Generalized Linear Models

For Over-Dispersed Data

Basics of Generalized Linear Models (GLMs)

- GLMs are a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.
- **Components of GLMs**:
  - **Random Component**: Specifies the distribution of the response variable (e.g., Normal, Binomial, Poisson).
  - **Systematic Component**: A linear predictor, a combination of explanatory variables (predictors).
  - **Link Function**: Connects the mean of the response variable to the linear predictor (e.g., identity, log, logit).

Common link functions and GLMS

- **Identity Link**: Used for normally distributed data (linear regression).
- **Logit Link**: Used for binary outcome data (logistic regression).
- **Log Link**: Used for count data (Poisson regression).

Over-Dispersed Data

- Over-dispersion in general refers to having variance greater than that assumed for the theoretical data model.
- Over-dispersion can also refer to having a variance that is greater than the mean
  - Similarly equi-dispersion would refer to having variance equal to the mean.

Poisson regression

Poisson regression is the most popular method for modeling count data. The Poisson distribution brings with it the assumption of equi-dispersion that is often unsatisfied.

- Count data in Biology
- Epidemiology
- Finance
- Insurance claims
- Environmental studies
- etc.

Almost any real world count data is subject to the possibility of over-dispersion.

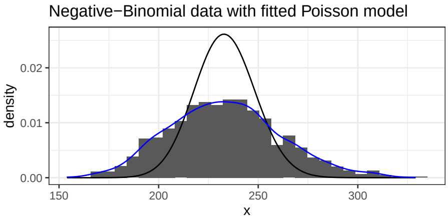
Causes

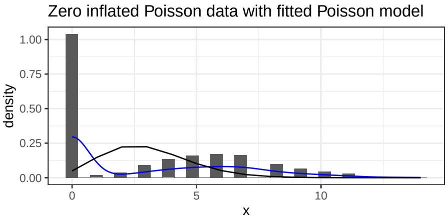
- Increased variability of counts
- Event clustering
- Increased number of 0
- Interaction effects
- Measurement error
- Environmental effects

Candidate Distributions

1. Negative-Binomial
2. Generalized Poisson
3. Double Poisson
4. Conway-Maxwell-Poisson (CMP)

- Zero inflated disitributions
  - ZIP
  - ZINB
  - ZIDP/ZIGP





Negative-Binomial

- Parameters: mean: μ, dispersion: k[1]
- Variance: $\mu + \mu^2/k$
  - Function of mean and dispersion parameter
  - Clearly captures over-dispersion.

[1] The classic negative-binomial is parameterized as number of success and probability of success, r and p.

Generalized Poisson

- Parameters: λ, θ
- Mean: $\lambda/(1-\theta)$
- Variance: $\lambda/(1-\theta)^3$

$$f(Y=y) = \frac{\lambda(\lambda+\theta y)^{y-1} e^{-(\lambda+\theta y)}}{y!}, \quad \lambda > 0, \theta \in \mathbb{R}$$

- Designed to extend to overdispersed and underdispersed count data

$\theta = 0$	$\theta > 0$	$\theta < 0$
reduces to Poisson	Models overdispersion	Models underdispersion