A new regression model for overdispersed binomial data accounting for outliers and an excess of zeros

Roberto Ascari - roberto.ascari@unimib.it
Joint work with Sonia Migliorati

Milano-Bicocca

November 17, 2021

A motivating example

Introduction

Let us consider a real dataset from Otake and Prentice regarding the survivors of the atomic bombs in Hiroshima and Nagasaki:

- Y: number of cells with chromosomal abnormalities among 100 cells.
- X: estimated radiation exposure levels (in rads).

RADIATION RESEARCH **98,** 456–470 (1984)

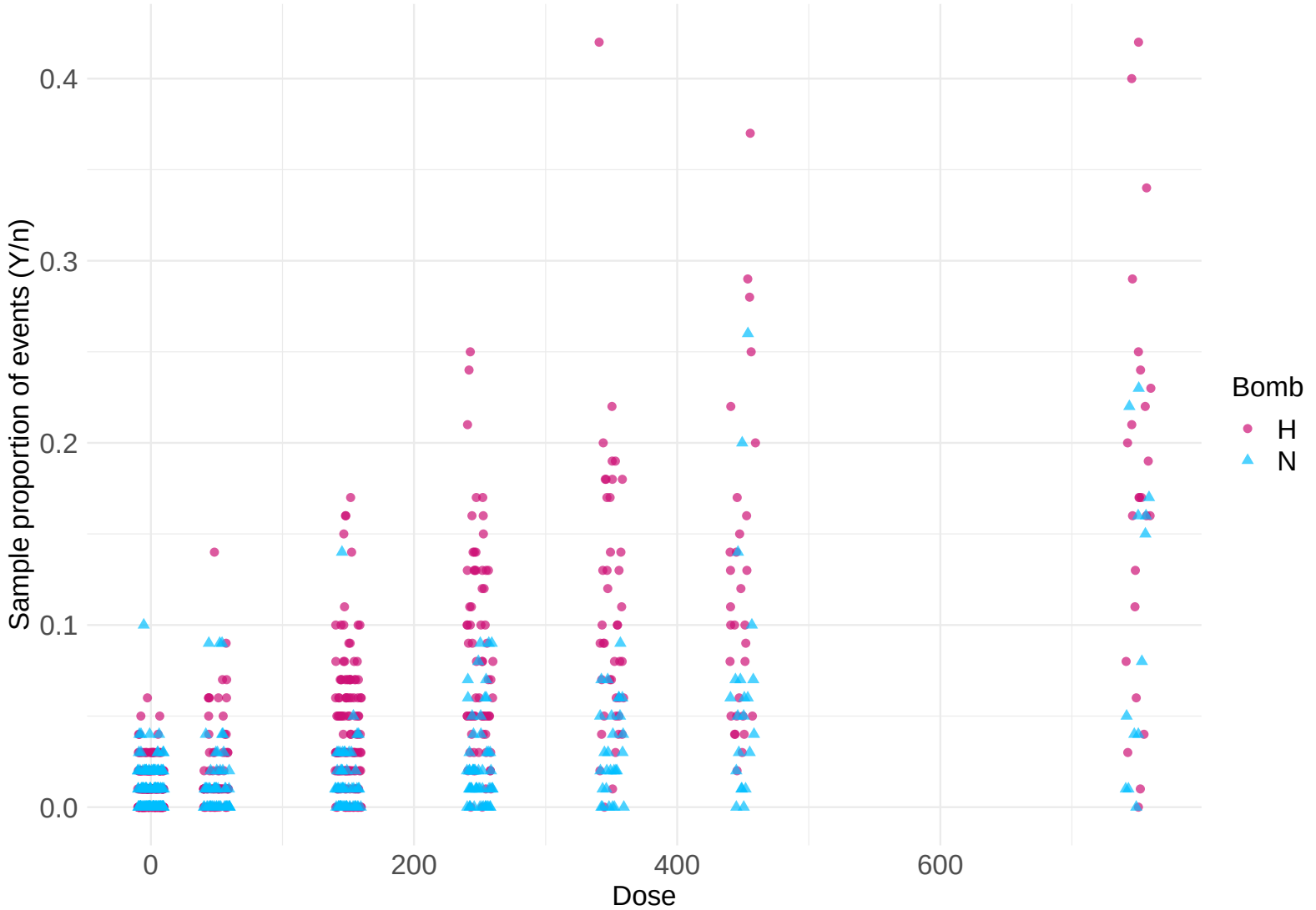
The Analysis of Chromosomally Aberrant Cells Based on Beta-Binomial Distribution[1]

MASANORI OTAKE* AND ROSS L. PRENTICE†

*Department of Epidemiology and Statistics, Radiation Effects Research Foundation, Hiroshima, Japan
†Department of Biostatistics, University of Washington, Seattle, Washington 98195

OTAKE, M., AND PRENTICE, R. L. The Analysis of Chromosomally Aberrant Cells Based on Beta-Binomial Distribution. *Radiat. Res.* **98,** 456–470 (1984).

Introduction



Introduction

In a (Bayesian) regression perspective, we can assume that:

- $Y_i|\beta_0, \beta_1 \sim \text{Bin}(n_i = 100, \pi_i)$, $i = 1, \ldots, N$.
- $\pi_i = \text{inv.logit}(\beta_0 + \beta_1 x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$,
- β_0 and β_1 follow independent Normal$(0, \tau^2)$ **prior** distributions.

A sample of size B from the **posterior** distribution of (β_0, β_1) can be obtained through the Hamiltonian Monte Carlo (HMC) algorithm.

Parameter	Post. Mean	Post. SD	95% CS
β_0	-4.1300	0.0256	(-4.184, -4.081)
β_1	0.0037	0.0001	(0.0036, 0.0039)

Table: Regression results: Posterior Mean, Standard Error (SD), and 95% Credible Set (CS) for each parameter.

Introduction

In a (Bayesian) regression perspective, we can assume that:

- $Y_i|\beta_0, \beta_1 \sim \text{Bin}(n_i = 100, \pi_i)$, $i = 1, \ldots, N$.
- $\pi_i = \text{inv.logit}(\beta_0 + \beta_1 x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$,
- β_0 and β_1 follow independent Normal$(0, \tau^2)$ **prior** distributions.

A sample of size B from the **posterior** distribution of (β_0, β_1) can be obtained through the Hamiltonian Monte Carlo (HMC) algorithm.

Parameter	Post. Mean	Post. SD	95% CS
β_0	-4.1300	0.0256	(-4.184, -4.081)
β_1	0.0037	0.0001	(0.0036, 0.0039)

Table: Regression results: Posterior Mean, Standard Error (SD), and 95% Credible Set (CS) for each parameter.

Model checking

Posterior Predictive Checks

Model checking: Posterior Predictive Checks

$$
\begin{array}{ccccc}
(\beta_0, \beta_1)^{(1)} & \longrightarrow & \mathbf{y}^{(1)} & \longrightarrow & T\left(\mathbf{y}^{(1)}\right) \\
(\beta_0, \beta_1)^{(2)} & \longrightarrow & \mathbf{y}^{(2)} & \longrightarrow & T\left(\mathbf{y}^{(2)}\right) \\
\vdots & & \vdots & & \vdots \\
(\beta_0, \beta_1)^{(B)} & \longrightarrow & \mathbf{y}^{(B)} & \longrightarrow & T\left(\mathbf{y}^{(B)}\right)
\end{array}
$$

where:

- $\mathbf{y}^{(b)} = \left(y_1^{(b)}, \ldots, y_N^{(b)}\right)^{\mathsf{T}}$ is a **replicated dataset**,
- $y_i^{(b)} \sim \text{Bin}\left(n_i, \text{inv.logit}\left(\beta_0^{(b)} + x_i \beta_1^{(b)}\right)\right)$,
- $T(\cdot)$ is a *discrepancy measure*, that is a function of data and (eventually) model parameters.

Posterior Predictive Checks

Model checking: Posterior Predictive Checks

$$
\begin{array}{ccccc}
(\beta_0, \beta_1)^{(1)} & \longrightarrow & \mathbf{y}^{(1)} & \longrightarrow & T\left(\mathbf{y}^{(1)}\right) \\
(\beta_0, \beta_1)^{(2)} & \longrightarrow & \mathbf{y}^{(2)} & \longrightarrow & T\left(\mathbf{y}^{(2)}\right) \\
\vdots & & \vdots & & \vdots \\
(\beta_0, \beta_1)^{(B)} & \longrightarrow & \mathbf{y}^{(B)} & \longrightarrow & T\left(\mathbf{y}^{(B)}\right)
\end{array}
$$

where:

- $\mathbf{y}^{(b)} = \left(y_1^{(b)}, \ldots, y_N^{(b)}\right)^{\mathsf{T}}$ is a **replicated dataset**,
- $y_i^{(b)} \sim \text{Bin}\left(n_i, \text{inv.logit}\left(\beta_0^{(b)} + x_i \beta_1^{(b)}\right)\right)$,
- $T(\cdot)$ is a *discrepancy measure*, that is a function of data and (eventually) model parameters.

Posterior Predictive Checks

Model checking: Posterior Predictive Checks

$$
\begin{array}{ccccc}
(\beta_0, \beta_1)^{(1)} & \longrightarrow & \mathbf{y}^{(1)} & \longrightarrow & T\left(\mathbf{y}^{(1)}\right) \\
(\beta_0, \beta_1)^{(2)} & \longrightarrow & \mathbf{y}^{(2)} & \longrightarrow & T\left(\mathbf{y}^{(2)}\right) \\
\vdots & & \vdots & & \vdots \\
(\beta_0, \beta_1)^{(B)} & \longrightarrow & \mathbf{y}^{(B)} & \longrightarrow & T\left(\mathbf{y}^{(B)}\right)
\end{array}
$$

where:

- $\mathbf{y}^{(b)} = \left(y_1^{(b)}, \ldots, y_N^{(b)}\right)^{\mathsf{T}}$ is a **replicated dataset**,
- $y_i^{(b)} \sim \text{Bin}\left(n_i, \text{inv.logit}\left(\beta_0^{(b)} + x_i \beta_1^{(b)}\right)\right)$,
- $T(\cdot)$ is a *discrepancy measure*, that is a function of data and (eventually) model parameters.

Posterior Predictive Checks

Posterior predictive checks compare the empirical distribution of $T(\mathbf{y}^{(b)})$ to the value of $T(\mathbf{y})$:

- Plots (histograms, density plots, boxplots, etc.),
- Posterior predictive p-values: $P(T(\mathbf{y}^{(b)}) \geq T(\mathbf{y})|\mathbf{y})$ (the closer to 0.5, the better).

Posterior Predictive Checks

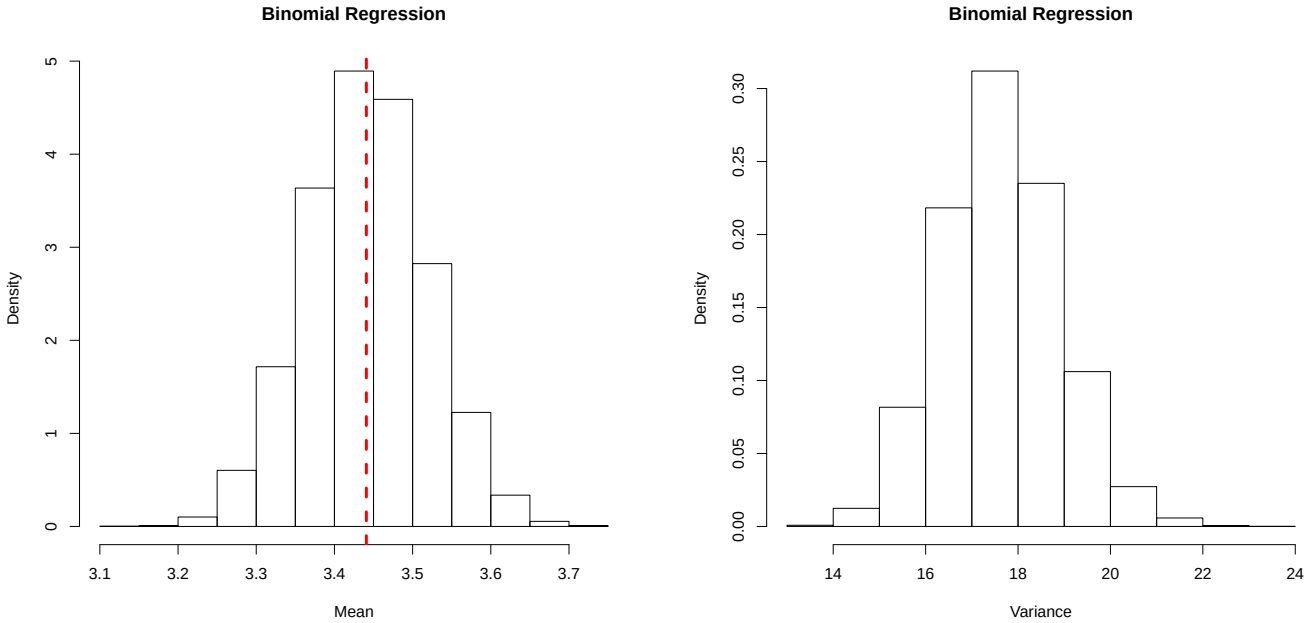


Figure: Posterior Predictive Checks.

	Mean	Variance
Observed	3.4408	30.2101
Post. pred. p-value	0.5025	0

Table: Observed test statistics and associated posterior predictive p-values.

The beta-binomial distribution

The beta-binomial distribution

The most popular model for overdispersed data is the **beta-binomial** (BB) distribution, that allows the success probability parameter π to vary according to a Beta distribution, that is:

$$\begin{aligned} Y|\Pi = \pi &\sim Bin(n, \pi) \\ \Pi &\sim Beta(\mu, \phi), \end{aligned}$$

where $\mathbb{E}[\Pi] = \mu$ and $\text{Var}(\Pi) = \frac{\mu(1-\mu)}{\phi+1}$. Then, Y is marginally distributed according to a $\text{BB}(n, \mu, \phi)$.

The beta-binomial distribution

Let $Y \sim \text{BB}(n, \mu, \phi)$, then its p.m.f. is:

$$\begin{aligned} f_{BB}(y; n, \mu, \phi) &= \int_0^1 f_{Bin}(y; n, \pi) f_B(\pi; \mu, \phi) d\pi \\ &= \binom{n}{y} \frac{B(\phi\mu + y, \phi(1-\mu) + n - y)}{B(\phi\mu, \phi(1-\mu))}, \end{aligned}$$

where $y \in \{0, 1, \ldots, n\}$, $\mu \in (0, 1)$, $\phi > 0$, and $B(\cdot, \cdot)$ is the Beta function.

The beta-binomial distribution

Let $Y \sim \text{BB}(n, \mu, \phi)$, then:

- $\mathbb{E}[Y] = n\mu \Longrightarrow \mu = \mathbb{E}[Y/n]$
- $\text{Var}(Y) = n\mu(1-\mu)[1 + \theta(n-1)]$,

where $\theta = \frac{1}{1+\phi} \in (0, 1)$.

When $\theta \to 0$, $\text{Var}(Y)$ approaches the binomial variance.

The parameter θ can be thought as an overdispersion parameter.

The beta-binomial distribution

Let us suppose that $U_1, \ldots, U_n$ are the n Bernoulli variables such that

- $Y = \sum_{r=1}^{n} U_r \sim \text{BB}(n, \mu, \phi)$,
- $U_1, \ldots, U_n$ are identically distributed but not independent.

Then $\text{Corr}(U_l, U_w) = \dfrac{1}{1+\phi} = \theta$ for any $l \neq w$.

The parameter θ is the **common** correlation between pairs of the Bernoulli variables forming the response count Y (i.e., it represents the **intraclass correlation coefficient** (ICC)).

The beta-binomial distribution

The binomial distribution assumes that $U_1, \ldots, U_n$ are identically distributed **and independent**.

Thus, under the binomial model, $\text{Corr}(U_l, U_w) = 0$ for any $l \neq w$.

A well known cause of overdispersion is the failure of the i.i.d. assumption of the individual binary responses (e.g., dependency due to belonging to the same litter, family, class, etc.)

$\Downarrow$

Introducing a specific parameter to handle this issue improves the fit!

The beta-binomial distribution

The binomial distribution assumes that $U_1, \ldots, U_n$ are identically distributed **and independent**.

Thus, under the binomial model, $\text{Corr}\,(U_l, U_w) = 0$ for any $l \neq w$.

A well known cause of overdispersion is the failure of the i.i.d. assumption of the individual binary responses (e.g., dependency due to belonging to the same litter, family, class, etc.)

$$\Downarrow$$

Introducing a specific parameter to handle this issue improves the fit!

Bayesian approach - BBReg

Let

- $Y_i|\boldsymbol{\beta}, \phi \sim \text{BB}(n_i, \mu_i, \phi)$ for $i = 1, \ldots, N$,
- $\mu_i = \text{inv.logit}(\mathbf{x}_i^\intercal \boldsymbol{\beta})$,

where $\mathbf{x}_i$ is a row of the design matrix $\mathbf{X}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)^\intercal$ is the vector of unknown regression coefficients.

Prior Elicitation:

- $\boldsymbol{\beta} \sim \text{Normal}_{K+1}\left(\mathbf{0}, \text{diag}(\tau^2, \ldots, \tau^2)\right)$
- $\dfrac{1}{1+\phi} \sim \text{Unif}(0, 1)$

Results: Posterior Predictive Checks

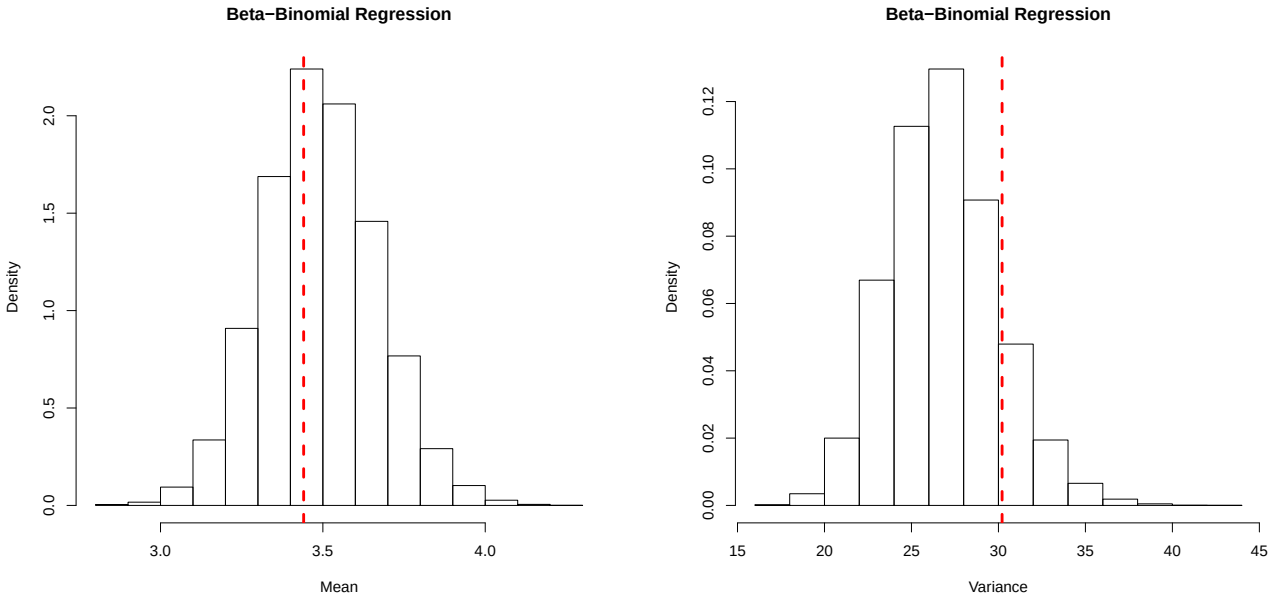


Figure: Posterior Predictive Checks.

	Mean	Variance
Observed	3.4408	30.2101
Post. pred. p-value	0.6075	0.1395

Table: Observed test statistics and associated posterior predictive p-values.

Overdispersion

The BBReg cannot handle the entire extra-variability in the data. Indeed, overdispersion may be due to several (possibly concomitant) causes:

- Excess of zeros
- Outliers
- Omission of important explanatory variables (latent groups)

$\Longrightarrow$ a more general distribution is needed!

The flexible beta-binomial distribution

The flexible beta-binomial distribution

$$Y|\Pi = \pi \sim \text{Bin}(n, \pi).$$

Let Π be distributed according to a flexible beta distribution[1], that is a mixture of particular beta components:

$$f_{FB}(\pi; \lambda_1, \lambda_2, \phi, p) = p f_B(\pi; \lambda_1, \phi) + (1 - p) f_B(\pi; \lambda_2, \phi),$$

where $\phi > 0$ is a common precision parameter, $p \in (0, 1)$ is the mixing weight and $0 < \lambda_2 < \lambda_1 < 1$ are the two component specific means.

[1] S. Migliorati, A. M. Di Brisco, and A. Ongaro. "A new regression model for bounded responses". In: *Bayesian Analysis* 13.3 (2018), pp. 845–872.

The flexible beta-binomial distribution

Let $Y|\Pi = \pi \sim Bin(n, \pi)$ and $\Pi \sim \text{FB}(\lambda_1, \lambda_2, \phi, p)$.
Then, Y follows a **flexible beta-binomial** (FBB) distribution, that is characterized by the following p.m.f.:

$$\begin{aligned} f_{FBB}(y; n, \lambda_1, \lambda_2, \phi, p) &= \int_0^1 f_{Bin}(y; n, \pi) f_{FB}(\pi; \mu, \phi, p, w) d\pi \\ &= p\, f_{BB}(y; n, \lambda_1, \phi) + (1-p) f_{BB}(y; n, \lambda_2, \phi), \end{aligned}$$

where $0 < \lambda_2 < \lambda_1 < 1$, $\phi > 0$, and $p \in (0, 1)$.

The flexible beta-binomial distribution

In a regression perspective we can re-parametrise the FBB distribution as:

- $\mu = p\lambda_1 + (1-p)\lambda_2 \; = \; \mathbb{E}\left[Y/n\right]$,
- $w = \dfrac{\lambda_1 - \lambda_2}{\min\left(\mu/p, (1-\mu)/(1-p)\right)} \in (0,1)$ represents a normalized distance between λ_1 and λ_2,
- $\phi = \phi$,
- $p = p$.

$$\lambda_1 = \mu + (1-p)\, w \min\left(\frac{\mu}{p}, \frac{1-\mu}{1-p}\right), \qquad \lambda_2 = \mu - p\, w \min\left(\frac{\mu}{p}, \frac{1-\mu}{1-p}\right).$$

The flexible beta-binomial distribution

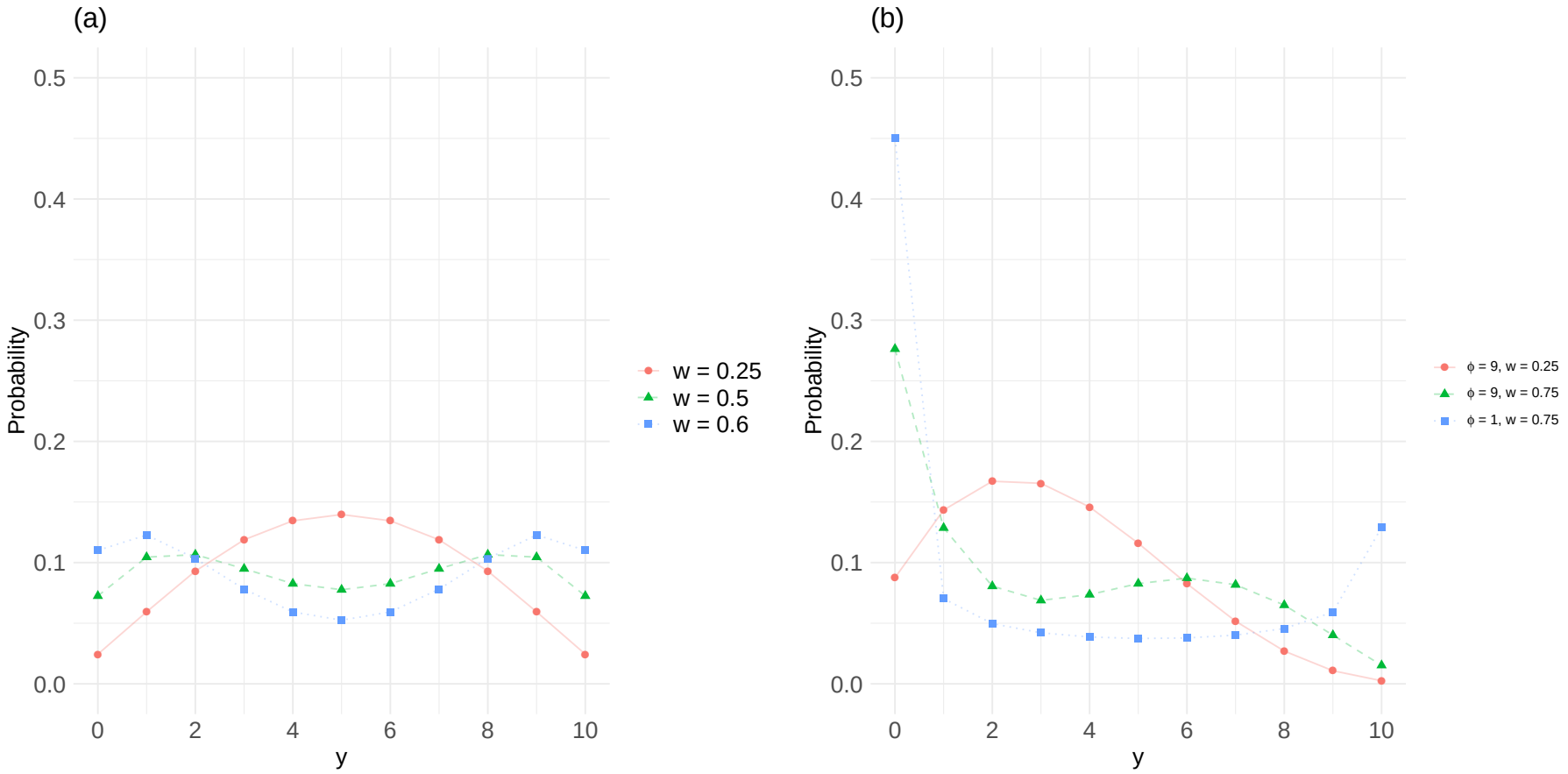


Figure: P.m.f. of the FBB distribution with $n = 10, \mu = 0.5, p = 0.5$, and $\phi = 9$ (panel (a)), and $n = 10, \mu = 1/3$, and $p = 0.5$ (panel (b)).

The flexible beta-binomial distribution

Let $Y \sim \text{FBB}(n, \mu, \phi, p, w)$, then:

- $\mathbb{E}[Y] = n\mu$
- $\text{Var}(Y) = n\mu(1-\mu)\left[1 + \frac{1}{1+\phi}(n-1) + \frac{\phi}{1+\phi}(n-1)w^2 m(\mu, p)\right]$, where $m(\mu, p) = \min\left(\frac{\mu(1-p)}{p(1-\mu)}, \frac{(1-\mu)p}{(1-p)\mu}\right)$
- if $U_1, \ldots, U_n$ are the n binary variables such that $Y = \sum_{r=1}^{n} U_r$, then $\text{Corr}(U_l, U_w) = \frac{1}{1+\phi} + \frac{\phi}{1+\phi} w^2 m(\mu, p)$ for any $l \neq w$.

The flexible beta-binomial distribution

Let $Y \sim \text{FBB}(n, \mu, \phi, p, w)$, then:

- $\mathbb{E}[Y] = n\mu$
- $\text{Var}(Y) = n\mu(1-\mu)\left[1 + \frac{1}{1+\phi}(n-1) + \frac{\phi}{1+\phi}(n-1)w^2 m(\mu, p)\right]$, where $m(\mu, p) = \min\left(\frac{\mu(1-p)}{p(1-\mu)}, \frac{(1-\mu)p}{(1-p)\mu}\right)$
- if $U_1, \ldots, U_n$ are the n binary variables such that $Y = \sum_{r=1}^{n} U_r$, then $\text{Corr}(U_l, U_w) = \frac{1}{1+\phi} + \frac{\phi}{1+\phi} w^2 m(\mu, p)$ for any $l \neq w$.

Bayesian approach - FBBReg

Let:

- $Y_i|\boldsymbol{\beta}, \phi, p, w \sim \text{FBB}(n_i, \mu_i, \phi, p, w)$ for $i = 1, \ldots, N$,
- $\mu_i = \text{inv.logit}(\mathbf{x}_i^\intercal \boldsymbol{\beta})$,

where $\mathbf{x}_i$ is a row of the design matrix $\mathbf{X}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)^\intercal$ is the vector of unknown regression coefficients.

The parametric space of the FBB is variation independent, meaning that we can assume prior independence among the parameters:

- $\boldsymbol{\beta} \sim \text{Normal}_{K+1}\left(\mathbf{0}, \text{diag}(\tau^2, \ldots, \tau^2)\right)$
- $\dfrac{1}{1+\phi} \sim \text{Unif}(0, 1)$
- $w \sim \text{Unif}(0, 1)$
- $p \sim \text{Unif}(0, 1)$

Results: Estimate of parameters

Param.	BinReg	BBReg	FBBReg
β_0	-4.130 (-4.184; -4.081)	-4.002 (-4.104; -3.898)	-4.122 (-4.223; -4.019)
β_1	0.0037 (0.0036; 0.0039)	0.0033 (0.0031; 0.0036)	0.0038 (0.0036; 0.0041)
ϕ	(—)	24.396 (21.071; 28.193)	32.975 (27.695; 38.958)
p	(—)	(—)	0.853 (0.745; 0.926)
w	(—)	(—)	0.763 (0.609; 0.879)
WAIC	6163.2	4418.1	4378.6

Table: Posterior means for each parameter and 95% Credible Sets.

WAIC is a fully Bayesian goodness of fit measure. The smaller the value of the WAIC, the better the model's fit.

Results: Posterior Predictive Checks

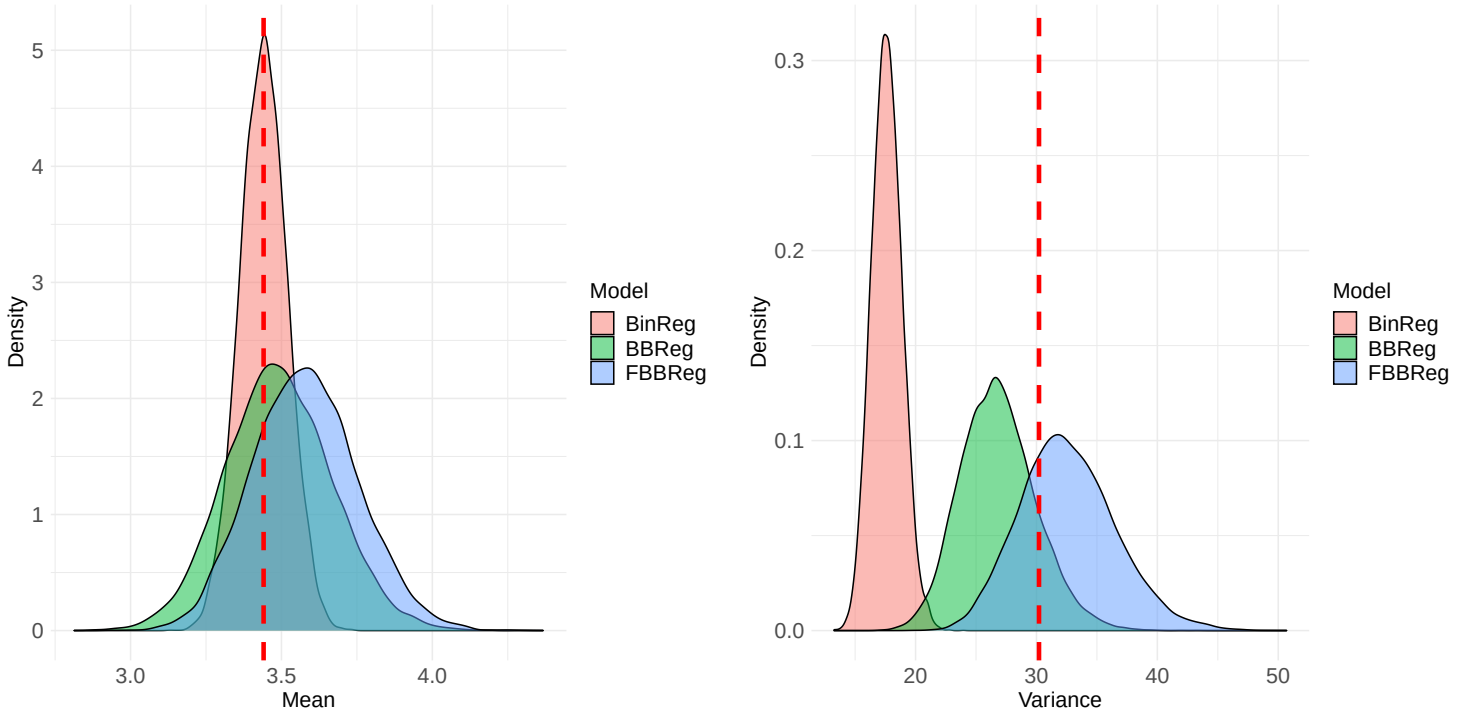


Figure: Posterior Predictive Checks.

	BinReg	BBReg	FBBReg
Mean	0.5025	0.6075	0.7829
Variance	0	0.1395	0.7161

Results: The FBB regression curves

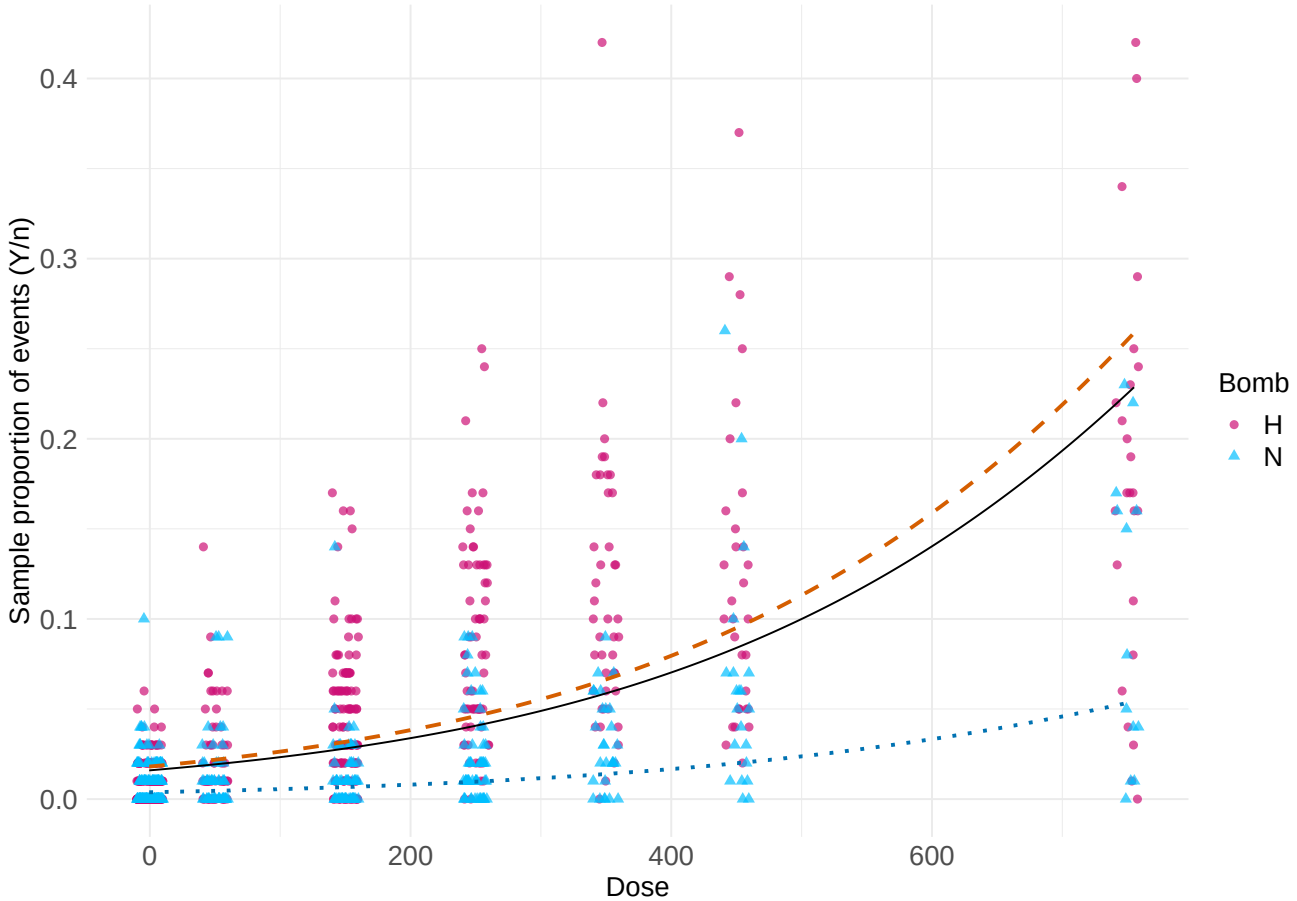


Figure: Atomic data. Black solid line represents the FBBReg curve μ whereas orange dashed and blue dotted lines represent λ_1 and λ_2, respectively.

Second application: Bacteria data

Bacteria data

During an experiment, different numbers of female parasitoids were allowed to attempt to infect $n = 128$ eggs of an alternative host.

Figure: Source: `https://en.wikipedia.org/wiki/Trichogramma`

The dataset provided by Demétrio et al. (2014) consists in the number Y of infected eggs as well as the number X of females.

Bacteria data

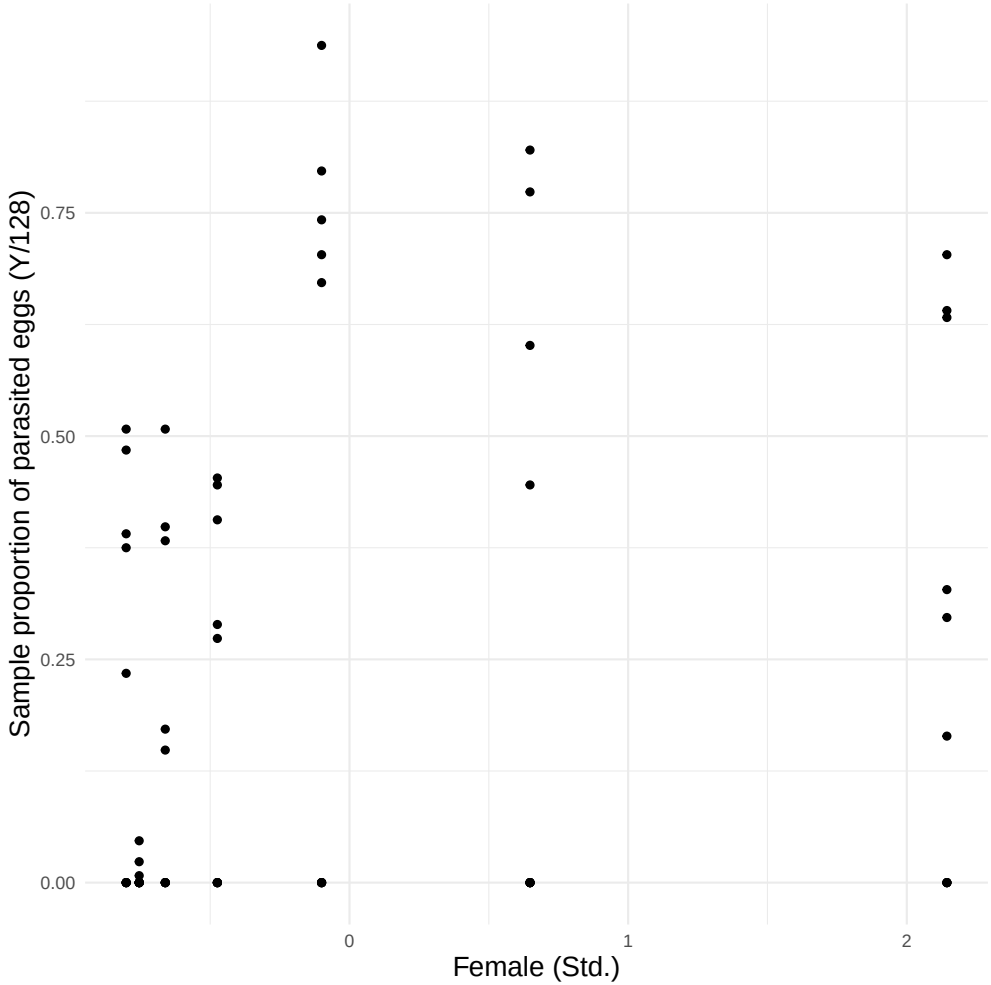


Figure: Bacteria data.

Bacteria data

The dataset is characterized by a large number of zero counts (37 out of 70, $\simeq 52.86\%$). Thus, we considered a zero-inflated binomial and BB models as competitors, that is

$$f_{ZI}(y; q, \cdot) = \begin{cases} q + (1-q)f(0; \cdot), & \text{if } y = 0 \\ (1-q)f(y; \cdot), & \text{if } y \in \{1, 2, \ldots, n\}. \end{cases}$$

We fitted all the regression models considering a quadratic function of the (standardized) number of females as the linear predictor:

$$\text{logit}\,(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \quad i = 1, \ldots, 70.$$

Bacteria data

The dataset is characterized by a large number of zero counts (37 out of 70, $\simeq 52.86\%$). Thus, we considered a zero-inflated binomial and BB models as competitors, that is

$$f_{ZI}(y; q, \cdot) = \begin{cases} q + (1-q)f(0; \cdot), & \text{if } y = 0 \\ (1-q)f(y; \cdot), & \text{if } y \in \{1, 2, \ldots, n\}. \end{cases}$$

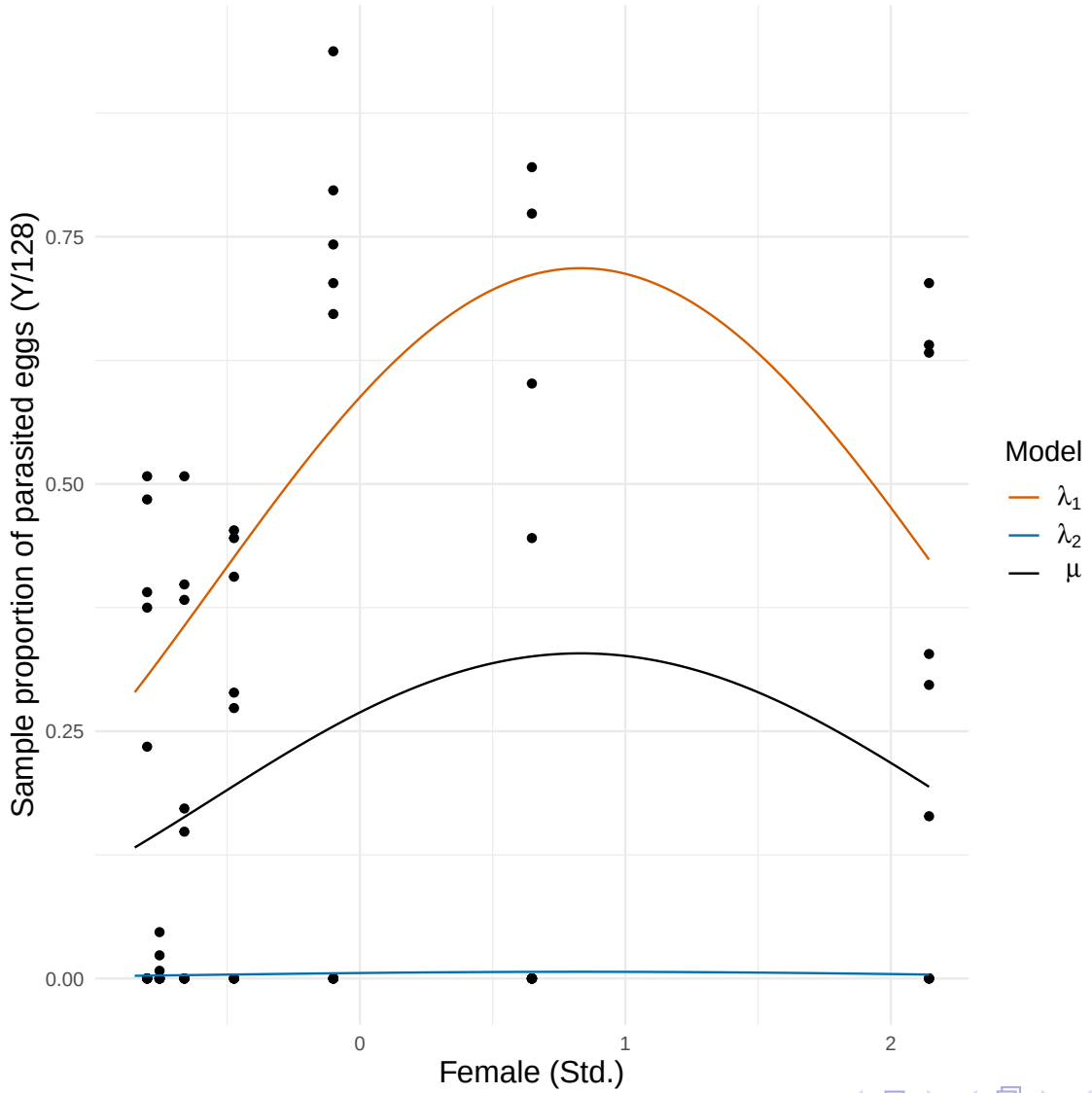
We fitted all the regression models considering a quadratic function of the (standardized) number of females as the linear predictor:

$$\text{logit}\,(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \quad i = 1, \ldots, 70.$$

Bacteria data: results

Param.	BinReg	BBReg	FBBReg	ZIBinReg	ZIBBReg
β_0	**-0.984**	**-1.688**	**-1.000**	**0.593**	**0.591**
	(-1.058, -0.910)	**(-2.374, -1.042)**	**(-1.457, -0.567)**	**(0.485, 0.704)**	**(0.041, 1.162)**
β_1	**0.823**	0.221	**0.687**	**1.393**	**1.530**
	(0.726, 0.920)	(-0.527, 0.957)	**(0.268, 1.130)**	**(1.259, 1.531)**	**(0.819, 2.298)**
β_2	**-0.394**	-0.021	**-0.413**	**-0.821**	**-0.891**
	(-0.455, -0.333)	(-0.508, 0.479)	**(-0.671, -0.161)**	**(-0.909, -0.737)**	**(-1.368, -0.437)**
w	(—)	(—)	0.979	(—)	(—)
			(0.936, 0.999)		
ϕ	(—)	0.769	5.216	(—)	4.827
		(0.481, 1.142)	(2.550, 9.347)		(2.698, 7.664)
p	(—)	(—)	0.453	(—)	(—)
			(0.330, 0.576)		
q	(—)	(—)	(—)	0.528	0.523
				(0.413, 0.641)	(0.408, 0.637)
WAIC	4613.8	446.5	408.7	1014.5	405.6

Bacteria data: FBB regression curves



Bacteria data

Due to the experimental nature of the trial, Demétrio et al. treat the number of females as a 7-level factor. Therefore, we also considered regression models with an intercept term and six dummy variables in the design matrix.

Results are coherent with the ones obtained by treating the number of females as numeric.

Under the FBB model, the ICC depends on μ. Thus, in a regression context, the ICC is naturally modeled as a function of covariates.

Bacteria data: ICC as a function of covariates

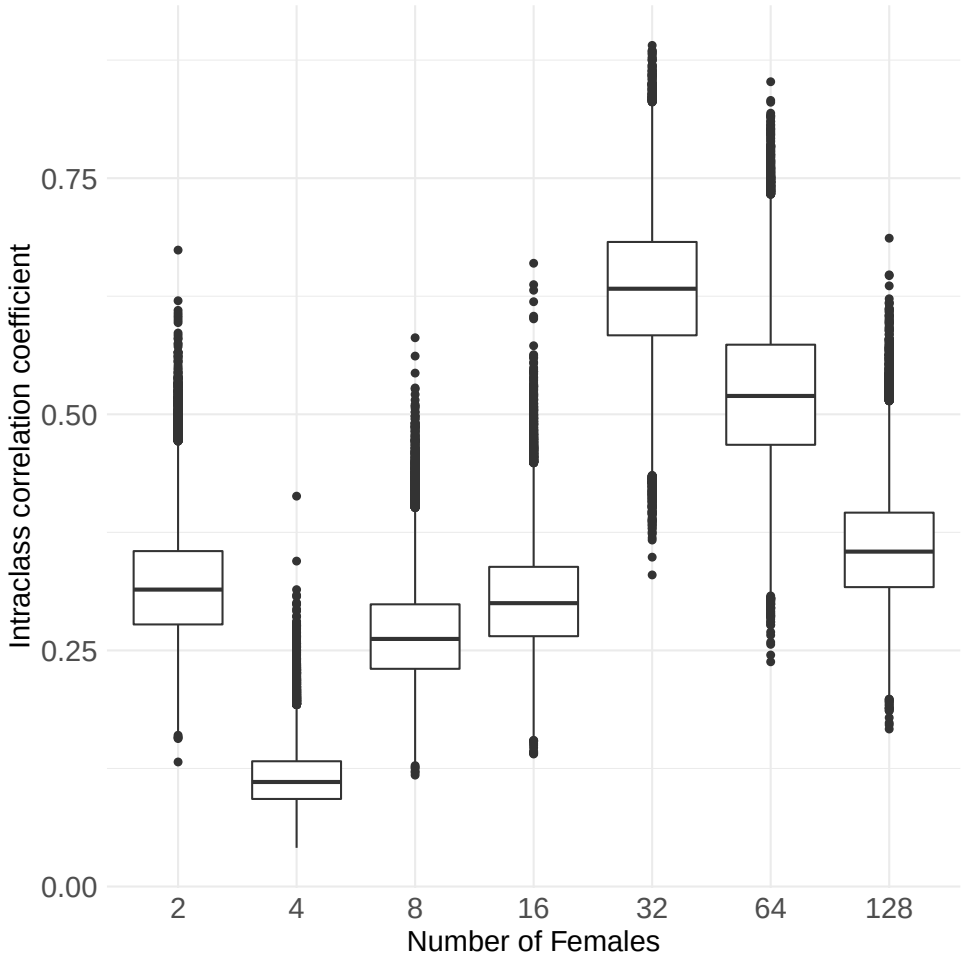


Figure: ICC under the FBBReg model (number of females treated as factor)

Conclusion

- The BBReg model does not always handle overdispersion properly.
- This can be due to multiple concomitant sources of extra variability.
- The FBBReg, thanks to its finite mixture structure, can handle extra variation due to missing covariates, excess of zeros and/or outliers.

Future developments:

- Compound the multinomial distribution with the **extended flexible Dirichlet**.
- The **FlexReg** package on CRAN is going to be updated to include a function to fit the FBBReg model.

References


R. Ascari and S. Migliorati. "A new regression model for overdispersed binomial data accounting for outliers and an excess of zeros". In: *Statistics in Medicine* 40.17 (2021), pp. 3895–3914.


A. Gelman et al. *Bayesian Data Analysis*. Third. CRC Press, 2014.


S. Migliorati, A. M. Di Brisco, and A. Ongaro. "A new regression model for bounded responses". In: *Bayesian Analysis* 13.3 (2018), pp. 845–872.


A. Ongaro, S. Migliorati, and R. Ascari. "A new mixture model on the simplex". In: *Statistics and Computing* 30 (2020), pp. 749–770.