Generalized Linear Models

For Over-Dispersed Data

Nathen Byford

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.

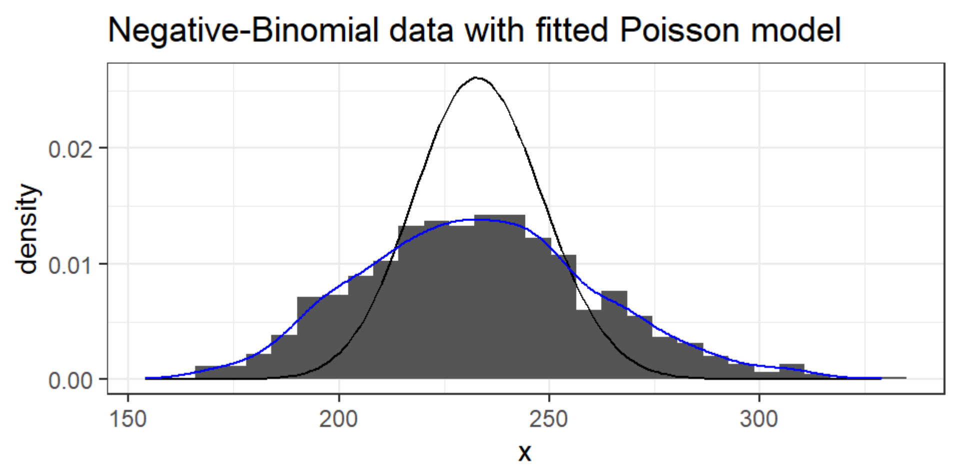
Causes

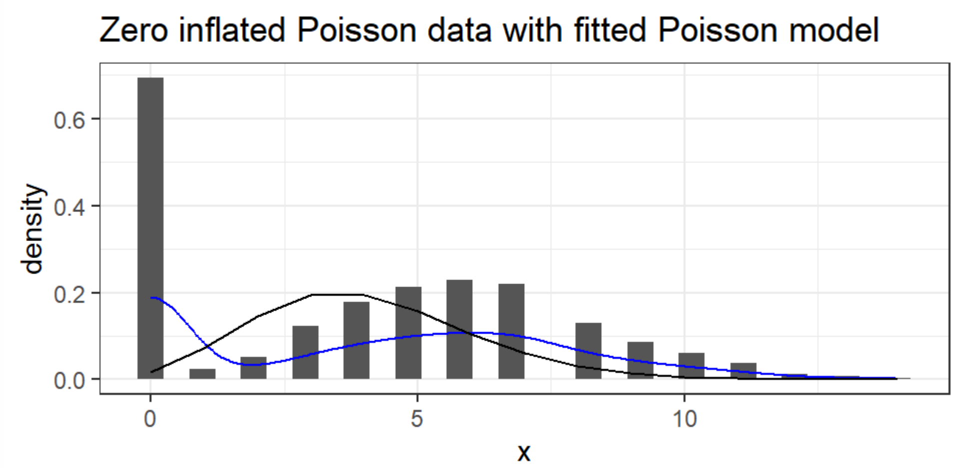
- Increased variability of counts
- Event clustering
- Increased number of 0
- Interaction effects
- Measurement error
- Environmental effects

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.

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:[1] 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 $\sum_{y=0}^{\infty} Pr(Y = y) \approx 1 + \frac{1-\theta}{12\mu\theta}(1 + \frac{1}{\mu\theta})$

Conway-Maxwell Poisson (CMP)

- Parameters:[1] λ, ν
- Mean: $\mu \approx \lambda + 1/2\nu - 1/2$ Variance:[2] $\sigma^2 \approx \lambda/\nu$
- 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 $\nu \to \infty$

Zero Inflated Distributions

- Zero inflated distributions are piece-wise distributions with components for 0s and non-0s
- For example the zero inflated Poisson (ZIP) has pmf

$$Pr(Y = y) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{if } y = 0 \\ (1-\pi)\frac{\lambda^y e^{-\lambda}}{y!} & \text{if } y = 1, 2, \ldots \end{cases}$$

- Zero inflated framework extends to other distributions to further capture over-dispersion and zero inflation
- Models can be fit using the `zeroinfl()` function in the package `pscl`

Model comparisons

- 2008 Bayesian paper compares the Generalized Poisson distribution, (Gschlößl and Czado 2008)

Model	Poisson	NB	GP
DIC	1,291.8	1,273.9	1,265.6

- Car crash analysis using the DP and CMP model (Zou et al. 2013)

Model	DP	NB	CMP
AIC	3,268.20	3,199.200	
MSPE	2.62	2.727	2.73

- Bayesian paper compared Poisson, Negative Binomial, and CMP for longitudinal counts using DIC to compare. (Alam et al. 2023)

Model	Poisson	NB	CMP
DIC	1,362.39	1,350.67	1,348.87

Further Comparison

- Another Bayesian paper compared using AIC and shows the following results (Sellers and Shmueli 2010)

Model	CMP	Poisson	Neg-Bin
AIC	5,073	5,589	5,077

- Mean parameterized CMP AIC and run time

Model	GP	CMP(Mean-param)	CMP
AIC	453.75	440.82	440.5
Run time (Sec)	0.33	8.50	31.5

- Zero inflated Poisson regression model comparison for occupational injuries (Wang et al. 2003)

Model	Poisson	ZIP
Log-Likelihood	-409.678	-397.704

Results

- It has been found and shown that modeling over-dispersed data with improper distributions leads to biased results.
- To prevent biased results from over dispersed data, using models such as the CMP, GP, or DP model can prove beneficial
- Zero inflated models have better fit when there are increased number of zeros observed and can easily be implemented
- Some of these models are easy to implement such as the CMP and CMP(mean-parameterized) in packages `COMPoissonReg` and `mpcmp`

References

Alam, M., Gwon, Y., and Meza, J. (2023), "Bayesian conway-maxwell-poisson (CMP) regression for longitudinal count data," *Communications for Statistical Applications and Methods*, 30, 291–309. https://doi.org/10.29220/CSAM.2023.30.3.291.

Efron, B. (1986), "Double exponential families and their use in generalized linear regression," *Journal of the American Statistical Association*, 81, 709–721. https://doi.org/10.2307/2289002.

Gschlößl, S., and Czado, C. (2008), "Modelling count data with overdispersion and spatial effects," *Statistical Papers*, 49, 531–552. https://doi.org/10.1007/s00362-006-0031-6.

Huang, A. (2017), "Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts," *Statistical Modelling*, 17, 359–380. https://doi.org/10.1177/1471082X17697749.

Sellers, K. F., Borle, S., and Shmueli, G. (2012), "The COM-poisson model for count data: A survey of methods and applications," *Applied Stochastic Models in Business and Industry*, 28, 104–116. https://doi.org/10.1002/asmb.918.

Sellers, K. F., and Shmueli, G. (2010), "A flexible regression model for count data," *The Annals of Applied Statistics*, 4, 943–961.

Wang, K., Lee, A. H., Yau, K. K. W., and Carrivick, P. J. W. (2003), "A bivariate zero-inflated Poisson regression model to analyze occupational injuries," *Accident Analysis & Prevention*, 35, 625–629. https://doi.org/10.1016/S0001-4575(02)00036-2.

Zou, Y., Geedipally, S. R., and Lord, D. (2013), "Evaluating the double poisson generalized linear model," *Accident Analysis & Prevention*, 59, 497–505. https://doi.org/10.1016/j.aap.2013.07.017.