

Extensões e Aplicações do Modelo de Regressão Conway-Maxwell-Poisson para Modelagem de Dados de Contagem

Eduardo Elias Ribeiro Junior
Orientação: Prof. Dr. Walmes Marques Zeviani

Trabalho de Conclusão de Curso - Laboratório B
Departamento de Estatística (DEST)
Universidade Federal do Paraná (UFPR)

27 de junho de 2016

Sumário

1. Introdução
2. Objetivos
3. Materiais e Métodos
4. Resultados
5. Considerações finais

1

Introdução

Dados de contagem



São variáveis aleatórias aleatórias que representam o número de ocorrências de um evento em um domínio discreto ou contínuo.

Se Y é uma v.a. de contagem, $y = 0, 1, 2, \dots$

Exemplos:

- ▶ Número de filhos por casal;
- ▶ Número de indivíduos infectados por uma doença;
- ▶ Número de acidentes de trânsito em um mês;
- ▶ Número de *posts* em uma rede social durante um dia;
- ▶ Número de frutos produzidos;
- ▶ ...

Análise de dados de contagem

- ▶ Modelos de regressão Gaussianos com dados transformados
 - ▶ Dificultam a interpretação dos resultados;
 - ▶ Não contemplam a natureza discreta da variável;
 - ▶ Não contemplam a relação média e variância;
 - ▶ Transformação logarítmica é problemática para valores 0.
- ▶ Modelos de regressão Poisson (NELDER; WEDDERBURN, 1972)
 - ▶ Fiel a natureza dos dados;
 - ▶ Contempla a relação média e variância;
 - ▶ Suposição de equidispersão.

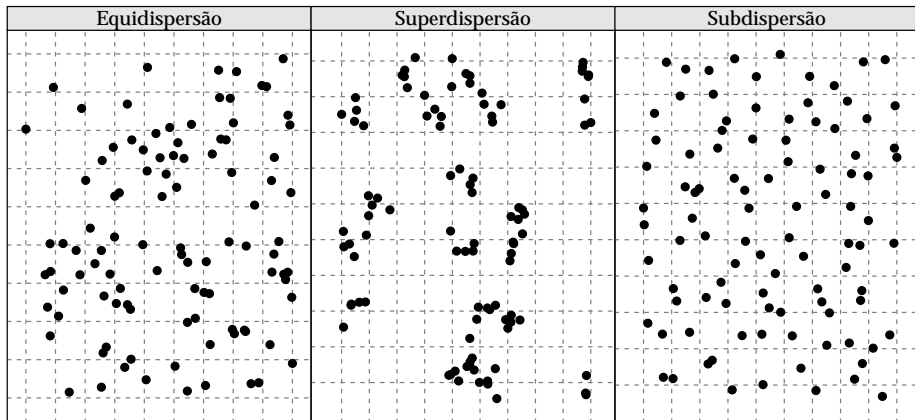


Figura 1: Ilustração de processos pontuais que levam a contagens com diferentes níveis de dispersão.

Distribuições de probabilidades para dados de contagem

Tabela 1: Distribuições de probabilidades para dados de contagem

Distribuição	Contempla a característica de		
	Equidispersão	Superdispersão	Subdispersão
Poisson	✓		
Binomial Negativa	✓	✓	
<i>Inverse Gaussian Poisson</i>	✓	✓	
<i>Compound Poisson</i>	✓	✓	
Poisson Generalizada	✓	✓	✓
<i>Gamma-Count</i>	✓	✓	✓
COM-Poisson	✓	✓	✓
Katz	✓	✓	✓
<i>Poisson Polynomial</i>	✓	✓	✓
<i>Double-Poisson</i>	✓	✓	✓
<i>Lagrangian Poisson</i>	✓	✓	✓

Distribuições de probabilidades para dados de contagem

Tabela 1: Distribuições de probabilidades para dados de contagem

Distribuição	Contempla a característica de		
	Equidispersão	Superdispersão	Subdispersão
Poisson	✓		
Binomial Negativa	✓	✓	
<i>Inverse Gaussian Poisson</i>	✓	✓	
<i>Compound Poisson</i>	✓	✓	
<i>Poisson Generalizada</i>	✓	✓	✓
<i>Gamma-Count</i>	✓	✓	✓
COM-Poisson	✓	✓	✓
<i>Katz</i>	✓	✓	✓
<i>Poisson Polynomial</i>	✓	✓	✓
<i>Double-Poisson</i>	✓	✓	✓
<i>Lagrangian Poisson</i>	✓	✓	✓

2

Objetivos

Objetivos gerais

Colaborar com a literatura estatística brasileira, no que diz respeito a dados de contagem:

- ▶ Apresentando e explorando o modelo de regressão COM-Poisson;
- ▶ Estendendo o modelo para modelagem de excesso de zeros e inclusão de efeitos aleatórios;
- ▶ Discutindo o desempenho do modelo via análise de dados reais;
- ▶ Disponibilizando os recursos computacionais para ajuste dos modelos, em formato de pacote R.

3

Materiais e Métodos

3.1

Materiais e Métodos

Materiais

Conjuntos de dados

Seis conjuntos de dados analisados:

- ▶ Capulhos de algodão sob desfolha artificial
- ▶ Produtividade de algodão sob infestação de Mosca-branca
- ▶ Produtividade de soja sob umidade e adubação potássica
- ▶ Ocorrência de ninfas de Mosca-branca em lavoura de soja
- ▶ Peixes capturados por visitantes de um parque Estadual
- ▶ Número de nematoides em raízes de feijoeiro

Conjuntos de dados

Seis conjuntos de dados analisados:

- ▶ Capulhos de algodão sob desfolha artificial
- ▶ **Produtividade de algodão sob infestação de Mosca-branca**
- ▶ Produtividade de soja sob umidade e adubação potássica
- ▶ **Ocorrência de ninfas de Mosca-branca em lavoura de soja**
- ▶ **Peixes capturados por visitantes de um parque Estadual**
- ▶ **Número de nematoides em raízes de feijoeiro**

Recursos Computacionais

Software R versão 3.3.0. Principais pacotes:

- ▶ MASS (modelo binomial negativo)
- ▶ pscl (modelagem de excesso de zeros)
- ▶ lme4 (modelo Poisson com efeito aleatório Normal)
- ▶ bbmle (ajuste de modelos via máxima verossimilhança)

3.2

Materiais e Métodos

Métodos

Estimação via máxima verossimilhança

- 1 Escreva a função de verossimilhança - $\mathcal{L}(\Theta | \underline{y})$
- 2 Tome seu logaritmo - $\ell(\Theta | \underline{y})$
- 3 As estimativas dos parâmetros são

$$\hat{\Theta} = \arg \max_{\Theta} \ell(\Theta | \underline{y})$$

- ▶ Algoritmo IWLS (*Interactive Weighed Least Squares*) para os modelos Poisson, Binomial Negativo e Quasi-Poisson.
- ▶ Método *BFGS* para os modelos COM-Poisson.

Verossimilhança do modelo COM-Poisson

- ▶ Reparametrizando $\phi = \log(\nu)$
 - ▶ $\phi < 0 \Rightarrow$ Superdispersão
 - ▶ $\phi = 0 \Rightarrow$ Equidispersão
 - ▶ $\phi > 0 \Rightarrow$ Subdispersão

log-verossimilhança

$$\ell(\phi, \beta \mid \underline{y}) = \sum_{i=1}^n y_i \log(\lambda_i) - e^{\phi} \sum_{i=1}^n \log(y_i!) - \sum_{i=1}^n \log(Z(\lambda_i, \phi)) \quad (1)$$

em que $\lambda_i = e^{X_i \beta}$, com X_i o vetor $(x_{i1}, x_{i2}, \dots, x_{ip})$ de covariáveis da i -ésima observação, e $(\beta, \phi) \in \mathbb{R}^{p+1}$.

Verossimilhança do modelo Hurdle COM-Poisson

- ▶ $\underline{\pi} = \frac{\exp(G\gamma)}{1+\exp(G\gamma)}$ a probabilidade de contagem nula.
- ▶ $\underline{\lambda} = \exp(X\beta)$ o parâmetro de locação da distribuição COM-Poisson truncada.

verossimilhança

$$\mathcal{L}(\phi, \beta, \gamma \mid \underline{y}) = \mathbb{1}[\underline{\pi}] \cdot (1 - \mathbb{1}) \left[(1 - \underline{\pi}) \left(\frac{\underline{\lambda}^y}{(y!)^{e\phi} Z(\underline{\lambda}, \phi)} \right) \left(1 - \frac{1}{Z(\underline{\lambda}, \phi)} \right) \right] \quad (2)$$

em que $\mathbb{1}$ é uma função indicadora para $y = 0$

Verossimilhança do modelo misto COM-Poisson

$$Y_{ij} \mid b_i, X_{ij} \sim \text{COM-Poisson}(\mu_{ij}, \phi)$$

$$g(\mu_{ij}) = X_{ij}\beta + Z_i b_i$$

$$b \sim \text{Normal}(0, \Sigma)$$

Verossimilhança

$$\mathcal{L}(\phi, \Sigma, \beta \mid \underline{y}) = \prod_{i=1}^m \int_{\mathbb{R}^q} \left(\prod_{j=1}^{n_i} \frac{\lambda^y}{(y!) e^\phi Z(\underline{\lambda}, \phi)} \right) \cdot (2\pi)^{q/2} |\Sigma| \exp \left(-\frac{1}{2} b^t \Sigma^{-1} b \right) db_i \quad (3)$$

sendo m o número de grupos que compartilham do mesmo efeito aleatório, q o número de efeitos aleatórios (intercepto aleatório, inclinação e intercepto aleatórios, etc.) e n_i o número de observações no i -ésimo grupo.

4

Resultados

4.1

Resultados Pacote R

cmpreg: Ajuste de Modelos de Regressões COM-Poisson

Implementação em R de um *framework* para ajuste dos modelos de regressão COM-Poisson, pacote cmpreg

```
## Pode ser instalado do GitHub
devtools::install_git("https://github.com/JrEduardo/cmpreg.git")
library(cmpreg)

## Regressão (efeitos fixos)
cmp(y ~ predictor, data = data)

## Regressão com componente de barreira
hurdlecmp(y ~ count_pred | zero_pred, data = data)

## Regressão (efeitos aleatórios)
mixedcmp(y ~ count_pred + (1 | ind.ranef), data = data)
```

4.2

Resultados Produtividade de algodão

Experimento

Conduzido na UFGD em casa de vegetação (MARTELLI et al., 2008).

- ▶ Delineamento: inteiramente casualizado com cinco repetições
- ▶ Objetivo: avaliar o impacto da praga Mosca-branca na produção de algodão.
- ▶ Unidade amostral: vaso com duas plantas.
- ▶ Covariável experimental:
 - ▶ Tempo de exposição das plantas à praga, em dias. (dexp)
- ▶ Variáveis resposta:
 - ▶ Número de capulhos produzidos
 - ▶ Número de estruturas reprodutivas
 - ▶ Número de nós

Modelagem

Preditores considerados:

- ▶ Preditor 1: $g(\mu_i) = \beta_0$
- ▶ Preditor 2: $g(\mu_i) = \beta_0 + \beta_1 \text{dexp}_i$
- ▶ Preditor 3: $g(\mu_i) = \beta_0 + \beta_1 \text{dexp}_i + \beta_2 \text{dexp}_i^2$

Modelos concorrentes:

- ▶ Poisson(μ_i)
- ▶ COM-Poisson(λ_i, ϕ)
- ▶ Quasi-Poisson(μ_i, σ^2)

Medidas de ajuste

Tabela 2: Medidas de ajuste para avaliação e comparação

np	Poisson			COM-Poisson			Quasi-Poisson	
	ℓ	AIC	$P(> \chi^2)$	ℓ	AIC	$P(> \chi^2)$	deviance	$P(> F)$
Número de capulhos produzidos								
1	-105,27	212,55	—	-92,05	188,09	—	20,80	—
2	-105,03	214,05	4,83E-01	-91,31	188,62	2,25E-01	20,31	2,23E-01
3	-104,44	214,88	2,78E-01	-89,47	186,95	5,52E-02	19,13	6,16E-02
Número de estruturas reprodutivas								
1	-104,74	211,49	—	-86,41	176,82	—	16,23	—
2	-104,27	212,54	3,32E-01	-84,59	175,18	5,66E-02	15,29	6,19E-02
3	-104,06	214,12	5,16E-01	-83,73	175,47	1,90E-01	14,87	2,07E-01
Número de nós da planta								
1	-143,79	289,59	—	-120,58	245,16	—	12,69	—
2	-143,48	290,95	4,25E-01	-119,03	244,06	7,87E-02	12,05	7,39E-02
3	-142,95	291,89	3,04E-01	-116,27	240,54	1,88E-02	11,00	2,23E-02

Valores preditos

----- Poisson

— COM-Poisson

----- Quasi-Poisson

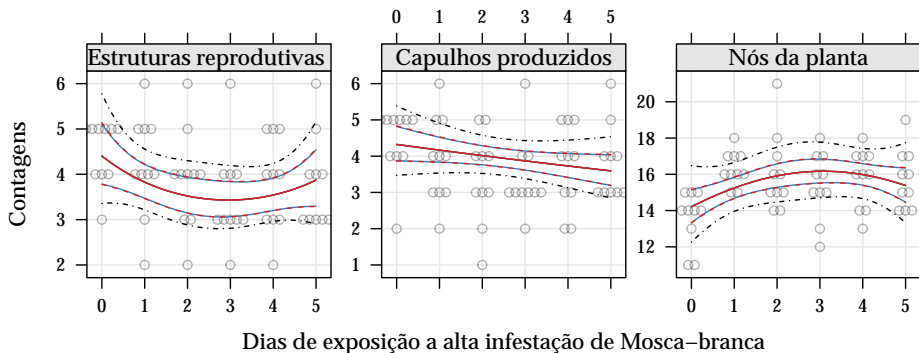


Figura 2: Curva dos valores preditos com intervalo de confiança de (95%) como função dos dias de exposição a alta infestação de Mosca-branca.

4.3

Resultados
**Ocorrência de ninfas de
Mosca-branca**

Experimento

Conduzido na UFGD em casa de vegetação (SUEKANE, 2011).

- ▶ Delineamento: blocos casualizados com quatro blocos.
- ▶ Objetivo: avaliar a propensão de cultivares de soja à praga Mosca-branca.
- ▶ Unidade experimental: dois vasos com duas plantas.
- ▶ Covariáveis experimentais:
 - ▶ Indicadora de bloco, I, II, III e IV, (bloco),
 - ▶ Dias decorridos após a primeira avaliação, 0, 8, 13, 22, 31 e 38 dias. (dias),
 - ▶ Indicadora de cultivar de soja, BRS 239, BRS 243 RR, BRS 245 RR, BRS246 RR, (cult).
- ▶ Variável resposta:
 - ▶ Número de ninfas de Mosca-branca nos folíolos dos terços superior, médio e inferior.

Modelagem

Preditores considerados:

- ▶ $\eta_1 = g(\mu_{ijk}) = \beta_0 + \tau_i + \gamma_j + \delta_k$
- ▶ $\eta_2 = g(\mu_{ijk}) = \beta_0 + \tau_i + \gamma_j + \delta_k + \alpha_{jk}$

τ_i é o efeito do i-ésimo bloco, $i = 1, 2, 3, 4$

γ_j o efeito da j-ésima cultivar, $j = 1, 2, 3, 4$

δ_k o efeito do k-ésimo nível de dias, $k = 1, 2, \dots, 6$ e

α_{jk} o efeito da interação entre a j-ésima cultivar e o k-ésimo nível de dias

Modelos concorrentes:

- ▶ Poisson(μ_{ijk})
- ▶ COM-Poisson(λ_{ijk}, ϕ)
- ▶ Binomial Negativo(μ_{ijk}, θ)
- ▶ Quasi-Poisson(μ_{ijk}, σ^2)

Medidas de ajuste

Tabela 3: Medidas de ajuste para avaliação e comparação

Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$	
Preditor 1	12	-922,98	1869,96				
Preditor 2	27	-879,23	1812,46	87,50	15	2,90E-12	
COM-Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$	$\hat{\phi}$
Preditor 1	13	-410,44	846,89				-3,08
Preditor 2	28	-407,15	870,30	6,59	15	9,68E-01	-2,95
Binomial Neg.	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$	$\hat{\theta}$
Preditor 1	13	-406,16	838,31				3,44
Preditor 2	28	-400,55	857,10	11,21	15	7,38E-01	3,99
Quase-Poisson	np	deviance	AIC	F	diff np	$P(>F)$	$\hat{\sigma}^2$
Preditor 1	12	1371,32					17,03
Preditor 2	27	1283,82		0,31	15	9,93E-01	19,03

Valores preditos

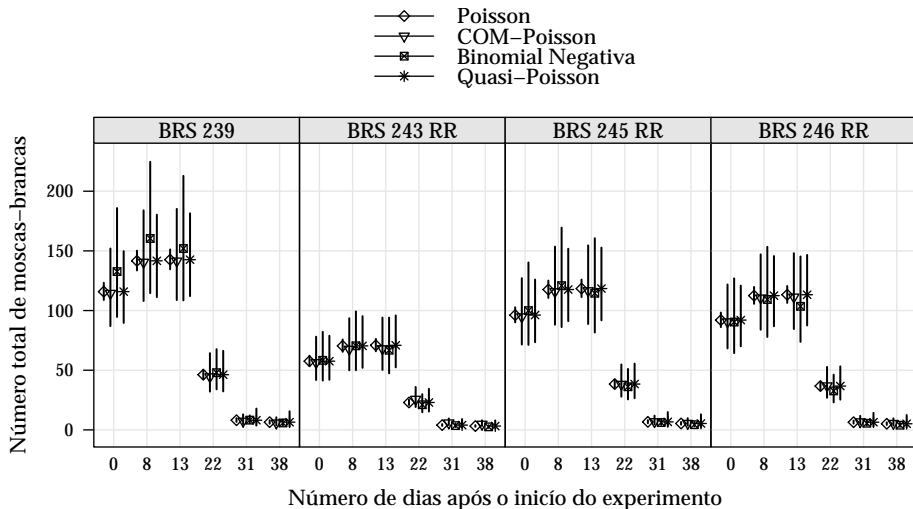


Figura 3: Valores preditos com intervalos de confiança (95%).

4.4

Resultados **Peixes capturados**

Estudo

Observacional conduzido por biólogos em um Parque Estadual (UCLA, 2015).

- ▶ Delineamento: amostragem aleatória.
- ▶ Objetivo: modelar o número de peixes capturados pela atividade de pesca esportiva.
- ▶ Unidade experimental: grupos de pescadores visitantes do parque.
- ▶ Covariáveis mensuradas:
 - ▶ Número de pessoas, (n_p),
 - ▶ Número de crianças. (n_c),
 - ▶ Indicador de campista no grupo, (ca).
- ▶ Variável resposta:
 - ▶ Número de peixes capturados pelo grupo.

Modelagem

Preditores considerados:

- ▶ Preditor 1:
$$g(\mu_i) = \beta_0 + \beta_1 ca_i + \beta_2 np_i$$
$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 ca_i + \gamma_2 np_i + \gamma_3 nc_i$$
- ▶ Preditor 2:
$$g(\mu_i) = \beta_0 + \beta_1 ca_i + \beta_2 np_i + \beta_3 nc_i + \beta_4 (np_i \cdot nc_i)$$
$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 ca_i + \gamma_2 np_i + \gamma_3 nc_i + \gamma_4 (np_i \cdot nc_i)$$

Modelos concorrentes:

- ▶ Hurdle Poisson(π_i, μ_i)
- ▶ Hurdle COM-Poisson(π_i, λ_i, ϕ)
- ▶ Hurdle Binomial Negativo(π_i, μ_i, θ)

Medidas de ajuste

Tabela 4: Medidas de ajuste para avaliação e comparação

Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$	
Preditor 1	7	-857,48	1728,96				
Preditor 2	10	-744,58	1509,17	225,79	3	1,12E-48	
Binomial Neg.	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$	$\hat{\theta}$
Preditor 1	8	-399,79	815,58				0,20
Preditor 2	11	-393,72	809,44	12,14	3	6,91E-03	0,37
COM-Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$	$\hat{\phi}$
Preditor 1	8	-409,85	835,71				-8,77
Preditor 2	11	-402,30	826,59	15,12	3	1,72E-03	-3,77

Valores preditos

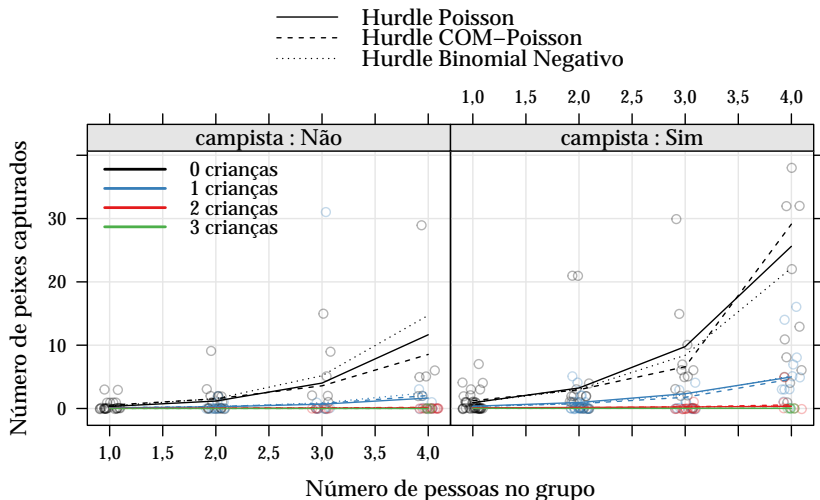


Figura 4: Valores preditos do número de peixes capturados.

4.5

Resultados

Número de nematoides

Experimento

Conduzido no IAPAR em casa de vegetação.

- ▶ Delineamento: inteiramente casualizado com cinco repetições.
- ▶ Objetivo: avaliar a resistência à nematoides de linhagens de feijoeiro.
- ▶ Unidade amostral: alíquota de 1ml da solução de raízes lavadas, trituradas, peneiradas e diluídas em água provida por um vaso com duas plantas.
- ▶ Covariáveis:
 - ▶ Indicador de linhagem de feijoeiro, A, B, C, ..., S (cult)
 - ▶ Concentração de raiz na solução. (sol)
- ▶ Variáveis resposta:
 - ▶ Número de nematoides.

Modelagem

Preditores considerados:

- ▶ Preditor 1: $g(\mu_{ij}) = \beta_0 + b_j$
 - ▶ Preditor 2: $g(\mu_{ij}) = \beta_0 + \beta_1 \log(\text{sol})_i + b_j$
- $$b_j \sim \text{Normal}(0, \sigma^2)$$

Modelos concorrentes:

- ▶ Poisson(μ_{ij})
- ▶ COM-Poisson(λ_{ij}, ϕ)

Medidas de ajuste

Tabela 5: Medidas de ajuste para avaliação e comparação

Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$		
Preditor 1	2	-237,20	478,40					
Preditor 2	3	-234,66	475,32	5,07	1	2,43E-02		
COM-Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	$P(> \chi^2)$	$\hat{\phi}$	$P(> \chi^2)$
Preditor 1	3	-236,85	479,71				0,15	4,06E-01
Preditor 2	4	-233,86	475,72	5,99	1	1,44E-02	0,23	2,05E-01

Valores preditos

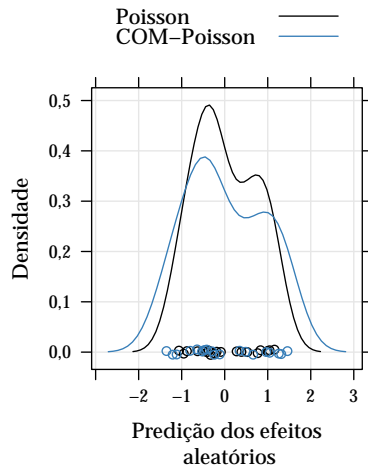
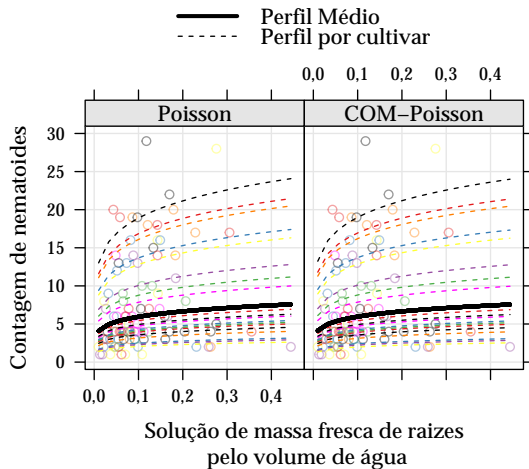


Figura 5: Valores preditos nos modelos de efeitos mistos.

4.6

Resultados Discussões

- ▶ Similaridade entre inferências via modelo Quasi-Poisson e COM-Poisson;
- ▶ Desempenho do modelo Binomial Negativo;
- ▶ Interpretação dos parâmetros nos modelos baseados na COM-Poisson
- ▶ Problemas numéricos para determinação da matriz hessiana no modelo Hurdle COM-Poisson;
- ▶ Procedimentos computacionalmente intensivos na avaliação da verossimilhança no caso COM-Poisson de efeitos aleatórios;
- ▶ Não ortogonalidade observada (empírica) entre os parâmetros de locação e de precisão no modelo COM-Poisson;
- ▶ Comportamento simétrico dos perfis de log-verossimilhança para o parâmetro ϕ da COM-Poisson.

5

Considerações finais

Conclusões

Aplicação do modelo COM-Poisson:

- ▶ Resultados similares aos providos pela abordagem semi-paramétrica via quasi-verossimilhança;
- ▶ A não ortogonalidade entre os parâmetros de locação e precisão no modelo COM-Poisson se mostra como característica da distribuição;
- ▶ A simetria nos perfis de verossimilhança do parâmetro de precisão também.
- ▶ A avaliação da constante de normalização é problemática no modelo.

Conclusões

Análise de dados de contagem:

- ▶ Modelo Poisson inadequado na maioria das aplicações, mostrando que a suposição de equidispersão é de fato restritiva;
- ▶ Modelos alternativos ao Poisson devem ser empregados na análise de dados de contagem;
- ▶ Sugere-se o modelo COM-Poisson como alternativa totalmente paramétrica e bastante flexível.

Trabalhos futuros

- ▶ Estudar reparametrizações do modelo COM-Poisson;
- ▶ Avaliar aproximações da constante de normalização;
- ▶ Realizar estudos de simulação para avaliar a robustez do modelo;
- ▶ Implementar o modelo COM-Poisson inflacionado de zeros;
- ▶ Implementar o modelo COM-Poisson com efeitos aleatórios dependentes.

Publicização



<https://github.com/JrEduardo/cmpreg>

<https://github.com/JrEduardo/tccDocument>



Referências

- MARTELLI, T. et al. **Influência do ataque de mosca-branca Bemisia tabaci Biotipo B, nos índices de produtividade do algodoeiro**Uberlândia- MGXXII Congresso Brasileiro de Entomologia, 2008.
- 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.
- SUEKANE, R. **DISTRIBUIÇÃO ESPACIAL E DANO DE MOSCA-BRANCA Bemisia tabaci (GENNADIUS, 1889) BIÓTIPO B NA SOJA**. PhD thesis—[s.l.] Universidade Federal da Grande Dourados, 2011.
- UCLA, S. C. G. **Data Analysis Examples**, 2015. Disponível em:
<<http://www.ats.ucla.edu/stat/dae/>>