Introduction to Generalized Linear Models
Part 3 of 3

Dr Rafael de Andrade Moral
Associate Professor of Statistics, Maynooth University

rafael.deandrademoral@mu.ie
https://rafamoral.github.io

Outline

- ~~The normal model: A recap~~
- ~~Models for binary data~~
- ~~Models for binomial data~~
- ~~Models for multinomial data~~
- ~~Models for count data~~
- Extensions: overdispersion models
- Extensions: zero-inflated models

Extensions: Overdispersion Models

The Mean-Variance Relationship

- We assume homogeneity of variances when fitting the normal model

$$
\begin{aligned}
Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\
\mathsf{E}[Y_i] &= \mu_i \\
\mathsf{Var}(Y_i) &= \sigma^2
\end{aligned}
$$

- As seen above the variance *is not dependent on the mean*
- If we wish to accommodate heterogeneity of variances we must do so explicitly using predictor variables
- However, this is not the case for all GLMs

The Mean-Variance Relationship

- For the Poisson model we have

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(\mu_i) \\
\text{E}[Y_i] &= \mu_i \\
\text{Var}(Y_i) &= \mu_i
\end{aligned}
$$

- Therefore, the variance is proportional to the mean
- There is *no homogeneity of variances*
- When the mean increases, so does the variance

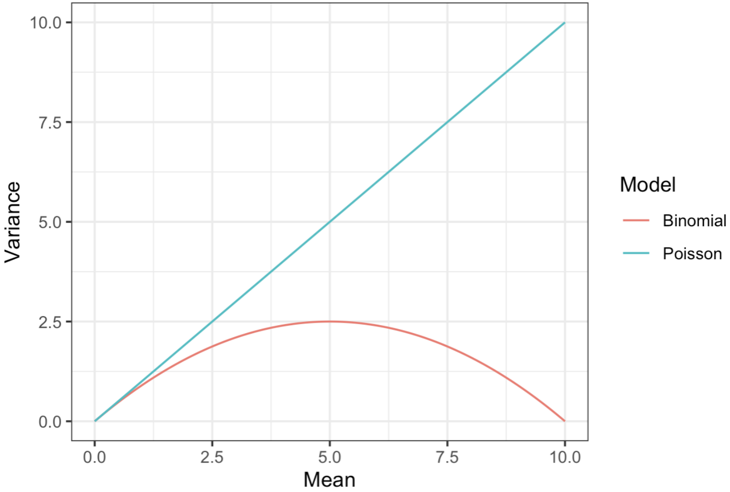
The Mean-Variance Relationship

- For the binomial model we have

$$
\begin{aligned}
Y_i &\sim \text{Binomial}(m_i, \pi_i) \\
\mathsf{E}[Y_i] &= m\pi_i \\
\mathsf{Var}(Y_i) &= m\pi_i(1-\pi_i) = \frac{\mu_i}{m}(m-\mu_i)
\end{aligned}
$$

- Again, the variance is a function of the mean
- There is *no homogeneity of variances*
- This is a quadratic funcion, represented by a parabola with negative concavity

The Mean-Variance Relationship



Overdispersion

- Example: Maize weevil progeny

Overdispersion: Causes

- Variability of the experimental material
- Correlation between individual responses
- Cluster and multistage sampling leading to complex dependencies
- Aggregation
- Omitted unobserved variables

Overdispersion: Consequences

- Underestimation of the standard errors of estimated regression coefficients
- Incorrect significance of effects
- Example: Maize weevil progeny analysis

Quasi-Likelihood

- We will focus on models for count data initially
- One alternative model that can be used to accommodate overdispersion when the Poisson model assumption for the mean-variance relationship fails is the *Quasi-Poisson model*
- The Quasi-Poisson model is not a true probability model, it is what we call a *marginal model*, because it makes assumptions for the mean and variance but is not based on a probability mass function
- The assumptions are

$$\begin{aligned} \mathrm{E}[Y_i] &= \mu_i \\ \mathrm{Var}(Y_i) &= \phi\mu_i \end{aligned}$$

- When $\phi > 1$ this model incorporates overdispersion

Quasi-Likelihood: Estimation and Inference

- The β estimates are exactly the same as for the Poisson model
- The dispersion parameter ϕ is estimated via the Pearson $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mathsf{V}(\hat{\mu}_i)}$ statistic:

$$\hat{\phi} = \frac{X^2}{n-p}$$

- Assessing significance of effects is now done via $F-$tests for *scaled* deviances ($\frac{D}{\phi}$), rather than χ^2 tests:

$$F_{p,n-p} = \frac{\frac{D_1}{\hat{\phi}} - \frac{D_2}{\hat{\phi}}}{n-p}$$

  where D_1 is the deviance for the reduced model and D_2 is the deviance for the full model
- Example: Maize weevil progeny dataset

The Negative Binomial Model

- Overdispersion models can also be achieved through the assumptions of compound processes, through a mixture of distributions and/or hierarchical formulations
- A commonly used model for overdispersed count data is the *negative binomial*
- The negative binomial distribution can be derived in many different ways
- One way to derive it is through a two-stage approach:

$$\begin{aligned} Y_i|M_i &\sim \text{Poisson}(M_i) \\ M_i &\sim \text{Gamma}(\theta, \delta_i) \end{aligned}$$

- Unconditionally, we have that

$$Y_i \sim \text{NegBin}(\mu_i, \theta)$$

where $\mu_i = \frac{\theta}{\delta_i}$ is the mean and θ is the dispersion parameter

The Negative Binomial Model

$$Y_i \sim \mathsf{NegBin}(\mu_i, \theta)$$

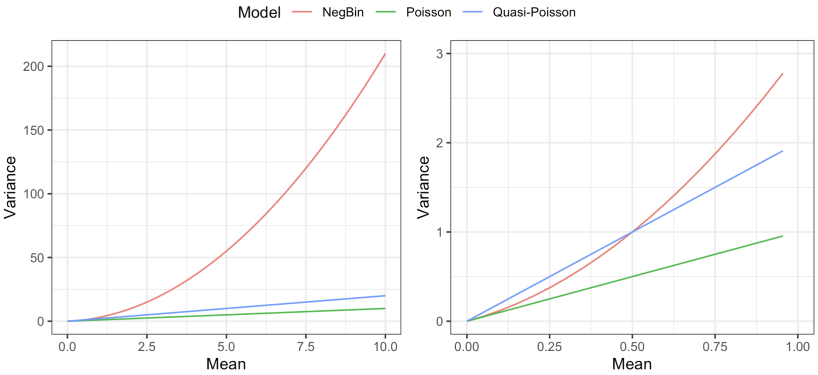
- Under this parameterisation, the mean is

$$\mathsf{E}[Y_i] = \mu_i$$

and the variance is

$$\mathsf{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\theta}$$

The Negative Binomial Model: Variance Function



$$\phi = 2, \theta = \frac{1}{2}$$

The Negative Binomial Model

- Example 1: effects of agricultural oils on *Diaphorina citri* oviposition
- Example 2: number of articles published by 915 biochemistry Ph.D. researchers

Overdispersion Models for Discrete Proportion Data

- For discrete proportions, when the variance is larger than expected by the binomial model, there are also alternative models
- Quasi-binomial model:

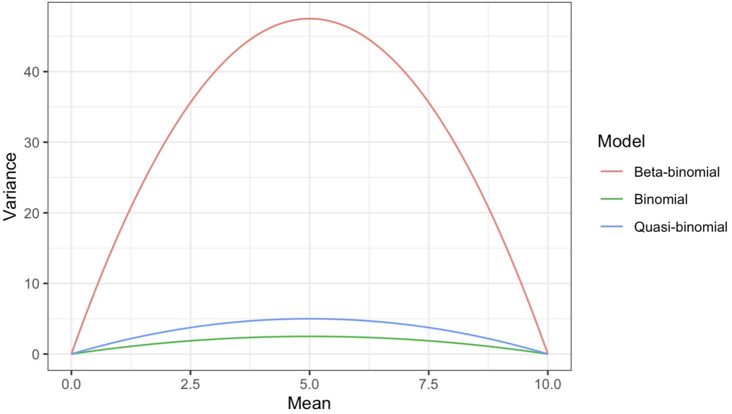
$$\text{Var}(Y_i) = \phi m \pi_i (1 - \pi_i)$$

- Beta-binomial model:

$$\begin{aligned} Y_i | P_i &\sim \text{Binomial}(m_i, P_i) \\ P_i &\sim \text{Beta}(a_i, b_i) \\ \text{Var}(Y_i) &= m_i \pi_i (1 - \pi_i)(1 + \phi(m_i - 1)) \end{aligned}$$

where $\pi_i = \frac{a_i}{a_i + b_i}$ and $\phi = \frac{1}{a_i + b_i + 1}$

Overdispersion Models for Discrete Proportion Data



$$m = 10, \phi = 2$$

Overdispersion Models for Discrete Proportion Data

- Example revisited: *Diaphorina citri* mortality data

Extensions: Zero-Inflated Models

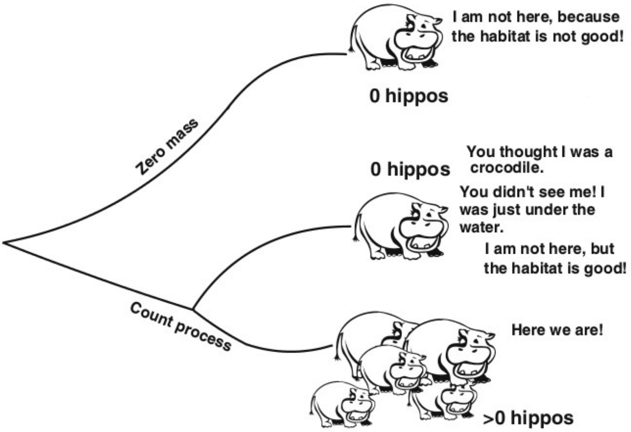
Zero-Inflation

- If $Y_i \sim \text{Poisson}(\mu_i)$, the probability of a zero count is

$$\mathsf{P}(Y_i = 0) = e^{-\mu_i}$$

- However, in many practical problems, an excess number of zero counts arises
  - e.g.[1] cohort of non-smokers when surveying number of cigarettes smoked per day
  - e.g.[2] non-suitable habitats when counting numbers of animals of a certain species at different locations
- Some overdispersed distributions can accommodate more zero counts than the standard Poisson (e.g. the negative binomial)
- It is also possible to explicitly model the excess zero counts by using a *zero-inflated* distribution

Zero-Inflated Process



The Zero-Inflated Poisson Model

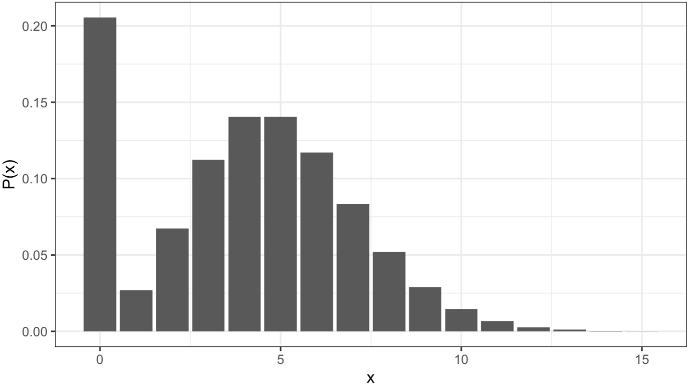
- We can write the ZIP model using a hierarchical formulation:

$$\begin{aligned} Y_i \mid Z_i = z_i &\sim \begin{cases} \text{Poisson}(\lambda_i), & \text{if } z_i = 0 \\ 0, & \text{if } z_i = 1 \end{cases} \\ Z_i &\sim \text{Bernoulli}(\omega_i) \end{aligned}$$

- We have, then, an inflated probability of a zero:

$$\mathsf{P}(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)e^{-\lambda_i}, & \text{if } y_i = 0 \\ (1-\omega_i)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, & \text{if } y_i = 1, 2, \ldots \end{cases}$$

The Zero-Inflated Poisson Model



$$\mu = 5, \omega = 0.2$$

The Zero-Inflated Poisson Model

- We can model both λ_i and ω_i with covariates:

$$\begin{aligned}\log \lambda_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \\ \log\left(\frac{\omega_i}{1-\omega_i}\right) &= \gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_q x_{qi}\end{aligned}$$

- It is possible to allow for $p \neq q$

The Zero-Inflated Poisson Model

- Example: Hunting spider abundance

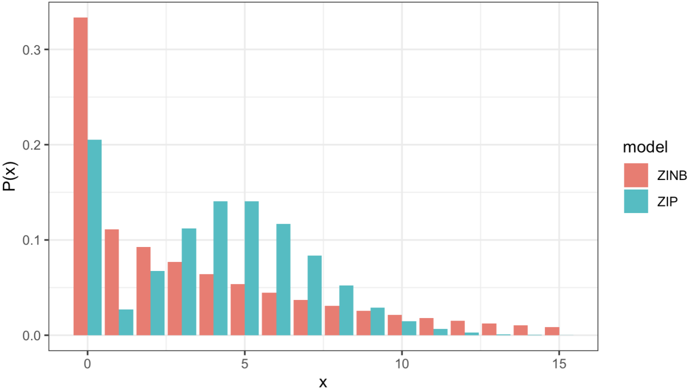
Overdispersion *and* Zero-Inflation

- It is possible to combine extensions for overdispersion *and* zero-inflation
- This is especially useful when there is an excess of zero counts combined with extra variability
- The zero-inflated negative binomial (ZINB) model can be written as:

$$
\begin{aligned}
Y_i | Z_i = z_i &\sim \begin{cases} \text{NegBin}(\lambda_i, \theta), & \text{if } z_i = 0 \\ 0, & \text{if } z_i = 1 \end{cases} \\
Z_i &\sim \text{Bernoulli}(\omega_i)
\end{aligned}
$$

- There are other ways of accommodating excess zero counts (e.g. hurdle models), and overdispersion (e.g. using random effects)

ZIP vs. ZINB



$$\mu = 5, \omega = 0.2, \theta = 1$$