

XXIXTH INTERNATIONAL BIOMETRIC CONFERENCE
Barcelona, Spain, 8-13 July 2018

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

Eduardo Elias Ribeiro Junior^{1 2}
Walmes Marques Zeviani¹
Wagner Hugo Bonat¹
Clarice Garcia Borges Demétrio²
John Hinde³

¹Statistics and Geoinformation Laboratory (LEG-UFPR)

²Department of Exact Sciences (ESALQ-USP)

³School of Mathematics, Statistics and Applied Mathematics (NUI-Galway)

10th July 2018

jreduardo@usp.br | edujrrib@gmail.com

Outline

1. Background
2. Reparametrization
3. Simulation study
4. Case studies
5. Final remarks

1

Background

Count data

Number of times an event occurs in the observation unit.

Random variables that assume non-negative integer values.

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

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;
- ▶ others.

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, $E(Y) = \text{Var}(Y)$;

Main limitations

- ▶ Overdispersion (more common), $E(Y) < \text{Var}(Y)$
- ▶ Underdispersion (less common), $E(Y) > \text{Var}(Y)$

COM-Poisson distribution

- Probability mass function (Shmueli et al. 2005) takes the form

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

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

- Moments are not available in closed form;
- Expectation and variance can be closely approximated by

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

with accurate approximations for $\nu \leq 1$ or $\lambda > 10^\nu$ (Shmueli et al. 2005, Sellers et al. 2012).

COM-Poisson regression models

Model definition

- ▶ Modelling the relationship between $E(Y_i)$ and \mathbf{x}_i indirectly (Sellers & Shmueli 2010);

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

Main goals

- ▶ Study distribution properties in terms of i) modelling real count data and ii) inference aspects.
- ▶ Propose a reparametrization in order to model the expectation of the response variable as a function of the covariate values directly.

2

Reparametrization

Reparametrized COM-Poisson

Reparametrization

- ▶ Introduced new parameter μ , using the mean approximation

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

- ▶ Precision parameter is taken on the log scale to avoid restrictions on the parameter space

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

Probability mass function

- ▶ Replacing λ and ν as function of μ and ϕ in the pmf of COM-Poisson

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

Study of the moments approximations

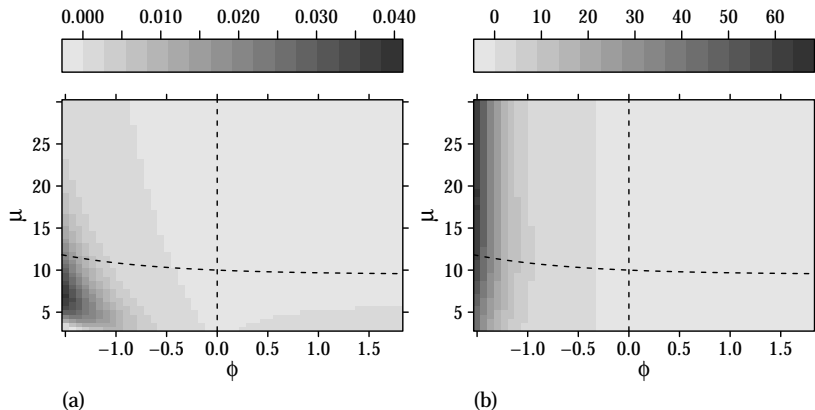


Figure: Quadratic errors for the approximation of the (a) expectation and (b) variance. Dotted lines represent the restriction for suitable approximations given by Shmueli et al. (2005).

COM-Poisson $_{\mu}$ distribution

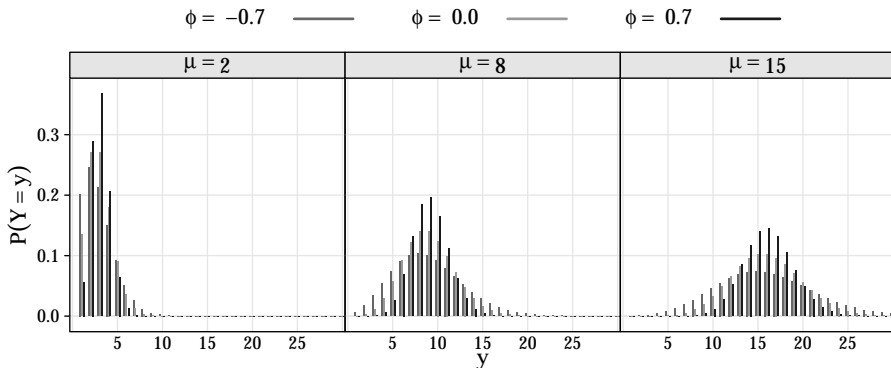


Figure: Shapes of the COM-Poisson distribution for different parameter values.

Properties of COM-Poisson distribution

To explore the flexibility of the COM-Poisson distribution, we consider the follow indexes:

- ▶ **Dispersion index:** $DI = \text{Var}(Y)/E(Y)$;
- ▶ **Zero-inflation index:** $ZI = 1 + \log \Pr(Y = 0)/E(Y)$;
- ▶ **Heavy-tail index:** $HT = \Pr(Y = y + 1) / \Pr(Y = y)$, for $y \rightarrow \infty$.

These indexes are interpreted in relation to the Poisson distribution:

- ▶ over- ($DI > 1$), under- ($DI < 1$) and equidispersion ($DI = 1$);
- ▶ zero-inflation ($ZI > 0$) and zero-deflation ($ZI < 0$) and
- ▶ heavy-tail distribution for $HT \rightarrow 1$ when $y \rightarrow \infty$.

Properties of COM-Poisson distribution

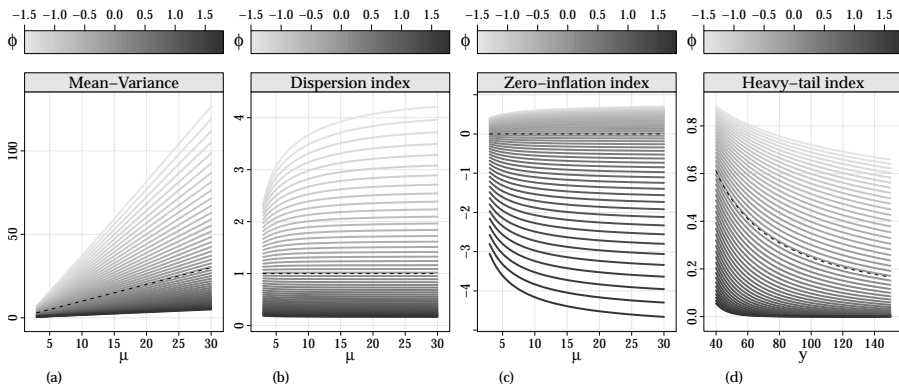


Figure: Indexes for COM-Poisson distribution. (a) Mean and variance relationship, (b–d) dispersion, zero-inflation and heavy-tail indexes for different parameter values. Dotted lines represents the Poisson special case.

COM-Poisson $_{\mu}$ regression models

Let y_i a set of independent observations from the COM-Poisson and $\mathbf{x}_i^{\top} = (x_{i1}, x_{i2}, \dots, x_{ip})$ is a vector of known covariates, $i = 1, 2, \dots, n$.

Model definition

- Modelling relationship between $E(Y_i)$ and \mathbf{x}_i directly

$$Y_i \mid \mathbf{x}_i \sim \text{COM-Poisson}_{\mu}(\mu_i, \phi)$$

$$\log(E(Y_i \mid \mathbf{x}_i)) = \log(\mu_i) = \mathbf{x}_i^{\top} \boldsymbol{\beta}$$

Log-likelihood function ($\ell = \ell(\boldsymbol{\beta}, \phi \mid \mathbf{y})$)

- $$\ell = e^{\phi} \left[\sum_{i=1}^n y_i \log \left(\mu_i + \frac{e^{\phi} - 1}{2e^{\phi}} \right) - \sum_{i=1}^n \log(y_i!) \right] - \sum_{i=1}^n \log(Z(\mu_i, \phi))$$

where $\mu_i = \exp(\mathbf{x}_i^{\top} \boldsymbol{\beta})$

Estimation and inference

The estimation and inference is based on the method of maximum likelihood. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$ the model parameters.

- ▶ Parameter estimates are obtained by numerical maximization of the log-likelihood function (by BFGS algorithm);
 $\ell(\hat{\boldsymbol{\theta}}) = \max \ell(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^{p+1};$
- ▶ Standard errors for regression coefficients are obtained based on the observed information matrix;
 $\text{Var}(\hat{\boldsymbol{\theta}}) = -\mathcal{H}^{-1}$, where \mathcal{H} is the matrix of second partial derivatives at $\hat{\boldsymbol{\theta}}$;
- ▶ Confidence intervals for $\hat{\mu}_i$ are obtained by delta method.
 $\text{Var}[g(\hat{\boldsymbol{\theta}})] \doteq \mathbf{G} \text{Var}(\hat{\boldsymbol{\theta}}) \mathbf{G}^\top$, where $\mathbf{G}^\top = (\partial g / \partial \beta_1, \dots, \partial g / \partial \beta_p)^\top$;
- ▶ The Hessian matrix \mathcal{H} is obtained numerically by finite differences.

3

Simulation study

Definitions on the simulation study

Objective: assess the properties of maximum likelihood estimators and orthogonality in the reparametrized model;

Simulation: we consider counts generated according a regression model with a continuous and categorical covariates and different dispersion scenarios.

Algorithm 1: Steps in simulation study.

```

for  $n \in \{50, 100, 300, 1000\}$  do
    set  $x_1$  as a sequence, with  $n$  elements, between 0 and 1;
    set  $x_2$  as a repetition, with  $n$  elements, of three categories;
    compute  $\mu$  using  $\mu = \exp(\beta_0 + \beta_1 x_1 + \beta_{21} x_{21} + \beta_{22} x_{22})$ ;
    for  $\phi \in \{-1.6, -1.0, 0.0, 1.8\}$  do
        repeat
            simulate  $y$  from COM-Poisson distribution with  $\mu$  and  $\phi$  parameters;
            fit COM-Poisson $_{\mu}$  regression model to simulated  $y$ ;
            get  $\hat{\theta} = (\hat{\phi}, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_{21}, \hat{\beta}_{22})$ ;
            get confidence intervals for  $\hat{\theta}$  based on the observed information matrix.
        until 1000 times;
  
```

Definitions on the simulation study

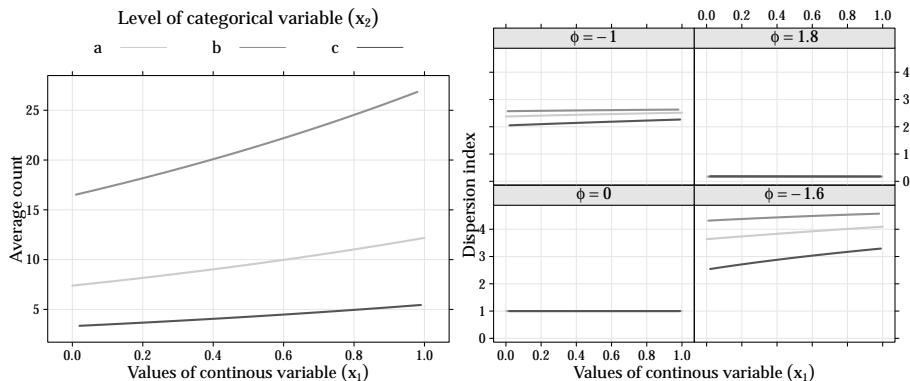
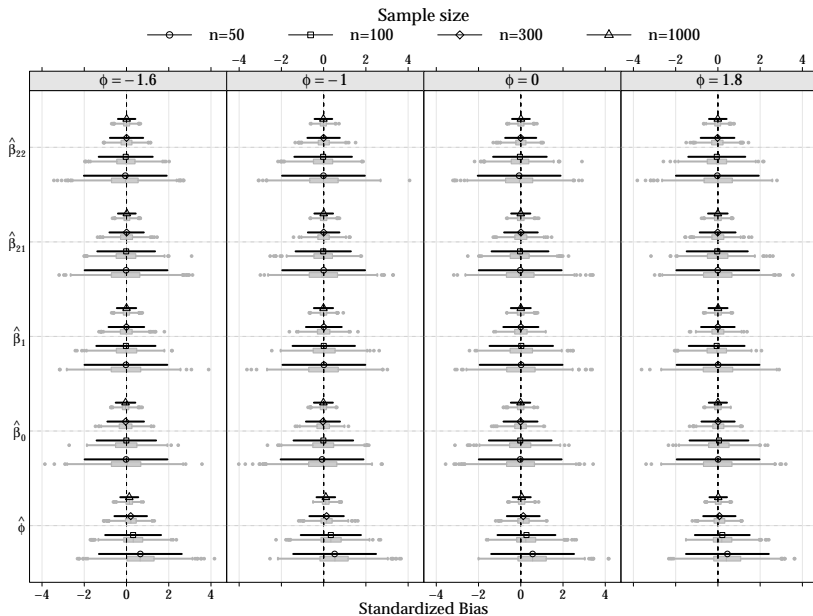


Figure: Average counts (left) and dispersion indexes (right) for each scenario considered in the simulation study.

Bias of the estimators



Coverage rate of the confidence intervals

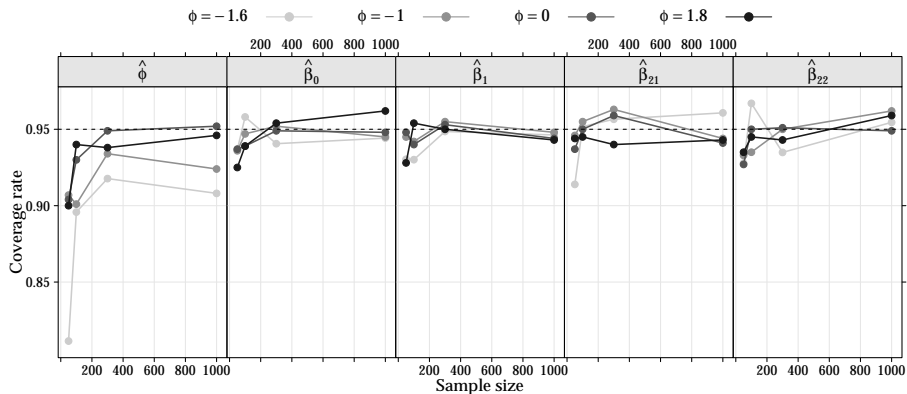


Figure: Coverage rate based on confidence intervals obtained by quadratic approximation for different sample sizes and dispersion levels.

Orthogonality property of the MLEs

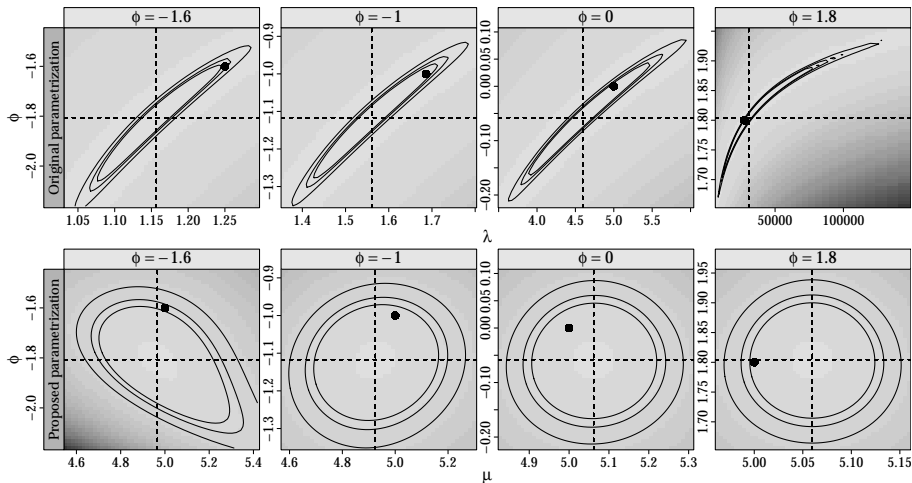


Figure: Deviance surfaces contour plots under original and proposed parametrization. The ellipses are confidence regions (90, 95 and 99%), dotted lines are the maximum likelihood estimates, and points are the real parameters used in the simulation.

4

Case studies

Motivating data sets and data analysis

- ▶ Three illustrative examples of count data analysis are reported.
 - ▶ Assessing toxicity of nitrofen in aquatic systems, an equidispersed example;
 - ▶ Soil moisture and potassium doses on soybean culture, an overdispersed example; and
 - ▶ Artificial defoliation in cotton phenology, an underdispersed example.
- ▶ In the data analysis, we consider the COM-Poisson model in the two forms (original and new parametrization) and the quasi-Poisson regression model as alternative models for the standard Poisson regression model.

4.1

Case studies
**Artificial defoliation in cotton
phenology**

Cotton bolls data



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 plot with 2 plants;

Factors:

- ▶ Artificial defoliation (des):
- ▶ Growth stage (est):

Response variable: Total number of cotton bolls;

Model specification

Linear predictor: following Zeviani et al. (2014)

- ▶ $\log(\mu_{ij}) = \beta_0 + \beta_1 \text{def}_i + \beta_2 \text{def}_i^2$
 i varies in the levels of artificial defoliation;
 j varies in the levels of growth stages.

Alternative models:

- ▶ Poisson (μ_{ij});
- ▶ COM-Poisson ($\lambda_{ij} = \eta(\mu_{ij}), \phi$)
- ▶ COM-Poisson _{μ} (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 _{μ}		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		—	

Fitted curves

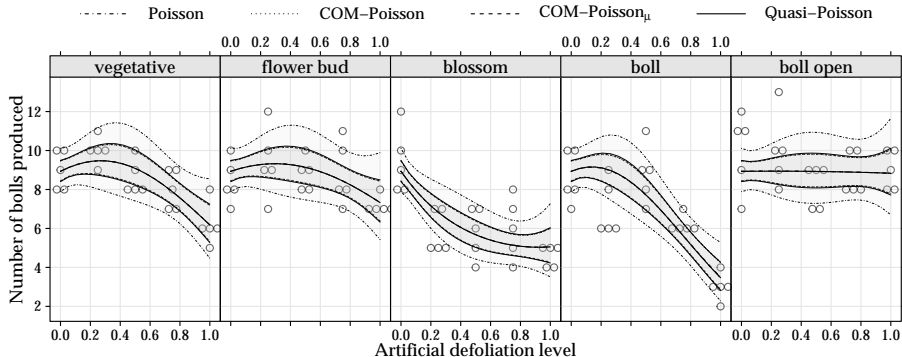


Figure: Scatterplots of the observed data and curves of fitted values with 95% confidence intervals as functions of the defoliation level for each growth stage.

4.2

Case studies
Additional results

Computational times and orthogonality property

To compare the computational times on the two parametrizations we repeat the fitting 50 times.

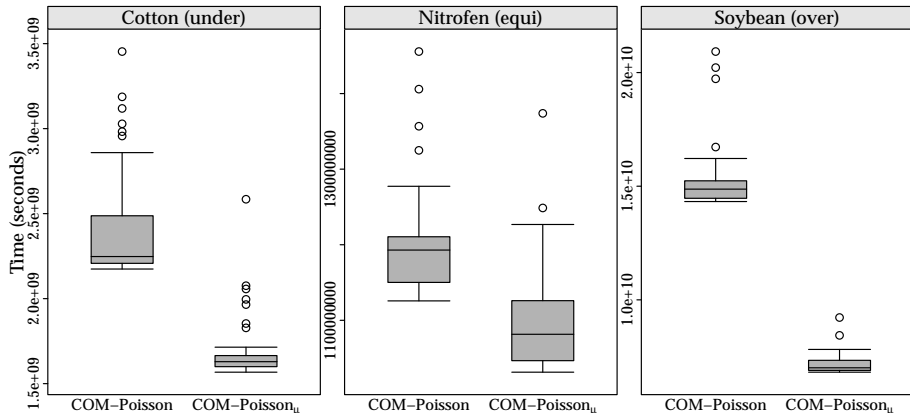


Figure: Computational times to fit the models under original and reparametrized versions based on the fifty repetitions.

5

Final remarks



Concluding remarks

Summary


- ▶ Over/under-dispersion needs caution;
- ▶ COM-Poisson is a suitable choice for these situations;
- ▶ The proposed reparametrization, COM-Poisson_μ has some advantages:
 - ▶ Simple transformation of the parameter space;
 - ▶ Leads to the orthogonality of the parameters (seen empirically);
 - ▶ Full parametric approach;
 - ▶ Empirical correlation between the estimators was practically null;
 - ▶ Faster for fitting;
 - ▶ Allows interpretation of the coefficients directly (like GLM-Poisson model).

Future work

- ▶ Simulation study to assess model robustness against distribution miss specification;
- ▶ Assess theoretical approximations for $Z(\lambda, \nu)$ (or $Z(\mu, \phi)$), in order to avoid the selection of sum's upper bound;
- ▶ Propose a mixed GLM based on the COM-Poisson_μ model.

- ▶  Full-text article is available on arXiv
<https://arxiv.org/abs/1801.09795>
- ▶  All codes (in R) and source files are available on GitHub
<https://github.com/jreduardo/article-reparcmp>

Acknowledgements

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

References

- Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384.
- Sellers, K. F., Borle, S. & Shmueli, G. (2012), 'The com-poisson model for count data: a survey of methods and applications', *Applied Stochastic Models in Business and Industry* **28**(2), 104–116.
- Sellers, K. F. & Shmueli, G. (2010), 'A flexible regression model for count data', *Annals of Applied Statistics* **4**(2), 943–961.
- 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.
- Zeviani, W. M., Ribeiro Jr, P. J., Bonat, W. H., Shimakura, S. E. & Muniz, J. A. (2014), 'The Gamma-count distribution in the analysis of experimental underdispersed data', *Journal of Applied Statistics* pp. 1–11.