BAYESIAN SEMI-PARAMETRIC ANALYSIS OF DEVELOPMENTAL TOXICOLOGY DATA

Francesca Dominici

Department of Biostatistics

Giovanni Parmigiani

Departments of Oncology and Biostatistics

Johns Hopkins University, Baltimore MD

http://biosun01.biostat.jhsph.edu/∼fdominic/

OUTLINE

- **Developmental Toxicology**

  # of malformations among live off-spring is not well modeled by the Binomial-logistic model

- **Motivation**

  greater flexibility in the distribution of the response given the dose

- **Semiparametric Extensions of GLM's**

  The cdf of the response given the dose follows a Dirichlet Process mixture

- **Data analyses of two toxicology studies**

- **Posterior and Predictive Bayesian inferences**

DEVELOPMENTAL TOXICOLOGY

- **Pregnant laboratory animals (dams) are exposed to varying doses of a toxic compound**
- **After they have given birth, the number of birth defects among the offsprings is recorded**
- **Goals:**
  1. **Investigate the relationship between the exposure to a potentially toxic compound and the frequency of birth defects**
  2. **Estimate the effective dose $ED(k)$ defined as the dose level at which the probability of malformation reaches a certain value k**

STANDARD APPROACH

- **Logistic-Binomial Model**

  $y_j^d \sim$ **Binomial**$(N_j^d, \theta^d),\ j = 1, \ldots, M^d$

  logit$(\theta^d) = \beta_0 + \beta_1 d,\ d = 1, \ldots, D$

  where:

- **y_j^d is the number of malformations among N_j^d live pups of dam j exposed to dose d**
- **θ^d is the probability of malformation**

PROBLEMS WITH THE STANDARD APPROACH

- data frequently display evidence of departure from the logistic-binomial model (Catalano, Ryan, 1994)
- clustered data - within-dam correlation need to be take into account
- hard-to model response distributions, displaying 0-inflation, n-inflation (an excessive number of dams with birth defects in n out of n pups)

EXTENSIONS

- A classical extension of the Binomial-logistic model is the Beta-Binomial model (Williams 1975,1983) where the dam-specific random effects are modeled as a Beta distribution
- Random Effects and Mixed-Effects GLMS (Cox, 1983, Prenctice, 1986)
- Bayesian Hierarchical Models with parametric and non parametric distributions of the dam-specific random effects (Wong and Mason 1985, Muller and Rosner, 1997)

SEMIPARAMETRIC APPROACH

- The distribution of the response is modeled in a general way
- The degree to which the distribution of the response adapts non-parametrically to the observations is driven by the data
- Logistic regression model, beta-binomial model, finite mixture models are special cases

BAYESIAN NON PARAMETRIC

- Let $f_j^d(y)$ with $0 \leq y \leq N_j^d$, the probability distribution functions (pdf) of y_j^d
- We treat $f_j^d(y)$ as an unknown parameter and we assume

$$f \sim \mathcal{D}(A, f_0) \text{ eg}$$
$$p(f \mid A, f_0) \propto \prod_{y=0}^{N} f(y)^{Af_0(y)-1}$$

- f_0 is the mean of the random pdf f
- A is the precision parameter, controlling the amount of variation of f around f_0 (also called the total mass parameter)

BAYESIAN NON PARAMETRIC

- Let Y be the # pups with malformations, we assume

$$Y \sim F$$
$$F \sim \mathcal{D}(A, F_0) \text{ where } F_0 = \textbf{Bin}(\cdot \mid \theta, N)$$

- A and θ are
  - Random
  - Depend on Covariates

SEMIPARAMETRIC LOGISTIC REGRESSION

- Dose $d \rightarrow M^d$
- Dam $j \rightarrow N_j^d$
- Live Pups $N_j^d \rightarrow y_j^d$
- y_j^d has an unknown cdf F_j^d

$$\begin{aligned} F_j^d \mid A^d, \theta^d &\sim \mathcal{D}(A^d, \textbf{Bin}(\theta^d, N_j^d)) \\ \textbf{logit}(\theta^d) &= \beta_0 + \beta_1 d \\ \log A^d &= \gamma_0 + \gamma_1 d \\ \boldsymbol{\beta}, \boldsymbol{\gamma} &\sim p(\boldsymbol{\beta}, \boldsymbol{\gamma}) \end{aligned}$$

- $A^d \rightarrow \infty$ then $y_j^d \sim$ Binomial-Logistic Regression

POSTERIOR DISTRIBUTIONS

- Analytical factorization into

$$\begin{aligned} p(F, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \textbf{data}) &\propto p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \textbf{data}) p(F \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \textbf{data}) \\ &\propto p(\boldsymbol{\beta}, \boldsymbol{\gamma}) p(\textbf{data} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) p(F \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \textbf{data}) \end{aligned}$$

- $p(\textbf{data} \mid \boldsymbol{\beta}, \boldsymbol{\gamma})$ is available in closed form (Antoniak 1974)
- $p(f_j^d, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \textbf{data}) \sim \mathcal{D}(A^d_{\textbf{post}}, f^d_{\textbf{post}})$ where
- $A^d_{\textbf{post}} = A^d + \sum_{j=1}^{M^d} \hat{f}_j^d$
- $f^d_{\textbf{post}} = w^d \textbf{Bin}(\theta^d, N_j^d) + (1 - w^d)\hat{f}_j^d$
- $w^d = A^d / (A^d + \sum_{j=1}^{M^d} \hat{f}_j^d)$
- $\hat{f}_j^d$ is the empirical frequency of y_j^d among all the dams having litter size N_j^d at dose d.

DATA ANALYSIS

- Effect of potential toxicology on birth defects
- Dams are treated; outcome in offspring
- 3 data sets
  - Simulated Binomial
  - DEHG
  - EG
- Goal: estimate dose-response parameters
- Units are clustered, extra-binomial variation, robust inference

PRIOR SPECIFICATION

- $\boldsymbol{\beta} \sim N_2(0, 3I)$ vague prior
- we assign a prior distribution on A which can be considered a "smoothness parameter"
- larger the values of A, the more the model will be closed to its parametric backbone
- we assign an uniform prior on the weights $A/(A+N)$ which control how close the posterior mean of the unknown pdf is close to the binomial pdf
- N indicate the number of dams in a litter-size/dose combination
- N is a prior-hyperparameter (sensitivity analysis)

POSTERIOR INFERENCE ON THE EFFECTIVE DOSE

- The effective dose at level k $ED(k)$ is defined as the dose level at which the probability of malformation is k times larger than it is at background or $ED(k)$
- Formally $\theta^{ED(k)} = k\theta^0$
- for $k > 1 + \exp(\beta_0)$

$$ED(k) = -\left\{\log(k^{-1}(1+\exp(-\beta_0)) - 1) + \beta_0)\right\}/\beta_1$$

- Posterior inferences on the $ED(k)$ are straightforward by using the MCMC runs

POSTERIOR PREDICTIVE INFERENCES

- Using $f_j^d, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \text{data} \sim \mathcal{D}(A^d_{\text{post}}, f^d_{\text{post}})$
- we can draw samples of the cdf function $\tilde{F}_j^d$ from a posterior predictive distribution given the samples values of β and A
- samples from $p(\tilde{F}_j^d \mid \text{data})$ useful to:

  prediction

  future data would be draws from such cdf

  goodness of fit

POSTERIOR PREDICTIVE INFERENCES

- we select a dose/litter size combination
- we choose $100\mu_g/m^3/11$ for the DEHP and $90\mu_g/m^3/12$ for the EG
- in the DEPH we have 6 dams total in the sample which receive dose 100 and have a litter size 11
- the empirical jumps are

$$0\ 1\ 1\ 0\ 0\ 1\ \mathbf{3}\ 0\ 0\ 0\ 0\ \ 0$$
$$0\ 1\ 2\ 3\ 4\ 5\ \mathbf{6}\ 7\ 8\ 9\ 10\ 11$$

- for example there are 3 dams with 6 malformed pups among the 11
- the cumulative sum of these frequencies is the empirical cdf $(\hat{F}_j^d)$

DISCUSSION

- **Dirichlet Process mixtures offer a PRACTICAL and FLEXIBLE approach to modeling binary data with difficult error distributions**
- **The degree to which the model adapts nonparametrically to the observations is DATA DRIVEN**
- **The marginal posterior distribution of the parameters of interest is available in closed form**
- **When there aren't enough observations they automatically collapse into a GLM's**

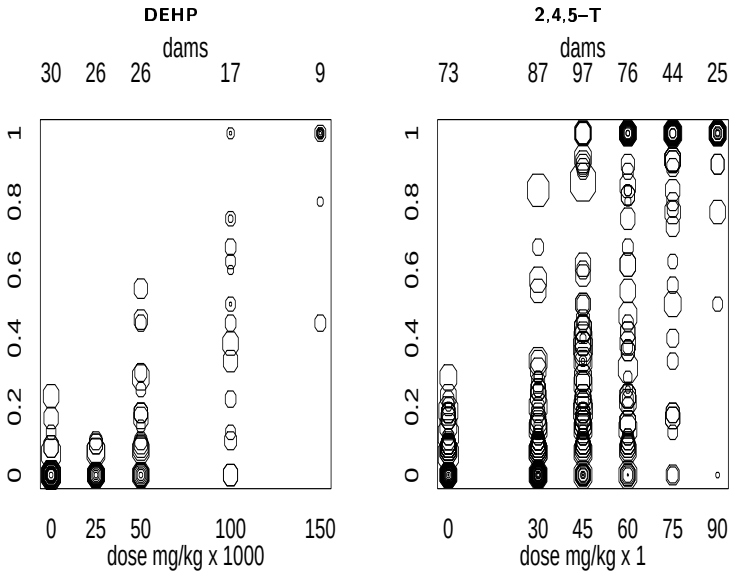

Figure 1: **The DEHP (left) and 2,4,5–T (right) data sets. The top panels display the raw data. Each circle corresponds to a dam. The circles' areas are proportional to the litter sizes; the circles' coordinates are the dose level and the observed relative frequency of malformations. In addition, the numbers of dams exposed to each dose level is displayed at the top.**

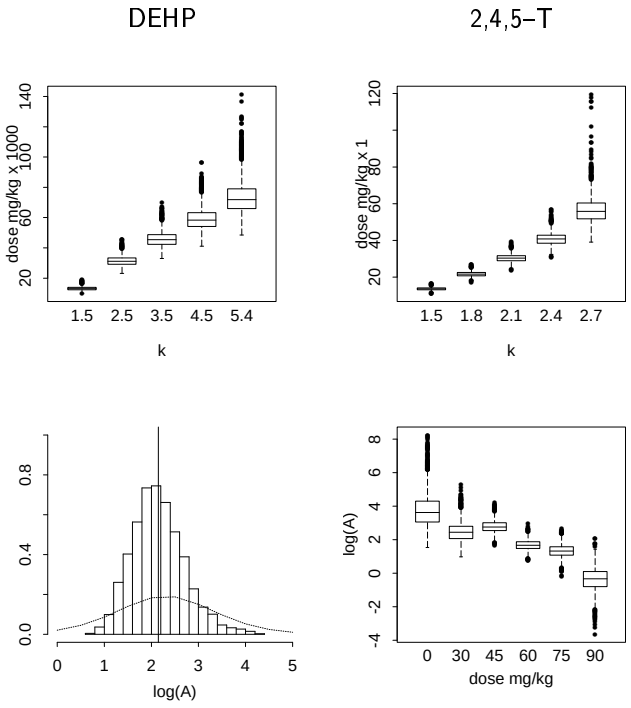

Figure 2: **Inference on the $ED(k)$ (top) and precision parameters (bottom) for the DEHP and 2,4,5,-T data. Top panels are boxplots of samples from the posterior distributions of effective doses $ED(k)$ corresponding to a k–fold increase of the probability of malformations above the background rate. The bottom left panel displays the posterior (histogram) and prior (solid line) distributions of the precision parameter $\log(A)$ for the DEHP data set. The bottom right panel shows boxplots of posterior samples of the precision parameters $\log(A^d)$ corresponding to the six doses for the 2,4,5-T data set.**

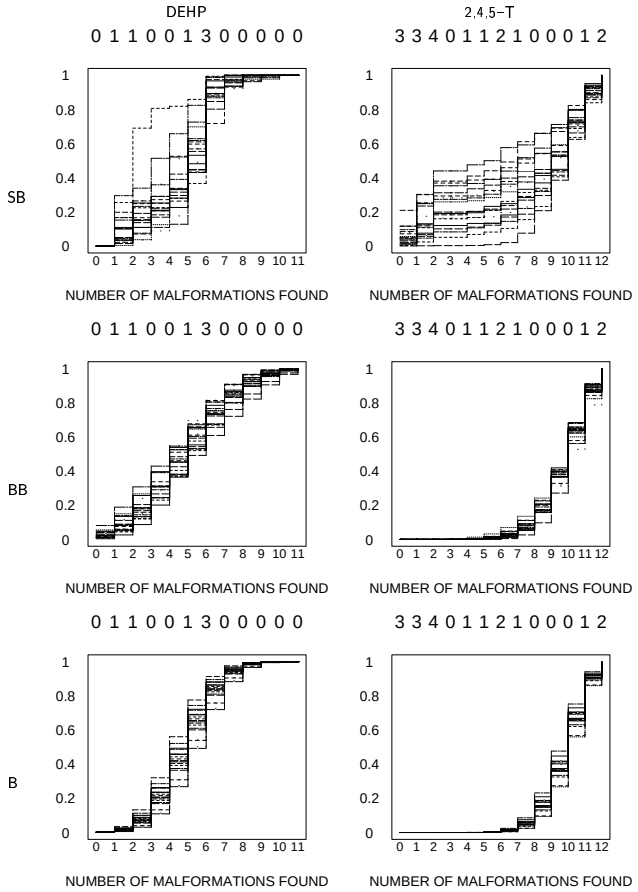

Figure 3: **Samples of 20 cumulative distributions functions $\tilde{F}_j^d$ obtained using the Semiparametric Bayesian (SB), the Beta-binomial (BB), and the Binomial (B) models. For the DEHP data set we choose a dose of $100\mu_g/m^3$ and a litter size of 11; for the 2,4,5-T we choose a dose of $90\mu_g/m^3$ and a litter size of 12. Empirical frequencies corresponding to the selected dose/litter size are displayed at the top of each**