Double COM-Poisson models: modelling mean and dispersion in the analysis of 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)

24th May 2018

jreduardo@usp.br | edujrrib@gmail.com

Outline

1

Introduction

Standard regression models

Generalized Linear Models (GLM) (Nelder & Wedderburn 1972):

Let $(y_i, \boldsymbol{x}_i)$ a cross-section data set where $y_i's$ are iid realizations of Y_i according to the exponential family (EF) distribution. The GLM is specified as follow

$$\begin{matrix} Y_i \sim \mathrm{EF}(\mu_i, \phi) \\ g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \end{matrix} \quad \Longrightarrow \quad \begin{matrix} \mathrm{E}(Y_i) = \mu_i \\ \mathrm{Var}(Y_i) = \phi V(\mu_i). \end{matrix}$$

Main limitations

- The exponential family is often restrictive (variance function);
- The only choice for count data analysis is the Poisson distribution;
- Only the mean parameter is allowed to depend on covariates.

1.1

Introduction

Motivating data set

Assessing toxicity of nitrofen in aquatic systems

Implication in Biology

- Nitrofen is a herbicide that was used extensively for the control of broad-leaved and grass weeds in cereals and rice;
- It is also acutely toxic and reproductively toxic to cladoceran zooplankton;
- Nitrofen is no longer in commercial use in the U.S.

Experimental study

- Assess the reproductive toxicity on a species of zooplankton (*Ceriodaphnia dubia*);
- Fifty animals were randomized into batches of ten;
- Each batch was put in a solution with a concentration level of nitrofen;
- The number of live offspring was recorded.

Descriptive analysis

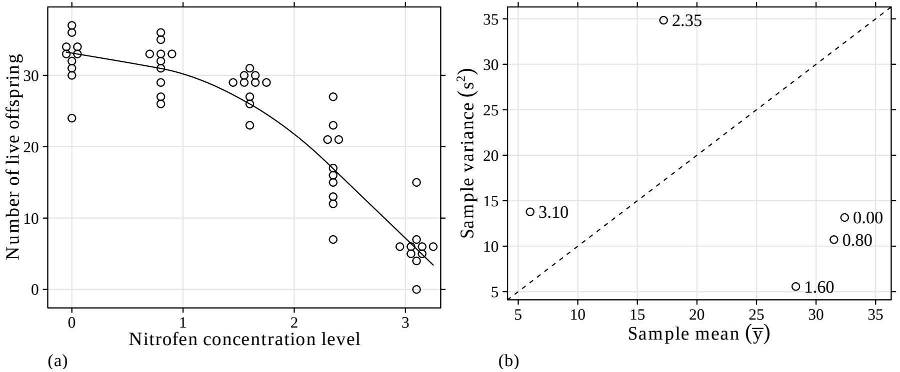

Figure: (a) Number of live offsprings observed for each nitrofen concentration level and (b) scatterplot of the sample means against sample variances.

2 Double COM-Poisson models

COM-Poisson distribution

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

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

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

- Moments are not available in closed form;

- Expectation and variance can be approximated by

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

Reparametrized COM-Poisson

Following Ribeiro Jr et al. (2018), we use the mean-parametrized COM-Poisson, introducing the new parameter μ by means of the approximation,

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

Model parameters:

- $\mu \in \mathbb{R}_+$, the mean parameter;
- $\nu \in \mathbb{R}_+$, the dispersion parameter ($\nu < 1 \implies$ over- and $\nu > 1 \implies$ underdispersion).

Orthogonality property

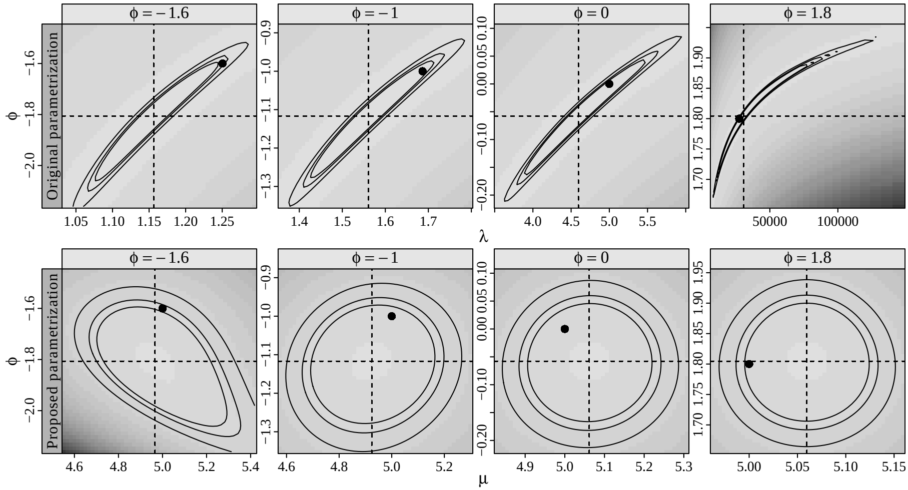

Figure: Deviance surface contour plots under original and proposed parametrization for four simulated data sets. $\phi = \log(\nu)$.

Regression models for mean and dispersion

Double COM-Poisson regression models (DCMP)

Let $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ a data set where $y_i's$ are iid realizations of Y_i according to the COM-Poisson distribution distribution and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are sub-vectors of the covariates vector. The DCMP is specified as follow

$$Y_i \sim \text{CMP}_\mu(\mu_i, \nu_i), \quad \text{where} \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad g(\nu_i) = \boldsymbol{z}_i^\top \boldsymbol{\gamma}.$$

Log-likelihood function

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \nu_i \log\left(\mu_i + \frac{\nu_i - 1}{2\nu_i} \right) - \nu_i \log(y_i) - \log[Z(\mu_i, \nu_i)], \right\}$$

where $\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ and $\nu_i = g^{-1}(\boldsymbol{z}_i^\top \boldsymbol{\gamma})$

Estimation and inference

- Parameters estimates are obtained by numerical maximization of the log-likelihood function (by BFGS algorithm)
- Standard errors for regression coefficients (for mean and dispersion) are obtained based on the observed information matrix
  $\boldsymbol{V}_{\beta|\gamma} = \boldsymbol{V}_\beta - (\boldsymbol{V}_{\beta,\gamma}\boldsymbol{V}_\gamma^{-1})^\top \boldsymbol{V}_{\beta,\gamma}$ and $\boldsymbol{V}_{\gamma|\beta} = \boldsymbol{V}_\gamma - (\boldsymbol{V}_{\gamma,\beta}\boldsymbol{V}_\beta^{-1})^\top \boldsymbol{V}_{\gamma,\beta}$.

Strategies

- **Joint:** Estimate $(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$ using the complete log-likelihood function;
- **Fixed:** Set the $\hat{\boldsymbol{\beta}}$ in the Poisson MLE, estimate $\boldsymbol{\gamma}$ (with fixed $\boldsymbol{\beta}$) and then estimate the Hessian matrix for $(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$.

3

Data analysis

Model specification

For mean:

Cubic: $\log(\mu_i) = \beta_0 + \beta_1 \mathsf{dose}_i + \beta_2 \mathsf{dose}_i^2 + \beta_3 \mathsf{dose}_i^3$

For dispersion:

Constant: $\log(\nu_i) = \gamma_0,$

Linear: $\log(\nu_i) = \gamma_0 + \gamma_1 \mathsf{dose}_i,$

Quadratic: $\log(\nu_i) = \gamma_0 + \gamma_1 \mathsf{dose}_i + \gamma_2 \mathsf{dose}_i^2$ e

Cubic: $\log(\nu_i) = \gamma_0 + \gamma_1 \mathsf{dose}_i + \gamma_2 \mathsf{dose}_i^2 + \gamma_3 \mathsf{dose}_i^3.$

Table: Estimates and standard errors.

Parameter	Estimate (Erro Padrão) Constant	Linear	Quadratic	Cubic
β_0	2.981 (0.035)[a]	2.978 (0.042)[a]	2.972 (0.049)[a]	2.975 (0.047)[a]
β_1	−3.952 (0.287)[a]	−3.980 (0.365)[a]	−4.041 (0.447)[a]	−4.013 (0.418)[a]
β_2	−2.131 (0.260)[a]	−2.161 (0.311)[a]	−2.218 (0.351)[a]	−2.197 (0.330)[a]
β_3	−0.543 (0.221)[a]	−0.573 (0.212)[a]	−0.604 (0.206)[a]	−0.597 (0.195)[a]
γ_0	0.048 (0.205)	0.295 (0.211)	0.243 (0.259)	0.353 (0.227)
γ_1	—	−5.244 (1.363)[a]	−7.013 (2.307)[a]	−5.729 (1.844)[a]
γ_2	—	—	−3.984 (2.444)	−2.918 (1.904)
γ_3	—	—	—	1.522 (1.412)

Est (EP)[a] indicates $|\text{Est}/\text{EP}| > 1,96$.

Table: Model fit measures and comparisons.

	D.f	Deviance	AIC	χ^2	$\Pr(> \chi^2)$
Constant	45	288.127	298.127	—	—
Linear	44	274.111	286.111	14.0163	0.0002
Quadratic	43	270.493	284.493	3.6179	0.0572
Cubic	42	269.503	285.503	0.9898	0.3198

Fitted mean and dispersion values

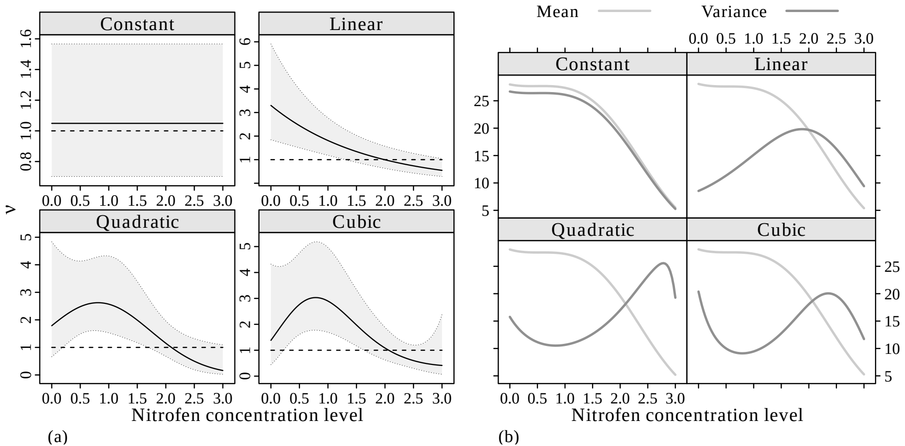

Figure: (a) Fitted values and confidence bands of 95% for de dispersion and (b) mean and variances obtained from the fitted model.

Comparison of the strategies for fitting

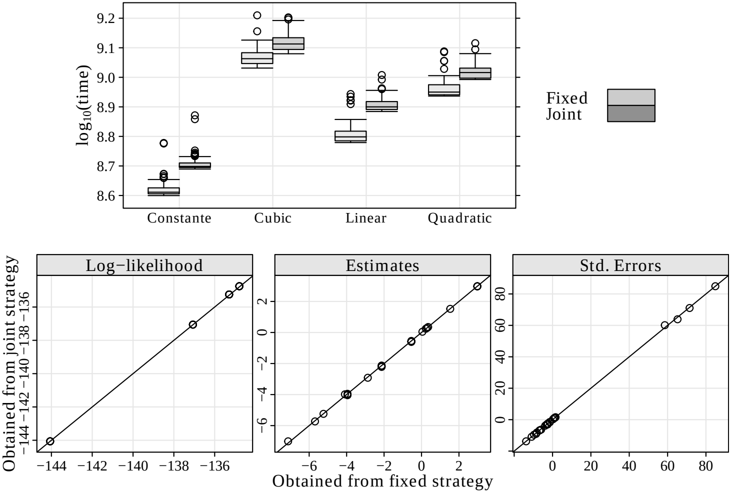

Figure: Comparison of (a) maximized likelihoods, estimates and standard errors and (b) computational times.

4

Final remarks

Concluding remarks

Summary

- We show how to allow mean and dispersion parameters to depend on covariates in the COM-Poisson regression model;;
- Estimation and inference can be done based on the likelihood paradigm;
- Using the orthogonality property in the fixed strategy for fitting is faster.

Future work

- Perform a simulation study to evaluate estimators properties;
- Compare the results with others approaches, DGLM's (Lee & Nelder 2006) and GAMLSS (Rigby & Stasinopoulos 2005).

- Extended abstract 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/rbras2018

Acknowledgments

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

References

Lee, Y. & Nelder, J. A. (2006), 'Double hierarchical generalized linear models (with discussion)', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **55**, 139–185.

Nelder, J. & Wedderburn, R. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **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)* .

Rigby, R. A. & Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape (with discussion)', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **54**, 507–554.

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.