

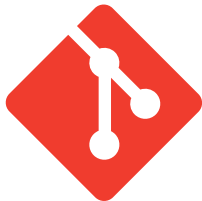
# Modelos de Regressão para Dados de Contagem com R

Prof. Dr. Walmes M. Zeviani  
Eduardo E. Ribeiro Jr  
Prof. Dr. Cesar A. Taconelli

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

23 de maio de 2016  
[edujrrib@gmail.com](mailto:edujrrib@gmail.com)

# Disponibilização



<https://github.com/leg-ufpr/MRDCr>

<https://gitlab.c3sl.ufpr.br/leg/MRDCr>

Modelos de Regressão para Dados de Contagem com `r` – MRDCr

# Conteúdo

1. Introdução
2. Modelo de Poisson
3. Estimação via Quase-Verossimilhança
4. Modelo Binomial Negativa
5. Modelos para Excesso de Zeros
  - 5.1 Modelos de Barreira *Hurdle*
  - 5.2 Modelos de Mistura (*Zero Inflated*)
6. Modelos Paramétricos Alternativos
  - 6.1 Modelo Poisson-Generalizada
  - 6.2 Modelo COM-Poisson
  - 6.3 Modelo Gamma-Count
7. Modelos com Efeitos Aleatórios

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 de 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 bem 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 variância e média, característica de dados de contagens.

2

# Modelo de Poisson

# A distribuição de Poisson

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

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

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

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

# Propriedades da distribuição de Poisson

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

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

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

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

# Distribuição Poisson para diferentes valores de $\lambda$

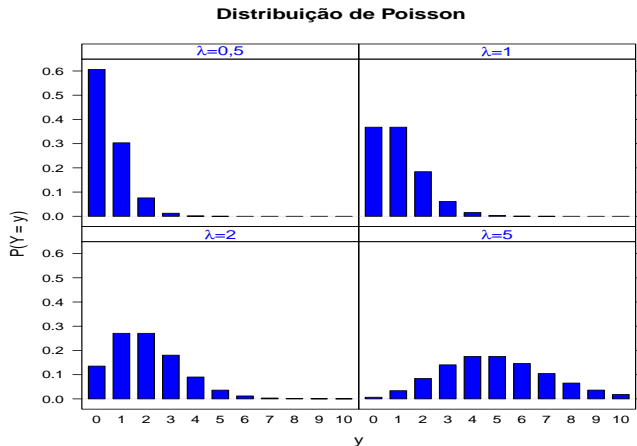


Figura : Distribuição de Poisson para diferentes valores de  $\lambda$

# Motivações para a distribuição de Poisson

- Caso limite da distribuição binomial( $n, \pi$ ) quando  $n \rightarrow \infty$  e  $\pi \rightarrow 0$ , fixado  $\lambda = n\pi$ , ou seja:

$$\lim_{n \rightarrow \infty \pi \rightarrow 0} \left[ \binom{n}{k} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{n-k} \right] = \frac{e^{-\lambda} \lambda^k}{k!}.$$

- Resultado do processo estocástico de Poisson, em que os eventos contados ocorrem **aleatoriamente** ao longo do tempo, espaço,...

# Motivações para a distribuição de Poisson

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

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

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

# Diferentes comportamentos para $\lambda(t)$

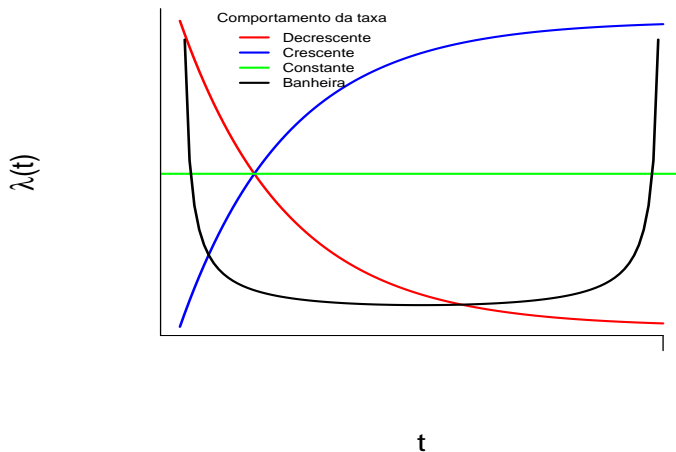


Figura : Diferentes comportamentos para  $\lambda(t)$



# O 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:

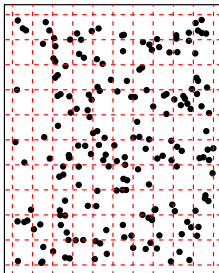
- 1  $Y(t)$  é inteiro e não negativo;
- 2 Para  $s < t$ ,  $Y(s) \leq Y(t)$ ;
- 3  $Y(t) - Y(s)$  é o número de eventos que ocorrem no intervalo  $(s, t]$ ;
- 4 O processo é estacionário:

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

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

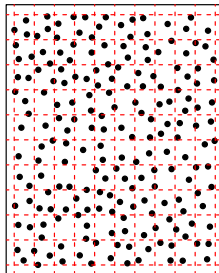
# Diferentes padrões em processos de contagens

Padrão aleatório



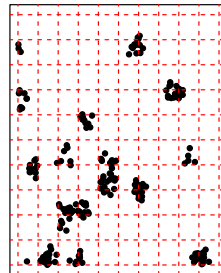
Equidispersão  
 $\text{Var}(Y) = E(Y)$

Padrão uniforme



Subdispersão  
 $\text{Var}(Y) < E(Y)$

Padrão agregado



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

Figura : Ilustração de diferentes tipos de processos pontuais

# Regressão Poisson

- ▶ O modelo de regressão Poisson (ou modelo log linear de Poisson) é o mais usado para a análise de dados de contagens.
- ▶ A regressão Poisson baseia-se nos pressupostos inerentes ao processo e à distribuição de Poisson.
- ▶ Caso tais pressupostos não sejam atendidos, a regressão Poisson ainda pode ser uma alternativa apropriada, desde que usada com os cuidados necessários.

# Regressão Poisson - Especificação do modelo

- Sejam  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias condicionalmente independentes, dado o vetor de covariáveis  $\mathbf{x}_i' = (x_{i1}, x_{i2}, \dots, x_{ip})$ ,  $i = 1, 2, \dots, n$ . A regressão Poisson é definida pela distribuição de Poisson:

$$f(y_i | \mathbf{x}_i) = \frac{e^{-\mu_i} (\mu_i)^{y_i}}{y_i!}, \quad y = 0, 1, 2, \dots,$$

sendo as covariáveis inseridas ao modelo por:

$$\ln(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta},$$

em que  $\boldsymbol{\beta}$  é o vetor de parâmetros de regressão.

# Regressão Poisson - Propriedades

- ▶  $f(y_i|\mathbf{x}_i) = \frac{e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})} \exp(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}}{y_i!}$
- ▶  $E[y_i|\mathbf{x}_i] = \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) ;$
- ▶  $Var[y_i|\mathbf{x}_i] = \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) .$

# Regressão Poisson - Estimação por máxima verossimilhança

Para a regressão Poisson:

- ▶ Log-verossimilhança:  $l(\beta) = \sum_{i=1}^n \{y_i x_i' \beta - \exp(x_i' \beta)\} - \ln y_i!$ ;
- ▶ Vetor escore:  $S(\beta) = \frac{\partial l(\beta; y)}{\partial \beta} = \sum_{i=1}^n (y_i - \exp(x_i' \beta)) x_i$ ;
- ▶ Matriz Informação:  $I(\beta) = \sum_{i=1}^n \mu_i x_i x_i' = \exp(x_i' \beta) x_i x_i'$ ;
- ▶ Distribuição assintótica:  $\hat{\beta} \stackrel{a}{\sim} N\left(\beta, \left[\sum_{i=1}^n \mu_i x_i x_i'\right]^{-1}\right)$ .

# Regressão Poisson - Modelo Linear Generalizado

A Regressão Poisson é um caso particular dos Modelos Lineares Generalizados (MLG). Algumas propriedades dessa classe de modelos:

- ▶ Os estimadores são consistentes ainda que a distribuição especificada seja incorreta, mas desde que a média condicional de  $Y$  seja declarada corretamente;
- ▶ Os erros padrões, intervalos de confiança e testes de hipóteses, no entanto, ficam comprometidos;
- ▶ O ajuste de um MLG requer apenas a especificação:
  - ▶ Da esperança de  $Y$  condicional às covariáveis, mediante especificação do preditor linear e da função de ligação;
  - ▶ Da variância condicional, mediante especificação da função de variância  $V(\mu)$ , possível inclusão do parâmetro de dispersão ( $\phi$ ), ou sua estimação por métodos robustos (abordagem de Quase-Verossimilhança).

# Estudos de caso

*Vignette* `v01_poisson.html`



3

# Estimação via Quase-Verossimilhança

# Regressão Poisson - Quase-Verossimilhança

- ▶ Para o ajuste de um modelo alternativo via Quase-Verossimilhança, definimos:

$$g(E(y_i|x_i)) = x_i'\beta;$$

$$Var(y_i|x_i) = \phi V(\mu_i).$$

- ▶ A obtenção dos estimadores se dá pela maximização da função de quase-verossimilhança:

$$Q(\mu) = \int_y^\mu \frac{y-t}{\phi V(t)} dt$$

.

- ▶ As funções de quase-verossimilhança, quase-escore e quase-informação compartilham propriedades comuns às correspondentes funções no caso paramétrico, para MLGs.

# Estimação via Quase-Verossimilhança

- Distribuição assintótica:

$$\hat{\beta}_{QL} \overset{a}{\sim} N(\beta, Var(\hat{\beta}_{QL}))$$

- Para o modelo Quase-Poisson, assume-se:

$$\ln(E(y_i|x_i)) = x_i'\beta;$$

# Estimação via Quase Verossimilhança

- ▶ A matriz de covariâncias assintótica para  $\hat{\beta}_{QL}$  fica dada por:

$$Var(\hat{\beta}_{QL}) = \left[ \sum_{i=1}^n x_i x_i' \mu_i \right]^{-1} \sum_{i=1}^n x_i x_i' \omega_i \left[ \sum_{i=1}^n x_i x_i' \mu_i \right]^{-1},$$

com  $\mu_i = \exp(x_i' \beta)$  e  $\omega_i = Var(y_i | x_i)$ .

- ▶ Podemos considerar  $\omega_i = V(\mu_i; \phi)$ , como  $\omega_i = \phi \mu_i$ ,  $\omega_i = \phi \mu_i^2$  ou, simplesmente, o estimador robusto, baseado em  $\omega_i = (y_i - \mu_i)^2$ .
- ▶ Um estimador para  $\phi$ :

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}.$$

# Estudos de caso

*Vignette* [Ovelhas.html](#)

4

# Modelo Binomial Negativa

# Distribuição binomial negativa

- Função de probabilidades:

$$P(Y = k) = \frac{\Gamma(\alpha + k)}{\Gamma(k + 1)\Gamma(\alpha)} \left(\frac{\lambda}{\lambda + \alpha}\right)^k \left(\frac{\alpha}{\lambda + \alpha}\right)^\alpha, k = 0, 1, 2, \dots; \alpha > 0, \lambda > 0$$

- Propriedades:

$$E(Y) = \lambda; \quad Var(Y) = \lambda + \alpha^{-1}\lambda^2$$

- Assim, para qualquer  $\alpha > 0$ , temos  $Var(Y) > \lambda$ .
- A distribuição binomial negativa tem como caso limite distribuição Poisson, quando  $\alpha \rightarrow \infty$ .

# Distribuição binomial negativa

- Uma parametrização alternativa:

$$P(Y = k) = \binom{r+k-1}{r-1} (1-p)^r p^k, \quad k = 0, 1, 2, \dots,$$

sendo  $r = \alpha$  e  $p = \lambda / (\lambda + \alpha)$ , com  $0 < p < 1$  e  $r > 0$ .

- Modelagem do número de "sucessos" até o  $r$ -ésimo "fracasso" ( $r = 1, 2, 3, \dots$ ), configurando uma generalização da distribuição geométrica (para  $r = 1$ ).
- Modelagem de alguns tipos de processos pontuais envolvendo contágio.



# Distribuição binomial negativa

- ▶ A principal motivação para a distribuição binomial negativa baseia-se num processo de contagem heterogêneo, em que  $Y \sim \text{Poisson}(\theta)$  e  $\theta$  tem distribuição  $\text{Gama}(\alpha, \beta)$  :

$$g(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \quad \alpha, \beta, \nu > 0,$$

com  $E(\theta) = \theta = \alpha/\beta$  e variância  $\text{Var}(\theta) = \alpha/\beta^2$ .

- ▶ Como resultado, temos uma mistura Poisson-Gamma, resultando, marginalmente (em relação a  $\theta$ ), na distribuição binomial negativa.

# Distribuição binomial negativa para diferentes parâmetros

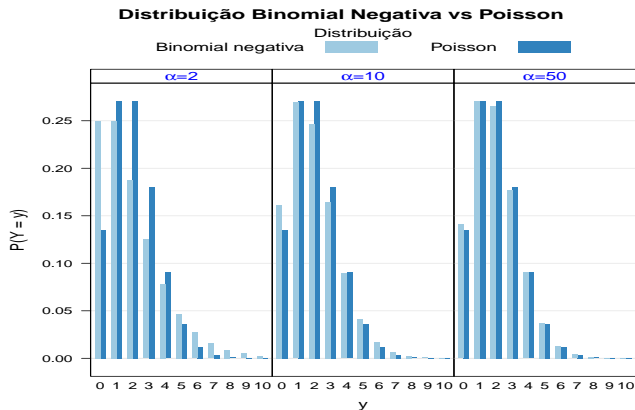


Figura : Distribuição binomial negativa para  $\lambda = 2$  e diferentes valores de  $\alpha$ .

# Distribuição binomial negativa

- ▶ O modelo de regressão com resposta binomial negativa pode ser especificado fazendo  $E(y|x) = \exp(x'\beta)$ .
- ▶ Para valores fixados de  $\alpha$ , a distribuição binomial negativa fica expressa na forma da família exponencial linear, contemplada pela teoria de MLG.
- ▶ A estimação dos parâmetros do modelo se dá numericamente, segundo um algoritmo em duas etapas, em que  $\alpha$  e  $\beta$  são estimados condicionalmente até convergência.

# Estudos de caso

*Vignette* Sinistros.html

5

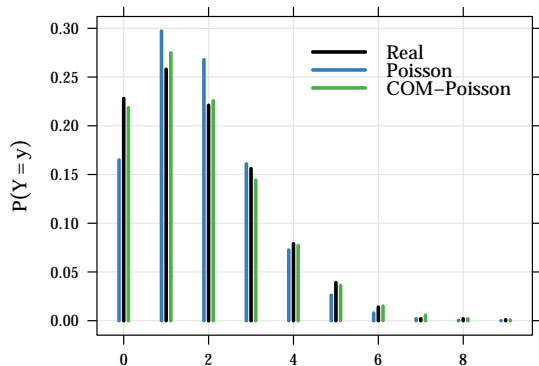
# Modelos para Excesso de Zeros

# Excesso de Zeros

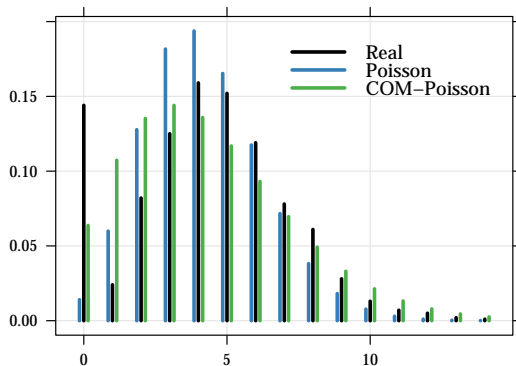
- ▶ Casos em que a proporção de valores nulos na amostra é superior àquela estimada por um modelo de contagem. No caso Poisson  $e^{-\lambda}$
- ▶ Geralmente contagens com um número excessivo de valores nulos apresentam superdispersão (ocasionada pelo excesso de zeros).
- ▶ Os modelos mais flexíveis abordados não capturam esse excesso de zeros e não se ajustam adequadamente.

# Excesso de Zeros

$$\mu_{\text{count}} = 2, \pi_{\text{zero extra}} = 0.1$$



$$\mu_{\text{count}} = 5, \pi_{\text{zero extra}} = 0.15$$



# Gerador de excesso de zeros

- ▶ Uma limitação das abordagens estudadas é que as contagens nulas e não nulas são provenientes do mesmo processo gerador dos dados.
- ▶ Para dados com excesso de zeros, é razoável a suposição da haver mais de um processo gerador atuando na geração dos dados.
- ▶ Assim a ocorrência de valores nulos podem ser caracterizada como:
  - ▶ **zeros amostrais:** Ocorrem segundo um processo gerador de contagens (e.g Processo Poisson)
  - ▶ **zeros estruturais:** Ausência de determinada característica da população.



# Gerador de excesso de zeros

Exemplo. Um estudo que visa avaliar a quantidade de produtos comprados em um mercado por uma família na última semana. A variável de interesse é o número de itens comprados.

*zeros estruturais*: Se a família não foi ao mercado na última semana. Inevitavelmente o número de produtos será 0.

*zeros amostrais*: A família foi ao mercado, porém não adquiriu nenhum produto.

# Modelando contagens com excesso de zeros

Como há dois processos que geram os valores da população, na modelagem deve-se considerar ambos. As principais abordagens nestes casos são via:

- ▶ **Modelos de barreira (*Hurdle Models*):** que desconsidera os zeros amostrais e modela os zeros estruturais e as contagens positivas (seção ??); e
- ▶ **Modelos de mistura (*Zero Inflated Models*):** que modela os zeros (estruturais e amostrais) em conjunto com as contagens positivas (??).

## 5.1

Modelos para Excesso de Zeros  
**Modelos de Barreira *Hurdle***

# Modelos *Hurdle*

- ▶ A variável de interesse é particionada em contagens nulas e não nulas;
- ▶ Consideram somente os zeros estruturais;
- ▶ São chamados também de modelos condicionais, hierárquicos ou de duas partes;
- ▶ Esta abordagem combina um modelo de contagem truncado à esquerda do ponto  $y = 1$  e um modelo censurado à direita no mesmo ponto  $y = 1$

# Modelos *Hurdle*

## Distribuição de probabilidades

$$Pr(Y = y) = \begin{cases} f_z(0) & \text{se } y = 0, \\ (1 - f_z(0)) \frac{f_c(Y = y)}{1 - f_c(Y = 0)} & \text{se } y = 1, 2, \dots \end{cases}$$

em que  $f_z$  é uma função de probabilidades degenerada no ponto 0 e  $f_c$  um função de probabilidades de uma variável  $Y^*$ , como a Poisson.

## Momentos da distribuição

### Média

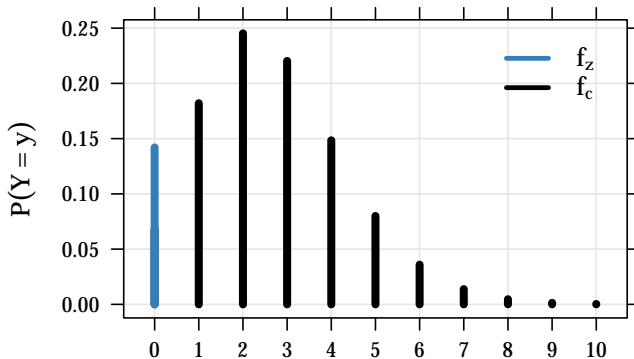
$$E(Y) = \frac{E(Y^*)(1 - f_z(0))}{1 - f_c(Y = 0)}$$

### Variância

$$V(Y) = \frac{1 - f_z(0)}{1 - f_c(Y = 0)} \left[ E(Y^*) \frac{(1 - f_z(0))}{1 - f_c(Y = 0)} \right]$$

## Distribuição *Hurdle*

- ▶  $f_z$  é uma função de probabilidades degenerada no ponto  $y = 0$ , ou seja, tem toda massa no ponto 0.
- ▶  $f_c$  é uma função de probabilidades tradicional, que no modelo é truncada em  $y = 1$ .
- ▶ Os modelos de barreira combinam  $f_z$  e  $f_c$  para descrever  $Y$
- ▶ Para a parte positiva os dados ainda podem apresentar sub, superdispersão ou excesso de valores em outro ponto.



## Combinações comuns

Pode-se propor diferentes distribuições para  $f_z$  e  $f_c$ . Uma escolha natural para  $f_z$  é a Bernoulli e para  $f_c$  a Poisson. Assim

$$\begin{array}{l} f_z \sim \text{Bernoulli}(\pi) \\ f_c \sim \text{Poisson}(\lambda) \end{array} \quad \Rightarrow \quad P(Y = y) = \begin{cases} 1 - \pi & \text{se } y = 0, \\ \pi \left( \frac{e^{-\lambda} \lambda^y}{y! (1 - e^{-\lambda})} \right) & \text{se } y = 1, 2, \dots \end{cases}$$

Embora essa escolha de modelo seja o que tem o maior suporte computacional, ressalta-se que outras distribuições podem ser escolhidas para ambas as partes  $f_z$  e  $f_c$ .

# Modelos de regressão *Hurdle*

- ▶ Incorporando covariáveis em  $f_z$  e  $f_c$  na forma  $h(Z\gamma)$  e  $g(X\beta)$ , respectivamente.
- ▶ As funções  $h(\cdot)$  e  $g(\cdot)$ , são as funções de ligação escolhidas conforme modelos  $f_z$  e  $f_c$ .
- ▶ O modelo de regressão *Hurdle* terá, portanto, os vetores de parâmetros  $\beta$ ,  $\gamma$  e potencialmente  $\phi$  (caso um modelo com parâmetro de dispersão for considerado)
- ▶ Se os modelos para  $f_z$  e  $f_c$  e as respectivas matrizes  $Z$  e  $X$  forem as mesmas, o teste  $H_0 : \beta = \gamma$  avalia a necessidade do modelo *Hurdle*.



# Modelos de regressão *Hurdle*

## Função de verossimilhança

$$L(\underline{\theta}; \underline{y}) = \prod_{i=1}^n (1 - \mathbb{1}(f_{z_i}(0))) \cdot \prod_{i=1}^n \mathbb{1} \left( (1 - f_{z_i}(0)) \left( \frac{f_{c_i}(y_i)}{1 - f_{c_i}(0)} \right) \right)$$

## Função de log-verossimilhança

$$l(\underline{\theta}; \underline{y}) = \sum_{i=1}^n (1 - \mathbb{1}(\log(f_{z_i}(0)))) + \sum_{i=1}^n \mathbb{1}(\log(1 - f_{z_i}(0)) + \log(f_{c_i}(y_i)) - \log(1 - f_{c_i}(0)))$$

Sendo  $\mathbb{1}$  a função indicadora que assume o valor 1 se  $y > 0$  e  $\beta$ ,  $\gamma$  e  $\phi$ , se houver). 0 se  $y = 0$  e  $\underline{\theta}$  o vetor de parâmetros do modelo.

# Modelos *Hurdle* no R

Neste minicurso utilizaremos principalmente pacote o `pscl` (*Political Science Computational Laboratory, Stanford University*)

```
library(pscl)
hurdle(y ~ fc_predictor | fz_predictor, dist = "poisson", zero.dist = "poisson")
```

# Modelos *Hurdle* no R

Um outro pacote que proporciona diversas funções e podemos adaptar para o ajuste desses modelos é o VGAM (*Vector Generalized Linear and Additive Models*)

```
library(VGAM)
vglm(y ~ predictor, family = zapoisson)

## ou ajustando as partes
vglm(y ~ fc_predictor, family = pospoisson, data = subset(data, y > 0))
vglm(SurvS4(cy, st) ~ fz_predictor, cens.poisson,
     data = transform(data, cy = pmin(1, y), st = ifelse(y >= 1, 0, 1)))
```

## 5.2

Modelos para Excesso de Zeros  
**Modelos de Mistura (*Zero Inflated*)**

# Modelo *Zero Inflated*

- ▶ Consideram uma mistura de modelos;
- ▶ Os zeros agora são caracterizados em amostrais e estruturais;
- ▶ Há contribuição para estimação da probabilidade em zero de duas funções de probabilidade;
- ▶ São chamados de modelos de mistura ou inflacionados de zero (*ZI*);
- ▶ Esta abordagem “mistura” um modelo de contagem sem restrição e um modelo censurado à direita no ponto  $y = 1$ .

# Modelo Zero Inflated

## Distribuição de probabilidades

$$Pr(Y = y) = \begin{cases} f_z(0) + (1 - f_z(0))f_c(Y = y) & \text{se } y = 0, \\ (1 - f_z(0))f_c(Y = y) & \text{se } y = 1, 2, \dots \end{cases}$$

## Momentos da distribuição

### Média

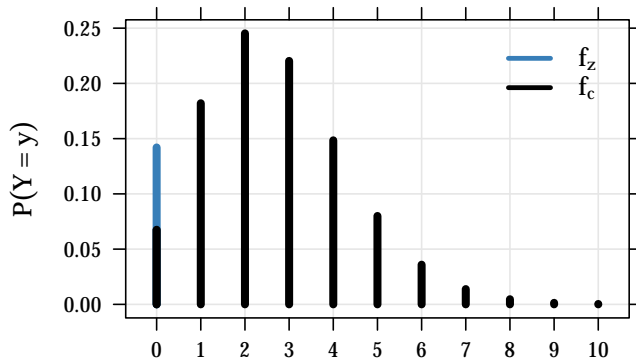
$$E(Y) = (1 - f_z(0))E(Y^*)$$

### Variância

$$V(Y) = (1 - f_z(0))E(Y^*)[E(Y^{*2}) - (1 - f_z(0))E^2(Y^*)]$$

# Distribuição Zero Inflated

- ▶  $f_z$  é uma função de probabilidades degenerada no ponto  $y = 0$ , ou seja, tem toda massa no ponto 0.
- ▶  $f_c$  é uma função de probabilidades para dados de contagem.
- ▶ Os modelos de mistura misturam  $f_z$  e  $f_c$  para descrever  $Y$
- ▶ Para a parte  $f_c$  os dados ainda podem apresentar sub, superdispersão ou excesso de valores em outro ponto.



# Misturas comuns

Pode-se propor diferentes distribuições para  $f_z$  e  $f_c$ . Uma escolha natural para  $f_z$  é a Bernoulli e para  $f_c$  a Poisson. Assim

$$\begin{array}{l} f_z \sim \text{Bernoulli}(\pi) \\ f_c \sim \text{Poisson}(\lambda) \end{array} \quad \Rightarrow \quad P(Y = y) = \begin{cases} (1 - \pi) + \pi e^{-\lambda} & \text{se } y = 0, \\ \pi \left( \frac{e^{-\lambda} \lambda^y}{y!} \right) & \text{se } y = 1, 2, \dots \end{cases}$$

Embora essa escolha de modelo seja o que tem o maior suporte computacional, ressalta-se que outras distribuições podem ser escolhidas para ambas as partes  $f_z$  e  $f_c$ .



# Modelos de regressão *Zero Inflated*

- ▶ Incorporando covariáveis em  $f_z$  e  $f_c$  na forma  $h(Z\gamma)$  e  $g(X\beta)$ , respectivamente.
- ▶ As funções  $h(\cdot)$  e  $g(\cdot)$ , são as funções de ligação escolhidas conforme modelos  $f_z$  e  $f_c$ .
- ▶ O modelo de regressão *Hurdle* terá, portanto, os vetores de parâmetros  $\beta$ ,  $\gamma$  e potencialmente  $\phi$  (caso um modelo com parâmetro de dispersão for considerado)
- ▶ Como agora são modelos misturados a comparação entre  $\beta$  e  $\gamma$  não tem a mesma interpretabilidade.
- ▶ Para comparação de modelos tradicionais contra os modelos de mistura, o teste de Vuong para modelos não aninhados pode ser aplicado.

# Modelos de regressão *Zero Inflated*

## Função de verossimilhança

$$L(\underline{\theta}; \underline{y}) = \prod_{i=1}^n \mathbb{1}((1 - f_{z_i}(0))f_{c_i}(y_i)) \cdot \prod_{i=1}^n (1 - \mathbb{1}(f_{z_i}(0) + (1 - f_{z_i}(0))f_{c_i}(0)))$$

## Função de log-verossimilhança

$$l(\underline{\theta}; \underline{y}) = \sum_{i=1}^n \mathbb{1}(\log(1 - f_{z_i}(0)) + \log(f_{c_i})) + \sum_{i=1}^n (1 - \mathbb{1}(\log(f_{z_i}(0) + (1 - f_{z_i}(0))f_{c_i}(0))))$$

Sendo  $\mathbb{1}$  a função indicadora que assume o valor 1 se  $y > 0$  e 0 se  $y = 0$  e  $\underline{\theta}$  o vetor de parâmetros do modelo ( $\beta$ ,  $\gamma$  e  $\phi$ , se houver).

# Modelos *Zero Inflated* no R

Usando o `pscl` (*Political Science Computational Laboratory, Stanford University*)

```
library(pscl)
zeroinfl(y ~ fc_predictor | fz_predictor, dist = "poisson", link = "logit")
```

Usando o VGAM (*Vector Generalized Linear and Additive Models*)

```
library(VGAM)
vglm(y ~ predictor, family = zipoisson)
```

# Estudos de caso

*Vignette* `v07_excesso-zeros.html`

`peixe` : número de peixes capturados por grupos em um parque estadual

`sinistros` : número de sinistros em uma seguradora de automóveis

6

# Modelos Paramétricos Alternativos

## 6.1

Modelos Paramétricos Alternativos

# **Modelo Poisson-Generalizada**

# A distribuição de probabilidade

- ▶ Introduzida por [?] e estudada em detalhes por [?]
- ▶ Modela casos de superdispersão e subdispersão.
- ▶ A Poisson é um caso particular.
- ▶ Se  $Y \sim \text{Poisson Generalizada}$ , sua função de probabilidade é

$$f(y) = \begin{cases} \theta(\theta + \gamma y)^{y-1} \exp\{-(\theta + \gamma y)\}, & y = 0, 1, 2, \dots \\ 0, & y > m \text{ quando } \gamma < 0. \end{cases}$$

- ▶  $\theta > 0$ .
- ▶  $\max\{-1, -\theta/m\} < \gamma < 1$ .
- ▶  $m$  é maior inteiro positivo para o qual  $\theta + m\gamma > 0$  quando  $\gamma$  é negativo.
- ▶ Note que o espaço paramétrico de  $\gamma$  é dependente do parâmetro  $\theta$ .

# Propriedades da Poisson Generalizada

## Média e variância

- ▶  $E(Y) = \theta(1 - \gamma)^{-1}$ .
- ▶  $V(Y) = \theta(1 - \gamma)^{-3}$ .

## Relação média-variância

- ▶ Superdispersa se  $\gamma > 0$ .
- ▶ Subdispersa se  $\gamma < 0$ .

Quando  $\gamma = 0$  a Poisson Generalizada reduz a distribuição Poisson e, portanto, apresenta equidispersão.



# Parametrização de média para modelo de regressão

Defina

$$\theta = \frac{\lambda}{1 + \alpha\lambda}, \quad \gamma = \alpha \frac{\lambda}{1 + \alpha\lambda}.$$

Ao substituir na função densidade, tem-se

$$f(y) = \left( \frac{\lambda}{1 + \alpha\lambda} \right)^y \frac{(1 + \alpha y)^{y-1}}{y!} \exp \left\{ -\lambda \frac{(1 + \alpha y)}{(1 + \alpha\lambda)} \right\}.$$

- ▶  $E(y) = \lambda$ ,
- ▶  $V(y) = \lambda(1 + \alpha\lambda)^2$ .
- ▶ Superdispersa se  $\alpha > 0$ ,
- ▶ Subdispersa se  $\alpha < 0$ .
- ▶ Poisson se  $\alpha = 0$ .

# Restrições no espaço paramétrico

- ▶  $\lambda > 0$ .
- ▶  $1 + \alpha\lambda > 0$ .
- ▶  $1 + \alpha y > 0$ .

Considerando uma amostra aleatória de  $y_i$  e valores conhecidos de  $\lambda_i$ ,  $i = 1, 2, \dots$ , as restrições combinadas sobre  $\alpha$  resultam em

$$\alpha > \min \left\{ \frac{-1}{\max(y_i)}, \frac{-1}{\max(\lambda_i)} \right\}, \quad \text{quando } \alpha < 0. \quad (1)$$

# Função de log-verossimilhança

Considerando uma amostra aleatória  $y_i, i = 1, 2, \dots, n$ , a verossimilhança é

$$L(y; \lambda, \alpha) = \prod_{i=1}^n \left( \frac{\lambda}{1 + \alpha\lambda} \right)^{y_i} \frac{(1 + \alpha y_i)^{y_i - 1}}{y_i!} \exp \left\{ -\lambda \frac{(1 + \alpha y_i)}{(1 + \alpha\lambda)} \right\}. \quad (2)$$

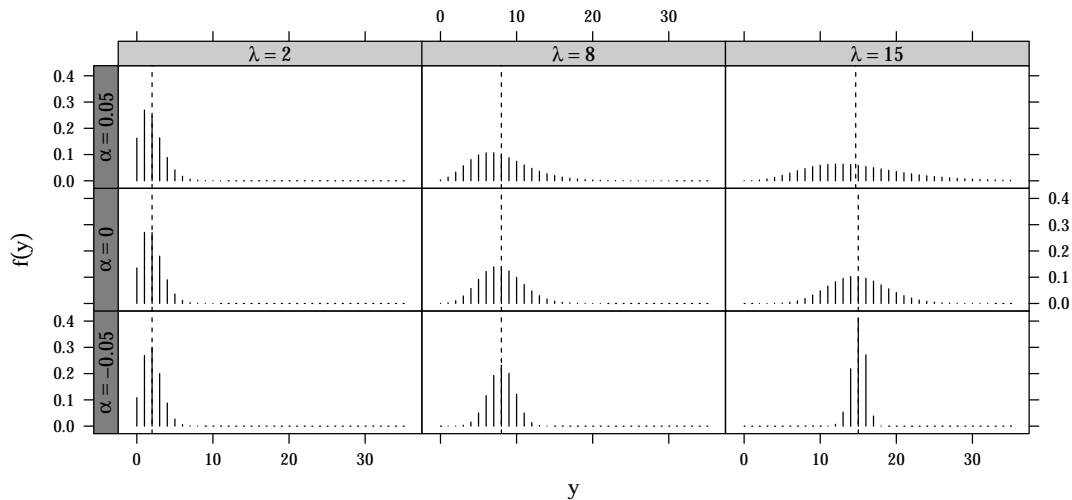
A função de log-verossimilhança é

$$\ell(y; \lambda, \alpha) = \sum_{i=1}^n y_i \ln(\lambda) - \ln(1 + \alpha\lambda) + (y_i - 1) \ln(1 + \alpha y_i) - \lambda \frac{(1 + \alpha y_i)}{(1 + \alpha\lambda)} - \ln(y_i!) \quad (3)$$

# Implementação da log-verossimilhança

```
## library(MRDCr)
devtools::load_all()
llpgnz

## function(params, y, X, offset = NULL) {
##   # params: vetor de parâmetros;
##   #   params[1]: parâmetro de dispersão (alpha);
##   #   params[-1]: parâmetro de locação (lambda);
##   # y: variável resposta (contagem);
##   # X: matriz do modelo linear;
##   # offset: tamanho do domínio onde y foi medido (exposição);
##   #-----
##   if (is.null(offset)) {
##     offset <- 1L
##   }
##   alpha <- params[1]
##   lambda <- offset * exp(X %*% params[-1])
```



# Estudos de caso

[poisson\\_generalizada.html](#)

**soja** : Número de vagens, de grãos e de grãos por vagem.

**capdesfo** : Número de capulhos produzidos em algodão.

**nematoide** : Número de nematoides em raízes de linhagens de feijoeiro.

## 6.2

Modelos Paramétricos Alternativos  
**Modelo 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

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

## ▶ Distribuição COM-Poisson

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



# Distribuição COM-Poisson

## Distribuição de probabilidades

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

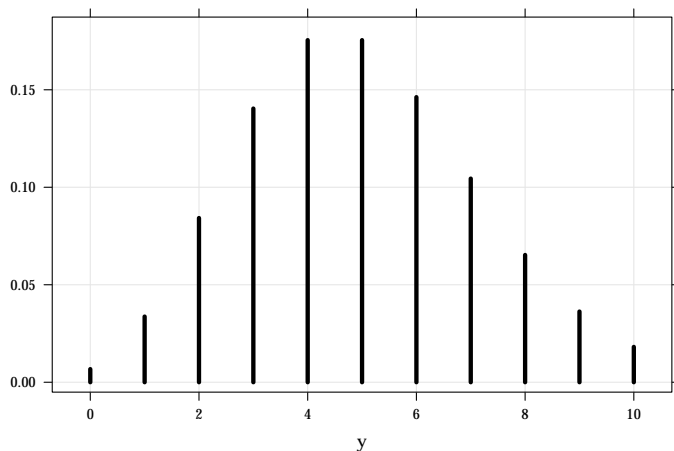
## Casos particulares

- ▶ 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  $\nu = 0, \lambda < 1$

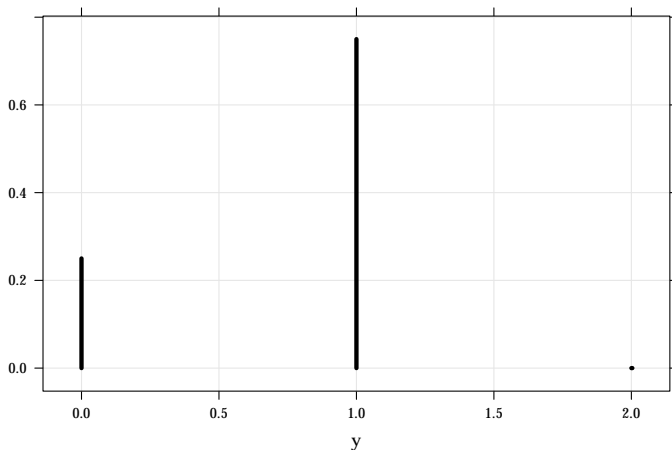
COM-Poisson ( $\lambda = 5, \nu = 1$ )



# Distribuição COM-Poisson

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

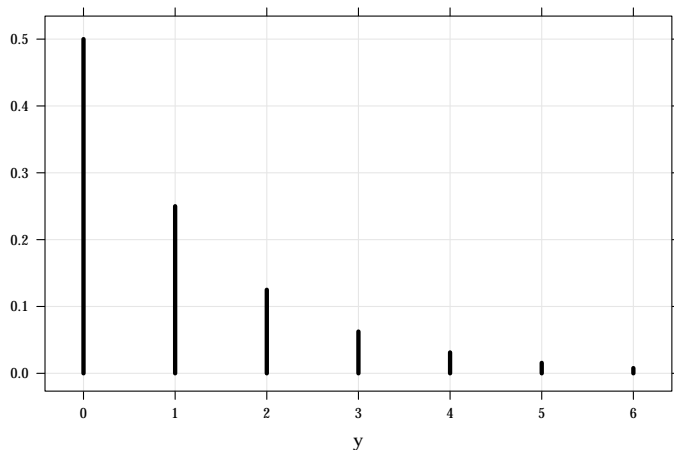
COM-Poisson ( $\lambda = 3, \nu = 20$ )

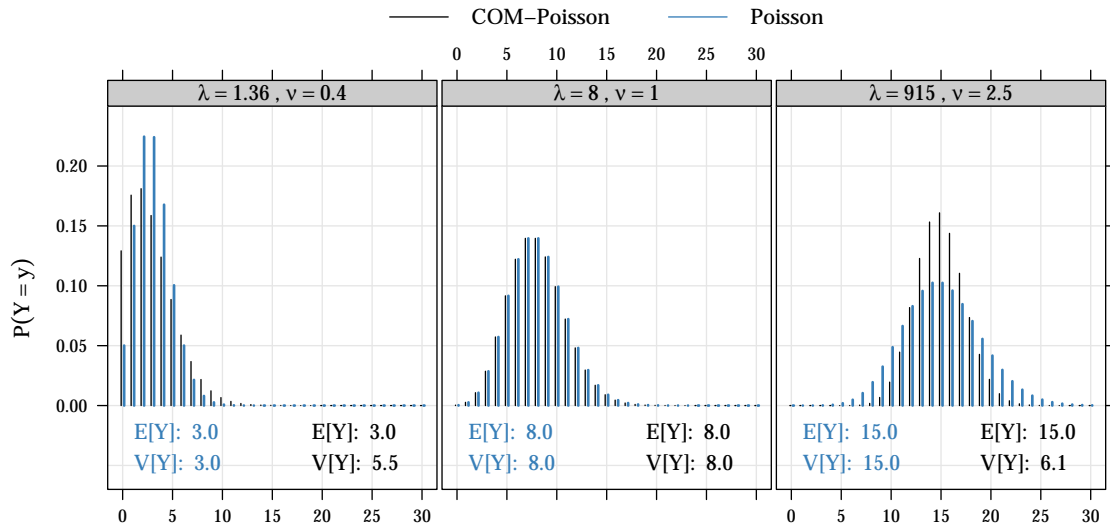


# Distribuição COM-Poisson

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

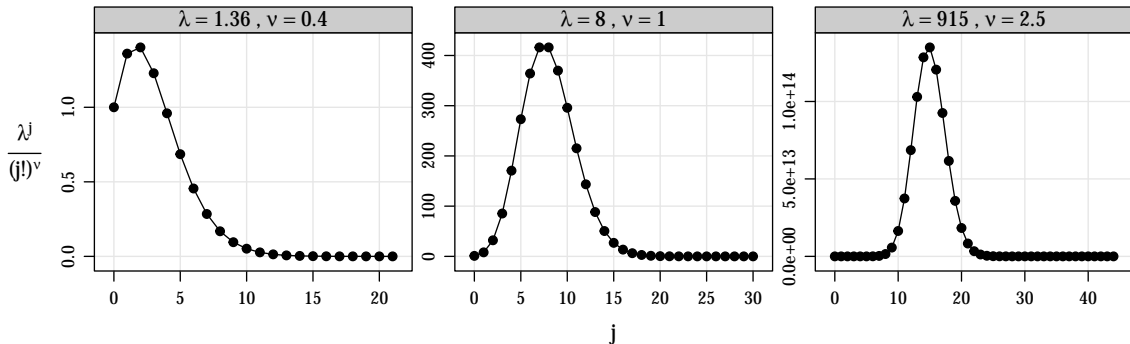
COM-Poisson ( $\lambda = 0.5, \nu = 0$ )





# 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;

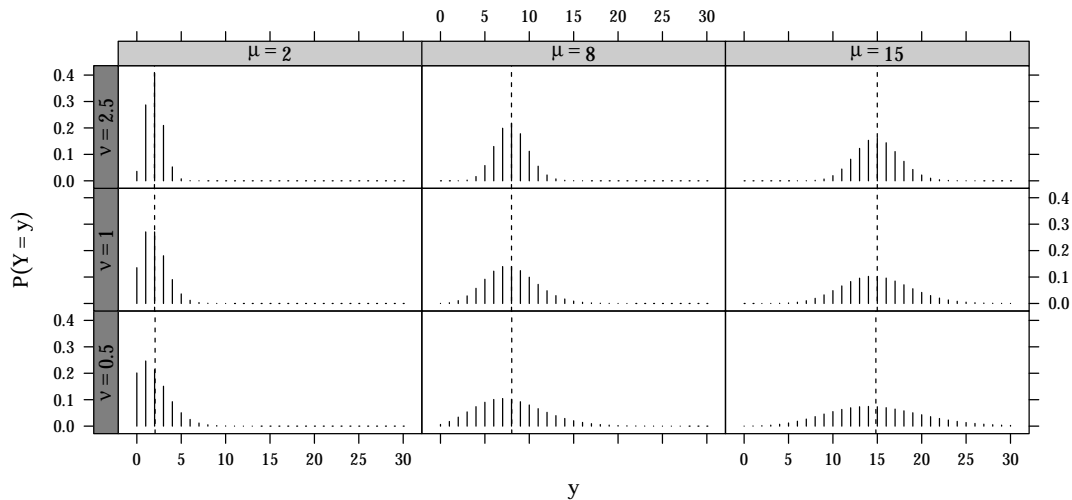
$$\blacktriangleright E(Y) = \sum_{y=0}^{\infty} y \cdot p(y)$$

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

Aproximação proposta por [?], boa aproximação para  $\nu \leq 1$  ou  $\lambda > 10^\nu$

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

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





# Modelo de Regressão COM-Poisson

- Incorporando covariáveis em  $\lambda$  da forma  $\lambda_i = \exp(X_i\beta)$ , em que  $X_i$  é o vetor de covariáveis do  $i$ -ésimo indivíduo e  $\beta$  o vetor de parâmetros.

## Função de verossimilhança

$$\begin{aligned} L(\beta, \nu; \underline{y}) &= \prod_i^n \left( \frac{\lambda_i^{y_i}}{(y_i!)^\nu} Z(\lambda_i, \nu)^{-1} \right) \\ &= \lambda_i^{\sum_i^n y_i} \prod_i^n \frac{Z(\lambda_i, \nu)^{-1}}{(y_i!)^\nu} \end{aligned}$$

## Função de log-verossimilhança

$$\begin{aligned} \ell(\beta, \nu, \underline{y}) &= \log \left( \lambda_i^{\sum_i^n y_i} \prod_i^n \frac{Z(\lambda_i, \nu)^{-1}}{(y_i!)^\nu} \right) \\ &= \sum_i^n y_i \log(\lambda_i) - \nu \sum_i^n \log(y_i!) - \sum_i^n \log(Z(\lambda_i, \nu)) \end{aligned}$$

# Estudos de caso

*Vignette* [compoisson.html](#)

`capdesfo` : Número de capulhos em algodão sob efeito de desfolha (sub)

`capmosca` : Número de capulhos em algodão sob exposição à mosca branca (sub)

`ninfas` : Número de ninfas de mosca branca em plantas de soja (super)

`soja` : Número de vagens, de grãos por planta (equi e super).

## 6.3

Modelos Paramétricos Alternativos  
**Modelo Gamma-Count**

# 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 v.a.  $\tau_i$ , 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  $N_T$  como o número de eventos no intervalo  $[0, T)$ .
- ▶ Defina  $\vartheta_n = \sum_{i=1}^n \tau_i$  o tempo até o  $n$ -ésimo evento.
- ▶ A distribuição de  $\vartheta_n$  determina a distribuição de  $N_T$ , mas é baseada em covolução.
- ▶ São distribuições fechadas para covolução: normal, Poisson, binomial e gama.
- ▶ Destas, apenas a gama é contínua e positiva.

# Duração dependência

- ▶ Denote  $E(\tau) = \mu$ ,  $V(\tau) = \sigma^2$  e  $CV(\tau) = \sigma/\mu$ .
- ▶ Defina  $\lambda(\tau) = \frac{f(\tau)}{1-F(\tau)}$  como a função de risco e assuma que é monótona.
- ▶ Existe relação entre o tipo de duração dependência e o coeficiente de variância

$$\left. \frac{d\lambda(t)}{dt} \begin{matrix} < \\ = \\ > \end{matrix} \right\} 0 \Rightarrow v = \left. \begin{matrix} < \\ = \\ > \end{matrix} \right\} 1 \quad (4)$$

# Relação entre número de eventos e intervalo entre eventos

- ▶ Intervalos entre tempo  $\tau \sim \text{Gama}(\alpha, \beta)$ ,

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

$$E(\tau) = \frac{\alpha}{\beta}, \quad V(\tau) = \frac{\alpha}{\beta^2}.$$

- ▶ Tempo até o  $n$ -ésimo evento

$$\vartheta_n = \tau_1 + \cdots + \tau_n \sim \text{Gama}(n\alpha, \beta),$$

$$f_n(\vartheta, \alpha, \beta) = \frac{\beta^{n\alpha}}{\Gamma(n\alpha)} \cdot \vartheta^{n\alpha-1} \cdot \exp\{-\beta\vartheta\},$$

$$E(\vartheta) = \frac{n\alpha}{\beta}, \quad V(\vartheta) = \frac{n\alpha}{\beta^2}.$$

# Relação entre número de eventos e intervalo entre eventos

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

$$F_n(T) = \Pr(\vartheta_n \leq T) = \int_0^T \frac{\beta^{n\alpha}}{\Gamma(n\alpha)} \cdot t^{n\alpha-1} \cdot \exp\{-\beta t\} dt.$$

- ▶ Seja  $[0, T)$  um intervalo e  $N_T$  a v.a. número de eventos neste intervalo.
- ▶ Segue que  $N_T < n$  se e somente se  $\vartheta_n \geq T$ . Assim

$$\Pr(N_T < n) = \Pr(\vartheta_n \geq T) = 1 - F_n(T);$$

- ▶ Já que  $\Pr(N_T = n) = \Pr(N_T < n + 1) - \Pr(N_T < n)$ , então

$$\Pr(N_T = n) = F_n(T) - F_{n+1}(T).$$



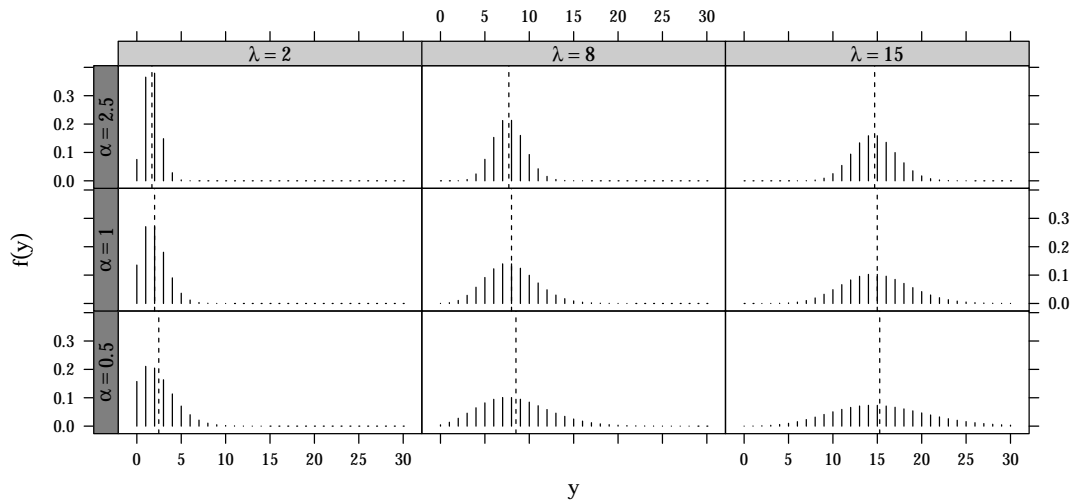
# Relação entre número de eventos e intervalo entre eventos

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

$$F_n(T) = G(n\alpha, \beta T) = \int_0^T \frac{\beta^{n\alpha}}{\Gamma(n\alpha)} t^{n\alpha-1} \cdot \exp\{-\beta t\} dt. \quad (5)$$

- ▶ Assim

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



# Parametrização para modelo de regressão

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

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

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

$$N(T) \sim \text{Normal} \left( \frac{\beta}{\alpha}, \frac{\beta}{\alpha^2} \right).$$

# Parametrização para modelo de regressão

- Considere que

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

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

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

- 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  $E(N_i|x_i) = [E(\tau_i|x_i)]^{-1}$ .

# Função de log-verossimilhança

Considerando uma amostra aleatória  $y_i, i = 1, 2, \dots, n$ , a verossimilhança é

$$L(y; \alpha, \beta) = \prod_{i=1}^n (G(y_i \alpha, \beta) - G((y_i + 1) \alpha, \beta)) \quad (6)$$

e a função de log-verossimilhança é

$$\ell(y; \alpha, \beta) = \sum_{i=1}^n \ln (G(y_i \alpha, \beta) - G((y_i + 1) \alpha, \beta)) \quad (7)$$

# Implementação da log-verossimilhança

```
library(MRDCr)
llgcnt

## function(params, y, X, offset = NULL) {
##   # params: vetor de parâmetros;
##   #   params[1]: parâmetro de dispersão (alpha);
##   #   params[-1]: parâmetro de locação (lambda);
##   # y: variável resposta (contagem);
##   # X: matriz do modelo linear;
##   # offset: tamanho do domínio onde y foi medido (exposição);
##   #-----
##   if (is.null(offset)) {
##     offset <- 1L
##   }
##   alpha <- exp(params[1])
##   eXb <- exp(X %*% params[-1])
##   alpha * eXb <- alpha * eXb
```

# Estudos de caso

[gamma\\_count.html](#)

`soja` : Número de vagens, de grãos e de grãos por vagem.

`capdesfo` : Número de capulhos produzidos em algodão.

`nematoide` : Número de nematoides em raízes de linhagens de feijoeiro.







`cambras` : Gols do Campeonato Brasileiro de 2010.

7

# Modelos com Efeitos Aleatórios



# Referências

-  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.
-  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).

# Referências



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.



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.