# Bayesian Latent Class Model for Predicting Gestational Age at Birth Using an Administrative Database

Shuang Wang[1], Manushi Welandawe[2], Jacob Strock[3], Khurshid Jahan[4], Gavino Puggioni[2], Xuerong Wen[1]

[1]University of Rhode Island College of Pharmacy, RI; [2]University of Rhode Island Department of Computer Science and Statistics;
[3]University of Rhode Island Department of Oceanography; [4]University of Rhode Island Department of Geosciences.

## Background

- Administrative Databases play an important role in research on medications or other exposures in pregnancy. Its usefulness may be limited by lacking an accurate estimate of gestational age at birth (GAB).
- As an important index for neonatal health care and the onset of pregnancy, GAB is typically assessed by ultrasound and reported in the Birth Certificate.
- In the absence of clinically reported GAB, an accurate and robust method for GAB estimation is needed.
- A number of algorithms have been proposed to estimate GAB using an administrative database or Electronic Health Records [1]-[8]. However, few allows for modeling heterogeneity in the population.

## Objective

To develop a Bayesian Latent Class Model (LCM) to predict gestational age at birth and to identify heterogeneity in the study population.

## Methods

- **Data source & study population**
  - We utilized data from Medicaid linked with birth certificate from Department of Health, RI.
  - The study cohort included all mother-child pairs with delivery date between January 2008 and December 2013. Pregnancies with invalid gestational age (missing gestational age, shorter than 20 weeks or longer than 44 weeks) were excluded.
  - The model is first estimated using 60% of the observations and then validated over the remaining 40%.

- **Proposed independent LCM: a mixture of multivariate Gaussian distributions**

  Let $y_i$ denote the gestational age of subject $i = 1, \ldots, n$. The LCM is given by,

$$f(y_i \mid \omega_k, \beta_k, \sigma_k^2; x_i) = \sum_{k=1}^{K} \omega_k N(y_i; x_i\beta_k, \sigma_k^2)$$

  where $K$ is the number of mixture components. The parameter $\omega_1, \ldots, \omega_k$ denote the probability of a categorical latent variable $C_k$ taking value $k(k = 1, \ldots, K)$.

- **Prior specification**

$$\beta_k \sim \mathcal{N}_p(0, \sigma_\beta^2 I_p)$$
$$\omega_k \sim \mathrm{Dir}(1/K, .., 1/K), for k = 1, ..., K$$
$$\sigma_k^2 \sim \mathcal{IG}(a, b)$$

- **Posterior computation**

$$\omega_1, \ldots, \omega_k \mid \cdot \sim \mathcal{D}ir(1/K + \sum_{j=1}^{n_1} 1(C_j = 1), \ldots, 1/K + \sum_{j=1}^{n_k} 1(C_j = K))$$
$$C_1, \ldots, C_k \mid \cdot \sim \mathcal{D}isc(\omega_1, \ldots, \omega_k)$$
$$\beta_k \mid \cdot \propto P(\mathcal{Y}_k \mid C_k, \beta_k, \sigma_k^2; x_k)\pi(\beta_k)$$
$$\sigma_k^2 \mid \cdot \propto P(\mathcal{Y}_k \mid C_k, \beta_k, \sigma_k^2; x_k)\pi(\sigma_k^2)$$

- **MCMC sampling algorithms**

  For independent mixture, we used Gibbs sampling to draw samples as closed forms are available for parameter $\beta_k, \sigma_k^2$, and $\omega_k$. The sampling steps are as follow:

  1. Assign initial values to $\beta_k, \sigma_k^2$, and $\omega_k$ for $k = 1, \ldots, K$.
  2. For $k = 1, \ldots, K$, update $\omega_k$ from a full-conditional distribution by using Gibbs sampling;
  3. Sample the latent class variable $C_k$ from a categorical distribution with probability vector $\omega_1, \ldots, \omega_k$;
  4. Given $C_k = k$, update $\beta_k$ from a full-conditional distribution by using Gibbs sampling;
  5. Given $C_k = k$, update $\sigma_k^2$ from a full-conditional distribution by using Gibbs sampling;

## Results

**Table 1. Population Characteristics, RI, U.S. 2008 - 2013**

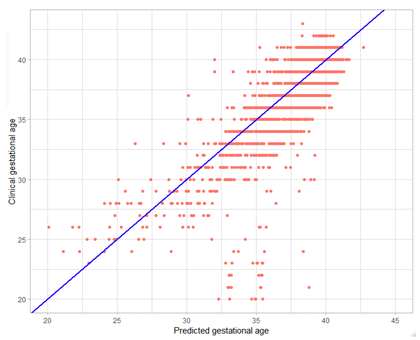
| | Mother-child Pairs for modelling (N = 28,636) Mean (SD) or N (%) |
|---|---|
| **Demographic** | |
| Age at delivery (years) | 25.7 (5.85) |
| Clinical gestational age at birth (weeks) | 38.6 (2.22) |
| Infant weight at birth (gram) | 3258.0 (598.07) |
| **Maternal and pregnancy characteristics** | |
| Multiple gestations | 190 (0.66) |
| Early or threatened labor | 1099 (3.84) |
| Disorders relating to short gestation and low birth weight | 225 (0.79) |
| Preexisting or gestational diabetes | 3791 (13.24) |
| Preexisting hypertensive disease/ gestational hypertension | 3760 (13.13) |
| Preeclampsia or eclampsia | 1689 (5.9) |
| Oligohydramnios | 1732 (6.05) |
| Polyhydramnios | 490 (1.71) |
| Placental abruption | 1439 (5.03) |
| Vaginal bleeding | 504 (1.76) |
| Anemia | 5345 (18.67) |
| Depression | 3973 (13.87) |
| Anxiety | 3810 (13.3) |
| Psychotic disorders | 3216 (11.23) |
| Obesity | 2430 (8.49) |
| **Procedures related to delivery** | |
| Cesarean delivery | 5012 (17.5) |
| Subsequent hospital care 1 | 4286 (14.97) |
| Subsequent hospital care 2 | 759 (2.65) |
| Subsequent hospital care 3 | 242 (0.85) |
| Admission to ICD | 2455 (8.57) |

Table 2. DIC and WAIC statistics

| Classes | Model description | pDIC | DIC | WAIC |
|---|---|---|---|---|
| 2 | Independent mixture | 64524.53 | 120786.60 | 56318.66 |
| 3 | | 62719.05 | 116656.30 | 54011.02 |
| 4 | | 62890.04 | 117064.20 | 54268.55 |
| 2 | Dependent mixture | 68066.41 | 128374.70 | 60342.82 |
| 3 | | 62507.07 | 116362.00 | 53905.92 |
| 4 | | 62206.51 | 115610.40 | 53475.88 |

### • 3-Class Independent Mixture

**Figure 1. Clinical gestational age against predicted gestational age**




### • 3-Class Independent Mixture Cont'd

**Figure 2. Histogram of clinical gestational age against posterior predictive distribution**

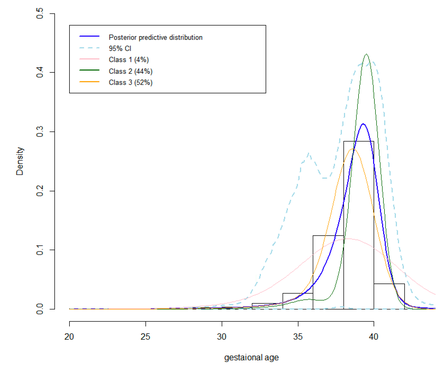



**Figure 3. Posterior density function of the regression coefficients**

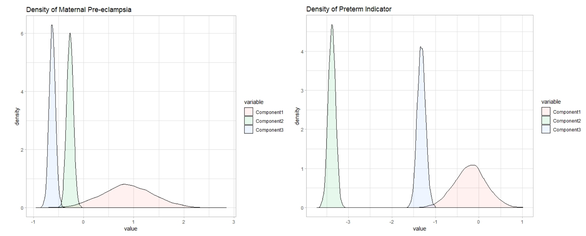
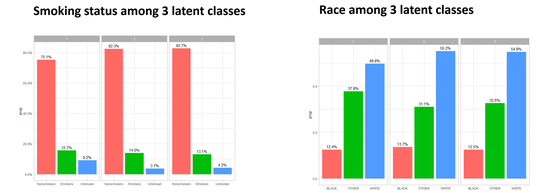



**Figure 4. Comparison of demographic characteristics among latent classes**




## Ongoing Work – Dependent Mixture

- **Proposed dependent LCM:**

$$f(y_i \mid \gamma_k, \beta_k, \sigma_k^2; x_i) = \sum_{k=1}^{K} \pi_k N(y_i; x_i\beta_k, \sigma_k^2)$$

  where $K$ is the number of mixture components.The parameter $\pi_1, \ldots, \pi_k$ denote the probability of a categorical latent variable $C_i$ taking value $k(k = 1, \ldots, K)$ if subject $i$ belongs to class $k$. To link the probability of latent class membership to individual-level characteristics, we assume the latent class variable $C_i$ has a categorical distribution with probability $\pi_{ik}$, where $\pi_{ik}$ incorporates a $r \times 1$ vector of individual covariates $z_i$ as follow:

$$\pi_{ik} = \frac{e^{z_i^T\gamma_k}}{\sum_{h=1}^{K} e^{z_i^T\gamma_h}}, \text{ with } \gamma_1 = 0$$

- **Posterior computation**

$$\gamma_k \mid \cdot \propto \prod_{i=1}^{N} [P(C_i = K \mid \gamma_k; z_i)]^{I(C_i=k)}\pi(\gamma_k)$$
$$C_i \mid \cdot \sim \mathcal{M}ultinom(\rho_{ik})$$
$$\beta_k \mid \cdot \propto P(\mathcal{Y}_k \mid C_i = k, \beta_k, \sigma_k^2; x_k)\pi(\beta_k)$$
$$\sigma_k^2 \mid \cdot \propto P(\mathcal{Y}_k \mid C_i = k, \beta_k, \sigma_k^2; x_k)\pi(\sigma_k^2)$$

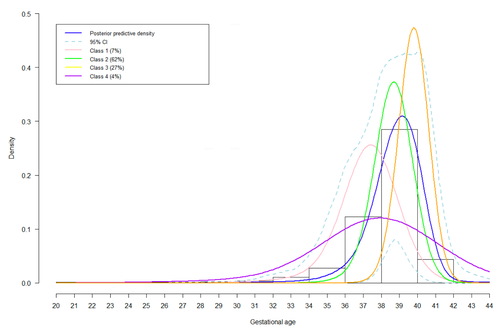
  $\rho_{ik}$ is the posterior probability of latent class variable $C_i$ that subject $i$ belongs to class $k$

$$\rho_{ik} = \frac{\pi_{ik}(\gamma_k)\mathcal{N}(y_i; x_i\beta_k, \sigma_k^2)}{\sum_{h=1}^{K} \pi_{ih}(\gamma_h)\mathcal{N}(y_i; x_i\beta_h, \sigma_h^2)}$$

- **MCMC sampling algorithms**

  1. Update $\gamma_2, \ldots, \gamma_k$ by using random walk Metropolis algorithm;
  2. Sample the latent class variable $C_i$ from a categorical distribution with probability vector of $\rho_{ik}$ for each subject;
  3. Given $C_i = k$, update $\beta_k$ by using Gibbs sampler;
  4. Given $C_i = k$, update $\sigma_k^2$ by using Gibbs sampler.

**Figure 5. Histogram of clinical gestational age against posterior predictive distribution using dependent mixture**




## Conclusions

- To our knowledge, this study is among the first to predict GAB using Bayesian LCM in an administrative database. We developed a flexible LCM to model gestational age trajectories.
- Within each latent class, we fit a multivariate regression model. Covariate coefficients are varied across latent classes suggesting that the risk factors of premature birth may have different impacts on gestational age in heterogeneous subgroups.
- As part of an ongoing work, we allow individual covariates to be related to the probability of latent class memberships. By doing so, subjects can be clustered into clinically relevant classes. We are still improving the MCMC chain to achieve better efficiency and mixing.
- An alternative approach to modeling latent classes is to adopt an infinite mixture by a Dirichlet process (DP) prior on the number of classes [9]. The DP prior shows the superiority by avoiding underfitting or overfitting the number of classes [10]. However, it may be difficult to interpret the number of classes greater than five or six from a clinical standpoint.

References:

1. Raebel MA, Ellis JL, Andrade SE. Evaluation of gestational age and admission date assumptions used to determine prenatal drug exposure from administrative data. Pharmacoepidemiol Drug Saf 2005; 14(12): 829–36. doi:10.1002/pds.1100.
2. Margulis AV, Setoguchi S, Mittleman MA, et al. Algorithms to estimate the beginning of pregnancy in administrative databases. Pharmacoepidemiol Drug Saf 2013; 22(1): 16–24. doi:10.1002/pds.3284.
3. Toh S, Mitchell AA, Werler MM, et al. Sensitivity and specificity of computer- ized algorithms to classify gestational periods in the absence of information on date of conception. Am J Epidemiol 2008; 167(6): 633–40. doi:10.1093/aje/ kwm367.
4. Devine ST, Hammad TA. Last menstrual period estimation techniques compared to physician recorded last menstrual period dates in the general practice research database [abstract]. Pharmacoepidemiol Drug Saf 2008; 17: S298.
5. Eworuke E, Hampp C, Saidi A, et al. An algorithm to identify preterm infants in administrative claims data. Pharmacoepidemiol Drug Saf 2012; 21(6): 640–50. doi:10.1002/pds.3264.
6. Li Q, Andrade SE, Cooper WO, et al. Validation of an algorithm to estimate ges- tational age in electronic health plan databases. Pharmacoepidemiol Drug Saf 2013; 22(5): 524–32. doi:10.1002/pds.3407.
7. Hardy JR, Leaderer BP, Holford TR, et al. Safety of medications prescribed be- fore and during early pregnancy in a cohort of 81,975 mothers from the UK General Practice Research Database. Pharmacoepidemiol Drug Saf 2006; 15(8): 555–64. doi:10.1002/pds.1269.
8. Kharbanda EO, Vazquez-Benitez G, Shi WX, et al. Assessing the safety of influ- enza immunization during pregnancy: the Vaccine Safety Datalink. Am J Obstet Gynecol 2012; 207(3 Suppl): S47–51. doi:10.1016/j.ajog.2012.06.073.
9. Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. The annals of statistics, 209-230.
10. Neelon, B., Swamy, G. K., Burgette, L. F., & Miranda, M. L. (2011). A Bayesian growth mixture model to examine maternal hypertension and birth outcomes. Statistics in Medicine, 30(22), 2721-2735.