

# Modelos de regressão para dados de contagem: além do modelo Poisson.

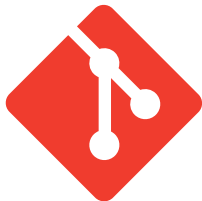
Prof. PhD. Wagner Hugo Bonat  
Prof. Dr. Walmes M. Zeviani  
Eduardo E. Ribeiro Jr

Laboratório de Estatística e Geoinformação  
Departamento de Estatística  
Universidade Federal do Paraná

22 de março de 2017

{wbonat,walmes,eduardo.jr}@ufpr.br

# Disponibilização



<https://github.com/leg-ufpr/rmcd>  
<http://cursos.leg.ufpr.br/rmcd/>

# Conteúdo

1. Introdução
2. Distribuições para contagens: propriedades e modelos de regressão
  - 2.1 Distribuição Poisson
  - 2.2 Distribuição Gamma-Count
  - 2.3 Distribuição Poisson-Tweedie
  - 2.4 Distribuição COM-Poisson
  - 2.5 Comparando distribuições para contagens
3. Método de máxima verossimilhança
4. Modelos especificados por suposições de momentos
  - 4.1 Especificação
  - 4.2 Estimação e Inferência
5. Aplicações

1

# Introdução

# Dados de contagens

Alguns exemplos de problemas envolvendo contagens:

- ▶ Número de acidentes em uma rodovia por semana;
- ▶ Número de automóveis vendidos por dia;
- ▶ Número de gols marcados por times de futebol por partida;
- ▶ Número de falhas por metro de fio de cobre produzido;
- ▶ Número de colônias de bactérias por  $0,01mm^2$  de uma dada cultura ...

# Modelos probabilísticos para dados de contagens

- ▶ Modelos probabilísticos para variáveis aleatórias discretas, com suporte no conjunto dos números inteiros não-negativos, são potenciais candidatos para a análise de dados de contagens.
- ▶ Algumas alternativas: Distribuição binomial, Poisson e generalizações; distribuições geradas por misturas, como a beta-binomial, binomial negativa; distribuições fundamentadas na modelagem do tempo entre eventos, na razão de probabilidades sucessivas ...

# Regressão para dados de contagens

- ▶ Modelos de regressão são utilizados para modelar a distribuição de uma variável aleatória  $Y$  condicional aos valores de um conjunto de variáveis explicativas  $x_1, x_2, \dots, x_p$ .
- ▶ Métodos para inferência e modelos de regressão para dados de contagem estão aquém, em quantidade e diversidade, em relação ao verificado para dados contínuos.
- ▶ A aplicação de modelos de regressão com erros normais na análise de contagens, embora frequente, em geral é desaconselhável.

# Regressão com erros normais na análise de dados de contagens

- ▶ O modelo linear com erros normais não considera a natureza discreta dos dados;
- ▶ Associa probabilidade nula a qualquer possível contagem;
- ▶ Admite probabilidades não nulas a valores negativos da variável;



# Regressão com erros normais na análise de dados de contagens

- ▶ O uso de transformações dificulta a interpretação dos resultados;
- ▶ O uso da transformação logarítmica apresenta problemas para contagens iguais a zero;
- ▶ Não se contempla a relação não constante entre média e variância, característica de dados de contagens.

# Distribuição de Poisson

- ▶ A distribuição de Poisson é a principal referência para a análise de dados de contagens.
- ▶ Função de probabilidades:

$$P(Y = k) = \frac{e^{-\mu} \mu^k}{k!}, \quad k = 0, 1, 2, \dots; \mu > 0.$$

- ▶ Se os eventos sob contagem ocorrem independentemente e sujeitos a uma taxa constante  $\mu > 0$ , sob o modelo Poisson, para um intervalo de exposição de tamanho  $t$  tem-se:

$$P(Y_t = k) = \frac{e^{-\mu t} (\mu t)^k}{k!}, \quad k = 0, 1, 2, \dots$$

# Propriedades da distribuição de Poisson

Dentre as principais propriedades da distribuição de Poisson, tem-se:

- ▶ Média:  $E(Y) = \mu$ ;
- ▶ Variância:  $\text{var}(Y) = \mu$  (equidispersão);
- ▶ Razão de probabilidades sucessivas:  $\frac{P(Y=k)}{P(Y=k-1)} = \frac{\lambda}{k}$ , gerando a relação de recorrência:

$$P(Y = k)k = P(Y = k - 1)\lambda;$$

- ▶ Se  $Y_1, Y_2, \dots, Y_n$  são va's independentes com  $Y_i \sim \text{Poisson}(\mu_i)$ , e  $\sum \mu_i < \infty$ , então  $\sum Y_i \sim \text{Poisson}(\sum \mu_i)$ .

# Motivações para a distribuição de Poisson

- ▶ Se o tempo decorrido entre sucessivos eventos é uma variável aleatória com distribuição exponencial de média  $\lambda = 1/\mu$ , então o número de eventos ocorridos em um intervalo  $t$  de tempo tem distribuição de Poisson com média  $\mu t$ .
  - ▶ A dualidade entre as distribuições Poisson e exponencial implica que a taxa de ocorrência do evento, definida por:

$$\mu(t) = \lim_{\Delta t \rightarrow 0} \frac{P \{ \text{evento ocorrer em } (t, t + \Delta t) \}}{\Delta t},$$

dado que o evento não ocorreu até o tempo  $t$ , **é constante** para todo  $t > 0$ .

# Diferentes comportamentos para $\mu(t)$

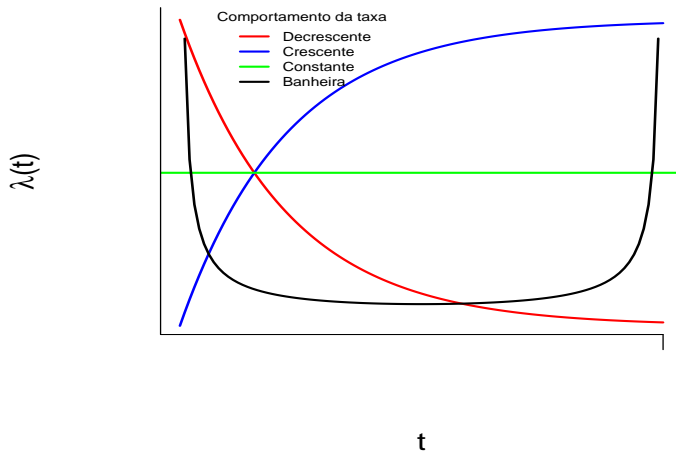


Figura : Diferentes comportamentos para  $\mu(t)$

# Processo de Poisson

O Processo de Poisson configura um processo de contagem em que  $Y(t), t \geq 0$ , representa o número de eventos que ocorrem até  $t$ , satisfazendo:

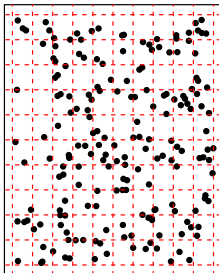
- 1  $Y(t)$  é inteiro e não negativo;
- 2 Para  $s < t$ ,  $Y(s) \leq Y(t)$ ;
- 3  $Y(t) - Y(s)$  é o número de eventos que ocorrem no intervalo  $(s, t]$ ;
- 4 O processo é estacionário:

$$Y(t_2 + s) - Y(t_1 + s) \stackrel{i.d.}{\sim} Y(t_2) - Y(t_1), \forall s > 0$$

- 5 O processo tem incrementos independentes, ou seja, os números de eventos verificados em intervalos disjuntos são independentes.

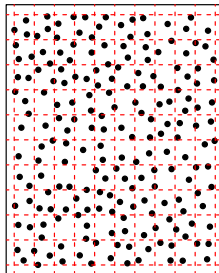
# Diferentes padrões em processos de contagens

**Padrão aleatório**



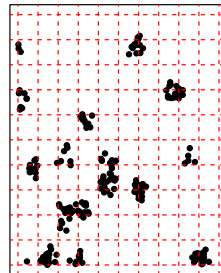
Equidispersão  
 $\text{Var}(Y)=E(Y)$

**Padrão uniforme**



Subdispersão  
 $\text{Var}(Y)<E(Y)$

**Padrão agregado**



Superdispersão  
 $\text{Var}(Y)>E(Y)$

**Figura :** Ilustração de diferentes tipos de processos de contagens.

# O desafio de dados de contagem

- ▶ Poisson implica equidispersão, ou seja,  $E(Y) = \text{var}(Y) = \mu$ .
- ▶ Na prática podemos ter
  - ▶ Subdispersão  $E(Y) > \text{var}(Y)$ ;
  - ▶ Superdispersão  $E(Y) < \text{var}(Y)$ .
- ▶ Desvios da equidispersão implicam:
  - ▶ Mais ou menos zeros e
  - ▶ Caudas mais leves ou mais pesadas que o modelo Poisson.



# Causas da não equidispersão

- ▶ Desvios do processo Poisson;
- ▶ Heterogeneidade entre unidades amostrais.
  
- ▶ O que acontece caso o modelo Poisson seja usada para dados não equidispersos?
  - ① Superdispersão: erros padrões associados aos coeficientes de regressão serão subestimados.
  - ② Subdispersão: erros padrões associados aos coeficientes de regressão serão superestimados.
  
- ▶ Ambos os casos o modelo Poisson resulta em erros padrões não-confiáveis o que implica em inferências incorretas.

# Como lidar com a não equidispersão

- ▶ Mudar a distribuição dos tempos entre eventos: Ex Gamma-Count.
- ▶ Incluir efeitos aleatórios ao nível das observações. Ex Poisson-Tweedie.
- ▶ Modificar a distribuição de Poisson incluindo um parâmetro extra de dispersão. Ex COM-Poisson.

2

# Distribuições para contagens: propriedades e modelos de regressão

## 2.1

Distribuições para contagens: propriedades e  
modelos de regressão  
**Distribuição Poisson**

# Distribuição Poisson

- Função de probabilidade

$$\begin{aligned}f(y; \mu) &= \frac{\mu^y}{y!} \exp\{-\mu\} \\&= \frac{1}{y!} \exp\{\phi y - \exp\{\phi\}\}, \quad y \in \mathbb{N}_0,\end{aligned}\tag{1}$$

onde  $\phi = \log\{\mu\} \in \mathbb{R}$  e  $\kappa(\phi) = \exp\{\phi\}$  denota a função cumulante.

- $E(Y) = \kappa'(\phi) = \exp\{\phi\} = \mu$ .
- $\text{var}(Y) = \kappa''(\phi) = \exp\{\phi\} = \mu$ .
- Em R temos `dpois()`.

# Regressão Poisson

- ▶ Considere  $(y_i, x_i), i = 1, \dots, n$ , onde  $y_i$ 's são iid realizações de  $Y_i$  de acordo com a distribuição Poisson.
- ▶ Modelo de regressão Poisson

$$Y_i \sim P(\mu_i), \quad \text{sendo} \quad \mu_i = g^{-1}(x_i^\top \beta),$$

onde  $x_i$  and  $\beta$  são vetores  $(p \times 1)$  de covariáveis conhecidas e parâmetros de regressão.

- ▶ Em R temos `glm(..., family = poisson)`.
- ▶  $g$  função de ligação (log link).

## 2.2

Distribuições para contagens: propriedades e  
modelos de regressão  
**Distribuição Gamma-Count**

© 1995 American Statistical Association

Journal of Business &amp; Economic Statistics, October 1995, Vol. 13, No. 4

# Duration Dependence and Dispersion in Count-Data Models

**Rainer WINKELMANN**

Department of Economics, University of Canterbury, Christchurch, New Zealand

This article explores the relation between nonexponential waiting times between events and the distribution of the number of events in a fixed time interval. It is shown that within this framework the frequently observed phenomenon of overdispersion—that is, a variance that exceeds the mean—is caused by a decreasing hazard function of the waiting times, whereas an increasing hazard function leads to underdispersion. Using the assumption of iid gamma-distributed waiting times, a new count-data model is derived. Its use is illustrated in two applications: the number of births and the number of doctor consultations.

**KEY WORDS:** Gamma distribution; Negative binomial distribution; Overdispersion; Poisson process; Renewal theory.

**WINKELMANN, R.**

Duration Dependence and Dispersion in Count-Data Models. *Journal of Business & Economic Statistics*, v.13, n.4, p.467–474, 1995.



# Duração dependência

- ▶ Considere um processo estocástico definido pela sequência da  $y$ -ésima  $\tau_k$ , intervalo de tempo entre eventos.
- ▶ Se  $\{\tau_1, \tau_2, \dots\}$  são independentes e identicamente distribuídos, todos com densidade  $f(\tau)$ , esse processo é chamado de *renewal process*.
- ▶ Defina a variável de contagem  $Y_T$  como o número de eventos no intervalo  $[0, T)$ .
- ▶ Defina  $\vartheta_y = \sum_{k=1}^y \tau_k$  o tempo até o  $y$ -ésimo evento.
- ▶ A distribuição de  $\vartheta_y$  determina a distribuição de  $Y_T$ , mas é baseada em convolução.
- ▶ São distribuições fechadas para convolução: normal, Poisson, binomial e gama.
- ▶ Destas, apenas a gama é contínua e positiva.

# Relação entre número de eventos e intervalo entre eventos

- ▶ Intervalos entre tempo  $\tau_k \sim G(\alpha, \gamma)$ , (omitindo  $k$ ) temos

$$f(\tau, \alpha, \gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)} \cdot \tau^{\alpha-1} \cdot \exp\{-\gamma\tau\},$$

$$E(\tau) = \frac{\alpha}{\gamma}, \quad \text{var}(\tau) = \frac{\alpha}{\gamma^2}.$$

- ▶ Tempo até o  $y$ -ésimo evento

$$\vartheta_y = \tau_1 + \cdots + \tau_y \sim G(y\alpha, \gamma),$$

$$f_y(\vartheta, \alpha, \gamma) = \frac{\gamma^{y\alpha}}{\Gamma(y\alpha)} \cdot \vartheta^{y\alpha-1} \cdot \exp\{-\gamma\vartheta\},$$

$$E(\vartheta) = \frac{y\alpha}{\gamma}, \quad \text{var}(\vartheta) = \frac{y\alpha}{\gamma^2}.$$

# Relação entre número de eventos e intervalo entre eventos

- ▶ A distribuição acumulada do tempo até  $\vartheta_y$  é

$$F_y(T) = \Pr(\vartheta_y \leq T) = \int_0^T \frac{\gamma^{y\alpha}}{\Gamma(y\alpha)} \cdot t^{y\alpha-1} \cdot \exp\{-\gamma t\} dt.$$

- ▶ Seja  $[0, T)$  um intervalo e  $Y_T$  a va número de eventos neste intervalo.
- ▶ Segue que  $Y_T < y$  se e somente se  $\vartheta_y \geq T$ . Assim

$$\Pr(Y_T < y) = \Pr(\vartheta_y \geq T) = 1 - F_y(T);$$

- ▶ Já que  $\Pr(Y_T = y) = \Pr(Y_T < y + 1) - \Pr(Y_T < y)$ , então

$$\Pr(Y_T = y) = F_y(T) - F_{y+1}(T).$$

# Relação entre número de eventos e intervalo entre eventos

- ▶ Portanto, distribuição de  $Y_T$  é resultado da diferença de acumuladas da distribuição Gama,

$$F_y(T) = G(y\alpha, \gamma T) = \int_0^T \frac{\gamma^{y\alpha}}{\Gamma(y\alpha)} t^{y\alpha-1} \cdot \exp\{-\gamma t\} dt. \quad (2)$$

- ▶ Assim

$$\begin{aligned} \Pr(Y_T = y) &= G(y\alpha, \gamma T) - G((y+1)\alpha, \gamma T) \\ &= \left[ \int_0^T \frac{\gamma^{y\alpha}}{\Gamma(y\alpha)} t^{y\alpha-1} \cdot \exp\{-\gamma t\} dt \right] \\ &\quad - \left[ \int_0^T \frac{\gamma^{(y+1)\alpha}}{\Gamma((y+1)\alpha)} t^{(y+1)\alpha-1} \cdot \exp\{-\gamma t\} dt \right]. \end{aligned}$$

# Função de probabilidade

- Em R temos

```
dgc <- function(y, gamma, alpha, log = FALSE) {  
  p <- pgamma(q = 1,  
              shape = y * alpha,  
              rate = alpha * gamma) -  
    pgamma(q = 1,  
          shape = (y + 1) * alpha,  
          rate = alpha * gamma)  
  if(log == TRUE) {p <- log(p)}  
  return(p)  
}
```

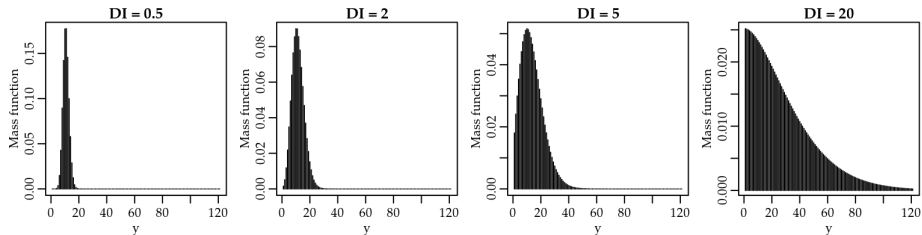


Figura : Função de probabilidade de acordo com valores do índice de dispersão - Gamma-Count.

- Índice de dispersão -  $DI = E(Y)/\text{var}(Y)$

# Parametrização para modelo de regressão

- ▶ A média da variável aleatória  $Y_T$  é resultado de

$$\begin{aligned} E(Y_T) &= \sum_{i=0}^{\infty} i \cdot \Pr(i) \\ &= \sum_{i=1}^{\infty} i \cdot \Pr(i) \\ &= \sum_{i=1}^{\infty} G(i\alpha, \gamma T). \end{aligned}$$

- ▶ Para um  $T$  cada vez maior, tem-se que

$$Y(T) \sim N\left(\frac{\gamma}{\alpha}, \frac{\gamma}{\alpha^2}\right).$$

# Parametrização para modelo de regressão

- Considere que

$$\frac{\gamma}{\alpha} = \exp\{\mathbf{x}^\top \beta\} \Rightarrow \gamma = \alpha \exp\{\mathbf{x}^\top \beta\}.$$

Essa parametrização produz um modelo de regressão para a média do tempo entre eventos definida por

$$E(\tau|\mathbf{x}) = \frac{\alpha}{\gamma} = \exp\{-\mathbf{x}^\top \beta\}.$$

- O modelo de regressão é para o tempo entre eventos ( $\tau$ ) e não diretamente para contagem porque, a menos que  $\alpha = 1$ , não é certo que  $E(Y_i|x_i) = [E(\tau_i|x_i)]^{-1}$ .
- $\alpha$  é um parâmetro de dispersão, assim  $\alpha > 1$  indica subdispersão,  $\alpha = 1$  equidispersão e  $\alpha < 1$  superdispersão.
- Em R temos `MrDCR::gcnt(formula, data)`.



## 2.3

Distribuições para contagens: propriedades e  
modelos de regressão  
**Distribuição Poisson-Tweedie**

# Distribuição Tweedie

- ▶ Distribuição Tweedie (Jørgensen, 1997)

$$f(z; \mu, \phi, p) = a(z, \phi, p) \exp\{(z\psi - k(\psi))/\phi\},$$

onde  $\mu = E(Z) = k'(\psi)$  é a média.

- ▶  $\phi > 0$  e  $\psi$  são os parâmetros de dispersão e canônico.
- ▶  $k(\psi)$  é a função cumulante e  $a(z, \phi, p)$  é a constante normalizadora.
- ▶  $\text{var}(Z) = \phi\mu^p$  onde  $p \in (-\infty, 0] \cup [1, \infty)$  é um index determinando a distribuição.
- ▶ Casos especiais: Normal ( $p = 0$ ), Poisson ( $p = 1$ ), não-central gamma ( $p = 1.5$ ), gamma ( $p = 2$ ), normal inversa ( $p = 3$ ) and distribuições estáveis ( $p > 2$ ).
- ▶ Notação  $Z \sim Tw_p(\mu, \phi)$ .

# Distribuição Poisson-Tweedie

- Especificação hierárquica:

$$\begin{aligned}Y|Z &\sim P(Z) \\ Z &\sim Tw_p(\mu, \phi).\end{aligned}$$

- Função de probabilidade ( $p > 1$ )

$$f(y; \mu, \phi, p) = \int_0^\infty \frac{z^y \exp^{-z}}{y!} a(z, \phi, p) \exp\{(z\psi - k(\psi))/\phi\} dz.$$

- Forma fechada está disponível apenas no caso especial - binomial negativa ( $p = 2$ ).
- Pode ser aproximada por integração Monte Carlo e/ou integração Gauss-Laguerre.

# Função de probabilidade

► Em R temos

```
require(tweedie)
# Integrand Poisson X Tweedie distributions
integrand <- function(x, y, mu, phi, power) {
  int = dpois(y, lambda = x)*dtweedie(x, mu = mu,
                                       phi = phi, power = power)

  return(int)
}

# Computing the pmf using Monte Carlo
dptw <- function(y, mu, phi, power, control_sample) {
  pts <- control_sample$pts
  norma <- control_sample$norma
  integral <- mean(integrand(pts, y = y, mu = mu, phi = phi,
                             power = power)/norma)

  return(integral)
}
```

# Função de probabilidade

## ► Exemplo

```
set.seed(123)
pts <- rtweedie(n = 1000, mu = 10, phi = 1, power = 2)
norma <- dtweedie(pts, mu = 10, phi = 1, power = 2)
control_sample <- list("pts" = pts, "norma" = norma)
dptw(y = c(0, 5, 10, 15), mu = 10, phi = 1, power = 2,
      control_sample = control_sample)

## [1] 0.09374152 0.05902132 0.03539478 0.02171625

dnbinom(x = c(0, 5, 10, 15), mu = 10, size = 1)

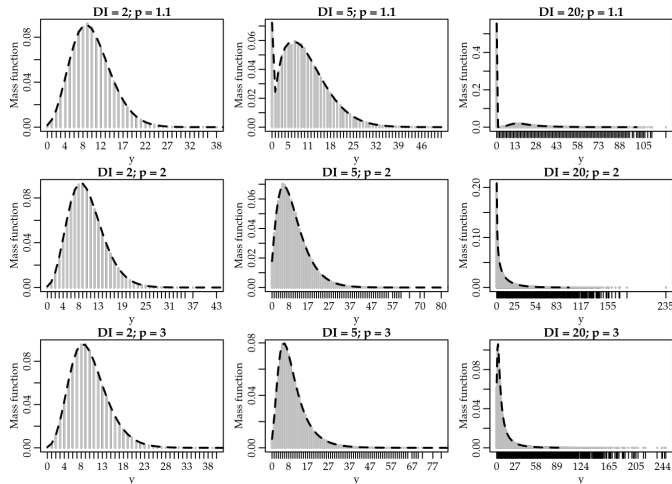
## [1] 0.09090909 0.05644739 0.03504939 0.02176291
```

# Momentos e casos especiais

- ▶ Média e variância marginal são facilmente obtidos

$$\begin{aligned}E(Y) &= \mu \\ \text{var}(Y) &= \mu + \phi\mu^p.\end{aligned}$$

- ▶ Casos especiais: Hermite ( $p = 0$ ), Neyman-Type A ( $p = 1$ ), Pólya-Aeppli ( $p = 1.5$ ), binomial negativa ( $p = 2$ ) e Poisson inversa-Normal ( $p = 3$ ).
- ▶ Cuidado! - Hermite é um caso limite.
- ▶  $p$  é um índice que distingue entre importantes distribuições.
- ▶ Espaço paramétrico de  $p$  é não trivial  $p \in 0 \cup [1, \infty)$ .
- ▶ Estimação de  $p$  funciona como uma seleção automática de distribuições.
- ▶ Notação  $Y \sim PTw_p(\mu, \phi)$ .



**Figura :** Distribuição de probabilidade empírica (cinza) e função de probabilidade aproximada (preta) por valores do índice de dispersão e valores do parâmetro de potência: Poisson-Tweedie.

# Regressão Poisson-Tweedie

- ▶ Considere  $(y_i, x_i)$ ,  $i = 1, \dots, n$ , onde  $y_i$ 's são iid realizações de  $Y_i$  de acordo com a distribuição Poisson-Tweedie.
- ▶ Modelo de regressão Poisson-Tweedie

$$Y_i \sim PTw_p(\mu_i, \phi), \quad \text{sendo} \quad \mu_i = g^{-1}(x_i^\top \beta),$$

onde  $x_i$  and  $\beta$  são vetores  $(p \times 1)$  de covariáveis conhecidas e parâmetros de regressão.

- ▶ Em R temos `dptw()`.
- ▶  $g$  função de ligação (log link).



## 2.4

Distribuições para contagens: propriedades e  
modelos de regressão  
**Distribuição COM-Poisson**

# Distribuição COM-Poisson

- ▶ Nome COM-Poisson, advém de seus autores **CO**nway e **MA**xwell (também é chamada de distribuição Conway-Maxwell-Poisson).
- ▶ Proposta em um contexto de filas, essa distribuição generaliza a Poisson com a adição de um parâmetro.
- ▶ Modifica a relação entre probabilidades consecutivas.

## ▶ Distribuição Poisson

$$\frac{Pr(Y = y - 1)}{Pr(Y = y)} = \frac{y}{\lambda}$$

## ▶ Distribuição COM-Poisson

$$\frac{Pr(Y = y - 1)}{Pr(Y = y)} = \frac{y^{\nu}}{\lambda}$$

# Distribuição COM-Poisson

## Distribuição de probabilidades

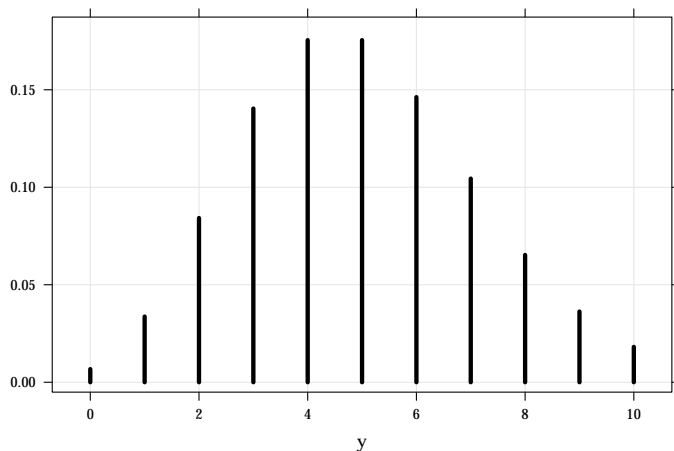
$$f(y; \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \quad \text{em que } Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}; \text{ e } \lambda > 0, \nu \geq 0$$

## Casos particulares

- ▶ Distribuição Poisson, quando  $\nu = 1$
- ▶ Distribuição Bernoulli, quando  $\nu \rightarrow \infty$
- ▶ Distribuição Geométrica, quando  $\nu = 0, \lambda < 1$

# Distribuição COM-Poisson

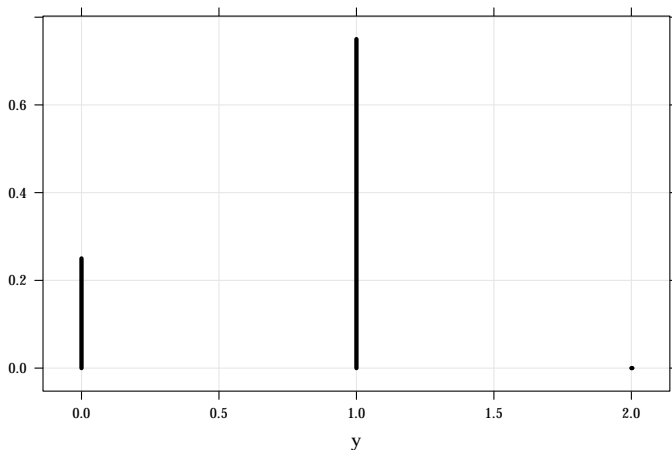
- Poisson  $\nu = 1$
- Bernoulli  $\nu \rightarrow \infty$
- Geométrica  $\nu = 0, \lambda < 1$

COM-Poisson ( $\lambda = 5, \nu = 1$ )

# Distribuição COM-Poisson

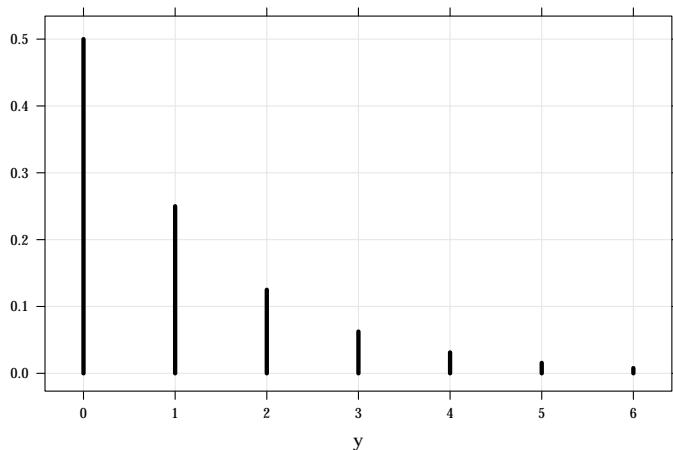
- ▶ Poisson  $\nu = 1$
- ▶ Bernoulli  $\nu \rightarrow \infty$
- ▶ Geométrica  $\nu = 0, \lambda < 1$

COM-Poisson ( $\lambda = 3, \nu = 20$ )



# Distribuição COM-Poisson

- ▶ Poisson  $\nu = 1$
- ▶ Bernoulli  $\nu \rightarrow \infty$
- ▶ Geométrica  $\nu = 0, \lambda < 1$

COM-Poisson ( $\lambda = 0.5, \nu = 0$ )

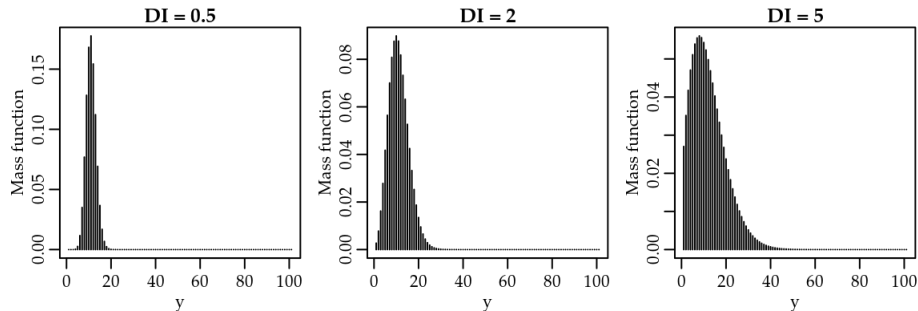
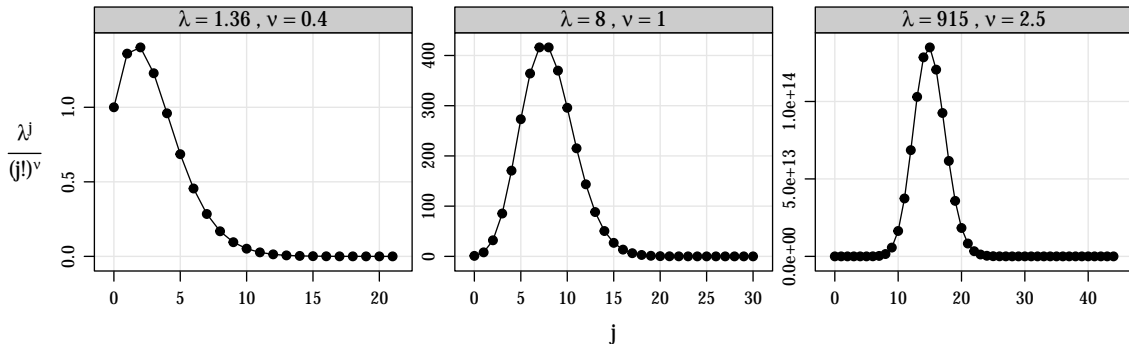


Figura : Distribuição de probabilidade por valores do índice de dispersão: COM-Poisson.

# Assintoticidade da função Z

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^{\nu}}$$





# Momentos da distribuição

Não tem expressão analítica, calculamos utilizando a definição de média e variância;

$$\blacktriangleright E(Y) = \sum_{y=0}^{\infty} y \cdot p(y)$$

$$\blacktriangleright \text{var}(Y) = \sum_{y=0}^{\infty} y^2 \cdot p(y) - E^2(Y)$$

► Regressão COM-Poisson:  $\lambda_i = \exp(x_i^\top \beta)$ , em que  $x_i$  é o vetor de covariáveis do i-ésimo indivíduo e  $\beta$  o vetor de parâmetros.

Aproximação proposta por Shimueli (2005), boa aproximação para  $\nu \leq 1$  ou  $\lambda > 10^\nu$

$$\blacktriangleright E(Y) \approx \lambda^{\frac{1}{\nu}} - \frac{\nu - 1}{2\nu}$$

$$\blacktriangleright \text{var}(Y) \approx \frac{1}{\nu} \cdot E(Y)$$

## 2.5

Distribuições para contagens: propriedades e  
modelos de regressão

**Comparando distribuições para  
contagens**

# Medindo propriedades das distribuições

- ▶ Índice de dispersão:

$$DI = \frac{\text{var}(Y)}{E(Y)}.$$

$DI < 1$  subdispersão,  $DI = 1$  equidispersão e  $DI > 1$  superdispersão.

- ▶ Índice de inflação de zeros:

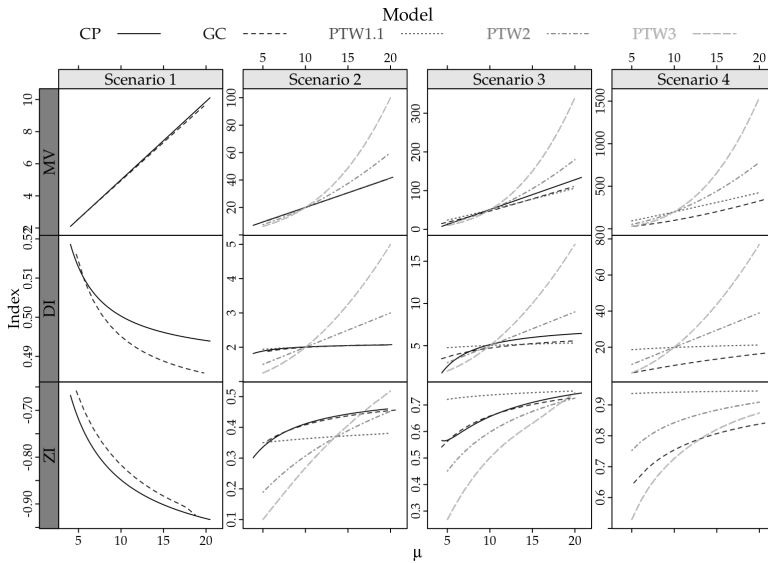
$$ZI = 1 + \frac{\log P(Y = 0)}{E(Y)}.$$

$ZI < 0$  zero deflacionado,  $ZI = 1$  não zero inflacionado e  $ZI > 1$  zero inflacionado.

- ▶ Índice de cauda pesada:

$$HT = \frac{P(Y = y + 1)}{P(Y = y)} \quad \text{for } y \rightarrow \infty.$$

$HT \rightarrow 1$  quando  $y \rightarrow \infty$  indica cauda pesada.



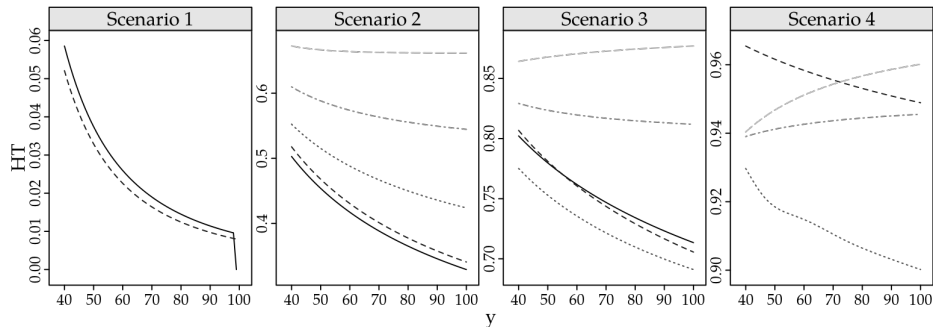


Figura : Índice de cauda pesada para alguns valores extremos da  $Y$ .

# Flexibilidade

**Tabela :** Reference models and dominant features by dispersion and power parameter values.

| Modelo de referência            | Fatos dominantes                   | Dispersão       | Power       |
|---------------------------------|------------------------------------|-----------------|-------------|
| Poisson                         | Equi                               | —               | —           |
| Gamma-Count                     | Sub, Equi, Super, deflação de zero | $\alpha \leq 1$ | —           |
| COM-Poisson                     | Sub, Equi, Super, deflação de zero | $\nu \leq 1$    | —           |
| Hermite                         | Super                              | $\phi > 0$      | $p = 0$     |
| Neyman Type A                   | Super, Zero-inflacionado           | $\phi > 0$      | $p = 1$     |
| <i>Poisson compound Poisson</i> | Super, Zero-inflacionado           | $\phi > 0$      | $1 < p < 2$ |
| Pólya-Aeppli                    | Super, Zero-inflacionado           | $\phi > 0$      | $p = 1.5$   |
| Negative binomial               | Super                              | $\phi > 0$      | $p = 2$     |
| <i>Poisson positive stable</i>  | Super, cauda pesada                | $\phi > 0$      | $p > 2$     |
| Poisson-inverse Gaussian        | Super, cauda pesada                | $\phi > 0$      | $p = 3$     |

3

# Método de máxima verossimilhança

# Método de máxima verossimilhança

- ▶ Conhecemos a distribuição que gerou os dados  $f(y; \theta)$ .
- ▶ Mas não seu finito vetor de parâmetros  $\theta \in \Theta$ .
- ▶  $\Theta$  em geral é um subconjunto do  $\mathbb{R}^n$ .
- ▶ Dado  $y$  uma observação da v.a.  $Y$  a função de verossimilhança

$$L(\theta; y) = f(y; \theta).$$

- ▶ Note que  $f(y; \theta)$  é uma função de probabilidade no espaço amostral.
- ▶ Porém,  $L(\theta; y) = f(y; \theta)$  é uma função no espaço paramétrico  $\Theta$ .
- ▶  $L(\theta; y)$  expressa a plausibilidade para diferentes valores dos parâmetros após observarmos  $y$  sem ter nenhuma outra informação sobre  $\theta$ .
- ▶ Para dados de contagem a verossimilhança é a probabilidade de observar o ponto  $y$  caso  $\theta$  seja o verdadeiro valor do parâmetro.



# Estimador de máxima verossimilhança

- ▶ Estimador de máxima verossimilhança (MLE)  $L(\hat{\theta}(y); y) = \max_{\theta \in \Theta} L(\theta; y)$ .
- ▶ Seja  $Y_i$  iid va com fp  $f(y; \theta)$  então

$$L(\theta; y) = \prod_{i=1}^n L(\theta; y_i) = \prod_{i=1}^n f(y_i; \theta).$$

- ▶ Log-verossimilhança

$$\ell(\theta) = \sum_{i=1}^n \log\{L(\theta; y_i)\}.$$

# Estimador de máxima verossimilhança

- ▶ MLE em geral pode ser obtido como a solução das equações de verossimilhança (ou escore)

$$\mathcal{U}(\boldsymbol{\theta}) = \left( \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p} \right)^\top = \mathbf{0}.$$

- ▶ Soluções analíticas são raras e métodos numéricos são necessários.
- ▶ A entrada  $(i, j)$  da matrix  $p \times p$  de informação de Fisher  $\mathcal{F}_\theta$  para o vetor  $\boldsymbol{\theta}$  é dada por

$$\mathcal{F}_{\theta_{ij}} = -\mathbb{E} \left\{ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\}.$$

- ▶ Algorithm Newton scoring

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mathcal{F}_\theta^{-1} \mathcal{U}(\boldsymbol{\theta}^{(i)}).$$

- ▶ Distribuição assintótica  $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathcal{F}_\theta^{-1})$ .

# Comentários MLE

- ▶ A verdadeira distribuição que gerou os dados é conhecida.
- ▶ É possível obter de forma analítica a primeira e segunda derivada da log-verossimilhança.
- ▶ A log-verossimilhança é
  - ① Poisson sem problemas!
  - ② Count-Gamma: diferença de duas integrais.
  - ③ Poisson-Tweedie: expressa como uma integral sem solução analítica.
  - ④ COM-Poisson: envolve uma soma infinita.
- ▶ Nestes casos não é possível obter expressões fechadas para as funções escore e matriz de informação de Fisher.
- ▶ Solução! Maximizar diretamente a log-verossimilhança usando algum método quase-Newton ou derivadas-free.
- ▶ Exemplos BFGS, Gradiente conjugado, Nelder-Mead.
- ▶ Ver função `optim()` em R.

# Exemplo: MLE distribuição Count-Gamma

## ► Log-verossimilhança

```
# Função de probabilidade
dgc <- function(y, gamma, alpha, log = FALSE) {
  p <- pgamma(q = 1,
              shape = y * alpha,
              rate = alpha * gamma) -
    pgamma(q = 1,
           shape = (y + 1) * alpha,
           rate = alpha * gamma)
  if(log == TRUE) {p <- log(p)}
  return(p)
}

# Função de log-verossimilhança
ll_gc <- function(gamma, alpha, y) {
  ll <- sum(dgc(y = y, gamma = gamma, alpha = alpha, log = TRUE))
  return(-ll)
}
```

# Exemplo: MLE distribuição Count-Gamma

- Maximização numérica e ajuste final

```
require(bbmle)
y <- rpois(100, lambda = 10)
fit_gc <- mle2(ll_gc, start = list("gamma" = 10, "alpha" = 1),
              data = list("y" = y))

fit_gc

##
## Call:
## mle2(minuslogl = ll_gc, start = list(gamma = 10, alpha = 1),
##      data = list(y = y))
##
## Coefficients:
##      gamma      alpha
## 9.5321032 0.8224768
##
## Log-likelihood: -263.21
```

4

## Modelos especificados por suposições de momentos

## 4.1

Modelos especificados por suposições de  
momentos  
**Especificação**

# Motivação

- ▶ Assumem que a distribuição de probabilidade é completamente conhecida a menos de um vetor de parâmetros.
- ▶ Na prática pode ser difícil escolher uma particular distribuição.
- ▶ Difícil de estimar usando métodos baseados em verossimilhança.
- ▶ Nem sempre a esperança é conhecida.
- ▶ O efeito das covariáveis não são diretamente relacionados a esperança da va.
- ▶ Abordagem mais geral que se adapte automaticamente a estrutura de dispersão dos dados.
- ▶ Fácil de implementar.
- ▶ SOLUÇÃO: Poisson-Tweedie estendida.



# Modelos de regressão

- ▶ Considere  $(y_i, x_i)$ ,  $i = 1, \dots, n$ , onde  $y_i$ 's são iid. va's.
- ▶ Especificação paramétrica completa:

$$Y_i \sim \text{PTw}_p(\mu_i, \phi).$$

- ▶ Especificação baseada em momentos:

$$\begin{aligned} E(Y_i) &= \mu_i \\ \text{var}(Y_i) &= \mu_i + \phi \mu_i^p \end{aligned}$$

onde  $g(\mu_i) = \eta_i = x_i^\top \beta$ ,  $x_i$  e  $\beta$  são  $(p \times 1)$  vetores de covariáveis conhecidas e parâmetros de regressão desconhecidos.

- ▶  $\text{var}(Y_i) > 0$ , assim  $\phi > -\mu_i^{(1-p)} \implies$  sub, equi e superdispersão.
- ▶  $g$  função de ligação (log link).

# Modelos de regressão

- Poisson-Tweedie estendida pode lidar com subdispersão  $\phi < 0$ .

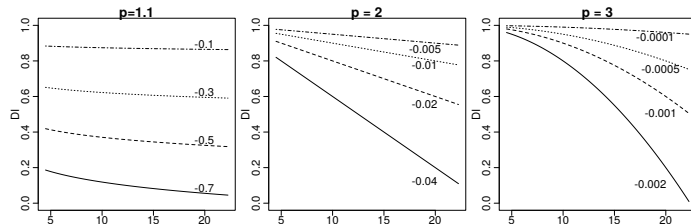


Figura : Índice de dispersão como uma função da média por valores dos parâmetros de dispersão e potência.

- Máxima verossimilhança precisa da especificação paramétrica completa.
- Funções de estimação (Bonat, et. al. 2016)
- Espaço paramétrico para o parâmetro de potência é livre ( $p \in \mathbb{R}$ ).

## 4.2

Modelos especificados por suposições de  
momentos  
**Estimação e Inferência**

## Parâmetros de regressão

- ▶ Seja  $\theta = (\beta^\top, \lambda^\top = (\phi, p)^\top)^\top$  o vetor de parâmetros.
- ▶ Função quasi-score para os parâmetros de regressão

$$\psi_\beta(\beta, \lambda) = \left( \sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_1} C_i^{-1} (y_i - \mu_i)^\top, \dots, \sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_Q} C_i^{-1} (y_i - \mu_i)^\top \right)^\top,$$

onde  $C_i = \mu_i + \phi \mu_i^p$  e  $\partial \mu_i / \partial \beta_j = \mu_i x_{ij}$  para  $j = 1, \dots, p$ .

- ▶ A entrada  $(j, k)$  da matriz  $p \times p$  de sensibilidade para  $\psi_\beta$  é dada por

$$S_{\beta_{jk}} = E \left( \frac{\partial}{\partial \beta_k} \psi_{\beta_j}(\beta, \lambda) \right) = - \sum_{i=1}^n \mu_i x_{ij} C_i^{-1} x_{ik} \mu_i. \quad (3)$$

- ▶ A entrada  $(j, k)$  da matriz  $p \times p$  de variabilidade para  $\psi_\beta$  é dada por

$$V_{\beta_{jk}} = \text{Var}(\psi_{\beta_{jk}}(\beta, \lambda)) = \sum_{i=1}^n \mu_i x_{ij} C_i^{-1} x_{ik} \mu_i. \quad (4)$$

# Parâmetros de dispersão

- Função de estimação de Pearson

$$\psi_{\lambda}(\lambda, \beta) = \left( \sum_{i=1}^n W_{i\phi} \left[ (y_i - \mu_i)^2 - C_i \right]^{\top}, \sum_{i=1}^n W_{ip} \left[ (y_i - \mu_i)^2 - C_i \right]^{\top} \right)^{\top},$$

onde  $W_{i\phi} = -\partial C_i^{-1} / \partial \phi$  e  $W_{ip} = -\partial C_i^{-1} / \partial p$ .

- A entrada  $(j, k)$  data matriz  $2 \times 2$  de sensibilidade é dada por

$$S_{\lambda_{jk}} = E \left( \frac{\partial}{\partial \lambda_k} \psi_{\lambda_j}(\lambda, \beta) \right) = - \sum_{i=1}^n W_{i\lambda_j} C_i W_{i\lambda_k} C_i, \quad (5)$$

onde  $\lambda_1$  e  $\lambda_2$  denota ambos  $\phi$  ou  $p$ .

# Matriz de sensibilidade cruzada

- A matriz de sensibilidade cruzada é dada por

$$S_{\beta_j \lambda_k} = E \left( \frac{\partial}{\partial \lambda_k} \psi_{\beta_j}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \right) = 0 \quad (6)$$

e

$$S_{\lambda_j \beta_k} = E \left( \frac{\partial}{\partial \beta_k} \psi_{\lambda_j}(\boldsymbol{\lambda}, \boldsymbol{\beta}) \right) = - \sum_{i=1}^n W_{i\lambda_j} C_i W_{i\beta_k} C_i, \quad (7)$$

onde  $W_{i\beta_k} = -\partial C_i^{-1} / \partial \beta_k$ .

- A matriz de sensibilidade conjunta para o vetor  $\boldsymbol{\theta}$  é dada por

$$S_{\boldsymbol{\theta}} = \begin{pmatrix} S_{\boldsymbol{\beta}} & \mathbf{0} \\ S_{\boldsymbol{\lambda}\boldsymbol{\beta}} & S_{\boldsymbol{\lambda}} \end{pmatrix},$$

cujas entradas são definidas por (3), (5), (6) e (7).

# Matriz de variabilidade

- ▶ A matriz de variabilidade para  $\theta$  tem a forma

$$V_{\theta} = \begin{pmatrix} V_{\beta} & V_{\lambda\beta}^{\top} \\ V_{\lambda\beta} & V_{\lambda} \end{pmatrix}$$

- ▶  $V_{\beta}$  foi definido em (4).
- ▶ As entradas para matriz de variabilidade empírica são dadas por

$$\tilde{V}_{\lambda_{jk}} = \sum_{i=1}^n \psi_{\lambda_j}(\lambda, \beta)_i \psi_{\lambda_k}(\lambda, \beta)_i \quad \text{and}$$

$$\tilde{V}_{\lambda_j\beta_k} = \sum_{i=1}^n \psi_{\lambda_j}(\lambda, \beta)_i \psi_{\beta_k}(\lambda, \beta)_i.$$

# Distribuição assintótica e algoritmo de ajuste

- ▶ Faça  $\hat{\theta}$  denotar o estimador função de estimação.
- ▶ A distribuição assintótica de  $\hat{\theta}$  é dada por

$$\hat{\theta} \sim N(\theta, J_{\theta}^{-1}), \quad \text{onde} \quad J_{\theta} = S_{\theta}^{-1} V_{\theta} S_{\theta}^{-1}$$

é a matriz de informação de Godambe.

- ▶ Algoritmo Chaser

$$\begin{aligned}\beta^{(i+1)} &= \beta^{(i)} - S_{\beta}^{-1} \psi_{\beta}(\beta^{(i)}, \lambda^{(i)}) \\ \lambda^{(i+1)} &= \lambda^{(i)} - \alpha S_{\lambda}^{-1} \psi_{\lambda}(\beta^{(i+1)}, \lambda^{(i)}).\end{aligned}$$







- ▶ Facilmente implementado em R através da função `mcglm()` do pacote `mcglm` (Bonat, 2015).
- ▶  $\alpha$  é um *tuning* constante para controlar o tamanho do passo.








5

# Aplicações

# Referências

-  Bonat, W. H., Jørgensen, B. (2016). Multivariate covariance generalized linear models. *Journal of Royal Statistical Society - Series C*, 65, 649–675.
-  Bonat, W. H. et.al, (2017). Extended Poisson-Tweedie: properties and regression models. *Statistical Modelling*, to appear.
-  Bonat, W. H. (2017). Multiple regression models in R. The mcglm package. *Journal of Statistical Software*, to appear.
-  Conway, R. W., Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12, 132–136.
-  Paula, G. A. (2013). *Modelos de regressão com apoio computacional*. IME-USP, São Paulo.
-  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.

# Referências

-  Zeileis, A., Kleiber, C., Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8), 1 - 25. doi:<http://dx.doi.org/10.18637/jss.v027.i08>
-  Winkelmann, R. (2008). *Econometric analysis of count data* (5th Ed.). Springer Science & Business Media.
-  SILVA, A. M.; DEGRANDE, P. E.; SUEKANE, R.; FERNANDES, M. G.; ZEVIANI, W. M.  
Impacto de diferentes níveis de desfolha artificial nos estádios fenológicos do algodoeiro.  
[Revista de Ciências Agrárias](#), v.35, n.1, 2012 (prelo).
-  WINKELMANN, R.; ZIMMERMANN, K.  
Count data models for demographic data.  
[Mathematical Population Studies](#), v.4, n.3, p.205–221, 1994.
-  WINKELMANN, R.  
Duration dependence and dispersion in count-data models.  
[Journal of Business & Economic Statistics](#), v.13, n.4, p.467–474, 1995.

# Referências



CONSUL, P. C. AND G. C. JAIN

A generalization of the Poisson distribution. **Technometrics**, v.15, n.4, p.791–799, 1973.



CONSUL, P. C

Generalized Poisson Distributions: Properties and Applications. **Statistics: Textbooks and Monographs**, New York: Marcel Dekker Inc. 1989.