Reparametrization of COM-Poisson Regression Models with Applications in the Analysis of Experimental Count Data

Eduardo Elias Ribeiro Junior [1 2]
Walmes Marques Zeviani [1]
Wagner Hugo Bonat [1]
Clarice Garcia Borges Demétrio [2]

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

25th July 2017

<jreduardo@usp.br> | <edujrrib@gmail.com>

Outline

1. Background
2. Reparametrization
3. Case studies
4. Final remarks

1

Background

Count data

Random variables that assume non-negative integer values, representing the number of times an event occurs in the observation unit.

Let Y a counting random variable, so $y = 0, 1, 2, \dots$

Examples in experimental researches:

- number of grains produced by a plant;
- number of fruits produced by a tree;
- number of insects on a particular cell;
- among others;

Poisson model and limitations

GLM framework (NELDER; WEDDERBURN, 1972)

- Suitable for a support of the random variable;
- Efficient algorithm for estimation and inference;
- Relantionship between mean and variance, $E(Y) = \text{var}(Y)$;
- Implemented in many software.

Main limitations

- Overdispersion (more commom), $E(Y) < \text{var}(Y)$
- Underdispersion (less commom), $E(Y) > \text{var}(Y)$

COM-Poisson distribution

- Allows for a non-linear decrease in ratios of successive probabilities, $\frac{\Pr(Y=y-1)}{\Pr(Y=y)} = \frac{y^\nu}{\lambda}$;
- Probability mass function (SHMUELI et al., 2005) takes the form

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \qquad Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}, \tag{1}$$

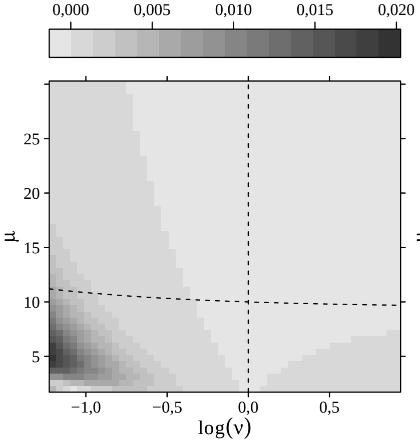
where $\lambda > 0$ and $\nu \geq 0$

- Moments are not available in closed form;
- Expected and variance values can be closely approximeted by

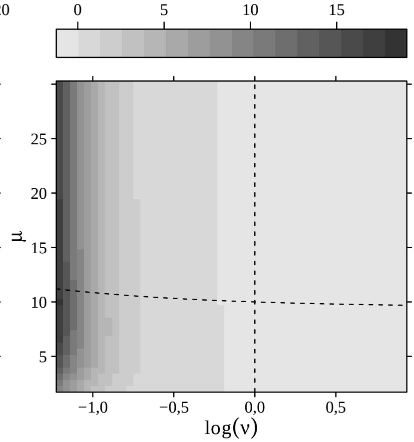
$$E(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \qquad \text{and} \qquad \mathrm{var}(Y) \approx \frac{\lambda^{1/\nu}}{\nu} \tag{2}$$

with accurate approximations for $\nu \leq 1$ or $\lambda > 10^\nu$ (SHMUELI et al., 2005; SELLERS; BORLE; SHMUELI, 2012)

Study of the moments approximations



(a) Errors for mean approximation



(b) Errors for variance approximation

Figure: Quadratic errors for moments approximation. Dotted lines representing the restriction for good approximations by (SHMUELI et al., 2005).

COM-Poisson regression models

Model definition

- Modelling relationship between $E(\boldsymbol{Y})$ and $\boldsymbol{x}$ indirectly (SELLERS; SHMUELI, 2010);

$$Y_i \mid \boldsymbol{x}_i \sim \text{COM-Poisson}(\lambda_i, \nu)$$
$$\eta(E(Y_i \mid \boldsymbol{x}_i)) = \log(\lambda_i) = \boldsymbol{x}_i^t \boldsymbol{\beta}$$

2

Reparametrization

Reparametrized COM-Poisson

Reparametrization

- Introduced new parameter μ, using the mean approximation

$$\mu = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad \Rightarrow \quad \lambda = \left(\mu + \frac{(\nu - 1)}{2\nu}\right)^{\nu}; \tag{3}$$

- Precision parameter is taken on the log scale, for avoid restrict parametric support

$$\phi = \log(\nu) \Rightarrow \phi \in \mathbb{R}$$

Probability mass function

- Replacing λ and ν as function of μ and ϕ in Equation 1

$$\Pr(Y = y \mid \mu, \phi) = \left(\mu + \frac{e^{\phi} - 1}{2e^{\phi}}\right)^{ye^{\phi}} \frac{(y!)^{-e^{\phi}}}{Z(\mu, \phi)}. \tag{4}$$

COM-Poisson$_\mu$ distribution

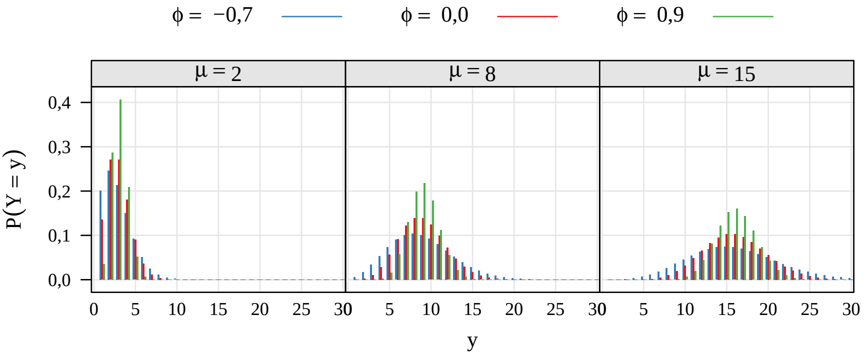


Figure: Probabilities from COM-Poisson distributions with different parameters.

COM-Poisson$_\mu$ regression models

Model definition

- Modelling relationship between $E(\boldsymbol{Y})$ and $\boldsymbol{x}$ directly

$$\begin{aligned} Y_i \mid \boldsymbol{x}_i &\sim \text{COM-Poisson}_\mu(\mu_i, \phi) \\ \log(E(Y_i \mid \boldsymbol{x}_i)) &= \log(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} \end{aligned}$$

Estimation and Inference

- Obtain parameters estimates by numerical maximization of likelihood;
- Maximization by BFGS algorithm, derivative-free (NOCEDAL; WRIGHT, 1995);
- Standard errors of coefficients are obtained by Wald method;
- Confidence intervals for $\hat{\mu}_i$ are obtained by delta method.

3

Case studies

Artificial defoliation in cotton phenology

Aim: to assess the effects of five defoliation levels on the bolls produced at five growth stages;

Design: factorial 5×5, with 5 replicates;

Experimental unit: a vase with 2 plants;

Factors:

- Artificial defoliation (`des`): 0, 0,25, 0,5, 0,75, 1
- Growth stage (`est`): vegetative, flower bud, blossom, fig, cotton boll

Response variable: Total number of cotton bolls;

Define model

Linear predictor: following Zeviani et al. (2014)

- $\log(\mu_{ij}) = \beta_0 + \beta_{1j}\text{def}_i + \beta_{2j}\text{def}_i^2$
  i varies in the levels of artificial defoliation;
  j varies in the levels of growth stages.

Models fitted:

- Poisson (μ_{ij});
- COM-Poisson (λ_{ij} , ϕ)
- COM-Poisson$_\mu$ (μ_{ij} , ϕ)
- Quasi-Poisson ($\text{var}(Y_{ij}) = \sigma\mu_{ij}$)

Parameter estimates

Table: Parameter estimates (Est) and ratio between estimate and standard error (SE)

	Poisson		COM-Poisson		COM-Poisson$_\mu$		Quasi-Poisson	
	Est	Est/SE	Est	Est/SE	Est	Est/SE	Est	Est/SE
ϕ, σ			1,585	12,417	1,582	12,392	0,241	
β_0	2,190	34,572	10,897	7,759	2,190	74,640	2,190	70,420
β_{11}	0,437	0,847	2,019	1,770	0,435	1,819	0,437	1,726
β_{12}	0,290	0,571	1,343	1,211	0,288	1,223	0,290	1,162
β_{13}	-1,242	-2,058	-5,750	-3,886	-1,247	-4,420	-1,242	-4,192
β_{14}	0,365	0,645	1,595	1,298	0,350	1,328	0,365	1,314
β_{15}	0,009	0,018	0,038	0,035	0,008	0,032	0,009	0,036
β_{21}	-0,805	-1,379	-3,725	-2,775	-0,803	-2,961	-0,805	-2,809
β_{22}	-0,488	-0,861	-2,265	-1,805	-0,486	-1,850	-0,488	-1,754
β_{23}	0,673	0,989	3,135	2,084	0,679	2,135	0,673	2,015
β_{24}	-1,310	-1,948	-5,894	-3,657	-1,288	-4,095	-1,310	-3,967
β_{25}	-0,020	-0,036	-0,090	-0,076	-0,019	-0,074	-0,020	-0,074
LogLik	-255,803		-208,250		-208,398		—	
AIC	533,606		440,500		440,795		—	
BIC	564,718		474,440		474,735		—	

Predictive curves

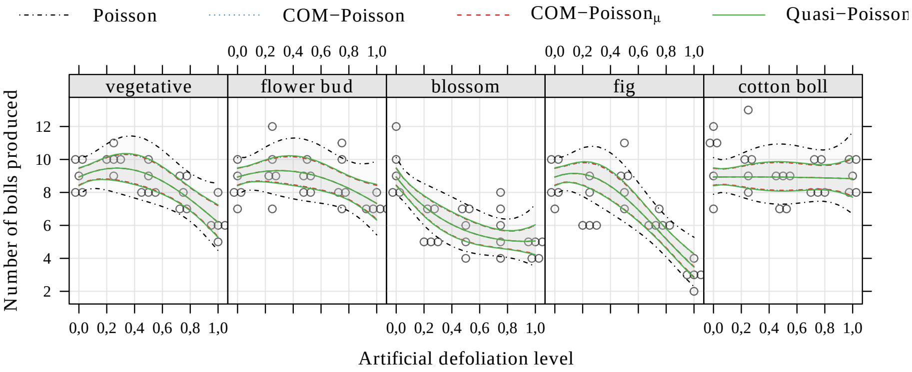


Figure: Curves of predicted values with 95% confidence intervals.

Additional results

- Empirical correlations between $\hat{\phi}$ and $\hat{\beta}$ estimators is approximately 0 for reparametrized model.

Table: Empirical correlations between dispersion and location parameters estimators.

	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\beta}_{23}$	$\hat{\beta}_{24}$	$\hat{\beta}_{25}$
COM-Poisson	0,995	0,223	0,153	-0,490	0,161	0,004	-0,350	-0,228	0,263	-0,458	-0,009
COM-Poisson$_\mu$	0,001	-0,000	-0,000	-0,001	-0,001	-0,000	0,000	0,000	0,001	0,002	0,000

- The computational time consuming to fit COM-Poisson was 1,258 times the time consuming by COM-Poisson$_\mu$.

4

Final remarks

Concluding remarks

Summary

- Over/under-dispersion need caution;
- COM-Poisson is a suitable distribution for this situations;
- The proposed reparametrization, COM-Poisson$_\mu$ has some advantages:
  - Simple transformation of the parametric space;
  - Full parametric approach (likelihood-based inference);
  - Correlation among the estimators was practically null;
  - Faster computational times;
  - Allows interpretation of the coefficients directly (like GLM-Poisson model).

Future work

- Simulation study for assessment model robustness in terms of bad specification of the distribution;
- Assessment of theoretical approximations for $Z(\lambda, \nu)$ (or $Z(\mu, \phi)$), may be useful in avoiding the requirement to select an upper bound sum;
- Modelling location and dispersion parameters jointly.

Provision

- Full-text article is available on ResearchGate (in portuguese) <https://www.researchgate.net/publication/316880329>
- All codes (in R) and source files are available on GitHub <https://github.com/jreduardo/rbras2017>

Acknowledgments

- National Council for Scientific and Technological Development (CNPq), for their support

References

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, p. 370–384, 1972.

NOCEDAL, J.; WRIGHT, S. J. *Numerical optimization*. [S.l.]: Springer, 1995. 636 p. ISSN 0011-4235. ISBN 0387987932.

SELLERS, K. F.; BORLE, S.; SHMUELI, G. The com-poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, v. 28, n. 2, p. 104–116, 2012.

SELLERS, K. F.; SHMUELI, G. A flexible regression model for count data. *Annals of Applied Statistics*, v. 4, n. 2, p. 943–961, 2010. ISSN 19326157.

SHMUELI, G. et al. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, v. 54, n. 1, p. 127–142, 2005.

ZEVIANI, W. M. et al. The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, p. 1–11, 2014.