

# Introdução à Análise de Dados de Contagem com R

Eduardo E. R. Junior

Curso de Estatística - UFPR  
**MEETUP USER-SP**

08 de abril de 2016  
[edujrrib@gmail.com](mailto:edujrrib@gmail.com)

# Introdução à Análise de Dados de Contagem com R

Eduardo E. R. Junior

Curso de Estatística - UFPR  
**MEETUP USER-SP**

08 de abril de 2016  
[edujrrib@gmail.com](mailto:edujrrib@gmail.com)

# Disponibilização



<https://github.com/jreduardo/meetup-iadcr>

Introdução à **Análise de Dados de Contagem com R** - **meetup-iadcr**

# Sumário

1. Introdução
2. Modelo Poisson
3. Modelos Alternativos

1

# Introdução

Representam o número de ocorrências de um evento de interesse em um domínio específico.

Se  $Y$  é uma v.a de contagem,  $y \in \mathbb{Z}_+$ , ou seja,  $y = 0, 1, 2, \dots$

Exemplos:

- ▶ Número de filhos por casal;
- ▶ Número de indivíduos infectados por uma doença;
- ▶ Número de insetos mortos após  $k$  dias da aplicação de inseticida;
- ▶ ...

2

# Modelo Poisson

# Distribuição Poisson

## Densidade de probabilidade

$$\Pr(Y = y \mid \lambda) = \frac{\lambda^y}{y!e^\lambda}, \quad y \in \mathbb{Z}_+ \quad (1)$$

## Propriedades

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



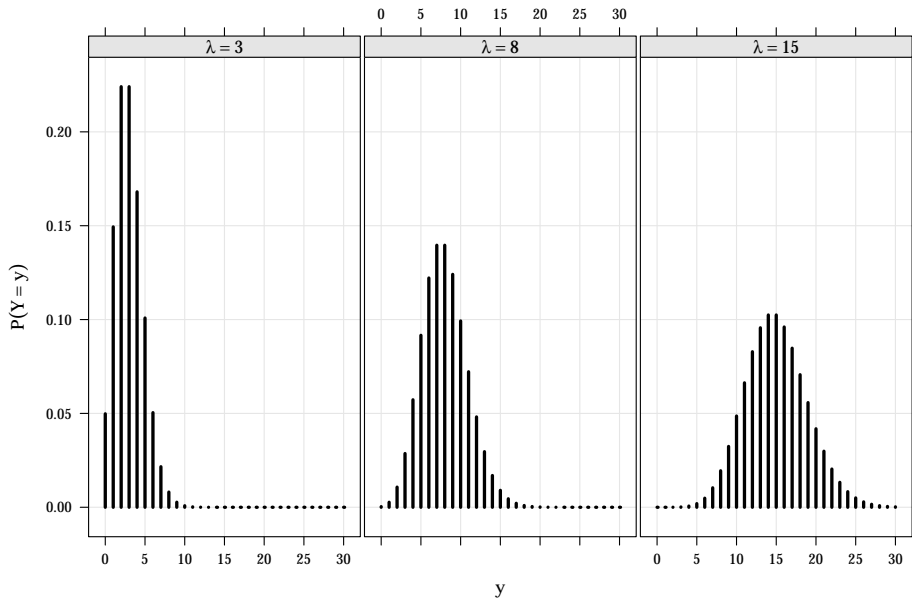


Figure 1: Probabilidades para modelos Poisson

# Regressão Poisson

$$Y_i | X_i, t_i \sim \text{Poisson}(\mu_i = \lambda_i t_i) \\ g(\lambda_i t_i) = \eta_i = X_i \beta$$

Sendo  $g$  uma função monótona que,

- ▶ Linearize a relação entre  $\mu$  e  $\eta$ ; e
- ▶ Confira valores válidos para  $\mu$  (pertencente ao espaço paramétrico)

As duas funções de ligação mais comuns são  $\log \mu$  e  $\sqrt{\mu}$

## Modelo log-linear Poisson

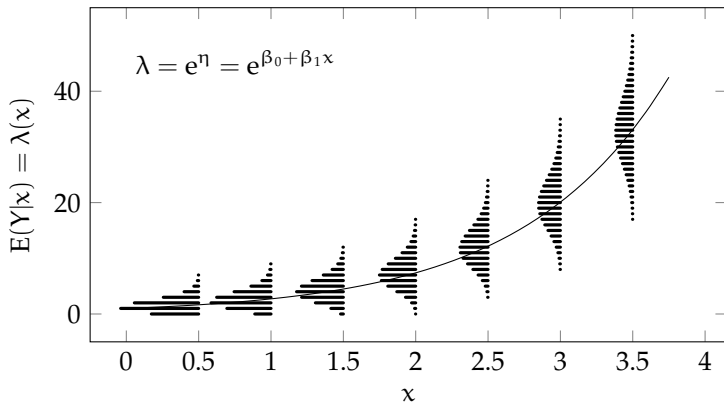


Figure 2: Representação esquemática de um modelo de regressão Poisson

## Estimação

- ▶ Via maximização da log-Verossimilhança
$$\ell(\beta | X) = yX\beta - X\beta - \log(X!)$$
- ▶ Via Mínimos Quadrados Ponderados Iterativamente  
Utiliza as propriedades da família exponencial
$$\beta^{m+1} = (X^t W^m X)^{-1} X^t W^m z^m$$
- ▶ No R

```
model <- glm(y ~ preditor, family = poisson)
```

# Estudos de caso

- ▶ [anomalias.html](#) - Caso equidisperso

3

# Modelos Alternativos

## 3.1

Modelos Alternativos

# Fuga de equidispersão

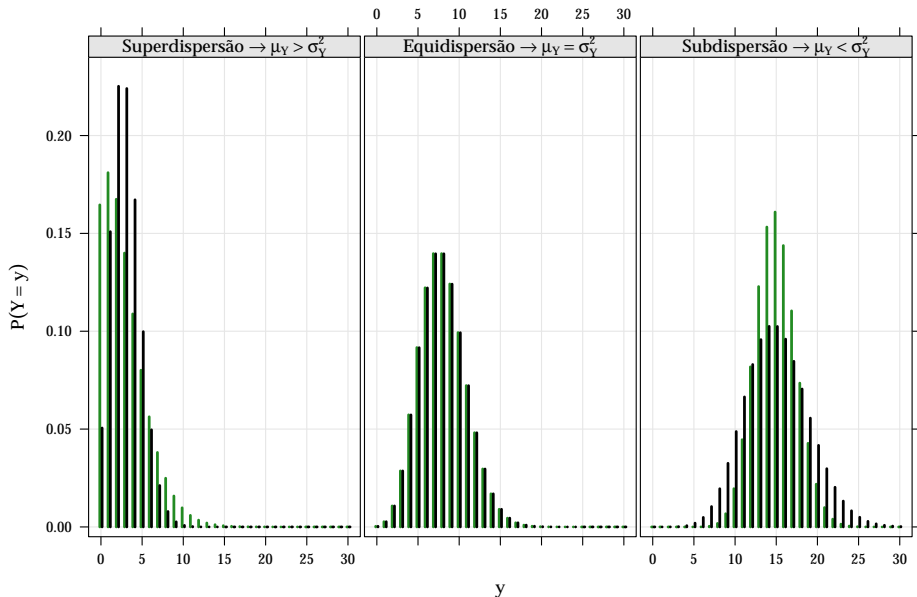


Figure 3: Probabilidades para a não verificação de equidispersão



## Modelo Binomial Negativo

Também chamado de Modelo de mistura *gamma-Poisson* (BN)

$$Y \sim \text{Poisson}(\lambda)$$

$$\lambda \sim \text{Gamma}(\mu, k = \phi\mu)$$

- ▶ Distribuição marginal Binomial Negativa ( $\pi = 1/(1 + \phi)$ ,  $k$ )
- ▶ Acomoda somente superdispersão
- ▶ Função de ligação canônica problemática No R

```
## Support Functions and Datasets for Venables and Ripley's MASS  
library(MASS)  
model <- glm.nb(y ~ predictor)
```

# Modelos de Quasi-Verossimilhança

- ▶ Estimação baseada em momentos

$$E(Y|X) = \exp X\beta$$

$$V(Y|X) = \phi \exp X\beta$$

- ▶ Interpretação de  $\phi$

$$0 < \phi < 1 \rightarrow \text{subdisperso}$$

$$\phi = 1 \rightarrow \text{equidisperso}$$

$$\phi > 1 \rightarrow \text{superdisperso}$$

- ▶ Estimação de  $\phi$

$$\hat{\phi} = \sum r_{pad}^2 / (n - p)$$

- ▶ No R

```
model <- glm(y ~ preditor, family = quasipoisson)
```

# Estudos de caso

- ▶ [ninfas.html](#) - Caso superdisperso
- ▶ [capulhos.html](#) - Caso subdisperso

# Outras abordagens

## 3.2

# Modelos Alternativos

## **Excesso de zeros**

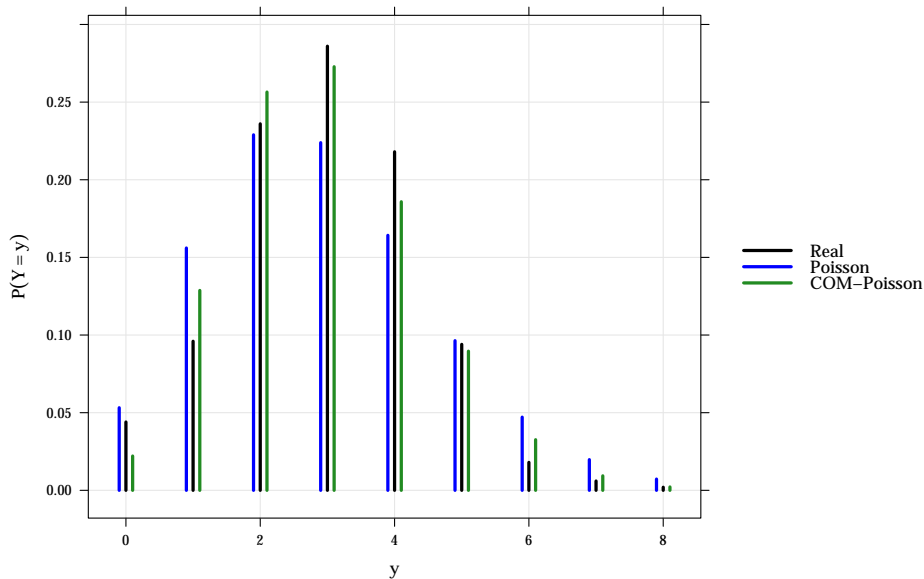


Figure 4: Contagens que apresentam excesso de zeros

## Modelos de Barreira *Hurdle models*

São chamados também de modelos condicionais ou truncados

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

Sendo  $P_Y$  uma distribuição de probabilidades associada às contagens e  $\pi$  a probabilidade associada às contagens 0.

```
## Political Science Computational Laboratory, Stanford University  
library(pscl)  
  
hurdle(resp ~ py_predictor | pi_predictor, dist = "poisson")
```

## Modelos de Inflação *Zero Inflated Models*

São chamados também de modelos de mistura

$$\Pr(Y = y) = \begin{cases} \pi + (1 - \pi)P_Y(0) & \text{se } y = 0, \\ (1 - \pi)P_Y(y) & \text{se } y = 1, 2, \dots \end{cases} \quad (3)$$

Sendo  $P_Y$  uma distribuição de probabilidades associada às contagens e  $\pi$  a probabilidade associada às contagens 0.

```
## Political Science Computational Laboratory, Stanford University  
library(pscl)  
  
zeroinfl(resp ~ py_predictor | pi_predictor, dist = "poisson")
```



## 3.3

# Modelos Alternativos

## Outras abordagens

# COM-Poisson

## Densidade de probabilidade

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \quad y \in \mathbb{Z}_+ \quad (4)$$

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

## Propriedades

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

## Casos particulares

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

# Modelo *Gamma-Count*

## Densidade de probabilidade

$$\Pr(Y = y \mid \alpha, \beta) = G(y, n\alpha, \beta T) - G(y, n\alpha + \alpha, \beta T), \quad y \in \mathbb{Z}_+ \quad (5)$$

onde  $G(y, \theta_1, \theta_2)$  é a distribuição de densidade acumulada até o ponto  $y$  da Gamma de parâmetros  $\theta_1$  e  $\theta_2$ ; e  $\alpha > 0$  e  $\beta > 0$

## Propriedades

- Generaliza a relação entre Poisson e Exponencial, considerando que o tempo entre eventos agora pode ser um Gamma com parâmetros estimados.

## Casos particulares

- Distribuição Poisson, quando  $\alpha = 1$

# Efeitos Aleatórios

$$Y | \mathbf{b} \sim f_*(\mu, \phi)$$

$$g(\mu) = \beta_0 + \mathbf{b}_i$$

$$\mathbf{b}_i \sim D(\Sigma)$$

$$\Pr(Y = y) = \int_{D_D} [Y | X, \mathbf{b}_i][\mathbf{b}_i] d\mathbf{b}_i \quad (6)$$

Onde  $D$  é a distribuição associada aos efeitos aleatórios e  $D_D$  é o suporte da distribuição.

- ▶ Necessários, métodos de integração numérica
  - ▶ Aproximação de Laplace
  - ▶ Quadratura Gaussiana
  - ▶ Monte Carlo (e.g. MCMC)

## Entre outros

- ▶ Modelos para dados censurados
- ▶ Modelos para respostas correlacionadas
- ▶ ...

# Participe!



**Minicurso:** Modelos de Regressão para Dados de Contagem com R - MRDCr

**Autores:** Walmes M. Zeviani, Eduardo E. R. Junior, Cesar A. Taconelli

# Obrigado!