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).
- **Reciprocal Link**: Used for increasing rate that levels off (Gamma 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 equa-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 equa-dispersion that is often unsatisfied.

Common applicataions

- 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.

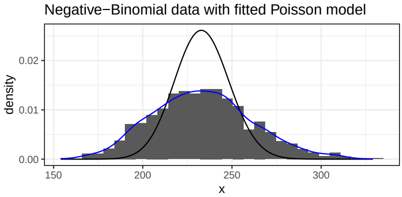
Causes

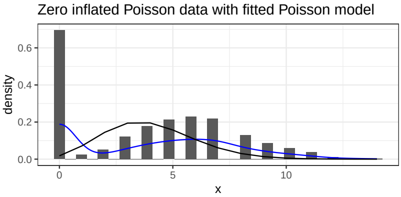
- Increased variability of counts
- Event clustering
- Increased number of 0
- Interaction effects
- Measurement error
- Environmental effects

Candidate Distributions

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

- Zero inflated distributions
  - 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: $\lambda/(1-\theta)^3$
- Model introduced by Consul 1989 as a way to modify the Poisson to handle over-dispersed and under-dispersed count data
- Probability distribution function

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

$\theta = 0$	$\theta < 0$	$\theta > 0$
Equa-dispersion (Poisson)	Under-dispersion	Over-dispersion

Double Poisson

- Parameters: μ, θ
- Mean: μ Variance:[2] μ/θ
- Extension of the double exponential family (Efron 1986) with approximate pmf

$$Pr(Y = y) = (\theta^{1/2} e^{-\theta\mu}) \left(\frac{e^y y^y}{y!}\right) \left(\frac{e\mu}{y}\right)^{\theta y}$$

- The exact double Poisson (DP) density includes a normalizing constant $\sum_{y=0}^{\infty} Pr(Y = y) \approx 1 + \frac{1-\theta}{12\mu\theta}(1 + \frac{1}{\mu\theta})$

[2]Over-dispersed for $(\theta < 1)$, under-dispersed for $(\theta > 1)$, Poisson$(\theta = 1)$

Conway-Maxwell Poisson (CMP)

- Parameters:[3] λ, ν
- Mean: $\mu \approx \lambda + 1/2\nu - 1/2$ Variance:[4] $\sigma^2 \approx \lambda/\nu$
- Weighted Poison distribution with pmf:

$$Pr(Y = y) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \quad Z(\lambda, \nu) = \sum_{y=0}^{\infty} \frac{\lambda^y}{(y!)^\nu}$$

- Includes spacial cases (Sellers et al. 2012) of Poison when $\nu = 1$, geometric when $\nu \to 0$, and Bernoulli when $\nu \to \infty$

[3] A mean parameterized CMP was introduced with better interpretation and computation (Huang 2017)

[4] Approximations only accurate under specific conditions $\lambda > 10^\nu$ or $\nu \leq 1$ (Sellers et al. 2012)