Modeling Overdispersion in Binomial Data with Regression Linked to a Finite Mixture Probability of Success

Andrew M. Raim and Nagaraj K. Neerchal
Department of Mathematics and Statistics,
University of Maryland, Baltimore County

Abstract
Logistic regression often cannot account for large variability seen in binomial data due to departures from standard assumptions. Many techniques have been considered to address this issue, commonly known as overdispersion. Finite mixture distributions may be used when the extra variation is explained by the presence of several latent subpopulations. For example, a finite mixture of regressions links the probability for each latent group to a seperate regression. Analogously to the usual logistic regression, we consider linking a regression to the mixture probability of success in a finite mixture of binomials. This can be seen as “marginal modeling” with respect to the latent groups, as opposed to the mixture of regressions which is seen as “conditional modeling” on the groups, and would allow more parsimonious models when only a single overall regression is desired. Our approach is likelihood-based, which may be considered an advantage over quasi-likelihood techniques often used to address overdispersion. This work presents the new model and an illustrative example.

Key Words: Overdispersion; GLM; Logistic Regression; Random Effects; Goodness-of-Fit.

1. Introduction

A common problem in the analysis of binomial data using standard logistic regression is that more variation is present in the data than can be expressed by the model; in this case, overdispersion is said to occur. This work considers a novel way of handling overdispersion in the binomial regression setting; that is, by linking a regression to the probability of success in a finite mixture of binomials. In the finite mixture, the overall probability of success $\pi_1 p_1 + \cdots + \pi_J p_J$ is a weighted sum of the probabilities p_j, for $1, \ldots, J$, from J latent subpopulations. Consequently, there are technical challenges must be overcome in expressing the model and carrying out even basic computations. This paper develops one possible implementation of the model. Although still in its early stages of development, we show that it provides a good fit for a real dataset with known overdispersion issues, faring well in comparison to several other binomial models with extra variation.

The rest of the paper proceeds as follows. Section 2 introduces the binomial regression problem and discusses some existing approaches to handling extra variation. Section 3 develops the new model, which is termed the Mixture Link distribution, and section 4 presents plots of its density. An illustrative example for the effect of radiation dose on chromosome aberrations is presented in section 5; several binomial models are compared using a goodness-of-fit test as well as AIC and BIC. Finally, section 6 concludes the paper.

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, 21250, U.S.A, Email: {araim1, nagaraj}@umbc.edu.

2. Background

Under the usual logistic regression model, T_i successes are observed in m_i trials for $i = 1, \dots, n$. The probability of success p_i for each observation is modeled on a covariate $\boldsymbol{x}_i \in \mathbb{R}^d$, which is taken to be fixed. It is assumed that $p_i = G(\boldsymbol{x}_i^T \boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathbb{R}^d$ and a prespecified inverse link function $G : \mathbb{R} \to (0, 1)$. For this paper, G will be taken to be the cumulative distribution function for the logistic distribution $G(x) = 1/(1 + e^{-x})$. The model just described may be written briefly as

$$T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, p_i), \quad p_i = G(\boldsymbol{x}_i^T \boldsymbol{\beta}).$$

In practice, T_i, $\boldsymbol{x}_i$, and m_i are observed, and statistical inference on the parameter $\boldsymbol{\beta}$ is a primary objective of analysis. Logistic regression is a special case of the generalized linear model (GLM) framework McCullagh and Nelder (1989), which allows non-normal, non-continuous outcomes to be modeled as responses to a regression. Although GLM accomplishes this naturally, a frequent problem is that the data exhibit more variation than the framework is capable of expressing (Morel and Neerchal, 2012). When this happens, we say that there is overdispersion with respect to the chosen model. Overdispersion may be caused, for example, when important covariates have not been included in the regression, or when the implicit assumption of independence within the m_i trials has been violated. The limitation in the amount of modeled variability can be seen by noting the relationship between the mean and variance

$$\text{E}(T_i) = m_i p_i \quad \text{and} \quad \text{Var}(T_i) = m_i p_i (1 - p_i);$$

therefore, the same regression used to model the probabilities of success of the T_i also must explain their mean and variance.

A simple workaround is to extend the model with a dispersion parameter ϕ so that $\text{Var}(T_i) = \phi m_i p_i (1 - p_i)$ (Agresti, 2002, Section 4.7). The resulting model is referred to as quasi-likelihood because it no longer correponds to a true distribution. For longitudinal data, a popular quasi-likelihood method is the generalized estimating equations (GEE) developed by Liang and Zeger (1986). GEE proposes that inference on $\boldsymbol{\beta}$ to be based on a score-like equation, and allows the analyst to assume a working correlation structure as a way to induce dependence for observations within a subject. This idea may be used when individual Bernoulli trials of a binomial experiment are observed. GEE has some desirable properties, such as consistency even under misspecification of the working correlation. However, being a quasi-likelihood method, it may not be based on a real likelihood.

There are also a variety of likelihood-based models that can be used to induce extra variation; we will mention several here. The zero-inflated binomial (ZIB) distribution discussed by Hall (2000),

$$\text{P}(T = t \mid m, p, \phi) = \phi I(t = 0) + (1 - \phi)\text{Bin}(t \mid m, p),$$

assumes a latent process that generates a zero with probability ϕ and a binomial random variable with probability $1 - \phi$. Similarly, any of the support values $0, 1, \dots, m$ may be selected by the analyst to be inflated. The random-clumped binomial (RCB) distribution (Morel and Nagaraj, 1993) may be used when the inflated value is not known ahead of time and is considered to be drawn randomly. An RCB distributed random variable $T = NY + (X \mid N)$ is obtained using

$$Y \sim \text{Ber}(p), \quad N \sim \text{Bin}(m, \phi), \quad (X \mid N) \sim \text{Bin}(m - N, p),$$

where Y represents success/failure of a leader, N is the number of trials that follows the leader, and $(X \mid N)$ are remaining trials that are selected independently. Here, $p \in (0, 1)$ is interpreted as the success probability for the trials, and $\phi \in (0, 1)$ is the probability of following the leader. Perhaps the most popular binomial distribution supporting extra variation is beta-binomial (BB), which assumes a hierarchy,

$$T \mid \mu \sim \text{Bin}(m, \mu), \quad \mu \sim \text{Beta}(\alpha, \beta),$$

where the probability of success is drawn from a beta distribution. BB may be reparameterized, as noted in (Morel and Neerchal, 2012, Section 4.2) and (Prentice, 1986) for example, using

$$\begin{aligned} \alpha = p\phi^{-1}(1-\phi) \quad &\text{and} \quad \beta = (1-p)\phi^{-1}(1-\phi) \\ \iff p = \frac{\alpha}{\alpha+\beta} \quad &\text{and} \quad \phi = \frac{1}{\alpha+\beta+1}, \end{aligned}$$

so that $p \equiv \text{E}(\mu) \in (0, 1)$ can be interpreted as a probability of success and

$$\text{Var}(T) = mp(1-p)\{1 + \phi(m-1)\}.$$

For the ZIB, RCB, and BB distributions as stated here, $\phi \in (0, 1)$ is seen as an overdispersion parameter with respect to the binomial distribution where $\phi \downarrow 0$ corresponds to "no overdispersion". Although ZIB, RCB, and BB are not exponential families, and therefore do not fall into the classical GLM framework, regressions may be linked to any of the individual parameters and inference for $\boldsymbol{\beta}$ may be carried out through the linked likelihood.

Adding random effects to the regression model of a GLM is a flexible way to model extra variation between observations or to group observations that naturally belong to the same cluster (c.f. Agresti, 2002; Morel and Neerchal, 2012). However, because random effects are unobserved and manifest themselves as integrals in the likelihood, computation quickly becomes difficult as random effect structures are allowed to become more elaborate. A compromise between flexibility and computation is found in the random intercept model, where only a random intercept is assumed. Logistic regression with a random intercept has been considered by Aitkin (1996), among others, who use nonparametric maximum likelihood (NPMLE) to avoid making assumptions about the distribution of the random intercept.

Finite mixture distributions are often used to model the situation of multiple latent subpopulations. In the basic finite mixture of binomials,

$$f(t \mid m, \boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j \text{Bin}(t \mid m, \mu_j), \tag{2.1}$$

it is assumed that there are J subpopulations, and a latent process Z is selecting from the labels $(1, \ldots, J)$ with corresponding probabilities $(\pi_1, \ldots, \pi_J)$. The finite mixture (2.1) can be extended to a finite mixture of regressions by linking regressions

$$p_j = G(\boldsymbol{x}^T \boldsymbol{\beta}_j), \quad \text{for } j = 1, \ldots, J.$$

This idea is discussed in Frühwirth-Schnatter (2006), and extended to more elaborate processes for the latent Z.

The remainder of this paper presents the Mixture Link model: a completely likelihood-based binomial model for extra variation which assumes a finite mixture distribution with a regression linked to the marginal probability of success. The finite mixture is used to handle heterogeneity in a robust way, but unlike the finite mixture of regressions model, the emphasis is on a single regression for the entire population. Therefore, the finite mixture of regressions can be thought of as "conditional modeling" with respect to latent subpopulations, while Mixture Link is "marginal modeling" on the entire population, with built in tolerance for heterogeneity.

3. Mixture Link Model

Consider a random variable T following the finite mixture of binomials distribution (2.1), which we will denote as $T \sim \text{BinMix}(m, \boldsymbol{\mu}, \boldsymbol{\pi})$. Without further restriction, the component probabilities of success $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$ naturally lie within the rectangle $[0,1]^J$, and the subpopulation proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ are within the J-dimensional probability simplex $\mathcal{S}^J = \{\boldsymbol{\mu} \in [0,1]^J : \sum_{j=1}^J \mu_j = 1\}$. Notice that

$$\text{E}(T) = \sum_{j=1}^J \pi_j m \mu_j = m \boldsymbol{\mu}^T \boldsymbol{\pi},$$

where $\boldsymbol{\mu}^T \boldsymbol{\pi}$ is the mixture probability of success. Analogously to logistic regression under the GLM framework, our goal is to link the regression $\boldsymbol{x}^T \boldsymbol{\beta}$ to the finite mixture by enforcing the constraint $\boldsymbol{\mu}^T \boldsymbol{\pi} = p$ where $p = G(\boldsymbol{x}^T \boldsymbol{\beta})$. The space of all $\boldsymbol{\mu}$ that honors the link is then

$$A(p, \boldsymbol{\pi}) = \{\boldsymbol{\mu} \in [0,1]^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = p\};$$

when there is no confusion, we will write A as shorthand.

Next consider an independent sample

$$T_i \overset{\text{ind}}{\sim} \text{BinMix}(m_i, \boldsymbol{\mu}_i, \boldsymbol{\pi}), \quad \boldsymbol{\mu}_i \in A_i, \quad i = 1, \ldots, n,$$

where $A_i = A(p_i, \boldsymbol{\pi})$ and $p_i = G(\boldsymbol{x}_i^T \boldsymbol{\beta})$. Here A_i and $\boldsymbol{\mu}_i$ vary with i to reflect that observations may have distinct covariates $\boldsymbol{x}_i$. When $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ are treated as fixed and unknown quantities, taking a maximum likelihood approach would mean maximizing

$$\prod_{i=1}^n \left\{ \sum_{j=1}^J \pi_j \text{Bin}(t_i \mid m_i, \mu_{ij}) \right\}, \quad \text{subject to} \quad \boldsymbol{\gamma}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n, \boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{3.1}$$

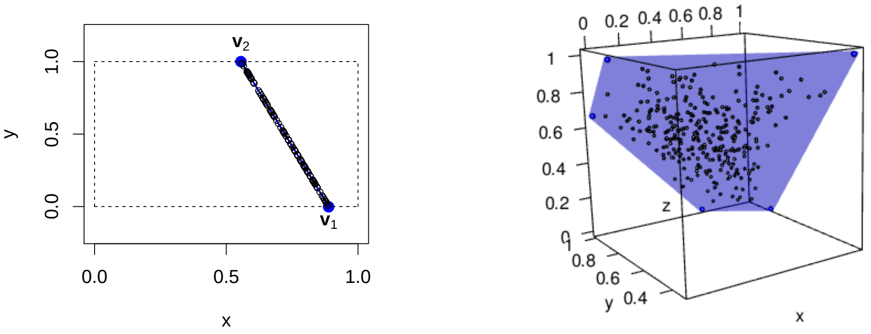
$$\boldsymbol{\gamma}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n, \boldsymbol{\pi}) = \begin{pmatrix} g(\boldsymbol{\mu}_1^T \boldsymbol{\pi}) \\ \vdots \\ g(\boldsymbol{\mu}_n^T \boldsymbol{\pi}) \end{pmatrix} : n \times 1, \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} : n \times d,$$

where $g = G^{-1}$, The parameter $\boldsymbol{\beta}$ only enters the optimization problem through the constraint, which suggests a that a profile likelihood approach such as

$$Q(\boldsymbol{\beta}) = \sup_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n, \boldsymbol{\pi}} \{\log L(\boldsymbol{\theta}) : \boldsymbol{\gamma}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n, \boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}\}$$

may be more natural to consider. However, the overall optimization problem is still on the space

$$\underbrace{[0,1]^J \times \cdots \times [0,1]^J}_{n} \times \mathcal{S}^J \times \mathbb{R}^d,$$



(a) (b)

Figure 1: A sample drawn from A: (a) $n = 100$ with $J = 2$, $\boldsymbol{\pi} = (\frac{3}{4}, \frac{1}{4})$, $p = \frac{2}{3}$, and (b) $n = 300$ with $J = 3$, $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, $p = \frac{2}{3}$.

whose dimension is increasing with the sample size n due to the nuisance parameters $\boldsymbol{\mu}_i$ for $i = 1, \ldots, n$, This is generally not a desirable quality for a model.

Instead of attacking the optimization problem (3.1) directly, we consider a hierarchical model where the $\boldsymbol{\mu}_i$ are unobservable random effects. The effects must be integrated out to obtain the likelihood of the observed data. This can be contrasted to the profile optimization which removes the $\boldsymbol{\mu}_i$ from consideration by an inner optimization. The tradeoff between having too many fixed nuisance parameters vs. unobservable random effects is traditionally seen in linear mixed models (McCulloch et al., 2008).

But what should be the distribution of the random effects? Because A_i is a bounded convex set, we can find vertices $\boldsymbol{v}_1^{(i)}, \ldots, \boldsymbol{v}_{k_i}^{(i)} \in \mathbb{R}^J$ such that

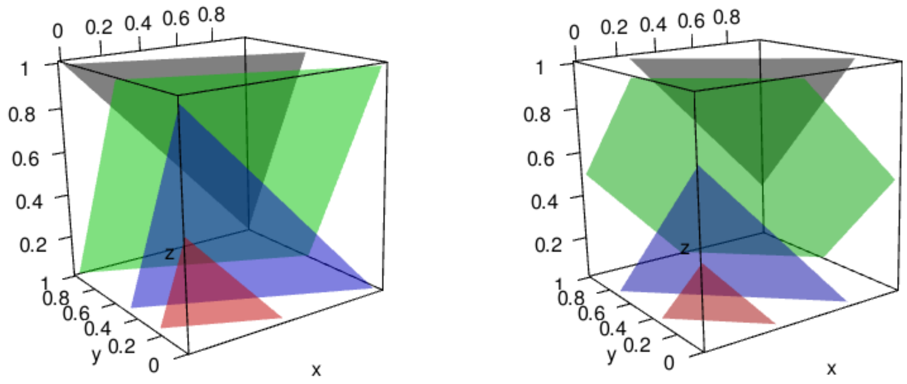
$$A_i = \operatorname{conv}(\boldsymbol{v}_1^{(i)}, \ldots, \boldsymbol{v}_{k_i}^{(i)}) = \left\{ \sum_{\ell=1}^{k_i} \lambda_\ell \boldsymbol{v}_\ell^{(i)} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \right\} = \left\{ \boldsymbol{V}^{(i)} \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathcal{S}^{k_i} \right\}, \tag{3.2}$$

where $\mathcal{S}^{k_i}$ is the k_i-dimensional probability simplex and $\boldsymbol{V}^{(i)} = (\boldsymbol{v}_1^{(i)} \cdots \boldsymbol{v}_{k_i}^{(i)}) \in \mathbb{R}^{J \times k_i}$. Note that $\boldsymbol{V}^{(i)}$ may be different for each observation when the set A_i depends on a covariate $\boldsymbol{x}_i$. The number of vertices k_i may also vary with each observation.

Now a natural way to place a distribution on the set A is to let $\boldsymbol{\lambda} \sim \text{Dirichlet}_k(\boldsymbol{\alpha})$, whose density is

$$f(\boldsymbol{\lambda} \mid \boldsymbol{\alpha}) = \frac{\lambda_1^{\alpha_1 - 1} \cdots \lambda_k^{\alpha_k - 1}}{\mathrm{B}(\boldsymbol{\alpha})} \cdot I(\boldsymbol{\lambda} \in \mathcal{S}^k), \quad \text{where } \mathrm{B}(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}{\Gamma(\alpha_1 + \cdots + \alpha_k)}.$$

Danaher et al. (2012) recently proposed priors based on the Minkowski-Weyl decomposition to enforce (biologically motivated) polyhedral constraints for parameters in Bayesian analysis. A Dirichlet prior is proposed for the simplex between the extreme points; the Dirichlet distribution is used in a similar manner in the present work. Figure 1 shows an example of the set A for $J = 2$ and $J = 3$, along with a random sample taken on the set using a $\text{Dirichlet}_k(\mathbf{1})$ distribution. Figure 2 shows how the set A changes as p is varied when $J = 3$. Note that the number of vertices k can change, as can the placement of the hyperplane segment. It is clear that for $J = 3$ it is possible for k to take on values at least in $\{3, 4, 5, 6\}$, and certainly $k = J$ need not hold.



(a) $\boldsymbol{\pi} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. (b) $\boldsymbol{\pi} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

Figure 2: The set $\left\{\boldsymbol{\mu} \in [0,1]^3 : \mu_1\pi_1 + \mu_2\pi_2 + \mu_3\pi_3 = p\right\}$ visualized with two different settings of $\boldsymbol{\pi}$. In each case, $p \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$ is shown (from front to back).

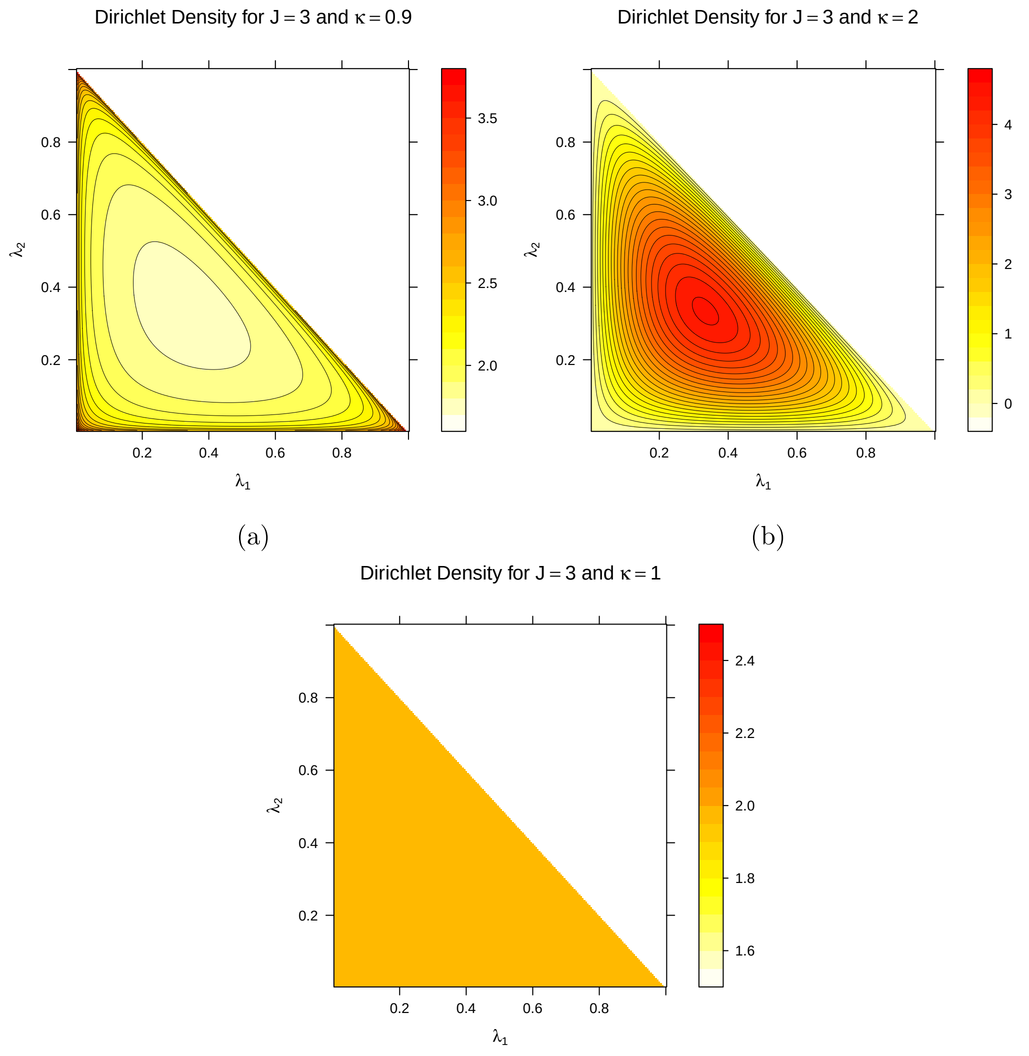
We can now write the hierarchical Mixture Link model as

$$\begin{aligned}
&T_i \mid \boldsymbol{\mu}_i, \boldsymbol{\pi} \overset{\text{ind}}{\sim} \text{BinMix}(m_i, \boldsymbol{\mu}_i, \boldsymbol{\pi}) \\
&\boldsymbol{\mu}_i = \boldsymbol{V}^{(i)}\boldsymbol{\lambda}^{(i)}, \quad \text{where } \boldsymbol{V}^{(i)} = (\boldsymbol{v}_1^{(i)} \cdots \boldsymbol{v}_{k_i}^{(i)}) \text{ are vertices of } A(p_i, \boldsymbol{\pi}) \\
&\boldsymbol{\lambda}^{(i)} \overset{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\boldsymbol{\alpha}^{(i)}).
\end{aligned} \tag{3.3}$$

Notice that the dimension of $\boldsymbol{\alpha}^{(i)} = (\alpha_1^{(i)}, \ldots, \alpha_{k_i}^{(i)})$ may vary between observations, depending on $\boldsymbol{\pi}$ and p_i. Because our main interest is when p_i vary due to a regression, we make the further assumption that $\boldsymbol{\alpha}^{(i)} = \kappa\mathbf{1}$ where $\mathbf{1} = (1, \ldots, 1)$ and $\kappa > 0$. There are also identifiability issues in letting the components of $\boldsymbol{\alpha}^{(i)}$ vary because the vertices in $\boldsymbol{V}^{(i)}$ are not strictly ordered, therefore it is difficult to maintain a correspondence between $\boldsymbol{v}_\ell^{(i)}$ and $\alpha_\ell^{(i)}$. Figure 3 shows Dirichlet distributions plotted for several settings of κ when $J = 3$. Notice that $\kappa = 1$ corresponds to the uniform distribution of $\boldsymbol{\lambda}^{(i)}$ on the simplex (and furthermore to a uniform distribution of $\boldsymbol{\mu}_i$ on A_i), while $0 < \kappa < 1$ results in more density focused toward the vertices than the interior, and $\kappa > 1$ yields more density in the interior of the simplex. The hierarchy (3.3) is parameterized by $\boldsymbol{\theta} = (p, \boldsymbol{\pi}, \kappa) \in \mathbb{R}^{1+(J-1)+1}$ if T_i are taken to be independent and identically distributed, or $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\pi}, \kappa) \in \mathbb{R}^{d+(J-1)+1}$ in the case of a regression. In a frequentist analysis, $\boldsymbol{\theta}$ will be a fixed but unknown parameter. A Bayesian analysis would put a prior on $\boldsymbol{\beta}$, the main parameter of interest, and perhaps $\boldsymbol{\pi}$ and κ as well. The density is given by

$$f(t \mid m, p, \boldsymbol{\pi}, \kappa) = \binom{m}{t} \sum_{j=1}^{J} \pi_j \int v^t (1-v)^{m-t} \cdot f_{A^{(j)}}(v)\, dv \tag{3.4}$$

where $f_{A^{(j)}}(v)$ is the marginal density of μ_j on the set A. The notation $T \sim \text{MixLink}_J(m, p, \boldsymbol{\pi}, \kappa)$ will be used to say that a random variable T is drawn from this distribution. The joint likelihood of the sample $T_i \overset{\text{ind}}{\sim} \text{MixLink}_J(m_i, p_i, \boldsymbol{\pi}, \kappa)$



(a)

(b)

(c)

Figure 3: Dirichlet$_3(\boldsymbol{\lambda} \mid \kappa\mathbf{1})$ density for several settings of κ. Only λ_1 and λ_2 are plotted, as $\lambda_3 = 1 - \lambda_1 - \lambda_2$ is redundant.

for $i = 1, \ldots, n$ is then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \binom{m_i}{t_i} \sum_{j=1}^{J} \pi_j \int v^{t_i}(1-v)^{m_i - t_i} \cdot f_{A_i^{(j)}}(v)\, dv, \right\}. \tag{3.5}$$

Computational details for the Mixture Link model, including identification of the vertices of A_i and evaluation of the density, are given in the Ph.D. thesis (Raim, 2013).

4. Mixture Link Density

To understand the utility of the Mixture Link distribution for modeling overdispersion in practice, we now examine some plots of the density. Here we consider the distribution MixLink$_J(m, p, \boldsymbol{\pi}, \kappa)$, i.e. without regression. Plotted in Figures 4 and 5 are the densities for the RCB and BB distributions, respectively, which were introduced in section 2. For each of $p \in \{0.25, 0.50\}$, the density is plotted for $m = 20$ trials and several settings of ϕ. Figure 6 shows corresponding plots for the Mixture Link Density letting $J = 2$ and $\kappa = 1$. Each shows the binomial density for reference. For beta-binomial, as the overdispersion parameter ϕ increases, the density moves from the standard binomial to one where most mass is at the extreme support values 0 and 20. Under RCB, increasing ϕ leads to the formation of a second mode. For the Mixture Link density, increasing π has the effect of fattening the tails compared to the standard binomial.

Figures 7, 8, and 9 show several more cases of the Mixture Link density, focusing only on the case $p = 0.5$ but varying $\kappa \in \{0.5, 1, 2\}$ and $J \in \{2, 3\}$. A variety of shapes can be seen for the limited settings of $\boldsymbol{\pi}$ that are shown. Expressing two modes is possible, as is inflating mass at the extreme support values 0 and 20.

5. Example: Hiroshima Data

Awa et al. (1971) and Sofuni et al. (1978) study the effects of radiation exposure on chromosome aberrations in survivors of the atomic bombs that were used in Hiroshima and Nagasaki. Subjects in the study consist of 649 residents in Hiroshima and 403 residents in Nagasaki for whom radiation dose estimates were available. Subjects are placed into exposed and control groups. Individuals in the control group were either not present in their city at the time of the bombings, or received an estimated dose of less than one rad. A chromosome analysis is carried out on m_i circulating lymphocytes for the ith subject, and of those, the number of chromosome aberrations t_i is recorded. Two types of radiation exposure are considered, neutron and gamma, where higher doses of neutron exposure in Hiroshima are suspected of leading to increased incidence of aberrations.

A subset of this data is featured in Morel and Neerchal (2012) as an illustrative example for goodness-of-fit in binomial models for extra variation. It is natural to suspect that overdispersion will be an issue in this data under standard logistic regression, as the presence or absence of aberrations within the m_i circulating lymphocytes of a particular subject may not be independent. Here, $n = 648$ observations from the Hiroshima portion of the original data are considered, and the covariate d_i represents the sum of neutron and gamma exposure for the ith subject. The total exposure is then normalized to

$$z_i = \frac{d_i - \bar{d}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})^2}}, \quad i = 1, \ldots n.$$

To compare several binomial models with extra variation on the smaller Hiroshima dataset, we consider the goodness-of-fit (GOF) test

$$H_0 : T_i \overset{\text{ind}}{\sim} f(t_i \mid m_i, \boldsymbol{\theta}) \text{ for some } \boldsymbol{\theta} \in \Theta \quad \text{vs.} \quad H_1 : \text{Not},$$

where f is fully specified up to a possibly unknown parameter $\boldsymbol{\theta}$, and $\Theta \subseteq \mathbb{R}^q$. For binomial data with m_i varying with observations, Neerchal and Morel (1998) proposed the following variation on the usual Pearson chi-square test statistic. Suppose

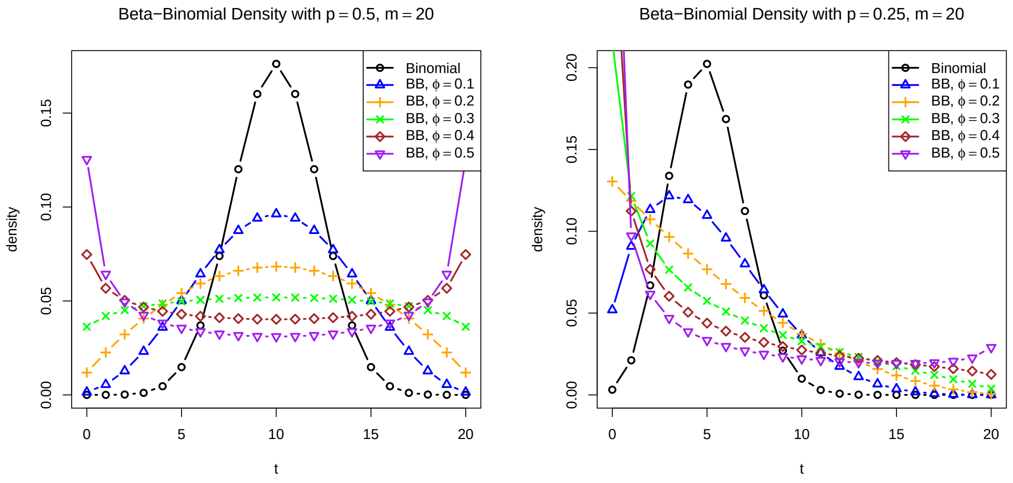


Figure 4: BB densities.

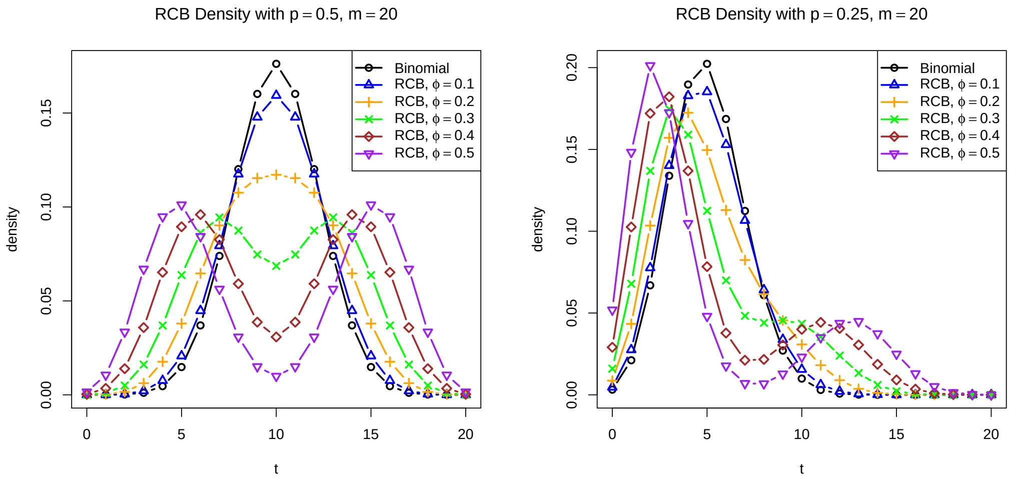


Figure 5: RCB densities.

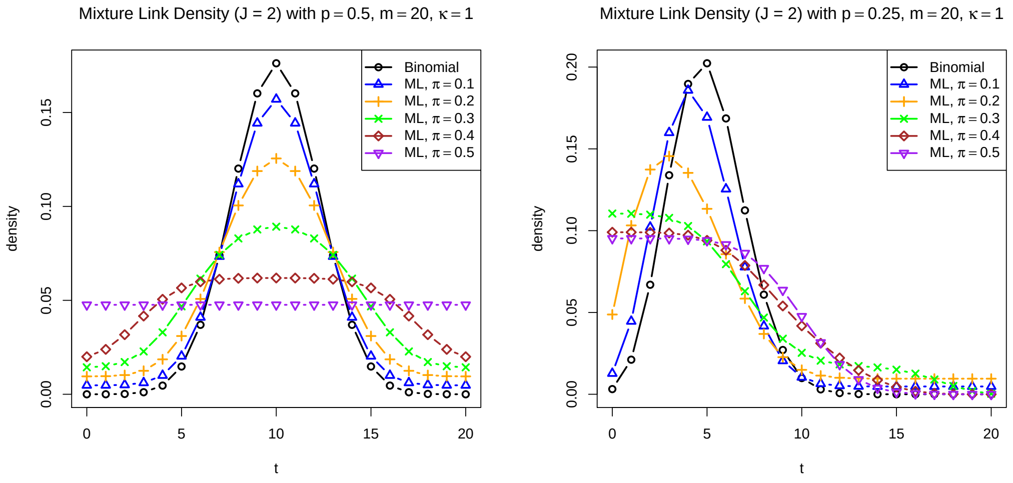


Figure 6: Mixture Link densities.

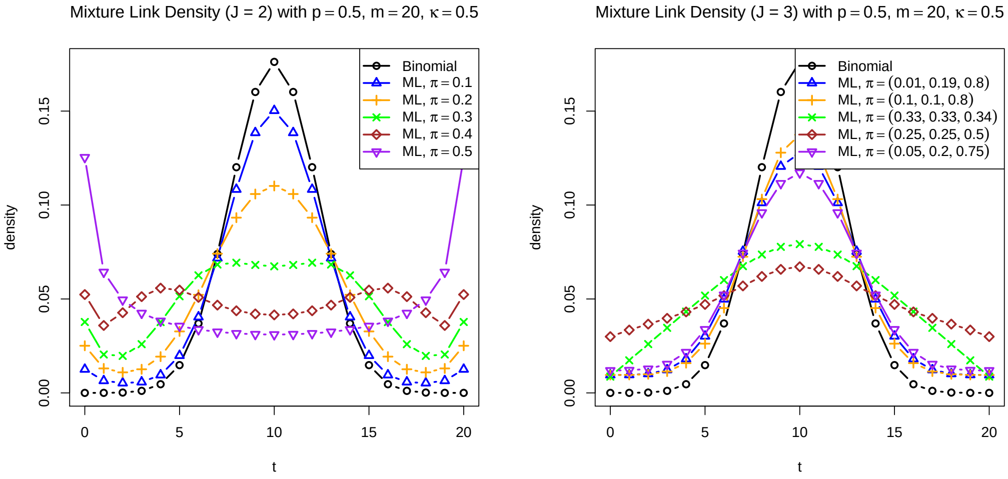


Figure 7: Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 0.5$.

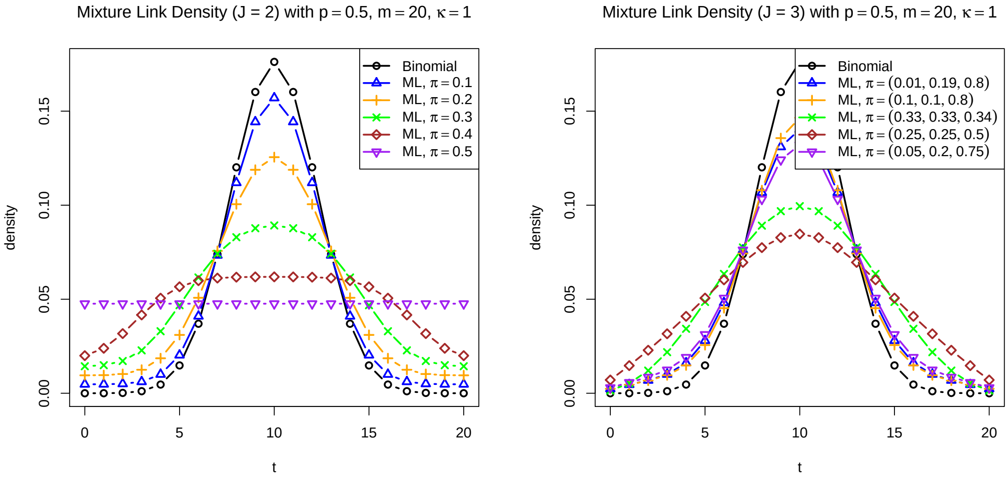


Figure 8: Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 1$.

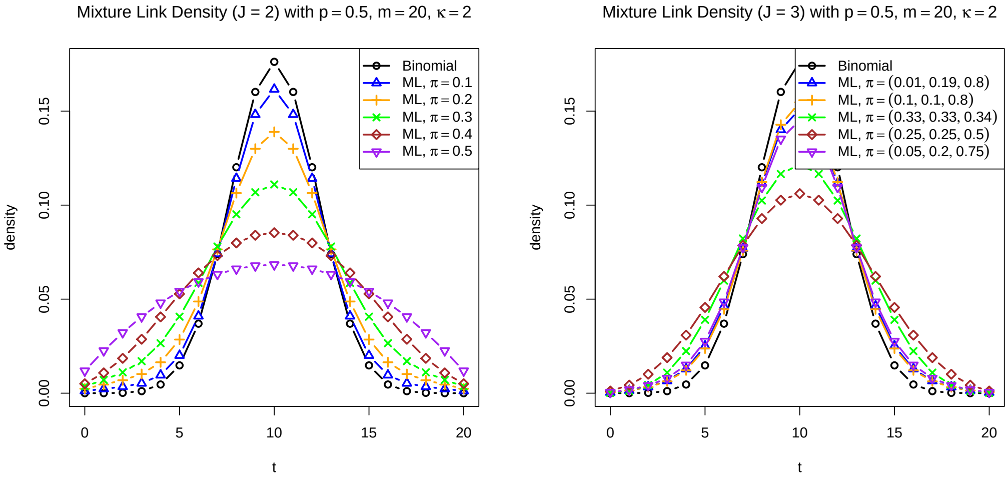


Figure 9: Compare Mixture Link densities for $J = 2$ and $J = 3$ when $\kappa = 2$.

$I_1, \ldots, I_r$ are disjoint intervals that cover $[0, 1]$, and let

$$X(\boldsymbol{\theta}) = \sum_{\ell=1}^{r} \frac{[O_\ell - E_\ell(\boldsymbol{\theta})]^2}{E_\ell(\boldsymbol{\theta})}, \quad \text{where} \tag{5.1}$$

$$E_\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{t=0}^{m_i} f(t \mid m_i, \boldsymbol{\theta}) I\left(\frac{t_i}{m_i} \in I_\ell\right) \quad \text{and} \quad O_\ell = \sum_{i=1}^{n} I\left(\frac{t_i}{m_i} \in I_\ell\right).$$

Sutradhar et al. (2008) shows that, when the null distribution f is RCB, $X(\boldsymbol{\theta}) \sim \chi^2_{r-1}$ when all parameters are known and $X(\hat{\boldsymbol{\theta}}) \sim \chi^2_{r-1-q}$ when $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^q$ is estimated by maximizing the *grouped* likelihood

$$L_g(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{\ell=1}^{r} \left[\mathrm{P}\left(\frac{t_i}{m_i} \in I_\ell \;\middle|\; m_i, \boldsymbol{\theta} \right)^{I\left(\frac{t_i}{m_i} \in I_\ell\right)} \right],$$

In practice, it is more natural to work with the *ungrouped* likelihood

$$L_u(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i \mid m_i, \boldsymbol{\theta})$$

of the observed T_i. There is a noted "recovery" of degrees of freedom in the GOF statistic when the ungrouped MLE is used, so that $X(\hat{\boldsymbol{\theta}})$ follows a χ^2_ν distribution with ν between $r - 1 - q$ and $r - 1$. Although the theory in (Sutradhar et al., 2008) is stated specifically for the RCB distribution, proofs are given for general binomial models with varying m_i. However, a number of regularity conditions are assumed to ensure, for example, first-order efficiency of the MLE. Our GOF studies use the ungrouped MLE and consider p-values based on $\nu = r - 1 - q$. Recall that smaller ν will result in a more right-skewed χ^2_ν distribution. Consequently, a fixed $X(\hat{\boldsymbol{\theta}})$ will count as stronger evidence against the hypothesis of adequate fit H_0, and therefore this choice of ν is conservative in declaring that a model fits adequately. The selection of intervals I_ℓ is left up to the analyst, but it is suggested to follow the rule of thumb that all $E_\ell(\boldsymbol{\theta}) > 5$. Some discussion on interval selection is given in (Kendall et al., 1991, Section 30.2); common choices include equal width intervals and those having equal probability.

To compute estimates and standard errors under the Mixture Link model, the `optim` function in `R` (R Core Team, 2013) is used to numerically maximize the logarithm of the likelihood (3.5). A suitable transformation $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\vartheta})$ allows the optimizer to work in $\mathbb{R}^q$ and ensures that $\boldsymbol{\theta}$ is restricted to the parameter space. Standard errors are obtained from the Hessian $\boldsymbol{H}$ evaluated at the solution $\hat{\boldsymbol{\theta}}$, after adjustment by the Jacobian of $\boldsymbol{h}$.

The following models are compared for goodness-of-fit for the Hiroshima dataset:

- Logistic: $T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, p_i)$,
- RCB: $T_i \overset{\text{ind}}{\sim} \text{RCB}(m_i, p_i, \phi)$,
- BB: $T_i \overset{\text{ind}}{\sim} \text{BB}(m_i, p_i, \phi)$,
- RCB-Reg: $T_i \overset{\text{ind}}{\sim} \text{RCB}(m_i, p_i, \phi_i)$,
- BB-Reg: $T_i \overset{\text{ind}}{\sim} \text{BB}(m_i, p_i, \phi_i)$,
- MixLinkJ2: $T_i \overset{\text{ind}}{\sim} \text{MixLink}_2(m_i, p_i, \boldsymbol{\pi}, \kappa)$,

where $g = G^{-1}$ is the logistic link function, $g(p_i) = \beta_0 + \beta_1 z_i + \beta_2 z_i^2$ for all models, and $g(\phi_i) = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2$ for the two "-Reg" models. The models RCB-Reg and

Table 1: Model comparison statistics.

Model	LogLik	q	AIC	BIC	GOF statistic	df range	p-value
Logistic	-1814.189	3	3634.400	3647.799	110.38	[17,20]	$< 10^{-13}$
RCB	-1567.499	4	3142.997	3160.893	68.25	[15,19]	$< 10^{-6}$
BB	-1487.923	4	2983.847	3001.742	93.79	[12,18]	$< 10^{-11}$
RCB-Reg	-1546.612	6	3105.224	3132.067	63.96	[18,22]	$< 10^{-5}$
BB-Reg	-1429.605	6	2871.211	2898.054	19.40	[17,23]	> 0.3063
MixLinkJ2	-1433.331	5	2876.662	2905.506	19.50	[18,23]	> 0.3615

Table 2: Maximum likelihood estimates for candidate models, with standard errors in parentheses.

	Logistic		RCB		BB
β_0	-3.0306 (0.0246)	β_0	-2.9901 (0.0352)	β_0	-2.9487 (0.0445)
β_1	1.3017 (0.0343)	β_1	1.2040 (0.0415)	β_1	1.1144 (0.0550)
β_2	-0.3071 (0.0158)	β_2	-0.3429 (0.0242)	β_2	-0.2676 (0.0276)
		ϕ	0.1511 (0.0080)	ϕ	0.1661 (0.0076)

	RCB-Reg		BB-Reg		MixLinkJ2
β_0	-3.0699 (0.0338)	β_0	-3.0145 (0.0445)	β_0	-3.0061 (0.0441)
β_1	1.3010 (0.0444)	β_1	1.3594 (0.0564)	β_1	1.3656 (0.0562)
β_2	-0.3705 (0.0244)	β_2	-0.3449 (0.0332)	β_2	-0.3383 (0.0314)
γ_0	-2.3526 (0.0965)	γ_0	-1.8611 (0.0737)	π_1	0.3297 (0.0175)
γ_1	0.9331 (0.1569)	γ_1	0.7993 (0.1109)	κ	1.6293 (0.2472)
γ_2	-0.2365 (0.0565)	γ_2	-0.1610 (0.0525)		

BB-Reg have been considered in (Morel and Neerchal, 2012). The quadratic effect on the regression model was previously suggested in (Sofuni et al., 1978). Morel and Neerchal (2012) consider linking the regression to the overdispersion parameter in RCB and BB, in addition to the probability of aberration, indicating that the amount of overdispersion also varies with radiation dose.

The MLEs and corresponding standard errors for the candidate models are given in Table 2. Model BB gives estimates that differ the most; these tend to be shrunken to zero a bit more than in the other models. However, for the most part, all models give similar estimates of $\boldsymbol{\beta}$. The standard errors for $\boldsymbol{\beta}$ given by the two best models BB-Reg and MixLinkJ2 are also very similar.

Table 1 shows the GOF test result for each model along with other standard fit statistics: $-2\,\text{LogLik}$, Akaike information criteria (AIC) and Bayesian information criterion (BIC). Here, LogLik is the maximized value of the log-likelihood so that $\text{AIC} = -2\,\text{LogLik} + 2q$ and $\text{BIC} = -2\,\text{LogLik} + q\log(n)$. We first consider the information theoretic (AIC and BIC) criteria; the smaller the AIC/BIC, the better the indication of model fit. As expected, Logistic results in the largest AIC/BIC because it does not account for the suspected overdispersion. The two RCB models have smaller AIC/BIC than Logistic, but not as small as in the two BB models. The BB-Reg model fares appears to fit significantly better than BB, indicating that the overdispersion parameter varies with radiation dose. The MixLinkJ2 model fits almost as well as BB-Reg, even without modeling $\boldsymbol{\pi}$ or κ as a function of radiation dose.

The GOF results give additional insight into the quality of the fits. For each model, the intervals I_ℓ were chosen by first considering $I_1 = [0, 0.0099], I_2 = (0.0099, 0.0198], \ldots, I_{r-1} = (0.2970, 0.3069], I_r = (0.3069, 1]$. This partitioning was selected so the results can be compared to (Morel and Neerchal, 2012). Using the (ungrouped) MLE for the model, expected counts for each I_ℓ were computed, and I_ℓ having expected counts less than 5 were merged with a neighboring interval. Table 1 gives a summary of the GOF test results and Figure 10 shows additional detail comparing observed and expected counts. The GOF comparison gives a similar ranking of models as the AIC/BIC comparison, except that RCB-Reg is seen to have significantly worse GOF that RCB. Also, the MixLinkJ2 model gives a slightly higher p-value than BB-Reg, indicating less evidence against adequate fit, due to having one less parameter. The BB-Reg and MixLinkJ2 models both give a statistically adequate fit and obtain similar plots in Figure 10. One feature which seems to be a challenge for the other models to capture is the large number of observations with a very low proportion of aberrations; these are counted in the first interval. We have not shown results for ZIB regression, which was designed to capture this feature. At best (with or without a regression on the inflation parameter) ZIB was able to fit the first interval well, but did not fit other features of the data as well as any of the RCB, BB, or Mixture Link models.

6. Conclusions

In this paper, we have presented a new binomial model with extra variation called Mixture Link, starting from the finite mixture of binomials and linking a regression to the mixture probability of success. This lead us to consider a random effects model on the set representing the link from the likelihood to the regression; a Dirichlet distribution was placed on the simplex between extreme points of the set. Plots of the Mixture Link density show that it takes on a variety of expressive shapes. As a promising first application, Mixture Link is shown to fit the Hiroshima data well in terms of AIC/BIC and goodness-of-fit. Many of the computational details for the model have been omitted in this paper, but are given in the Ph.D. thesis (Raim, 2013). Initial results for Mixture Link are encouraging, and the model appears worthy of further study as a tool for the analysis of binomial data.

Acknowledgements

The computational resources used for this work were provided by the High Performance Computing Facility at the University of Maryland, Baltimore County (`www.umbc.edu/hpcf`). The first author additionally thanks the facility for financial support as an RA.

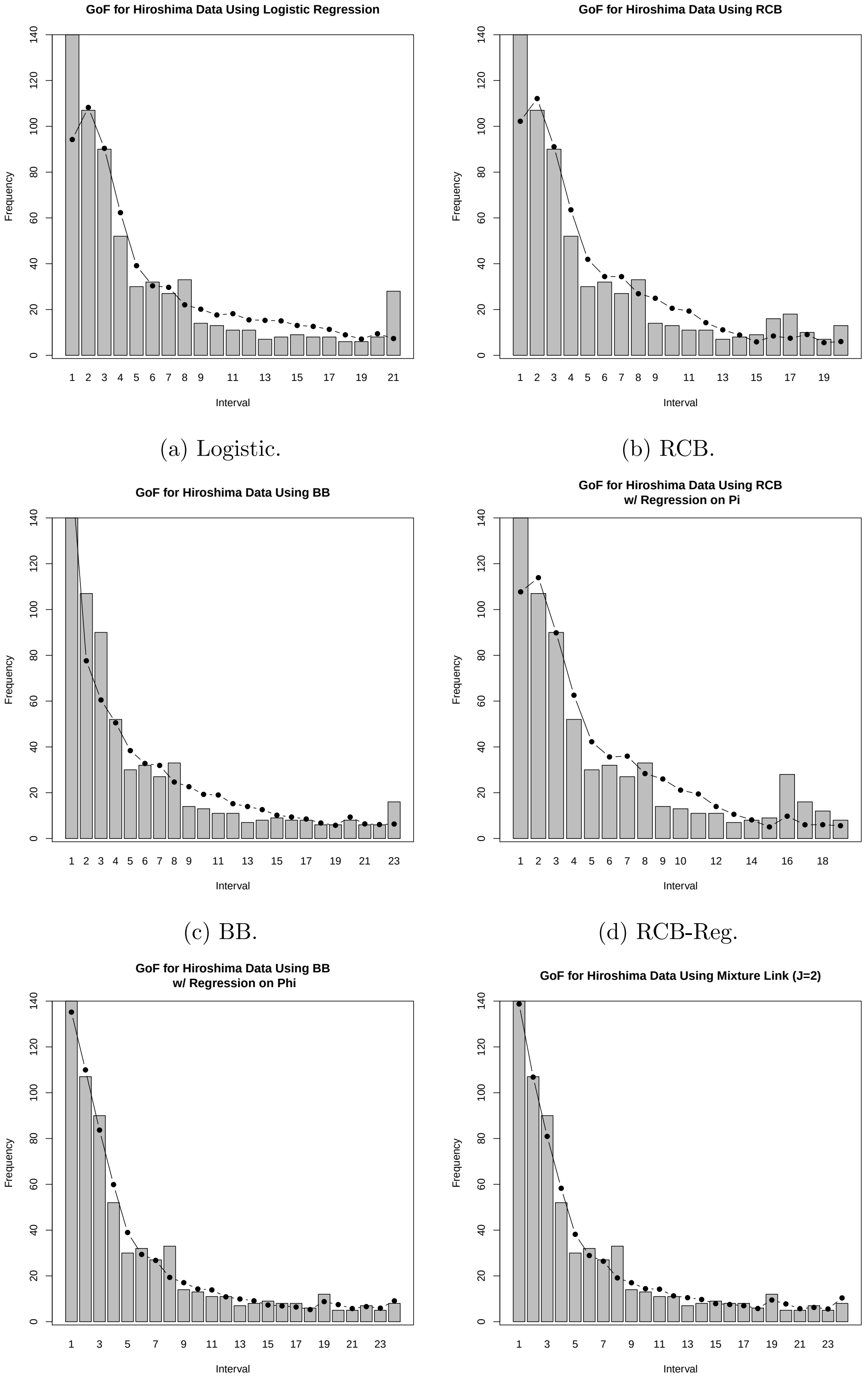
References

A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2nd edition, 2002.

M. Aitkin. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6:251–262, 1996.

A. Awa, T. Honda, T. Sofuni, S. Neriishi, M. Yoshida, and T. Matsui. Chromosome-aberration frequency in cultured blood-cells in relation to radiation dose of A-bomb survivor. *The Lancet*, 298(7730):903–905, 1971.

M. R. Danaher, A. Roy, Z. Chen, S. L. Mumford, and E. F. Schisterman. Minkowski-Weyl priors for models with parameter constraints: An analysis of the biocycle study. *Journal of the American Statistical Association*, 107(500):1395–1409, 2012.



(a) Logistic.

(b) RCB.

(c) BB.

(d) RCB-Reg.

(e) BB-Reg.

(f) MixLinkJ2.

Figure 10: GOF plots for observed vs. expected counts. The grey bars represent the observed counts for a given interval, and the black dots are the expected counts under the MLE. Note that the choice of intervals varies between models.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models.* Springer, 2006.

D. B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):1030–1039, 2000.

M. Kendall, A. Stuart, and K. Ord. *Kendalls Advanced Theory of Statistics*, volume 2. Oxford University Press, USA, 5th edition, 1991.

K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall/CRC, 2nd edition, 1989.

C. E. McCulloch, S. R. Searle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models*, volume 2. Wiley-Interscience, 2nd edition, 2008.

J. G. Morel and N. K. Nagaraj. A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80(2):363–371, 1993.

J. G. Morel and N. K. Neerchal. *Overdispersion Models in SAS.* SAS Institute, 2012.

N. K. Neerchal and J. G. Morel. Large cluster results for two parametric multinomial extra variation models. *Journal of the American Statistical Association*, 93(443): 1078–1087, 1998.

R. L. Prentice. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81(394):321–327, 1986.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

A. M. Raim. *Computational Methods for Finite Mixtures using Approximate Information and Regression Linked to the Mixture Mean.* PhD thesis, University of Maryland, Baltimore County, 2013. (In progress).

T. Sofuni, T. Honda, M. Itoh, S. Neriishi, and M. Otake. Relationship between the radiation dose and chromosome aberrations in atomic bomb survivors of hiroshima and nagasaki. *Journal of Radiation Research*, 19(2):126–140, 1978.

S. C. Sutradhar, N. K. Neerchal, and J. G. Morel. A goodness-of-fit test for overdispersed binomial (or multinomial) models. *Journal of Statistical Planning and Inference*, 138(5):1459–1471, 2008.