Strategies for analysis of under- and overdispersed count data

Eduardo Elias Ribeiro Junior [1, 2]
Clarice Garcia Borges Demétrio [2]

[1]Statistics and Geoinformation Laboratory (LEG-UFPR)
[2]Department of Exact Sciences (ESALQ-USP)

18th May 2018

jreduardo@usp.br | edujrrib@gmail.com

Outline

1. Background
2. Alternative Models
3. Data analysis
4. Final remarks

1

Background

Corn damage: *Sitophilus zeamais* progeny

Motivation:

- Study in Entomology;
- Major pest of stored maize in Brazil is *Sitophilus zeamais*.

Objective:

- Assess the insecticide action of organic extracts of Annonacae.

Experiment:

- Design: completely randomized experiment with 10 replicates;
- Experimental unit: petri dishes containing 10g of corn treated with extracts;
- Factor: Extracts prepared with different parts of the plant (seeds, leaves and branches) or just water (control).
- Response variable: Number of emerged insects (progeny) after 60 days.

Descriptive analysis

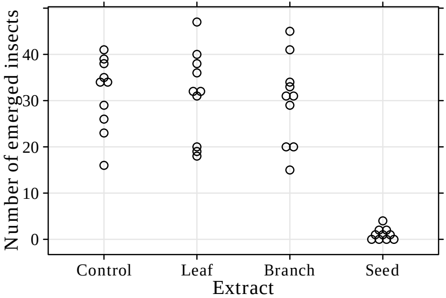

Figure: Number of emerged insects for each extract

Table: Sample mean and sample variance for the *Sitophilus zeamais* data.

Extract	Sample mean	Sample variance
Control	31.50	62.50
Leaf	31.30	94.01
Branch	29.90	88.77
Seed	1.10	1.66

Poisson model and limitations

GLM framework (Nelder & Wedderburn 1972)

- Provide suitable distribution for a counting random variables;
- Efficient algorithm for estimation and inference;
- Implemented in many software.

Poisson model

- Relationship between mean and variance, $\mathrm{E}(Y) = \mathrm{Var}(Y)$;

Main limitations

- Overdispersion (more common), $\mathrm{E}(Y) < \mathrm{Var}(Y)$;
- Underdispersion (less common), $\mathrm{E}(Y) > \mathrm{Var}(Y)$.

2

Alternative Models

Proposing of alternative models

The origin of such phenomena of under- and overdispersion can be interpreted as a failure of some basic assumptions of the model (Hinde & Demétrio 1998).

Examples of Poisson process failures:

- Variability of experimental material (Random-effects);
- Aggregate level data (Compound distributions);
- Non-constancy of the hazard function of waiting times (Duration dependence);

We shall present the genesis and definition of models:

- COM-Poisson;
- Gamma-Count;
- Generalized Poisson; and
- Poisson-Tweedie.

2.1

Alternative Models
COM-Poisson model

Weighted Poisson models

The family of weighted Poisson distributions (WPD) (Del Castillo & Pérez-Casany 1998), weights the Poisson probability function by a suitable function,

$$\Pr(Y = y) = \frac{w(y)\exp(-\lambda)\lambda^{y}}{y!\mathrm{E}_{\lambda}[w(Y)]}, \quad y \in \mathbb{N},$$

where $\mathrm{E}_{\lambda}(\cdot)$ denotes the mean value with respect to the Poisson random variable with parameter λ and $w(y)$ is a weight function.

The weight function may depend on extra parameter to ensure more flexibility to the distribution.

COM-Poisson distribution

The COM-Poisson (Shmueli et al. 2005) belongs to the family of weighted Poisson distributions and it is obtained when $w(y, \nu) = (y!)^{1-\nu}$. The probability mass function of Y a COM-Poisson random variable is

$$\Pr(Y = y) = \frac{\lambda^y \exp(-\lambda)}{(y!)^\nu \mathrm{E}_\lambda[(Y!)^{1-\nu}]} = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \quad Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu},$$

where ν is the dispersion parameter.

The moments of distribution are approximated by

$$\mathrm{E}(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \qquad \text{and} \qquad \mathrm{Var}(Y) \approx \frac{\lambda^{1/\nu}}{\nu}.$$

Mean-parametrized and regression models

Following Ribeiro Jr et al. (2018), we use the parametrization given by introducing μ based on mean approximation

$$\mu = h(\lambda, \nu) = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad \Rightarrow \quad \lambda = h^{-1}(\mu, \nu) = \left(\mu + \frac{(\nu - 1)}{2\nu}\right)^{\nu}$$

Regression models

$$Y_i \sim \text{CMP}(\nu, \mu_i), \quad \text{where} \quad \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

2.2

Alternative Models

Gamma-Count model

Renewal process

Following Winkelmann (1995),

- Let $\tau_k > 0$, $k \in \mathbb{N}^*$, denote the waiting times between the $(k-1)$ and the *k-ht* event;
- Let ϑ_n, denote the arrival time of the *n-th* event, so $\vartheta_n = \sum_{k=1}^{n} \tau_k$.
- Finally, denote Y_T the number of events within a $(0, T)$ interval.

$$
\begin{aligned}
Y_T < y &\iff \vartheta_y \geq T \\
\Pr(Y_T < y) &= \Pr(\vartheta_y \geq T) = 1 - \mathrm{F}_y(T), \\
\Pr(Y_T = y) &= \Pr(Y_T < y) - \Pr(Y_T < y + 1) \\
\Pr(Y_T = y) &= \mathrm{F}_{\vartheta_y}(T) - \mathrm{F}_{\vartheta_{y+1}}(T),
\end{aligned}
$$

where $\mathrm{F}_{\vartheta_n}(T)$ is the cumulative density function of ϑ_n and T is the interval of the counting.

Gamma-Count distribution

For the Gamma-Count distribution we assume τ_k are identically and independently Gamma(α, κ). So the reproductive property of Gamma random variables, leads to $\vartheta_y \sim \text{Gamma}(y\alpha, \kappa)$.

Consequently, the probability mass function of Y a Gamma-Count random variable is

$$\Pr(Y_T = y) = \int_0^T \frac{\kappa^{y\alpha} t^{y\alpha - 1}}{\Gamma(y\alpha)\exp(\kappa t)} dt - \int_0^T \frac{\kappa^{(y+1)\alpha} t^{(y+1)\alpha - 1}}{\Gamma[(y+1)\alpha]\exp(\kappa t)} dt,$$

a difference between two Gamma cumulative density functions, $\text{G}(y\alpha, \kappa) - \text{G}((y+1)\alpha, \kappa)$, where $\text{G}(\alpha, \kappa)$ is the cumulative function $\text{F}_y(T)$ for the Gamma variable with parameters α and κ.

Properties and regression models

Winkelmann (1995) showed for increasing T, i.e. high counts, it holds that

$$Y_T \overset{asy}{\sim} \mathcal{N}\left(\frac{\kappa T}{\alpha}, \frac{\kappa T}{\alpha^2}\right),$$

thus the limiting variance-mean ratio equals a constant $1/\alpha$.

Regression models

$$Y_i \sim \mathrm{GCT}(\alpha, \kappa_i), \quad \text{where} \quad \kappa_i = \alpha g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

2.3

Alternative Models

Generalized Poisson model

Generalized Poisson distribution

The distribution results from the limiting form of the generalized negative binomial distribution (Zamani & Ismail 2012).

The probability mass function is given by

$$\Pr(Y = y) = \begin{cases} \frac{\lambda(\lambda+y\gamma)^{y-1}\exp(-\lambda-y\gamma)}{y!}, & y = 0, 1, 2, \ldots \\ 0 & \text{for } y > m, \text{when } \gamma < 0, \end{cases}$$

where $\lambda > 0$, $\max(-1, -\lambda/4) \leq \gamma \leq 1$ and m is the largest positive integer for which $\theta + m\lambda > 0$ when λ is negative.

- $\mathrm{E}(Y) = \lambda(1-\gamma)^{-1}$;
- $\mathrm{Var}(Y) = \lambda(1-\gamma)^{-3}$.

Mean-parametrization and regression models

In order to specify regression models based on Generalized Poisson distribution, we use the mean-parametrization

$$\lambda = \frac{\mu}{1+\sigma\mu} \quad \text{and} \quad \gamma = \frac{\sigma\mu}{1+\sigma\mu}.$$

Under this parametrization, the moments are given by

- $\mathrm{E}(Y) = \mu$;
- $\mathrm{Var}(Y) = \mu(1+\mu\sigma)^2$.

Regression models

$$Y_i \sim \mathrm{GPo}(\mu_i, \sigma), \quad \text{where} \quad \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

2.4

Alternative Models

Poisson-Tweedie model

General two-stage models

The Poisson-Tweedie class of distributions is a general case of the two-stages models (Jørgensen & Kokonendji 2016). This family is given by the following hierarchical specification

$$Y \mid Z \sim \text{Po}(Z) \quad \text{where} \quad Z \sim \text{Tw}_p(\mu, \phi),$$

and Tw_p denotes a Tweedie distribution with power parameter p, $p \in (-\infty, 0] \cup [1, \infty)$, $\mu \in \Omega_p$ and $\phi > 0$.

The probability mass function for this distribution cannot be obtained in closed form apart from the special case corresponding to the Negative Binomial distribution, $p = 2$.

Second-moments assumptions and regression

Although the probability function cannot be obtained, the moments mean and variance can,

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}[\mathrm{E}(Y \mid Z)] = \mu \\ \mathrm{Var}(Y) &= \mathrm{Var}[\mathrm{E}(Y \mid Z)] + \mathrm{E}[\mathrm{Var}(Y \mid Z)] = \mu + \phi\mu^p. \end{aligned}$$

The Poisson-Tweedie has Hermite $(p = 0)$, Neymann tipo-A $(p = 1)$, Pólya-Aeppli $(p = 1,5)$, negative binomial $(p = 2)$ and Poisson–inverse Gaussian $(p = 3)$ as special cases (Bonat et al. 2018).

Regression models

$$Y_i \sim \mathrm{PTw}_p(\mu_i, \omega), \quad \text{where} \quad \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

2.5

Alternative Models

Comparison of the distributions

Summary of distributions

Table: Probabilistic models for analysis of count data.

	COM-Poisson	Gamma-Count	Generalized Poisson	Poisson-Tweedie
Notation	$\text{CMP}(\mu_i, \nu)$	$\text{GCT}(\kappa_i, \gamma)$	$\text{GPo}(\mu_i, \sigma)$	$\text{PTw}_p(\mu_i, \omega)$
Dispersion parameter	$\phi = \log(\nu)$ $\nu > 0$	$\gamma = \log(\alpha)$ $\alpha > 0$	σ $\sigma > c^*$	ω $\omega > 0$
Expectation	$\approx \mu_i$	$\overset{a}{\approx} \kappa_i/\alpha$	μ_i	μ_i
Variance	$\approx \mu_i/\nu$	$\overset{a}{\approx} \kappa_i/\alpha^2$	$\mu_i(1+\sigma\mu_i)^2$	$\mu_i(1+\omega\mu_i^{p-1})$
Dispersion index (DI)	$\approx 1/\nu$	$\overset{a}{\approx} 1/\alpha$	$(1+\sigma\mu_i)^2$	$1+\omega\mu_i^{p-1}$
Regression	$\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\kappa_i = \alpha g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$

$c^* = \min[-\max(y_i^{-1}), -\max(\mu_i^{-1})]$; $\overset{a}{\approx}$ asymptotically when $T \to \infty$.

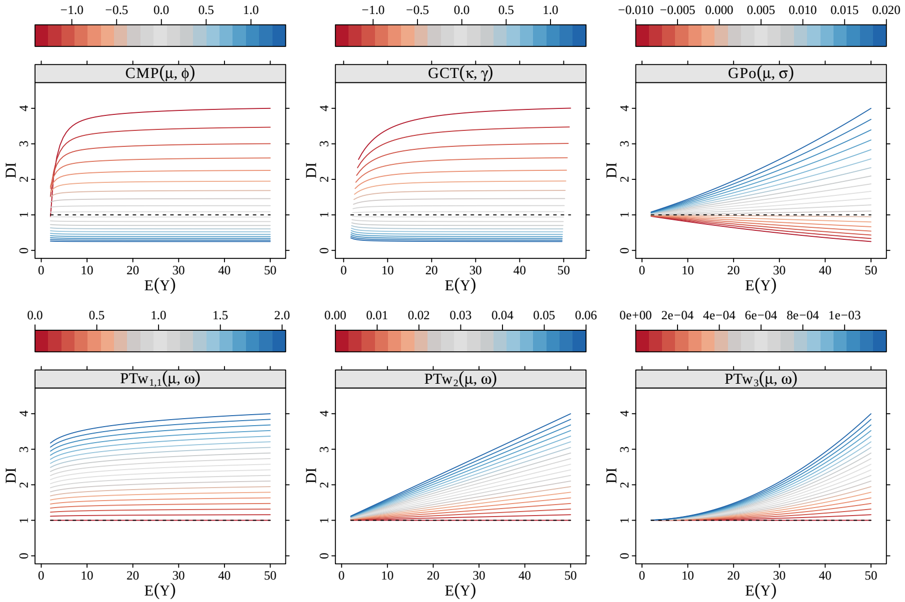

Figure: Dispersion indexes for different parameters of the CMP, GCT, GPo and PTW.

3

Data analysis

Model specification

- **Model:** $\log(\mu_{ij}) = \beta_0 + \tau_j$.
- $i = 1, 2, \ldots, 10; j = 1, 2, 3, 4;$
- Restriction $\tau_1 = 0$.

Table: Parameter estimates and standard errors for the model strategies.

	CMP	GCT	GPo	PTw
β_0	$3.477(-)$	$3.425(0.091)$	$3.450(0.091)$	$3.450(0.087)$
τ_{leaf}	$-0.008(-)$	$-0.007(0.128)$	$-0.006(0.128)$	$-0.006(0.124)$
τ_{branch}	$-0.061(-)$	$-0.054(0.130)$	$-0.052(0.129)$	$-0.052(0.125)$
τ_{seed}	$-3.204(-)$	$-4.016(0.663)$	$-3.355(0.321)$	$-3.355(0.362)$
Disp.	$-1.072(-)$	$-0.927(0.263)$	$0.019(0.007)$	$1.403(0.570)$
Power	–	–	–	$0.348(0.670)$
LogLik	-120.919	-121.651	-122.284	-121.847
AIC	251.839	253.302	254.568	255.693

Fitted values

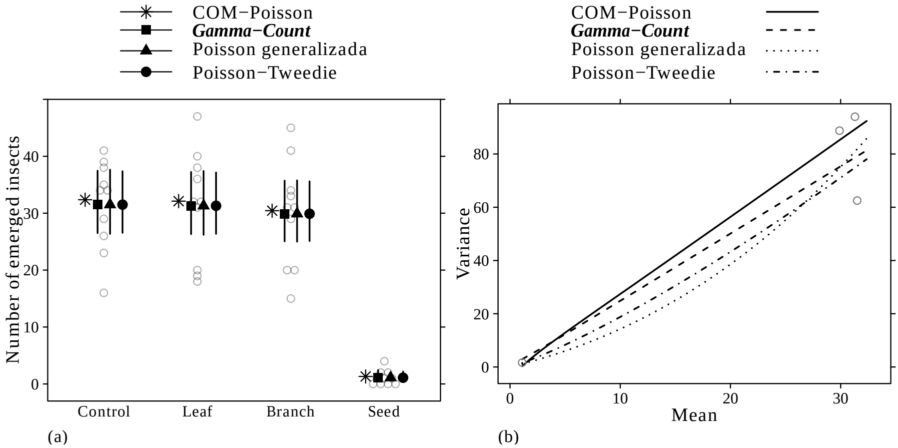

Figure: (a) Fitted values with confidence intervals (95%) and (b) Mean–variance relationship for the fitted models.

4

Final remarks

Concluding remarks

Summary

- Over- / underdispersion are phenomena that needs caution;
- Generalizations of Poisson distribution can be derived of Poisson process failures;
- For most practical problems, the models are similar in terms of fitted values and confidence intervals;
- The mean-variance relationship characterizes them.

Future work

- Perform a extensive simulation to assess the capacity of each model.

References

Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J. & Démetrio, C. G. B. (2018), 'Extended Poisson-Tweedie: properties and regression model for count data', *Statistical Modelling* **18**(1), 24–49.

Del Castillo, J. & Pérez-Casany, M. (1998), 'Weighted Poisson distributions for overdispersion and underdispersion situations', *Annals of the Institute of Statistical Mathematics* **50**(3), 567–585.

Hinde, J. & Demétrio, C. G. B. (1998), 'Overdispersion: models and estimation', *Computational Statistics & Data Analysis* **27**(2), 151–170.

Jørgensen, B. & Kokonendji, C. C. (2016), 'Discrete dispersion models and their tweedie asymptotics', *Advances ub Statistical Analysis* **100**, 43–78.

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384.

Ribeiro Jr, E. E., Zeviani, W. M., Bonat, W. H., Demétrio, C. G. B. & Hinde, J. (2018), 'Reparametrization of COM-Poisson regression models with applications in the analysis of experimental data', *arXiv (Statistics Applications and Statistics Methodology)* .

References

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. & Boatwright, P. (2005), 'A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution', *Journal of the Royal Statistical Society. Series C: Applied Statistics* **54**(1), 127–142.

Winkelmann, R. (1995), 'Duration Dependence and Dispersion in Count-Data Models', *Journal of Business & Economic Statistics* **13**(4), 467–474.

Zamani, H. & Ismail, N. (2012), 'Functional form for the generalized Poisson regression model', *Communication in Statistics – Theory and Methods* **41**, 3666–3675.