

# 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

9 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. Modelos Lineares Generalizados
3. Modelo de Regressão Poisson
4. Modelo de Quase-Verossimilhança
5. Modelos Paramétricos Alternativos
  - 5.1 Modelo Binomial Negativa
  - 5.2 Modelo Poisson-Generalizada
  - 5.3 Modelo COM-Poisson
  - 5.4 Modelo Gamma-Count
6. Modelos para Excesso de Zeros
  - 6.1 Modelos de Barreira *Hurdle*
  - 6.2 Modelos de Mistura (*Zero Inflated*)
7. Modelos com Efeitos Aleatórios

1

# Introdução

2

# Modelos Lineares Generalizados

3

# Modelo de Regressão Poisson

4

# Modelo de Quase-Verossimilhança

5

# Modelos Paramétricos Alternativos



## 5.1

Modelos Paramétricos Alternativos

# **Modelo Binomial Negativa**

## 5.2

Modelos Paramétricos Alternativos  
**Modelo Poisson-Generalizada**

## 5.3

Modelos Paramétricos Alternativos  
**Modelo COM-Poisson**

# Distribuição COM-Poisson I

- ▶ Nome COM-Poisson, advém de seus autores **C**onway e **M**axwell (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.

## Razão de probabilidades consecutivas

### ▶ Distribuição Poisson

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

### ▶ Distribuição COM-Poisson

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

# Distribuição COM-Poisson II

## Densidade de probabilidade

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

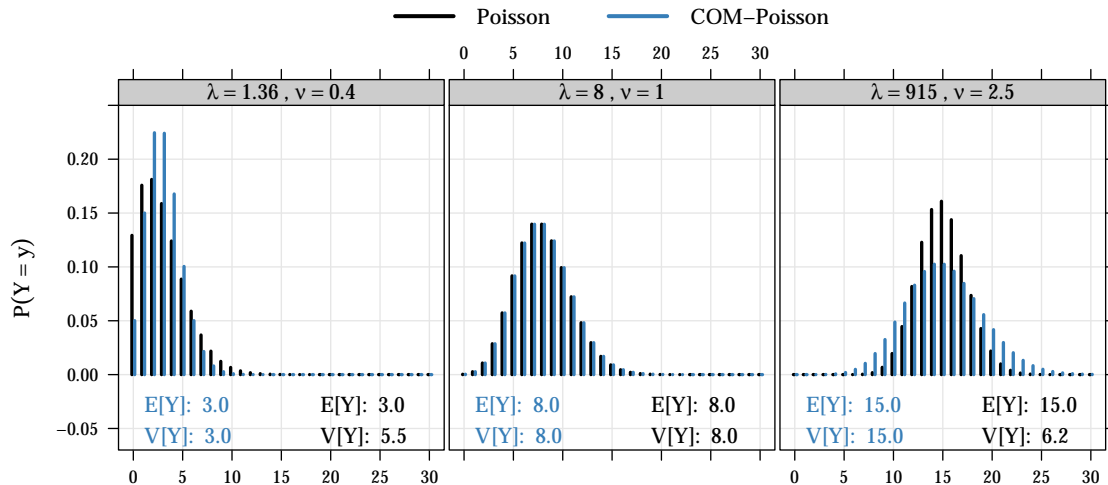
## Propriedades

- ▶  $\frac{P(Y=y-1)}{P(Y=y)} = \frac{y^\nu}{\lambda}$
- ▶  $E(Y) \approx \lambda^{\frac{1}{\nu}} - \frac{\nu-1}{2\nu}$
- ▶  $V(Y) \approx \frac{1}{\nu} E(Y)$

## Casos particulares

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

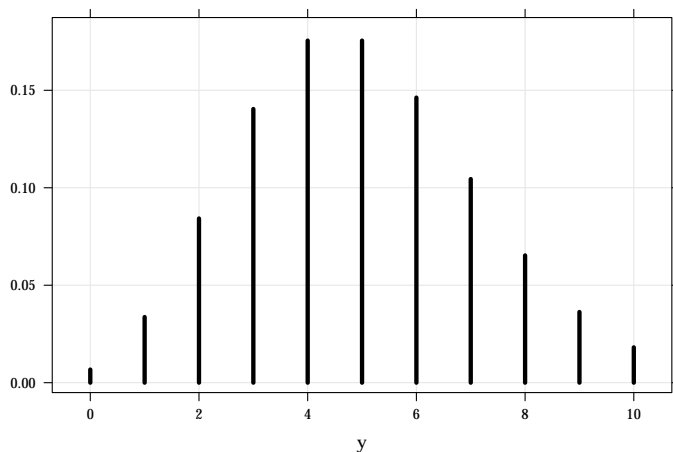
# Distribuição COM-Poisson III



# Casos Particulares

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

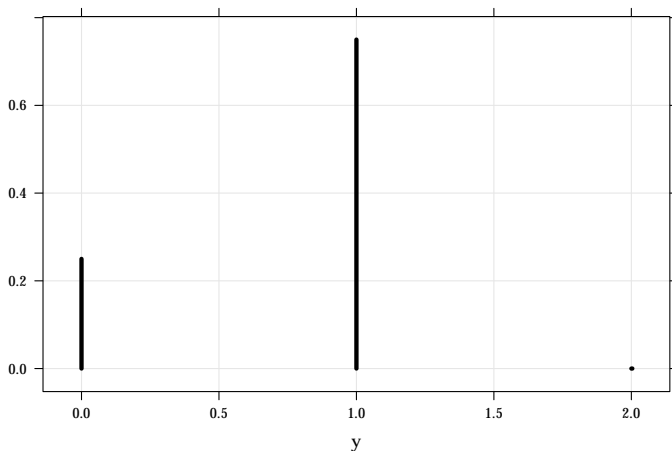
COM-Poisson ( $\lambda = 5, \nu = 1$ )



# Casos Particulares

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

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

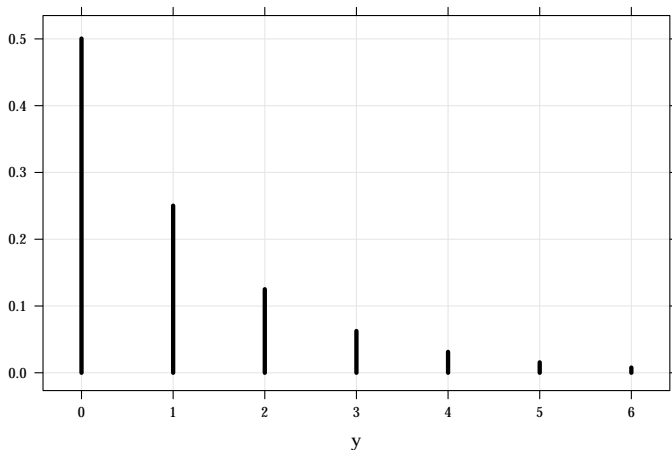




# Casos Particulares

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

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



# 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

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

## Função de log-verossimilhança

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

# Estudos de caso

*Vignette* [compoisson.html](#)

`capdesfo` : número de capulhos sob efeito de desfolha (sub)

`capmosca` : número de capulhos sob exposição à mosca branca (sub)

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

## 5.4

Modelos Paramétricos Alternativos  
**Modelo Gamma-Count**

6

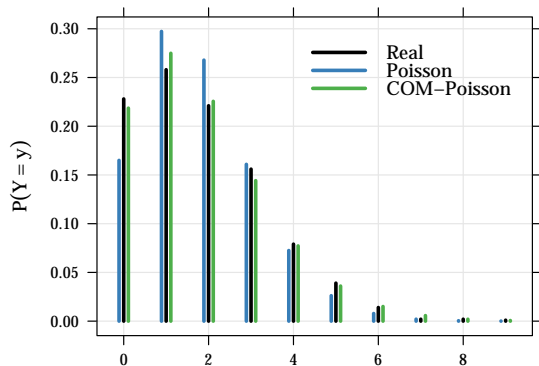
# Modelos para Excesso de Zeros

# Excesso de Zeros I

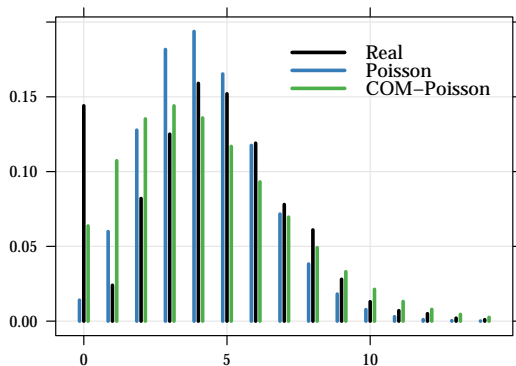
- ▶ Casos em que a proporção de valores nulos na amostra é superior àquela estimada por um modelo de contagem. No caso Poisson  $e^{-\lambda}$
- ▶ Geralmente contagens com um número excessivo de valores nulos apresentam superdispersão (ocasionada pelo excesso de zeros).
- ▶ Os modelos mais flexíveis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

# Excesso de Zeros II

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



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



# Gerador de excesso de zeros I

- ▶ 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 pode ser caracterizada como:
  - ▶ **zeros estruturais:** Ausência de determinada característica da população.
  - ▶ **zeros amostrais:** Ocorrem segundo um processo gerador de contagens (e.g Processo Poisson)



## Gerador de excesso de zeros II

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

## 6.1

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

# Modelo Hurdle I

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

# Modelo Hurdle II

## 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

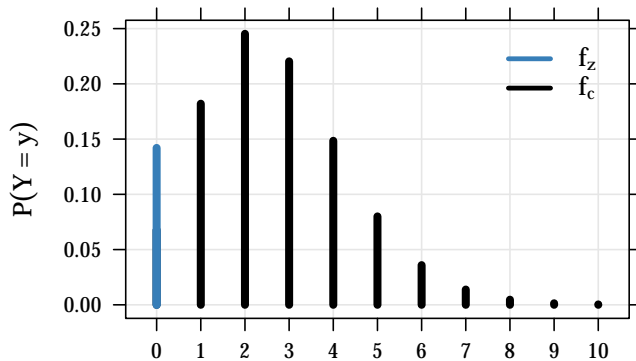
$$E(Y) = \frac{E(Y^*)(1 - f_z(0))}{1 - f_c(Y = 0)}$$

### Variância

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

# Modelos de barreira

- ▶  $f_z$  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* I

- ▶ 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* II

## 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 0 se  $y = 0$  e  $\underline{\theta}$  o vetor de parâmetros do modelo ( $\beta$ ,  $\gamma$  e  $\phi$ , se houver).

# Modelos *Hurdle* no R I

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 II

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 ~ preditor, family = zapoisson)

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

# Estudos de caso

*Vignette* [v07\\_hurdle.html](#)

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

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

## 6.2

Modelos para Excesso de Zeros

**Modelos de Mistura (*Zero Inflated*)**

# Modelo *Zero Inflated* I

- ▶ 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 II

## 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

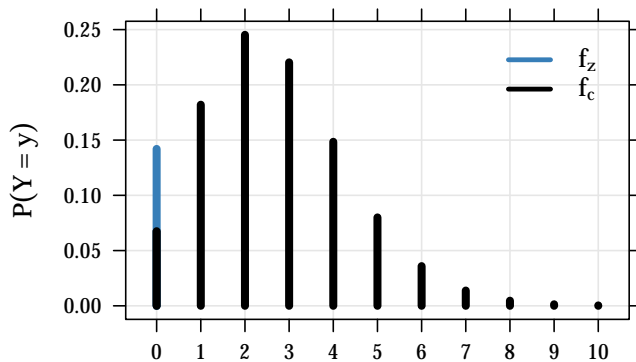
$$E(Y) = (1 - f_z(0))E(Y^*)$$

### Variância

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

# Modelos de mistura

- ▶  $f_z$  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.
- ▶ 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* I

- ▶ 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* II

## 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 I

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 = zapoisson)
```

# Estudos de caso

*Vignette* `v07_zeroinfl.html`






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

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

7

# Modelos com Efeitos Aleatórios

# Referências I

-  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](http://dx.doi.org/10.18637/jss.v027.i08)
-  Winkelmann, R. (2008). *Econometric analysis of count data* (5th Ed.). Springer Science & Business Media.