

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)

11th October, 2018 - UNESP/Botucatu

clarice.demetrio@usp.br | jreduardo@usp.br

Outline

1. Motivating data sets
2. Background
3. Reparametrization
4. Simulation study
5. Data analysis results
6. Final remarks

1

Motivating data sets

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;

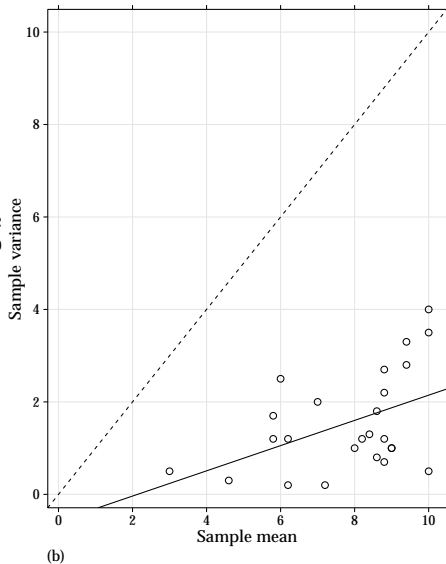
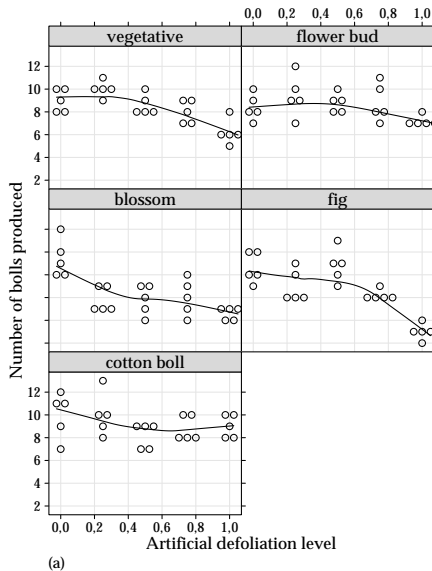
Experimental unit: a plot with 2 plants;

Factors:

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

Response variable: Total number of cotton bolls;

Artificial defoliation in cotton phenology



Potassium fertilization on soybean culture



Aim: evaluate the effects of potassium doses applied to soil in different soil moisture levels;

Design: factorial 5×3 in a randomized complete block design (5 blocks);

Experimental unit: a pot with a plant;

Factors:

- ▶ Potassium fertilization dose (K):
- ▶ Soil moisture level (umid):

Response variable: Total number of bean seeds per pot;

Potassium fertilization on soybean culture

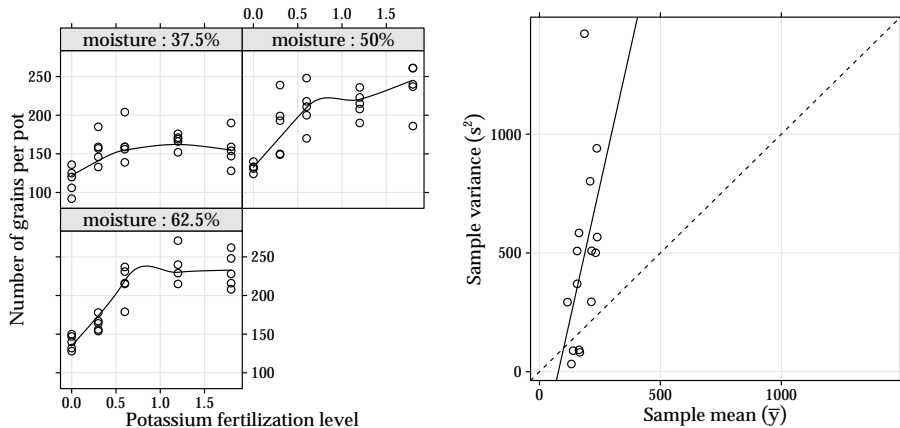


Figure: (a) Number of bean seeds per pot for each potassium dose and moisture level (left) and (b) sample mean against sample variance of the five replicates for each experimental treatment. Solid lines are smoothing curves and the least-squares line.

Assessing toxicity of nitrofen in aquatic systems



Aim: measure the reproductive toxicity of the herbicide nitrofen on a species of zooplankton (*Ceriodaphnia dubia*);

Design: completely randomized design, with 10 replicates;

Experimental unit: zooplankton animal;

Factors:

- herbicide nitrofen dose (dose);

Response variable: Total number of live offspring;

Assessing toxicity of nitrofen in aquatic systems

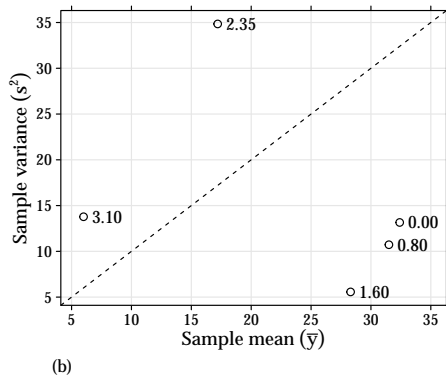
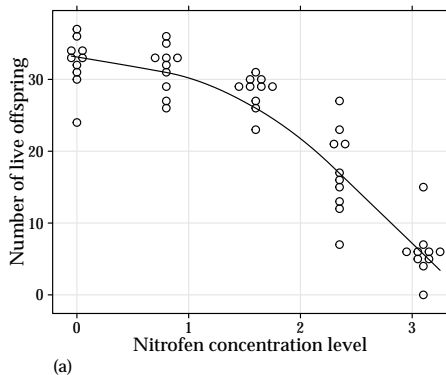


Figure: (a) Number of live offsprings observed for each nitrofen concentration level and (b) scatterplot of the sample means against sample variances.

2

Background

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.

3

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

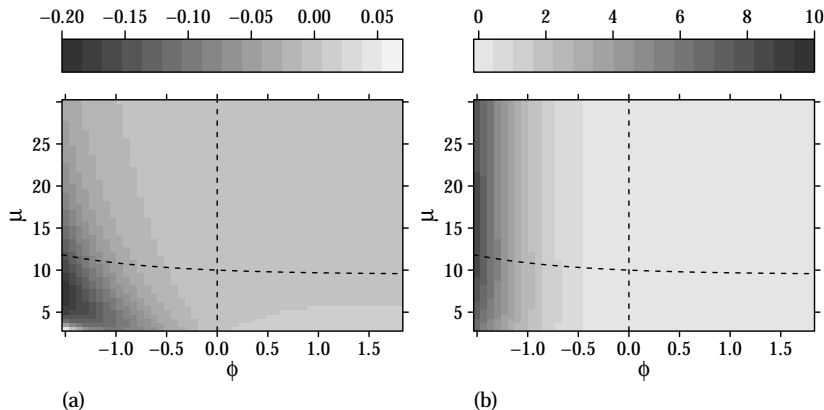


Figure: Errors for the approximation of the (a) expectation and (b) variance. Dotted lines represent the restriction for suitable approximations given by Shmueli et al. (2005).

COM-Poisson $_{\mu}$ distribution

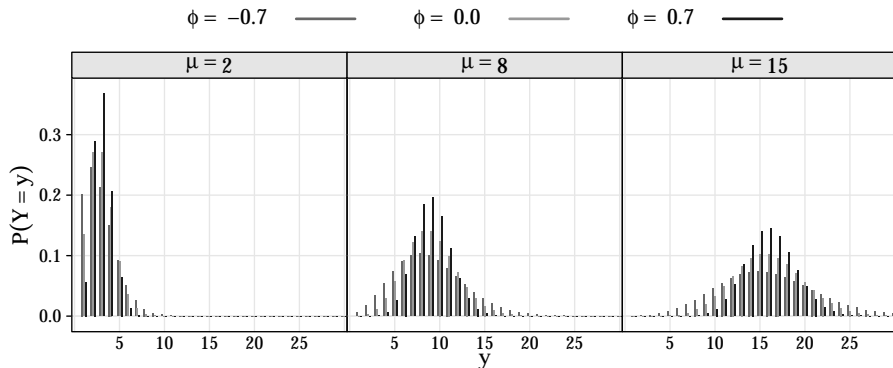


Figure: Shapes of the COM-Poisson distribution for different parameter values.

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

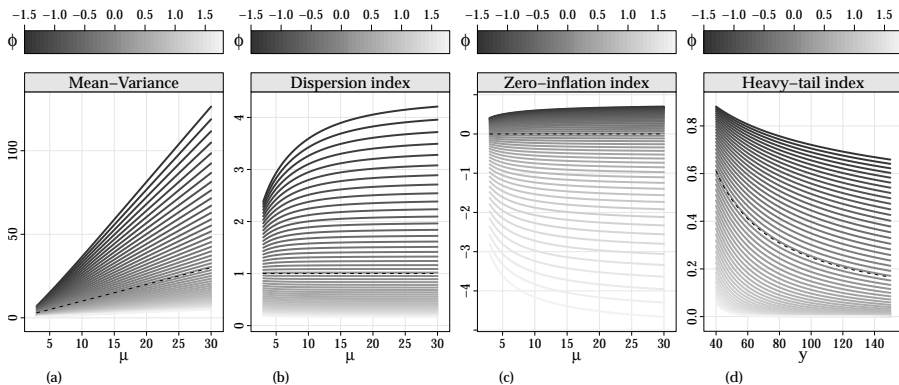


Figure: Indexes for COM-Poisson distribution. (a) Mean and variance relationship, (b–d) dispersion, zero-inflation and heavy-tail indexes for different parameter values. Dotted lines represents the Poisson special case.

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

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

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

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

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

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

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.

4

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

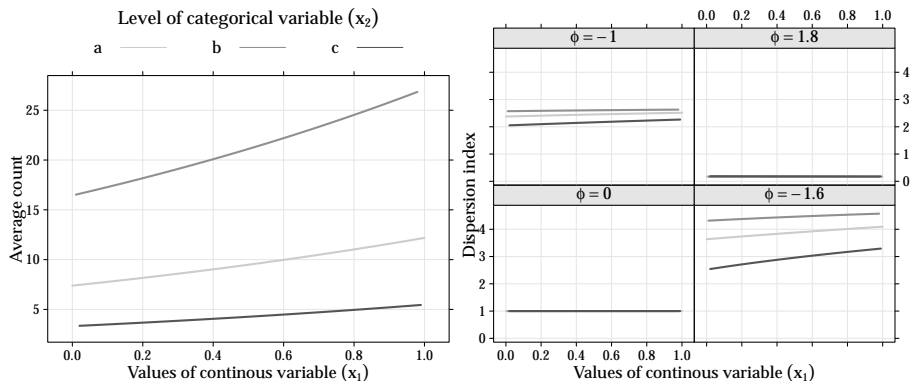
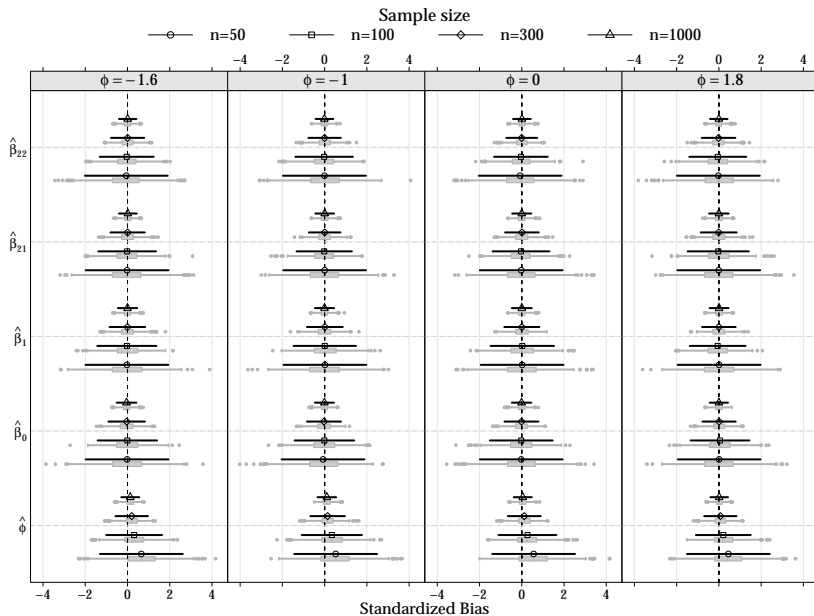


Figure: Average counts (left) and dispersion indexes (right) for each scenario considered in the simulation study.

Bias of the estimators



Coverage rate of the confidence intervals

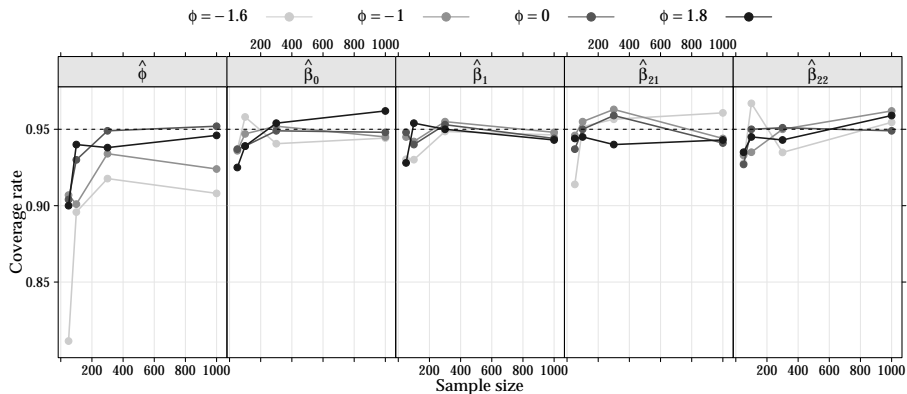


Figure: Coverage rate based on confidence intervals obtained by quadratic approximation for different sample sizes and dispersion levels.

Orthogonality property of the MLEs

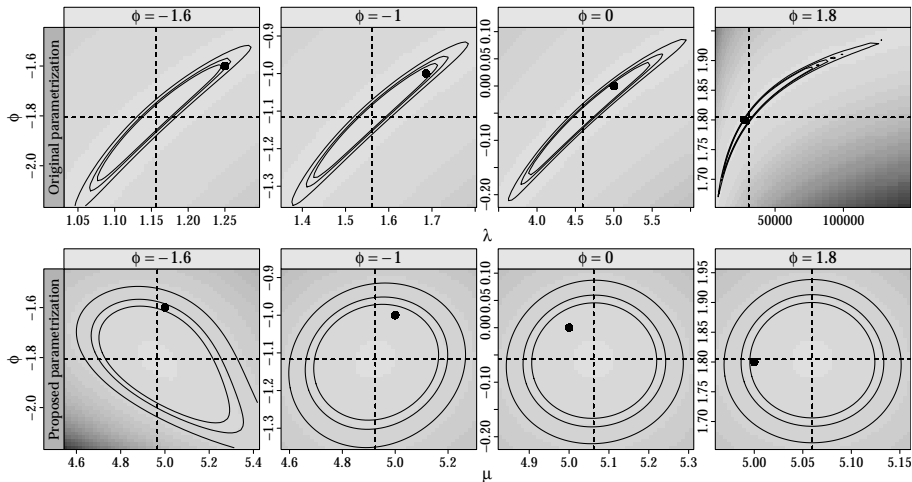


Figure: Deviance surfaces contour plots under original and proposed parametrization. The ellipses are confidence regions (90, 95 and 99%), dotted lines are the maximum likelihood estimates, and points are the real parameters used in the simulation.

5

Data analysis results

5.1

Data analysis results

**Artificial defoliation in cotton
phenology**

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

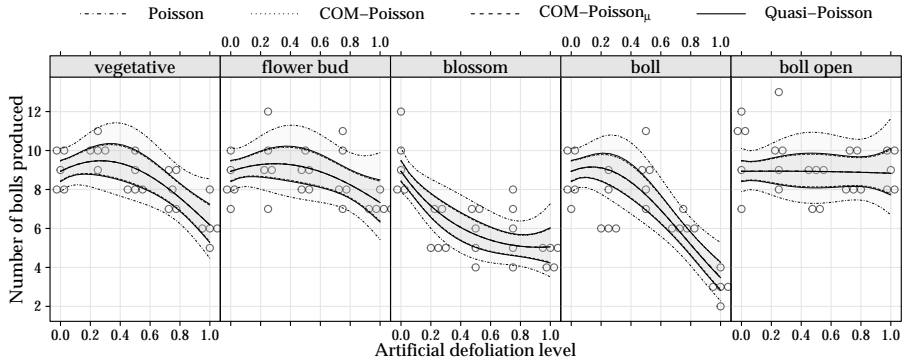


Figure: Scatterplots of the observed data and curves of fitted values with 95% confidence intervals as functions of the defoliation level for each growth stage.

5.2

Data analysis results

Soil moisture and potassium doses on soybean culture

Model specification

Linear predictor: based on descriptive analysis,

- ▶ $\log(\mu_{ijk}) = \beta_0 + \gamma_i + \tau_j + \beta_1 K_k + \beta_2 K_k^2 + \beta_3 j K_k$
 i varies according the blocks;
 j varies in the levels of soil moisture;
 k varies in the levels of potassium fertilization.

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 | | | -0.779 | -4.721 | -0.782 | -4.737 | 2.615 | |
| β_0 | 4.867 | 144.289 | 2.232 | 6.042 | 4.867 | 97.781 | 4.867 | 89.225 |
| γ_1 | -0.019 | -0.730 | -0.009 | -0.494 | -0.019 | -0.495 | -0.019 | -0.452 |
| γ_2 | -0.037 | -1.373 | -0.017 | -0.921 | -0.037 | -0.931 | -0.037 | -0.849 |
| γ_3 | -0.106 | -3.889 | -0.049 | -2.422 | -0.106 | -2.634 | -0.106 | -2.405 |
| γ_4 | -0.092 | -3.300 | -0.042 | -2.102 | -0.092 | -2.237 | -0.092 | -2.040 |
| τ_1 | 0.132 | 3.647 | 0.061 | 2.295 | 0.132 | 2.472 | 0.132 | 2.255 |
| τ_2 | 0.124 | 3.432 | 0.057 | 2.177 | 0.124 | 2.326 | 0.124 | 2.122 |
| β_1 | 0.616 | 11.014 | 0.284 | 4.729 | 0.616 | 7.464 | 0.616 | 6.811 |
| β_2 | -0.276 | -10.250 | -0.127 | -4.589 | -0.276 | -6.946 | -0.276 | -6.338 |
| β_{31} | 0.146 | 4.268 | 0.067 | 2.614 | 0.146 | 2.892 | 0.146 | 2.639 |
| β_{32} | 0.165 | 4.829 | 0.076 | 2.884 | 0.165 | 3.272 | 0.165 | 2.986 |
| LogLik | -340.082 | | -325.241 | | -325.233 | | — | |
| AIC | 702.164 | | 674.482 | | 674.467 | | — | |
| BIC | 727.508 | | 702.130 | | 702.116 | | — | |

Fitted curves

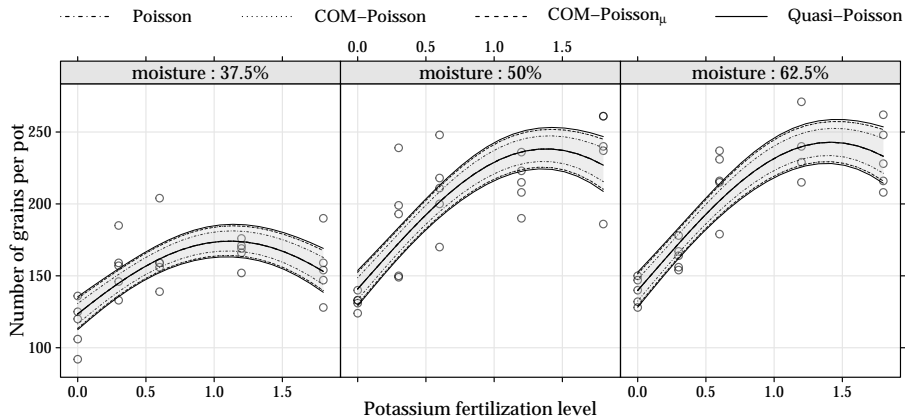


Figure: Dispersion diagrams of been seeds counts as function of potassium doses and humidity levels with fitted curves and confidence intervals (95%).

5.3

Data analysis results

**Assessing toxicity of nitrofen
in aquatic systems**

Model specification

Linear predictors:

Linear: $\log(\mu_i) = \beta_0 + \beta_1 \text{dose}_i,$
Quadratic: $\log(\mu_i) = \beta_0 + \beta_1 \text{dose}_i + \beta_2 \text{dose}_i^2$ and
Cubic: $\log(\mu_i) = \beta_0 + \beta_1 \text{dose}_i + \beta_2 \text{dose}_i^2 + \beta_3 \text{dose}_i^3.$

Alternative models:

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

Likelihood ratio tests

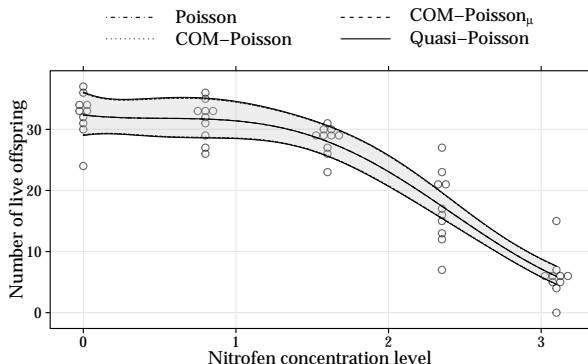
Table: Model fit measures and comparisons between linear predictors.

| Poisson | np | ℓ | AIC | 2(diff ℓ) | diff np | P(> χ^2) | |
|---|----|----------|---------|-----------------|---------|----------------|----------------|
| Linear | 2 | -180.667 | 365.335 | | | | |
| Quadratic | 3 | -147.008 | 300.016 | 67.319 | 1 | 2.31E-16 | |
| Cubic | 4 | -144.090 | 296.180 | 5.835 | 1 | 1.57E-02 | |
| COM-Poisson | np | ℓ | AIC | 2(diff ℓ) | diff np | P(> χ^2) | $\hat{\phi}$ |
| Linear | 3 | -167.954 | 341.908 | | | | -0.893 |
| Quadratic | 4 | -146.964 | 301.929 | 41.980 | 1 | 9.22E-11 | -0.059 |
| Cubic | 5 | -144.064 | 298.129 | 5.800 | 1 | 1.60E-02 | 0.048 |
| COM-Poisson _{μ} | np | ℓ | AIC | 2(diff ℓ) | diff np | P(> χ^2) | $\hat{\phi}$ |
| Linear | 3 | -167.652 | 341.305 | | | | -0.905 |
| Quadratic | 4 | -146.950 | 301.900 | 41.405 | 1 | 1.24E-10 | -0.069 |
| Cubic | 5 | -144.064 | 298.127 | 5.773 | 1 | 1.63E-02 | 0.047 |
| Quasi-Poisson | np | QDev | AIC | F | diff np | P(> F) | $\hat{\sigma}$ |
| Linear | 3 | 123.929 | | | | | 2.262 |
| Quadratic | 4 | 56.610 | | 60.840 | 1 | 5.07E-10 | 1.106 |
| Cubic | 5 | 50.774 | | 5.659 | 1 | 2.16E-02 | 1.031 |

Parameter estimates and fitted values

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 |
| β_0 | 3.477 | 62.817 | 3.649 | 4.850 | 3.477 | 64.308 | 3.477 | 61.860 |
| β_1 | -0.086 | -0.433 | -0.091 | -0.448 | -0.088 | -0.452 | -0.086 | -0.426 |
| β_2 | 0.153 | 0.863 | 0.161 | 0.878 | 0.155 | 0.894 | 0.153 | 0.850 |
| β_3 | -0.097 | -2.398 | -0.102 | -2.229 | -0.098 | -2.464 | -0.097 | -2.361 |



5.4

Data analysis results
Additional results

Computational times for fitting

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

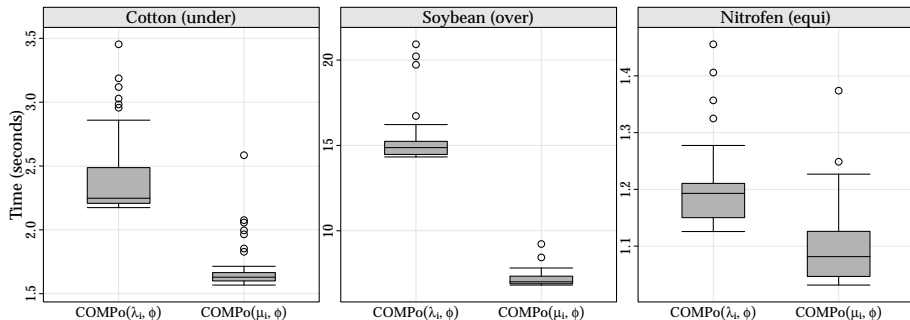


Figure: Computational times to fit the models under original and reparametrized versions based on the fifty repetitions.

6

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* **41**(12), 2616–2626.