	−0.005	0.001	−0.003	0.002	−0.006	0.015
β_{22}	−0.005	0.013	−0.005	0.025	−0.009	0.133
$\sigma^2_{u_1}$	0.031	0.026	0.012	0.016	0.009	0.013
$\sigma^2_{u_2}$	−0.003	0.025	−0.013	0.031	0.021	0.046

Table 3 Bias and mean square error (MSE) of REMQL estimators based on 500 replications of the two-component gamma mixture regression model with $m = 20$, $n_i = 30$, variance component parameters $\sigma^2_{u_1} = 0.3$, $\sigma^2_{u_2} = 0.5$, shape parameters $v_1 = 2$, $v_2 = 3$, and varying mixing proportions.

	$p = 0.3$		$p = 0.5$		$p = 0.8$	
	Bias	MSE	Bias	MSE	Bias	MSE
p	−0.004	0.001	−0.005	0.001	−0.003	0.001
β_{10}	−0.028	0.028	−0.033	0.028	−0.006	0.018
β_{11}	0.005	0.005	−0.001	0.003	0.001	0.001
β_{12}	−0.007	0.051	0.023	0.038	0.002	0.014
β_{20}	−0.012	0.031	−0.018	0.033	−0.068	0.057
β_{21}	−0.000	0.001	−0.005	0.002	−0.011	0.008
β_{22}	0.002	0.012	0.004	0.019	0.023	0.086
$\sigma^2_{u_1}$	0.022	0.017	0.003	0.013	0.002	0.011
$\sigma^2_{u_2}$	0.000	0.030	0.010	0.032	0.023	0.045

confirm the applicability of the EM algorithm for parameter estimation in the two-component gamma mixture regression model with random effects.

5 Application to Maternity LOS

5.1 Sample characteristics

Data were collected from $n = 909$ women hospitalized for vaginal delivery with multiple complicating diagnoses in Western Australia. Their maternity LOS ranged between one and 45 days with mean 6.22 days and SD 5.19 days. The empirical LOS distribution exhibits substantial heterogeneity (skewness = 3.44) because of disparities in patient characteristics and possibly hospital care and medical practice received during hospitalization. Figure 1 plots the observed LOS together with the fitted two-component gamma mixture distribution. The proportion of women with relatively long LOS is estimated to be 5.6%.

In addition to maternity LOS, concomitant information was also retrieved from the hospital morbidity database. The variables are age, number of diagnoses, number of procedures, martial status (0 = married, 1 = non-married), admission type (0 = elective, 1 = emergency), indigenous status (0 = non-Aboriginal, 1 = Aboriginal), location (0 = urban, 1 = rural) and payment classification (0 = public, 1 = private). These variables have been found to be associated with maternity LOS (Lee et al., 2001).

For this sample of women admitted to $m = 26$ public hospitals for child birth, their average age was 27.68 years (S.D. = 6.07), 27.3% were non-married, 34% were emergency cases, but only 5.6% had private medical insurance coverage. Aboriginals accounted for 11.8% of the sample, and the majority of women (85%) resided in urban areas. In terms of clinical factors, the average number of diagnoses recorded was 7.67 (SD 3.16), while the average number of surgical or obstetrical procedures performed during hospitalization was 2.87 (SD 1.94).

5.2 Gamma mixture regression analysis

To account for the clustering of observations within hospitals, a two-component gamma mixture regression model with random hospital effects is fitted to the heterogeneous LOS data using the three covariates and five binary variables above. Results are summarized in Table 4. The long-stay subgroup

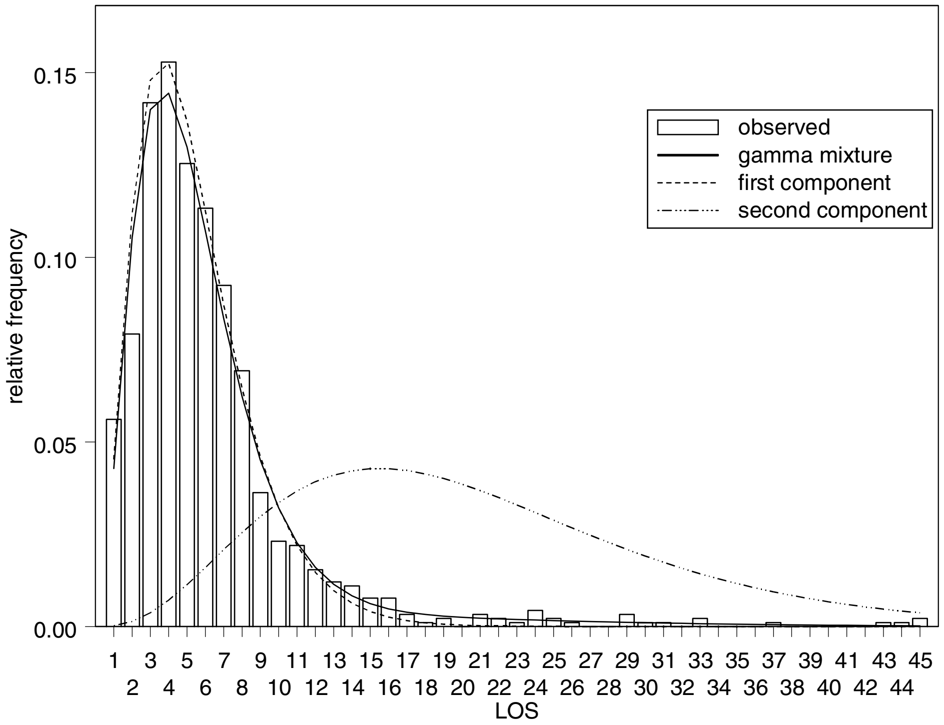

Figure 1 Empirical distribution of maternity LOS and fitted two-component gamma mixture model.

Table 4 Results from fitting the two-component gamma mixture regression with random effects to the maternity LOS data.

Parameter	Reference category	First component estimate (S.E.)	Second component estimate (S.E.)
Intercept		1.035 (0.141)[a]	1.486 (0.390)[a]
Age		0.001 (0.004)	0.005 (0.012)
Number of diagnoses		0.047 (0.007)[a]	0.087 (0.024)[a]
Number of procedures		0.075 (0.013)[a]	−0.023 (0.032)
Marital status: non-married	married	0.023 (0.052)	−0.161 (0.169)
Admission type: emergency	elective	0.132 (0.047)[a]	0.293 (0.148)[a]
Indigenous status: Aboriginal	non-Aboriginal	0.043 (0.072)	0.686 (0.200)[a]
Location: rural	urban	0.050 (0.108)	0.002 (0.212)
Payment classification: private	public	0.227 (0.092)[a]	0.180 (0.274)
v		3.860 (0.369)	2.968 (0.613)
Variance component σ		0.163	0.110
p		0.857 (0.033)	

[a] P-value < 0.05.

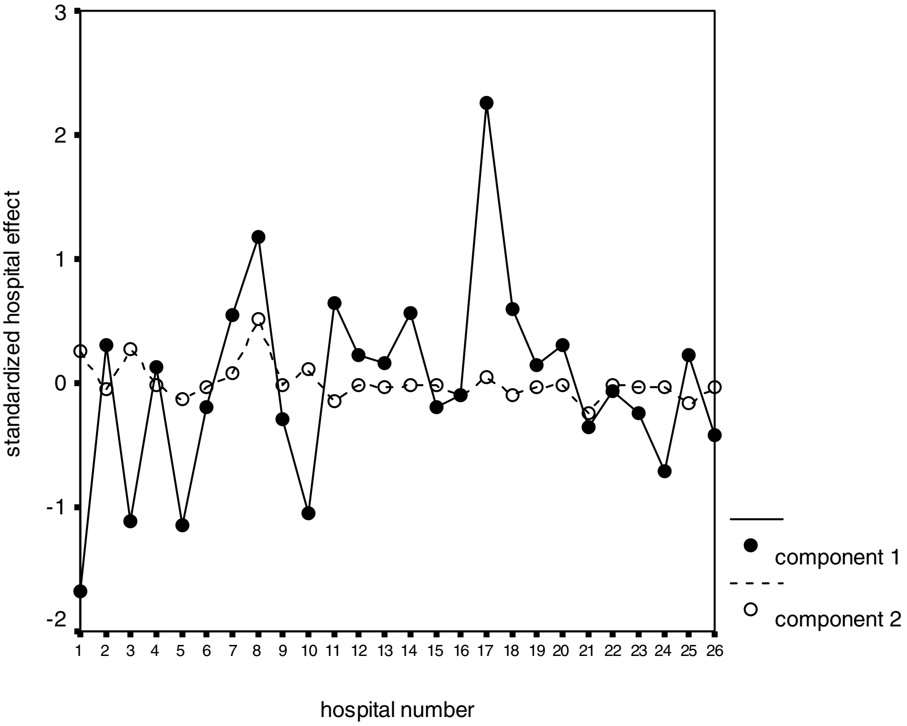

Figure 2 Estimated hospital effects from fitting the two-component gamma mixture regression with random effects to the maternity LOS data.

accounts for an estimated $1-\hat{p}=14.3\%$ of the hospitalizations, which is substantially different from the initial estimate of 5.6% derived by fitting the gamma mixture distribution alone. It reflects the importance of using concomitant information (both patient and health provision factors) and random hospital effects in characterizing LOS. The maternity LOS is positively associated with the number of diagnoses and admission type, which reflect the complications and co-morbidities that the patient experienced during hospitalization. A greater number of diagnoses and/or emergency admission would indicate a more severe condition warranting a longer stay in hospital. As expected, Aboriginal women incur a much prolonged stay relative to their non-Aboriginal counterparts, while the effects of private insurance coverage and number of procedures are only prominent for short-stay cases.

The random effects are used to capture discrepancies in clinical expertise, obstetrical care, and other unmeasurable characteristics of the hospitals. Standardized random effect predictions are plotted in Figure 2. For the first component, hospital number 17 exhibits large positive effect on prolonging hospitalization, whereas hospital number 1 is relatively more efficient in terms of maternity LOS. An inspection of the data reveals that hospital 1 corresponds to a large teaching hospital in the Perth metropolitan area while hospital 17 is a rural regional hospital. For the second component, the performances of hospitals are quite similar and no outstanding random effect predictions are found. Overall, considerable hospital variation in maternity LOS is evident from the two variance component estimates in Table 4, confirming the need for random effects adjustment in the gamma mixture regression model.

Finally, the Pearson statistic of 923.38 (881 degrees of freedom, P-value $=0.156$) provides no evidence of lack of fit for the two-component gamma mixture model. The adequacy of the specified model is also assessed by the Pearson residuals, but no obvious pattern in the residual plot is observed.

Table 5 Results from fitting the one-component gamma and log-normal mixed regression models to the maternity LOS data.

Parameter	Reference category	One-component gamma estimate (S.E.)	One-component log-normal estimate (S.E.)
Intercept		1.127 (0.153)[a]	0.913 (0.140)[a]
Age		0.002 (0.004)	0.003 (0.004)
Number of diagnoses		0.056 (0.009)[a]	0.050 (0.008)[a]
Number of procedures		0.047 (0.013)[a]	0.072 (0.012)[a]
Marital status: non-married	married	−0.058 (0.058)	−0.017 (0.051)
Admission type: emergency	elective	0.221 (0.054)[a]	0.152 (0.048)[a]
Indigenous status: Aboriginal	non-Aboriginal	0.218 (0.081)[a]	0.132 (0.071)
Location: rural	urban	0.046 (0.108)	0.032 (0.108)
Payment classification: private	public	0.211 (0.109)	0.189 (0.095)[a]
Scale parameter		$\hat{v} = 1.864$	$\hat{\phi} = 0.412$
Variance component σ		0.139	0.163

[a] P-value < 0.05.

5.3 Other model fits

We have also fitted one-component random effects models with a heavy tailed distribution to the maternity LOS data. Table 5 presents the results from fitting gamma and log-normal mixed regression models using SAS PROC GLIMMIX (Littell et al., 1996). We found that the regression coefficients are quite different, and different sets of significant risk factors are identified. Therefore, the same conclusions cannot be reached by fitting these simpler models. On the other hand, the effects of significant factors are component-specific in the proposed gamma mixture model.

To further justify the proposed gamma mixture model, an alternative two-component Gaussian mixture regression model with random effects is fitted to the log-transformed LOS data, results of which

Table 6 Results from fitting the two-component Gaussian mixture regression with random effects to the maternity LOS data.

Parameter	Reference category	First component estimate (S.E.)	Second component estimate (S.E.)
Intercept		0.744 (0.253)[a]	1.158 (0.173)[a]
Age		0.005 (0.007)	0.0002 (0.005)
Number of diagnoses		0.042 (0.014)[a]	0.060 (0.008)[a]
Number of procedures		0.104 (0.022)[a]	0.024 (0.013)
Marital status: non-married	married	−0.112 (0.094)	0.100 (0.055)
Admission type: emergency	elective	0.105 (0.088)	0.228 (0.064)[a]
Indigenous status: Aboriginal	non-Aboriginal	0.277 (0.134)[a]	−0.069 (0.129)
Location: rural	urban	−0.110 (0.182)	0.146 (0.125)
Payment classification: private	public	0.148 (0.170)	0.249 (0.107)[a]
Scale parameter ϕ		0.617	0.084
Variance component σ		0.053	0.043
p		0.553 (0.032)	

[a] P-value < 0.05.

are presented in Table 6. The set of significant covariates under the Gaussian mixture model is quite different from that under the gamma mixture model. It should be remarked that the estimated proportion of 44.7% for the long-stay subgroup does not appear to be reasonable when compared with the empirical LOS distribution in Figure 1. The proposed gamma mixture model is preferable in view of the clinical interpretations for its significant risk factors and the consistency with literature findings, as explained in the next section below.

6 Discussion

This study demonstrates the application of two-component gamma mixture regression modelling to maternity LOS. It highlights significant heterogeneity in the duration women are hospitalized for child birth. The method avoids arbitrary trimming and transformation of the data for a single component analysis (Quantin et al., 1999), that is, the assumption of a homogeneous patient population is no longer required. The two components have clinical interpretation in representing the short stay and long stay subgroups for women having obstetrical delivery with multiple complicating diagnoses. Moreover, the method extends the gamma mixture regression model to simultaneously account for the inherent correlation of patients clustered within hospitals.

For the empirical application, the long LOS in the second component corresponds to extremely complex cases being admitted into the public hospitals (Lee et al., 2001; Lee et al., 2002). Furthermore, the sets of significant factors affecting maternity LOS appear to be different between the short and long-stay subgroups. Although the observed heterogeneity is primarily linked to the number of diagnoses and admission type, we found little statistical association between maternity LOS and the age, marital status and location of the patient. Aboriginal mothers tend to have additional co-morbidities in conjunction with a delayed admission due to transportation difficulties, which in turn may lead to further complications and consequently a late discharge. The findings are consistent with the literature (Lee et al., 2001; Lee et al., 2002). However, it is interesting to note that the significant impact of private medical insurance and of number of procedures performed is limited to women with relatively short stays.

The management of maternity LOS has become an important issue for acute care facilities. Determination of relevant factors associated with the two components based on the gamma mixture regression model would enable an efficient throughput of patients, so that beds are available for both elective and emergency admissions. Development of an appropriate mathematical model for LOS can provide an evidence-based approach to guide discharge planning. Although the current study addresses LOS for obstetrical delivery, the proposed method is applicable to paediatric or other types of inpatient LOS. Similarly, the methodology can be applied to other units of clustering such as health regions or local districts within the state. The model can also be modified to analyze longitudinal data, where repeated episodes of some events such as discharge and readmissions for each patient are monitored, so that the response variable of interest represents recurrent times or repeated LOS.

The model provides information on inter-hospital variation and estimates on (random) hospital effects via an EM algorithm. Consequently, hospital performance may be compared after adjustment for patient characteristics and health provision factors. Finally, the two-component gamma mixture regression model can be generalized to a *K*-component mixture model, with the component density from the exponential family, hence providing a general framework for the development of finite mixture generalized linear mixed models. This will involve separate weighted maximizations in the M-step of the EM algorithm, with corresponding random effects to accommodate the dependent data.

Instead of adopting the EM algorithm in Section 3, PROC NLMIXED in SAS version 9.1 can be used to fit the gamma mixture regression model with random effects by specifying the MODEL and RANDOM statements, as follows:

```
MODEL y ~ general(ll);
RANDOM u1 u2 ~ normal ([0,0], [v11,0,v22]) subject=hospital;
```

Table 7 Results from fitting the two-component gamma mixture regression with random effects to the maternity LOS data using SAS PROC NLMIXED.

Parameter	Reference category	First component estimate (S.E.)	Second component estimate (S.E.)
Intercept		0.922 (0.133)[a]	1.488 (0.416)[a]
Age		0.003 (0.004)	0.003 (0.013)
Number of diagnoses		0.038 (0.007)[a]	0.083 (0.026)[a]
Number of procedures		0.083 (0.014)[a]	−0.019 (0.038)
Marital status: non-married	married	0.034 (0.055)	−0.188 (0.185)
Admission type: emergency	elective	0.121 (0.050)[a]	0.370 (0.175)[a]
Indigenous status: Aboriginal	non-Aboriginal	0.023 (0.078)	0.704 (0.204)[a]
Location: rural	urban	0.131 (0.068)	−0.003 (0.214)
Payment classification: private	public	0.177 (0.094)	0.202 (0.288)
v		3.723 (0.231)	2.987 (1.172)
Variance component σ		0.001	0.223
p		0.844 (0.045)	

[a] P-value < 0.05.

where general(ll) specifies the conditional log-likelihood for a two-component gamma mixture regression model.

The instructions to implement model fitting with covariates and random effects in SAS are given in Appendix A.3. However, it should be pointed out that REMQL estimation is not achievable in PROC NLMIXED. According to the SAS manual's section on *PROC NLMIXED Compared with Other SAS Procedures and Macros* (SAS Institute Inc., 2004), PROC NLMIXED can only perform maximum likelihood (ML) estimation because the analog to REMQL involves a high dimensional integral over all of the fixed-effect parameters, and this integral is typically not available in closed form. Consequently, regardless of the optimization algorithm and numerical approximation being implemented within PROC NLMIXED, even though the same set of initial values and stopping criteria is used, the parameter estimates from SAS will be different from those obtained under the proposed REMQL estimation method.

For illustration purpose, the two-component gamma mixture regression model with random effects is fitted to the maternity LOS data of Section 5 using PROC NLMIXED. By default, the Adpative Gaussian Quadrature method and Dual Quasi-Newton optimization technique are used. The initial values of the parameters are taken to be those estimates obtained from the REMQL fit. Table 7 shows that the resulting ML estimates are comparable to the corresponding REMQL estimates, apart from the small variance component estimate for σ_{u_1}. If initial values are not specified by the PARMS statement (Appendix A.3), then the parameters are assigned an initial value of 1. Although the specification of initial values is not required in PROC NLMIXED, the specification of accurate initial values is always useful to avoid the possible divergence problem in estimation.

Appendix

A.1 Derivatives for the M-step of the EM algorithm

Let η_k denote the vector $\eta_{k,ij}$, then the linear predictors of the two-component gamma mixture regression model with random effects can be written as

$$\eta_k = X_k\beta_k + Ru_k, \quad k = 1, 2,$$

where X_k and R are the design matrices for β_k and u_k ($k = 1, 2$), respectively. The following expressions are required for the M-step of the EM algorithm.

$$\mathfrak{I}_{\beta_k, u_k} = \begin{bmatrix} X_k^T \\ R^T \end{bmatrix} \left(-\frac{\partial^2 l_{\eta_k}}{\partial \eta_k \, \partial \eta_k^T} \right) [X_k \quad R] + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{u_k}^{-2} I_m \end{bmatrix}, \quad k = 1, 2$$

$$\frac{\partial l_{\eta_k}}{\partial \beta_k} = X_k^T \, \frac{\partial l_{\eta_k}}{\partial \eta_k}, \qquad \frac{\partial l_{\eta_k}}{\partial u_k} = R^T \, \frac{\partial l_{\eta_k}}{\partial \eta_k} - \sigma_{u_k}^{-2} u_k,$$

$$\frac{\partial l_{\eta_1}}{\partial \eta_{1,ij}} = z_{ij} v_1 (y_{ij}/\exp(\eta_{1,ij}) - 1),$$

$$\frac{\partial l_{\eta_2}}{\partial \eta_{2,ij}} = (1 - z_{ij}) \, v_2 (y_{ij}/\exp(\eta_{2,ij}) - 1).$$

Let y and z denote the corresponding vectors of y_{ij} and z_{ij}, we then have

$$\frac{\partial l_{\eta_1}}{\partial \eta_1} = z v_1 (y/\exp(\eta_1) - 1),$$

$$\frac{\partial l_{\eta_2}}{\partial \eta_2} = (1 - z) \, v_2 (y/\exp(\eta_2) - 1).$$

Hence,

$$-\frac{\partial^2 l_{\eta_1}}{\partial \eta_1 \, \partial \eta_1^T} = \text{Diag}\,[z v_1 y/\exp(\eta_1)],$$

$$-\frac{\partial^2 l_{\eta_2}}{\partial \eta_2 \, \partial \eta_2^T} = \text{Diag}\,[(1 - z) \, v_2 y/\exp(\eta_2)].$$

A.2 Information matrix and asymptotic standard errors

The first-order derivatives of the penalized log-likelihood $l = l_1 + l_2$ are:

$$\frac{\partial l}{\partial \beta_k} = X_k^T \, \frac{\partial l_k}{\partial \eta_k}, \qquad \frac{\partial l}{\partial u_k} = R^T \, \frac{\partial l_k}{\partial \eta_k} - \sigma_{u_k}^{-2} u_k, \quad k = 1, 2.$$

The observed information matrix is given by:

$$\mathfrak{I}_{p, \beta_1, \beta_2, u_1, u_2} = H + \text{blockdiag}\,[0,\ 0,\ 0,\ \sigma_{u_1}^{-2} I_m,\ \sigma_{u_2}^{-2} I_m],$$

$$H = -Q^T \begin{pmatrix} \dfrac{\partial^2 l_1}{\partial p^2} & \dfrac{\partial^2 l_1}{\partial p \, \partial \eta_1^T} & \dfrac{\partial^2 l_1}{\partial p \, \partial \eta_2^T} \\ \dfrac{\partial^2 l_1}{\partial \eta_1 \, \partial p^T} & \dfrac{\partial^2 l_1}{\partial \eta_1 \, \partial \eta_1^T} & \dfrac{\partial^2 l_1}{\partial \eta_1 \, \partial \eta_2^T} \\ \dfrac{\partial^2 l_1}{\partial \eta_2 \, \partial p^T} & \dfrac{\partial^2 l_1}{\partial \eta_2 \, \partial \eta_1^T} & \dfrac{\partial^2 l_1}{\partial \eta_2 \, \partial \eta_2^T} \end{pmatrix} Q,$$

and

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & X_1 & 0 & R & 0 \\ 0 & 0 & X_2 & 0 & R \end{pmatrix}.$$

The inverse of $\Im_{p,\beta_1,\beta_2,u_1,u_2}$ provides the asymptotic standard errors of the parameters. To obtain the derivatives, we first rewrite the k-th component density function in vector form as:

$$f_k(y) = \frac{1}{y\Gamma(v_k)} \left(\frac{v_k y}{\exp\,(\eta_k)}\right)^{v_k} \exp\left(-\frac{v_k y}{\exp\,(\eta_k)}\right), \quad k = 1, 2.$$

Then

$$f'_k(y) = f_k(y)\, v_k (y/e^{\eta_k} - 1),$$

$$f''_k(y) = f_k(y)\, [(v_k y/e^{\eta_k} - v_k)^2 - v_k y/e^{\eta_k}].$$

Note that

$$l_1 = \sum_y (\log\,(p f_1(y) + (1-p)\, f_2(y))).$$

The first- and second-order derivatives of l_1 are as follows:

$$\frac{\partial l_1}{\partial p} = \sum \frac{f_1(y) - f_2(y)}{p f_1(y) + (1-p)\, f_2(y)},$$

$$\frac{\partial l_1}{\partial \eta_1} = \frac{p f'_1(y)}{p f_1(y) + (1-p)\, f_2(y)},$$

$$\frac{\partial l_1}{\partial \eta_2} = \frac{(1-p)\, f'_2(y)}{p f_1(y) + (1-p)\, f_2(y)},$$

$$\frac{\partial^2 l_1}{\partial p^2} = -\sum \frac{[f_1(y) - f_2(y)]^2}{[p f_1(y) + (1-p)\, f_2(y)]^2},$$

$$\frac{\partial^2 l_1}{\partial p\, \partial \eta_1^T} = \left(\frac{f'_1(y)\, f_2(y)}{[p f_1(y) + (1-p)\, f_2(y)]^2}\right)^T,$$

$$\frac{\partial^2 l_1}{\partial p\, \partial \eta_2^T} = \left(-\frac{f_1(y)\, f'_2(y)}{[p f_1(y) + (1-p)\, f_2(y)]^2}\right)^T,$$

$$\frac{\partial^2 l_1}{\partial \eta_1\, \partial \eta_1^T} = \text{Diag}\left(\frac{p f''_1(y)\, [p f_1(y) + (1-p)\, f_2(y)] - [p f'_1(y)]^2}{[p f_1(y) + (1-p)\, f_2(y)]^2}\right),$$

$$\frac{\partial^2 l_1}{\partial \eta_1\, \partial \eta_2^T} = \text{Diag}\left(-\frac{p(1-p)\, f'_1(y)\, f'_2(y)}{[p f_1(y) + (1-p)\, f_2(y)]^2}\right),$$

$$\frac{\partial^2 l_1}{\partial \eta_2\, \partial \eta_2^T} = \text{Diag}\left(\frac{(1-p)\, f''_2(y)\, [p f_1(y) + (1-p)\, f_2(y)] - [(1-p)\, f'_2(y)]^2}{[p f_1(y) + (1-p)\, f_2(y)]^2}\right).$$

A.3 SAS model fitting instructions

Without loss of generality, three covariates and two random effects terms are considered when invoking PROC NLMIXED. Initial values of the parameters are specified by using the PARMS statement.

```
data los;
input y x1 x2 x3 hospital;
datalines;
.
.
run;
```

```
proc nlmixed data=los;
      bounds v11 > 0, v22 > 0, nu1 > 0, nu2 > 0;
      parms v11=0.0266 v22=0.0121 nu1=3.86 nu2=2.968 xi=1.79 b10=..........;
            mu1 = exp(b10 + b11*x1 + b12*x2 + b13*x3 + u1);
            mu2 = exp(b20 + b21*x1 + b22*x2 + b23*x3 + u2);
            ka1 = (nu1*y) / mu1;
            ka2 = (nu2*y) / mu2;
            f1 = ((ka1**nu1)*exp(-ka1)) / (y*gamma(nu1));
            f2 = ((ka2**nu2)*exp(-ka2)) / (y*gamma(nu2));
            p = exp(xi) / (1+exp(xi));
            ll = log(p*f1 + (1-p)*f2);
      model y ~ general(ll);
      random u1 u2 ~ normal([0,0],[v11,0,v22]) subject=hospital;
run;
```

Specifications:

Dependent Variable:	y
Covariates:	x1, x2, x3
Distribution for dependent variable:	2-component gamma mixture regression model
Random Effects:	u1, u2
Distributions for Random Effects	Normal, Normal
Subject Variable:	Hospital

Parameters:

Model Parameter	*SAS PROC NLMIXED*
μ_1, μ_2	mu1, mu2
v_1, v_2	nu1, nu2
β_1, β_2	b10 b11 b12 b13, b20 b21 b22 b23
$\sigma^2_{u_1}$, $\sigma^2_{u_2}$	v11, v22
p	exp (xi)/(1 + exp (xi))

Acknowledgements The authors would like to thank the associate editor and two anonymous referees for helpful comments. This research was supported by grants from the National Health and Medical Research Council of Australia, Australian Research Council, and the Research Grants Council of Hong Kong. The authors are grateful to the Health Information Centre, Department of Health Western Australia, for provision of the maternity length of stay data set. The model fitting procedure (REMQL estimation) via the EM algorithm has been implemented as a SPlus computer program available from http://fbstaff.cityu.edu.hk/mskyau/.

References

Quantin, C., Sauleau, E., Bolard, P., Mousson, C., Kerkri, M., Brinet Lecomte, P., Moreau, T., and Dusserie, L. (1999). Modeling of high-cost patient distribution within renal failure diagnosis related group. *Journal of Clinical Epidemiology* **52**, 251–258.

Lee, A. H., Ng, A. S. K., and Yau, K. K. W. (2001). Determinants of maternity length of stay: A gamma mixture risk-adjusted model. *Health Care Management Science* **4**, 249–255.

Lee, A. H., Xiao, J., Codde, J. P., and Ng, A. S. K. (2002). Public versus private hospital maternity length of stay: A gamma mixture modeling approach. *Health Service Management Research* **15**, 46–54.

Lee, A. H., Wang, K., Yau, K. K. W., Carrivick, P. J. W., and Stevenson, M. R. (2005). Modelling bivariate count series with excess zeros. *Mathematical Biosciences* **196**, 226–237.

Leung, K. M., Elashoff, R. M., Rees, K. S., Hasan, M. M., and Legorreta, A. P. (1998). Hospital- and patient-related characteristics determining maternity length of stay: a hierarchical linear model approach. *American Journal of Public Health* **88**, 377–381.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models*. SAS Institute Inc., Cary, NC, USA.

Marazzi, A., Paccaud, F., Ruffieux, C., and Beguin, C. (1998). Fitting the distributions of length of stay by parametric models. *Medical Care* **36**, 915–927.

McGilchrist, C. A. (1994). Estimation in generalised mixed models. *Journal of Royal Statistical Society B* **56**, 61–69.

Ng, S. K., McLachlan, G. J., Yau, K. K. W., and Lee, A. H. (2004). Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. *Statistics in Medicine* **23**, 2729–2744.

SAS Institute Inc. (2004). *SAS 9.1.3 Help and Documentation*. SAS Institute Inc., Cary, NC, USA.

Solomon, G. L. (1996). Length of the hospital stay for mothers and newborns. *New England Journal of Medicine* **334**, 1134.

Xiao, J., Douglas, D., Lee, A. H., and Vemuri, S. R. (1997). A Delphi evaluation of the factors influencing length of stay in Australian hospitals. *International Journal of Health Planning and Management* **12**, 207–218.

Xiao, J., Lee, A. H., and Vemuri, S. R. (1999). Mixture distribution analysis of length of hospital stay for efficient funding. *Socio-Economic Planning Sciences* **33**, 39–59.

Yau, K. K. W., Lee, A. H., and Ng, A. S. K. (2003). Finite mixture regression model with random effects: Application to neonatal hospital length of stay. *Computational Statistics & Data Analysis* **41**, 359–366.

Yau, K. K. W., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal* **45**, 437–452.