Mixture Link Models for Binomial Data with Overdispersion

Andrew M. Raim

Center for Statistical Research & Methodology
U.S. Census Bureau
andrew.raim@gmail.com

2015 Joint Statistical Meetings
Seattle, WA, U.S.A.

Joint work with Nagaraj K. Neerchal (UMBC) and Jorge G. Morel (UMBC)

Overview

- Overdispersion occurs when a given statistical model can not capture the variability observed in the data. It is commonly encountered in the analysis of categorical and count data.

- The Mixture Link Binomial distribution was proposed in Raim (2014, Ph.D. Thesis) as a model for overdispersed binomial data.

- We will discuss a motivating example, the model, and application to a classical dataset on chromosome aberrations in atomic bomb survivors.

United States™ Census Bureau

Regression in a Heterogeneous Population

- Suppose there are J possible regression functions

$$\boldsymbol{x}^T\boldsymbol{\beta}^{(1)}, \quad \ldots, \quad \boldsymbol{x}^T\boldsymbol{\beta}^{(J)}.$$

- Suppose $T_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, G(\boldsymbol{x}_i^T\boldsymbol{\beta}^{(Z_i)}))$, given a latent subpopulation label

$$Z_i = \begin{cases} 1 & \text{w.p. } \pi_1 \\ \vdots & \\ J & \text{w.p. } \pi_J. \end{cases}$$

  where G is an inverse link function such as the Logistic$(0, 1)$ CDF.

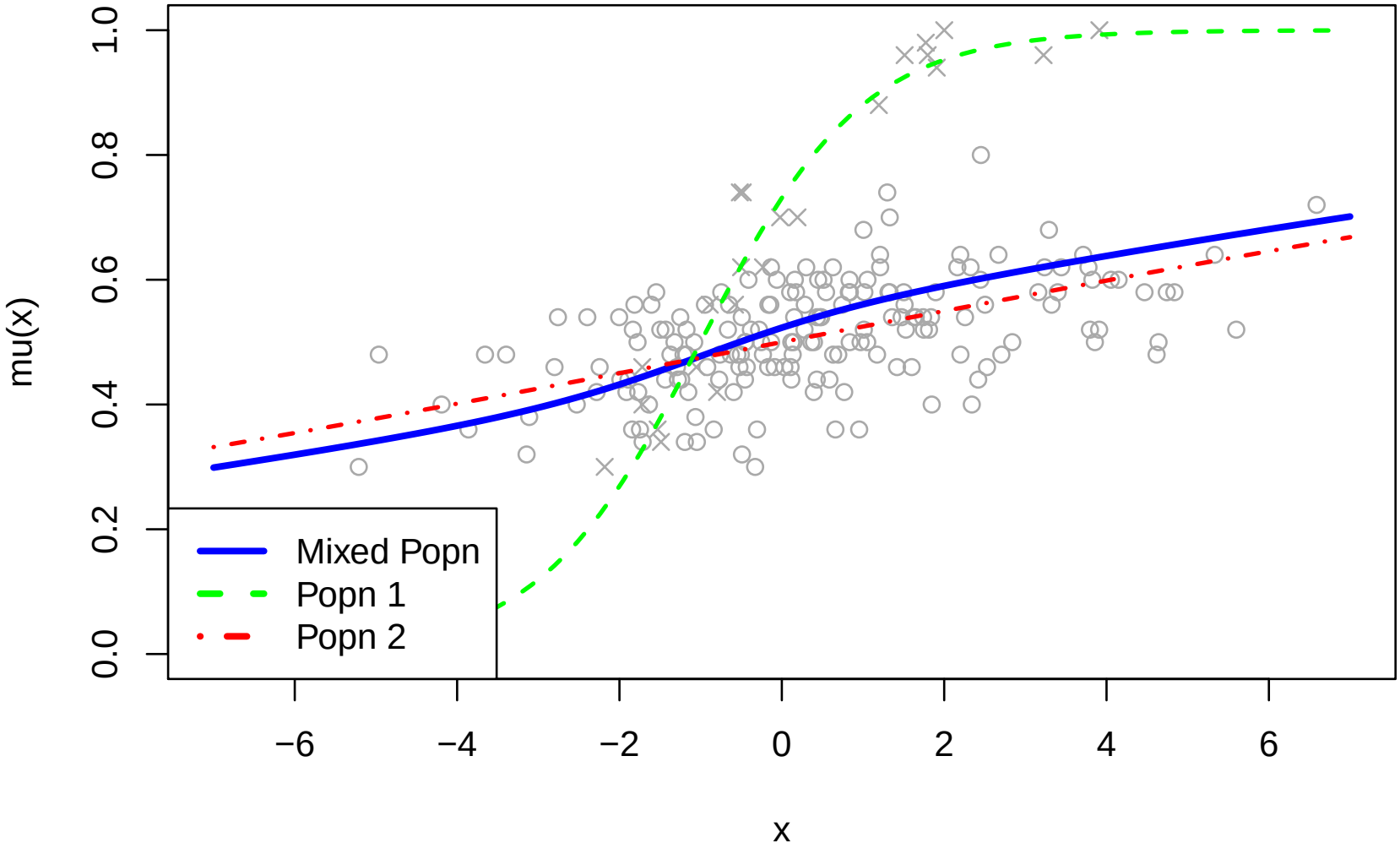
- The overall success probability of a single trial is

$$\text{E}\left(\frac{T}{m} \,\middle|\, \boldsymbol{x}\right) = \sum_{j=1}^{J} \pi_j G(\boldsymbol{x}^T\boldsymbol{\beta}^{(j)}).$$

Example

$$T_i \overset{\text{ind}}{\sim} \begin{cases} \text{Bin}[50, \mu_1(x_i)] & \text{w.p. } \pi_1 = 0.1, \\ \text{Bin}[50, \mu_2(x_i)] & \text{w.p. } \pi_2 = 0.9, \end{cases} \quad i = 1, \ldots, 200,$$

$$\mu_1(x) = G(1 + x), \quad \mu_2(x) = G(0 + 0.1x), \quad \mu(x) = \pi_1\mu_1(x) + \pi_2\mu_2(x)$$

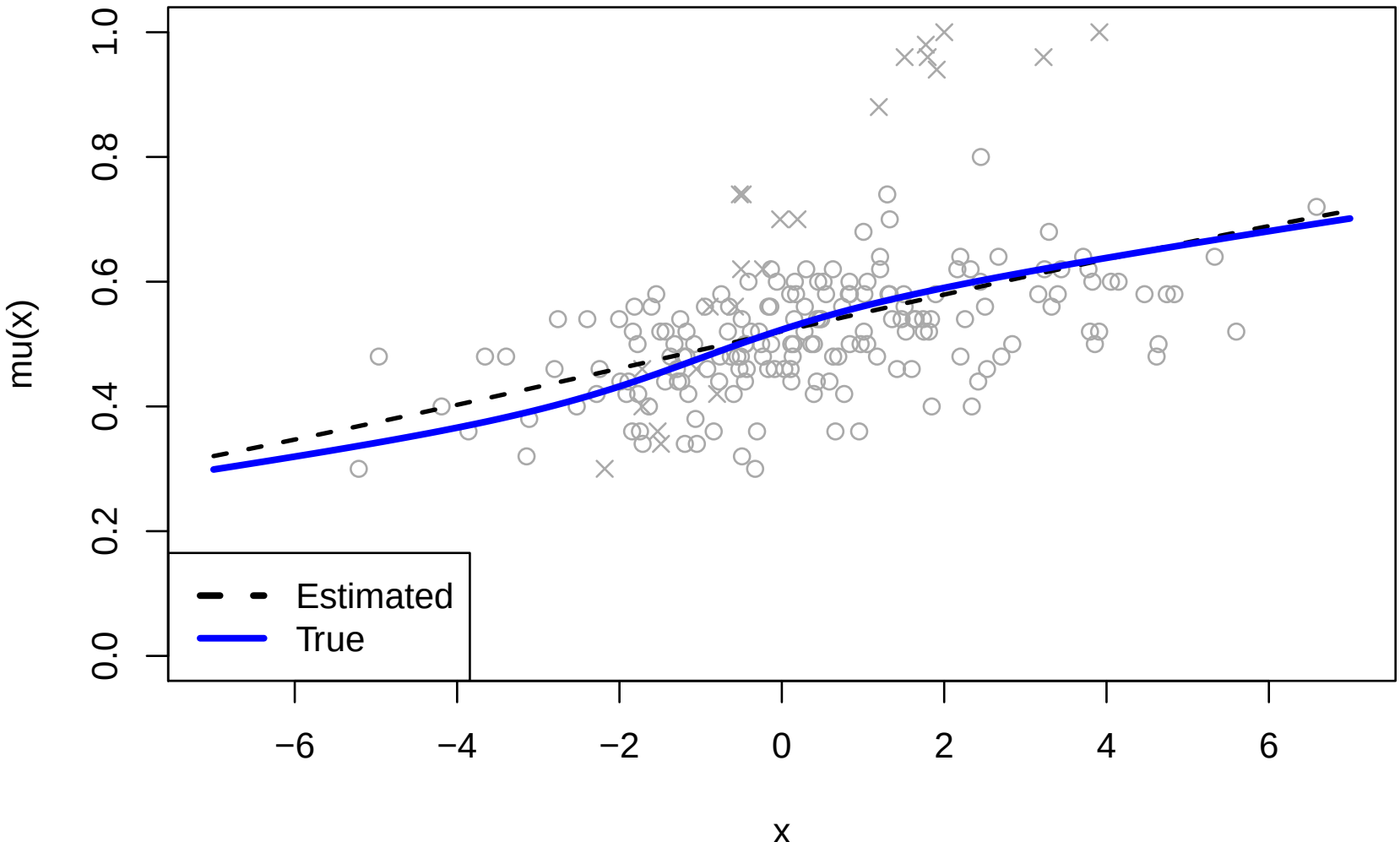


Example

Logistic Regression

	Estimate	SE	z-value	p-value
β_0	0.0817	0.0205	3.9890	< 0.0001
β_1	0.1191	0.0101	11.8010	< 0.0001

LogLik: -724.77 AIC: 1453.54 BIC: 1460.13



Randomized Quantile Residuals

- Dunn and Smyth (1996) propose randomized quantile residuals for diagnostics on GLMs and other non-normal models.
- Interpretation of residuals is similar to OLS residuals on a standard normal scale.
- For y_i independently drawn from a continuous distribution,

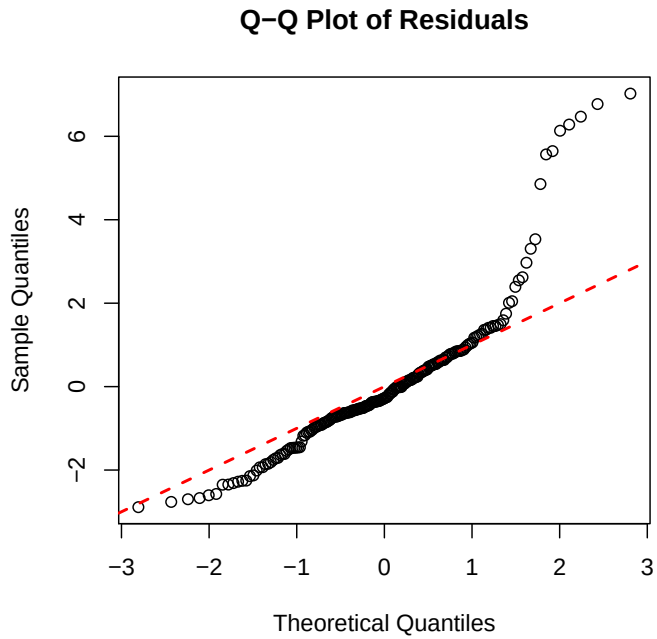
$$r_i = \Phi^{-1}\{F(y_i \mid \hat{\boldsymbol{\theta}})\}.$$

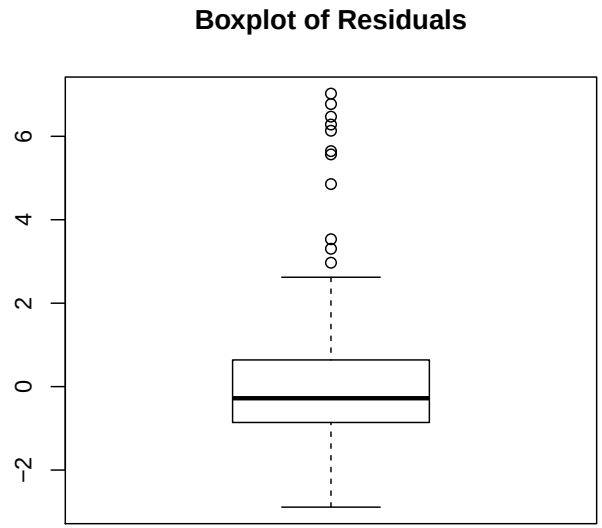
- For y_i independently drawn from a discrete distribution,

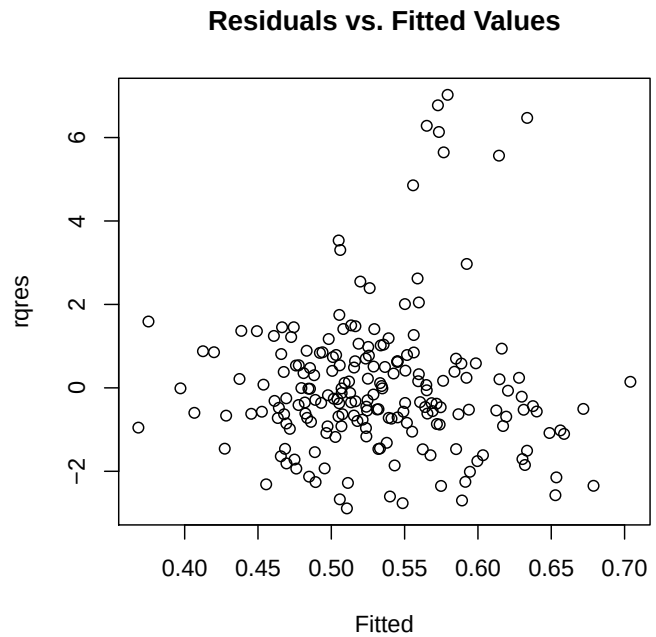
$$\begin{aligned} r_i &= \Phi^{-1}\{u_i\}, \\ u_i &\overset{\text{ind}}{\sim} U(a_i, b_i), \\ a_i &= \lim_{\varepsilon \downarrow 0} F(y_i - \varepsilon \mid \hat{\boldsymbol{\theta}}), \\ b_i &= F(y_i \mid \hat{\boldsymbol{\theta}}). \end{aligned}$$

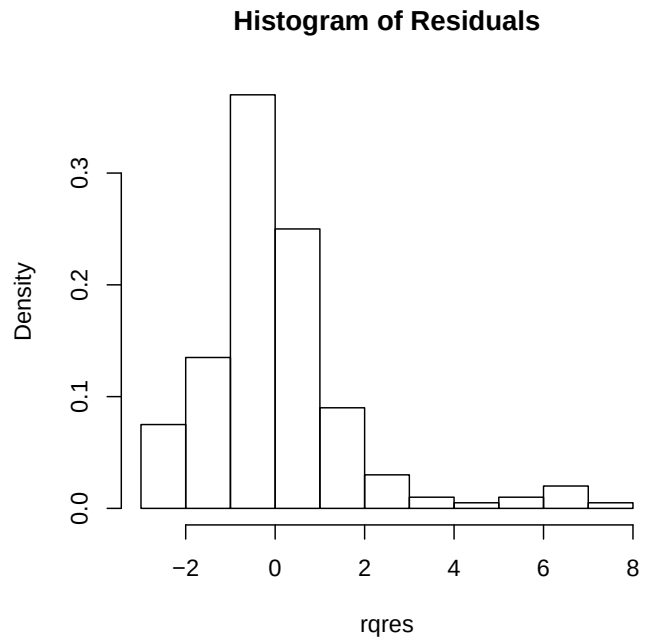
Example

Residuals









Binomial Regression Models for Overdispersion

Some Established Approaches

- Likelihoods which support overdispersion using latent random variables.
  1. Beta-Binomial (Otake and Prentice, 1984),
  2. Zero-Inflated Binomial (Hall, 2000)
  3. Random-Clumped Binomial (Morel and Nagaraj, 1993).
- Quasi-likelihood methods.
  1. Dispersion multiplier (Agresti, 2002, §4.7).
  2. Generalized Estimating Equations (Liang and Zeger, 1986).
- Generalized Linear Mixed Models (McCulloch, Searle, and Neuhaus, 2008).
- Finite mixtures of regressions (Frühwirth-Schnatter, 2006).

United States™ Census Bureau

Mixture Link Binomial Model

Formulation

- Start with a finite mixture of binomial densities,

$$T_i \overset{\text{ind}}{\sim} f(t \mid m_i, \boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j \binom{m_i}{t_i} \mu_{ij}^{t_i} (1-\mu_{ij})^{m-t},$$

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J) \in \mathcal{S}^J, \qquad \leftarrow \text{the probability simplex in } \mathbb{R}^J$$

$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iJ}) \in [0,1]^J, \qquad \leftarrow \text{the unit cube in } \mathbb{R}^J.$$

- Mixture success probability for a single trial is $\mathrm{E}(T_i/m_i) = \boldsymbol{\mu}_i^T \boldsymbol{\pi}$.

- **Objective**: Link $\boldsymbol{x}_i^T \beta$ to $\boldsymbol{\mu}_i^T \boldsymbol{\pi}$,

$$\boldsymbol{\mu}_i^T \boldsymbol{\pi} \overset{\text{link}}{=} p_i, \quad \text{where } p_i \overset{\text{def}}{=} G(\boldsymbol{x}_i^T \beta).$$

- To enforce the link, $\boldsymbol{\mu}_i$ must be in the set

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

United States Census Bureau

Mixture Link Binomial Model

Random Effects Approach

- $A_i = \{\boldsymbol{\mu} \in [0,1]^J : \boldsymbol{\mu}^T \boldsymbol{\pi} = p_i\}$ is a bounded convex set. Therefore we can find vertices $\boldsymbol{v}_1^{(i)}, \ldots, \boldsymbol{v}_{k_i}^{(i)} \in \mathbb{R}^J$ such that

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

- $\boldsymbol{V}^{(i)}$ can vary for each observation. Number of vertices (columns) k_i can also vary.

- If $\boldsymbol{\lambda}^{(i)} \overset{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\boldsymbol{\alpha})$, then $\boldsymbol{\mu}_i = \boldsymbol{V}^{(i)} \boldsymbol{\lambda}^{(i)}$ is a draw from A_i.

- A related approach was taken by Danaher et al. (2012). They use priors based on the Minkowski-Weyl decomposition to enforce (biologically motivated) polyhedral constraints for parameters in Bayesian analysis.

United States Census Bureau

Mixture Link Binomial Model

Hierarchical Model

We can write the model as

$$T_i \mid \boldsymbol{\mu}_i, \boldsymbol{\pi} \overset{\text{ind}}{\sim} \text{BinMix}(m_i, \boldsymbol{\mu}_i, \boldsymbol{\pi})$$

$$\boldsymbol{\mu}_i = \boldsymbol{V}^{(i)} \boldsymbol{\lambda}^{(i)}, \quad \text{where } \boldsymbol{V}^{(i)} = (\boldsymbol{v}_1^{(i)} \cdots \boldsymbol{v}_{k_i}^{(i)}) \text{ are vertices of } A(p_i, \boldsymbol{\pi})$$

$$\boldsymbol{\lambda}^{(i)} \overset{\text{ind}}{\sim} \text{Dirichlet}_{k_i}(\kappa, \ldots, \kappa).$$

Symmetric Dirichlet is assumed because:

- k_i can vary between observations.
- Difficult to identify vertices with distinct parameters.

$$\text{Density: } f(t \mid m, \boldsymbol{\theta}) = \binom{m}{t} \sum_{j=1}^{J} \pi_j \int w^t (1-w)^{m-t} \cdot f_{\boldsymbol{v}_{j\cdot}^T \boldsymbol{\lambda}}(w)\, dw$$

$$\text{Parameterized by: } \boldsymbol{\theta} = \begin{cases} (p, \boldsymbol{\pi}, \kappa) \in \mathbb{R}^{1+(J-1)+1}, & \text{no-regression case,} \\ (\boldsymbol{\beta}, \boldsymbol{\pi}, \kappa) \in \mathbb{R}^{d+(J-1)+1}, & \text{regression case.} \end{cases}$$

United States Census Bureau

Mixture Link Binomial Model

Details

- The vertices $\boldsymbol{V}^{(i)}$ of $A(p_i, \boldsymbol{\pi})$ can be enumerated in $O(J \cdot 2^{J-1})$ steps.
- The expectation and variance of $T \sim \text{MixLink}_J(m, p, \boldsymbol{\pi}, \kappa)$ can be obtained as

$$\text{E}(T) = mp,$$

$$\text{Var}(T) = mp\,(1 - mp) + m(m-1) \sum_{j=1}^{J} \pi_j \frac{\boldsymbol{v}_{j\cdot}^T \boldsymbol{v}_{j\cdot} + \kappa(k\bar{v}_{j\cdot})^2}{k(1+\kappa k)}.$$

- Moment-based estimators of p (no-regression case) and κ can be obtained from above. An estimator for $\boldsymbol{\beta}$ can be obtained by Gauss-Newton method.
- Computation of the density:
  1. Exact computation following Provost and Cheong (2000) to calculate linear combination of Dirichlet density.
  2. Beta approximation to linear combination of Dirichlet density by moment-matching.

Chromosome Aberration Data

An illustrative dataset used in (Morel and Neerchal, 2012), from Awa et al. (1978).

Chromosome aberrations were studied in Hiroshima atomic bomb survivors between Jan 1968 and Nov 1969

- $n = 648$ subjects
- m_i: number of circulating lymphocytes examined on the ith subject (between 30 and 100)
- t_i: count with chromosome aberrations
- d_i: total radiation dose (T65-gamma + T65-neutron, in rads) received by the ith subject
- $z_i = \frac{d_i - \bar{d}}{\sqrt{\frac{1}{n}\sum_{\ell=1}^{n}(d_\ell - \bar{d})^2}}$: standardized radiation dose

for $i = 1, \ldots, n$.

Qn: What is the effect of radiation dose on the probability of chromosome aberration?

Chromosome Aberration Data

Compare models for goodness-of-fit:

- 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: $T_i \overset{\text{ind}}{\sim} \text{MixLink}_2(m_i, p_i, \boldsymbol{\pi}, \kappa)$,

with regressions

- $\text{logit}(p_i) = \beta_0 + \beta_1 z_i + \beta_2 z_i^2$ for all models,
- $\text{logit}(\phi_i) = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2$ for the two "-Reg" models.

Numerical MLE used for all models in this study.

Chromosome Aberration Data

Maximum Likelihood Estimates

	Logistic
β_0	-3.0306 (0.0246)
β_1	1.3017 (0.0343)
β_2	-0.3071 (0.0158)

	RCB
β_0	-2.9901 (0.0352)
β_1	1.2040 (0.0415)
β_2	-0.3429 (0.0242)
ϕ	0.1511 (0.0080)

	BB
β_0	-2.9487 (0.0445)
β_1	1.1144 (0.0550)
β_2	-0.2676 (0.0276)
ϕ	0.1661 (0.0076)

	RCB-Reg
β_0	-3.0699 (0.0338)
β_1	1.3010 (0.0444)
β_2	-0.3705 (0.0244)
γ_0	-2.3526 (0.0965)
γ_1	0.9331 (0.1569)
γ_2	-0.2365 (0.0565)

	BB-Reg
β_0	-3.0145 (0.0445)
β_1	1.3594 (0.0564)
β_2	-0.3449 (0.0332)
γ_0	-1.8611 (0.0737)
γ_1	0.7993 (0.1109)
γ_2	-0.1610 (0.0525)

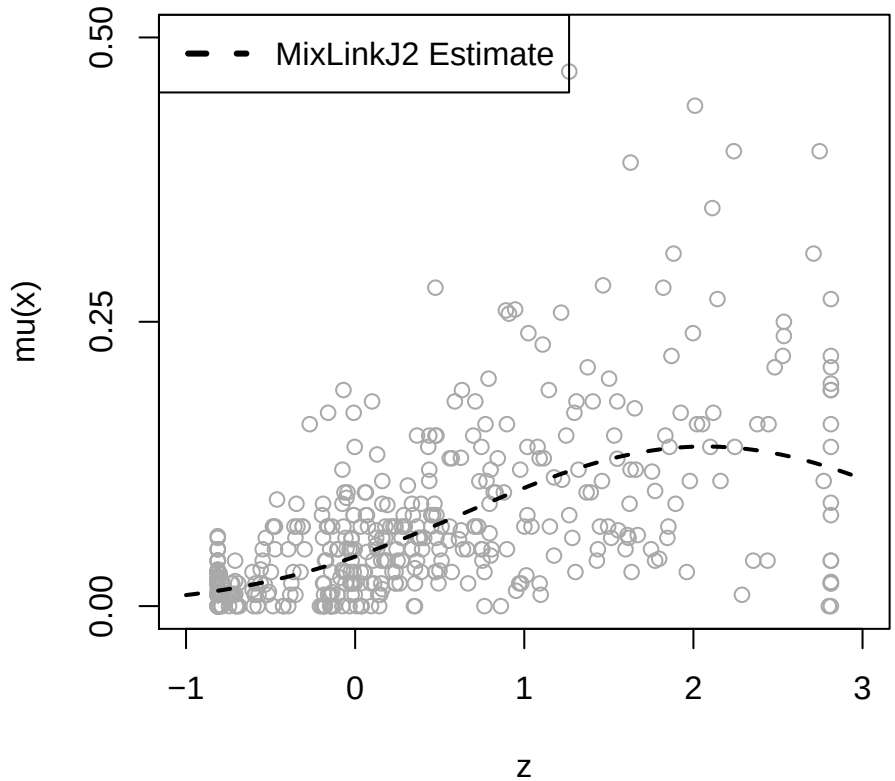
	MixLinkJ2
β_0	-3.0061 (0.0441)
β_1	1.3656 (0.0562)
β_2	-0.3383 (0.0314)
π_1	0.3297 (0.0175)
κ	1.6293 (0.2472)

(Standard errors using Hessian are in parentheses.)

United States Census Bureau

Chromosome Aberration Data

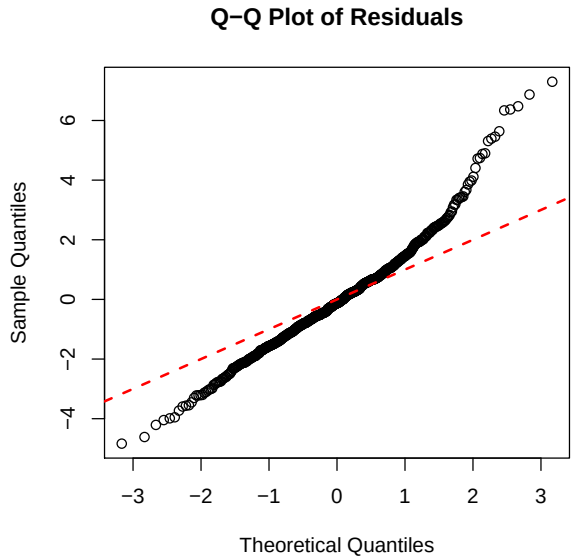
Model Comparison

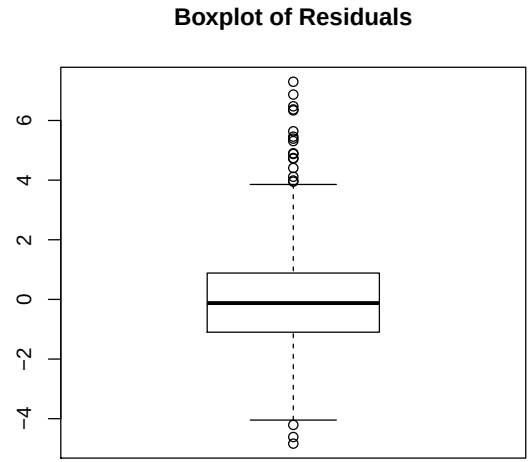


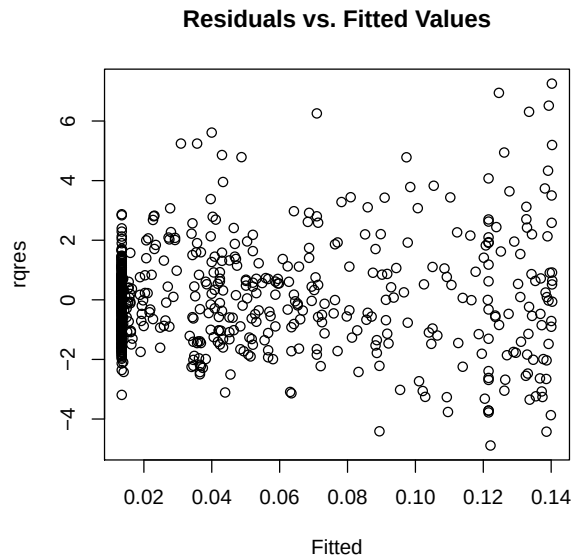
Model	LogLik ↑	dim θ	AIC	BIC
Logistic	-1814.19	3	3634.40	3647.80
RCB	-1567.50	4	3143.00	3160.90
RCB-Reg	-1546.61	6	3105.22	3132.07
BB	-1487.92	4	2983.85	3001.74
MixLinkJ2	-1433.33	5	2876.66	2905.51
BB-Reg	-1429.61	6	2871.21	2898.05

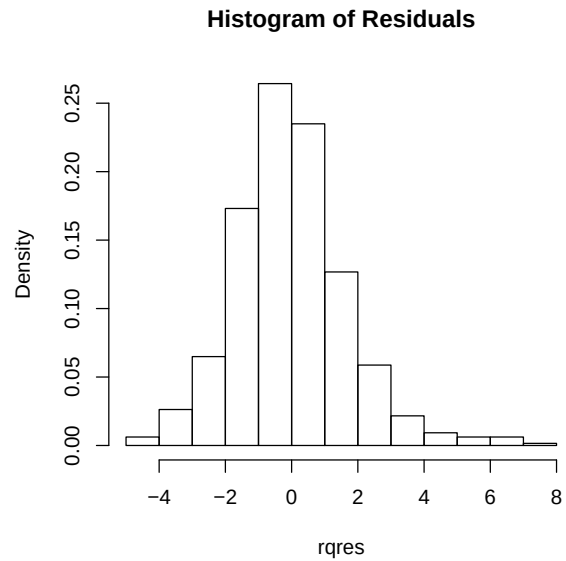
Chromosome Aberration Data

Quantile Residuals for Logistic



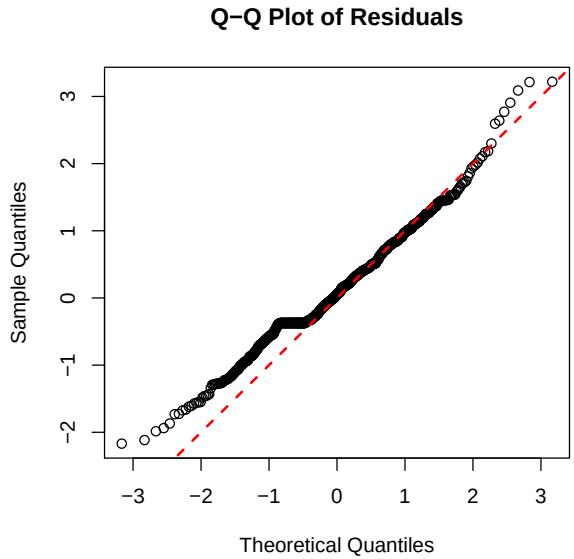


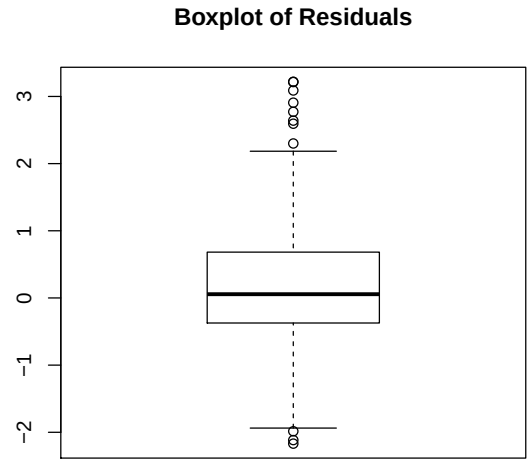


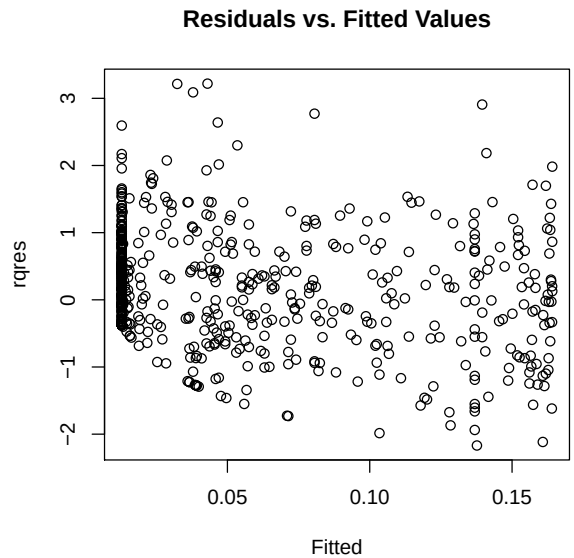


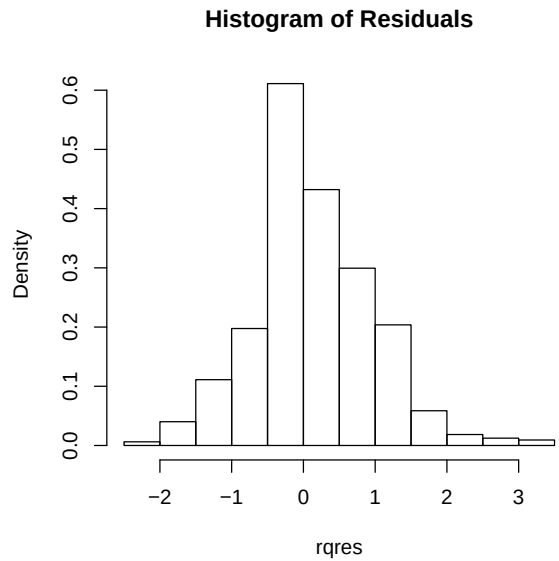
Chromosome Aberration Data

Quantile Residuals for MixLinkJ2









Conclusions

Conclusions

- Overdispersion can have an especially detrimental effect on model-dependent quantities such as quantile residuals.
- Mixture Link Binomial is able to capture some of the large variation observed in the Hiroshima dataset.
- Mixture Link can be considered among the likelihood-based models for overdispersed binomial data.

Future Work

- Bayesian inference.
- Effect of increasing J.
- Other outcome types: Normal, Poisson, etc.

United States™ Census Bureau

References I

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

A. Awa, 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.

Michelle R. Danaher, Anindya Roy, Zhen Chen, Sunni L. Mumford, and Enrique 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.

Peter K. Dunn and Gordon K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

Dean A. Follmann and Diane Lambert. Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84(405): 295–300, 1989.

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

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

United States™ Census Bureau

References II

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Charles E. McCulloch, Shayle R. Searle, and John M. Neuhaus. *Generalized, Linear, and Mixed Models*, volume 2. Wiley-Interscience, 2nd edition, 2008.

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

Jorge G. Morel and Nagaraj K. Neerchal. *Overdispersion Models in SAS*. SAS Institute, 2012.

Masanori Otake and Ross L. Prentice. The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Radiation Research*, 98(3):456–470, 1984.

Serge B. Provost and Young-Ho Cheong. On the distribution of linear combinations of the components of a dirichlet random vector. *Canadian Journal of Statistics*, 28(2):417–425, 2000.

Andrew M. Raim. Computational methods in finite mixtures using approximate information and regression linked to the mixture mean. Ph.D. Thesis, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 2014.

United States Census Bureau