

**MBA
USP
ESALQ**

Supervised Machine Learning: Modelos para Dados de Contagem

Prof. Dr. Luiz Paulo Fávero

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização.

Lei nº 9610/98



MODELOS PARA DADOS DE CONTAGEM

- Fundamentação teórica, conceitos e aplicações;
- Especificação dos modelos;
- Modelos dos tipos Poisson e binomial negativo;
- Estimação dos parâmetros por máxima verossimilhança;
- Identificação do fenômeno de superdispersão nos dados;
- Modelos inflacionados de zeros;
- Estimações em Python.

Modelos Lineares Generalizados (GLM)

$$Y = f(X_1, X_2, X_3, \dots, X_k)$$

Modelo de Regressão	Característica da Variável Dependente	Distribuição
Linear	Quantitativa	Normal
Com Transformação de Box-Cox	Quantitativa	Normal Após a Transformação
Logística Binária	Qualitativa com 2 Categorias (<i>Dummy</i>)	Bernoulli
Logística Multinomial	Qualitativa M ($M > 2$) Categorias	Binomial
Poisson	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson
Binomial Negativo	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson-Gama

Siméon Denis Poisson



(1781-1840)

Modelos para Dados de Contagem

Os modelos de regressão Poisson e binomial negativo fazem parte do que é conhecido por modelos de regressão para dados de contagem, e têm por objetivo analisar o comportamento, em função de variáveis preditoras, de determinada variável dependente que se apresenta na forma quantitativa, com valores discretos e não negativos. Deve ser definida também a exposição (unidade temporal, espacial, social, etc.).

Modelos para Dados de Contagem: Exemplos e Aplicações

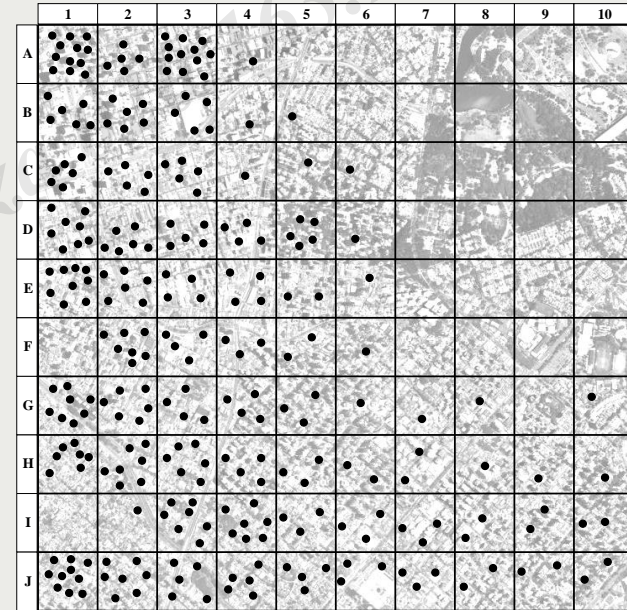
- Avaliação da quantidade de vezes que um grupo de pacientes idosos vai ao médico por ano, em função da idade de cada um deles, do sexo e das características dos seus planos de saúde.
- Estudo sobre a quantidade de ofertas públicas de ações que são realizadas em uma amostra de países emergentes em determinado ano, com base em seus desempenhos econômicos, como inflação, taxa de juros, produto interno bruto e taxa de investimento estrangeiro.

Note que a quantidade de visitas ao médico ou a quantidade de IPOs são as variáveis dependentes nos dois casos, sendo representadas por dados quantitativos, discretos, não negativos e com exposição anual. Ou seja, são dados de contagem!

Modelos para Dados de Contagem: Exemplos e Aplicações



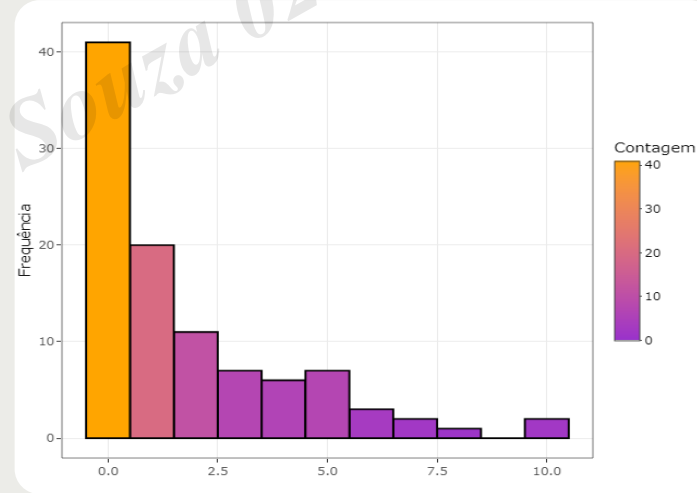
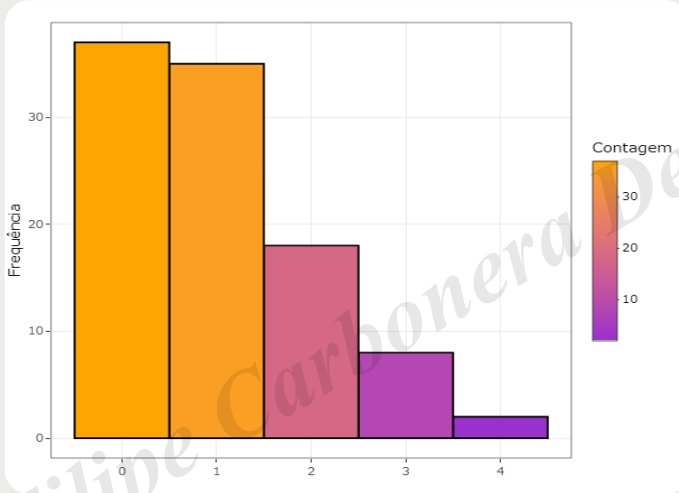
Ecologia



Mercado Imobiliário

Distribuições Poisson e Binomial Negativa

$$\ln(\hat{Y}_i) = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \dots + \beta_k.X_{ki}$$

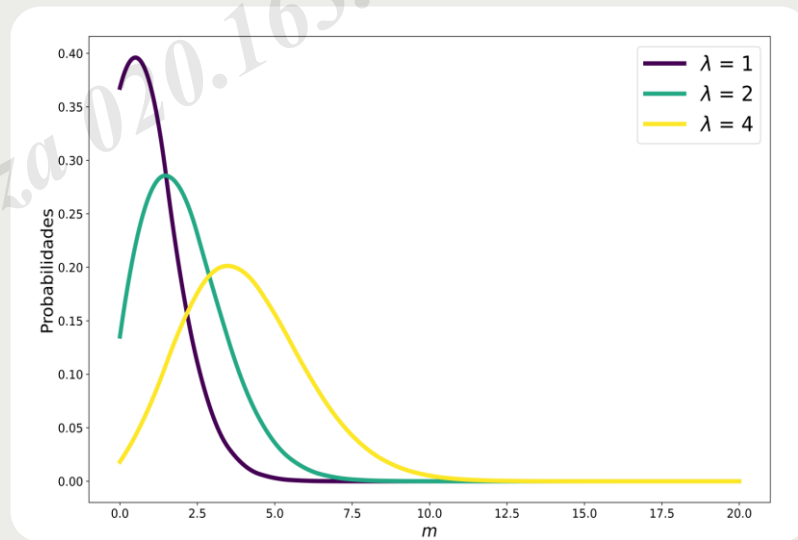


A Distribuição Poisson

Determinada observação i ($i = 1, 2, \dots, n$, em que n é o tamanho da amostra) possui a seguinte probabilidade de ocorrência de uma contagem m em uma determinada exposição (período, área, região, entre outros exemplos):

$$p(Y_i = m) = \frac{e^{-\lambda_i} \cdot \lambda_i^m}{m!}$$

em que λ é o número esperado de ocorrências ou a taxa média estimada de incidência do fenômeno em estudo para uma dada exposição.



A Distribuição Poisson e o Modelo Poisson

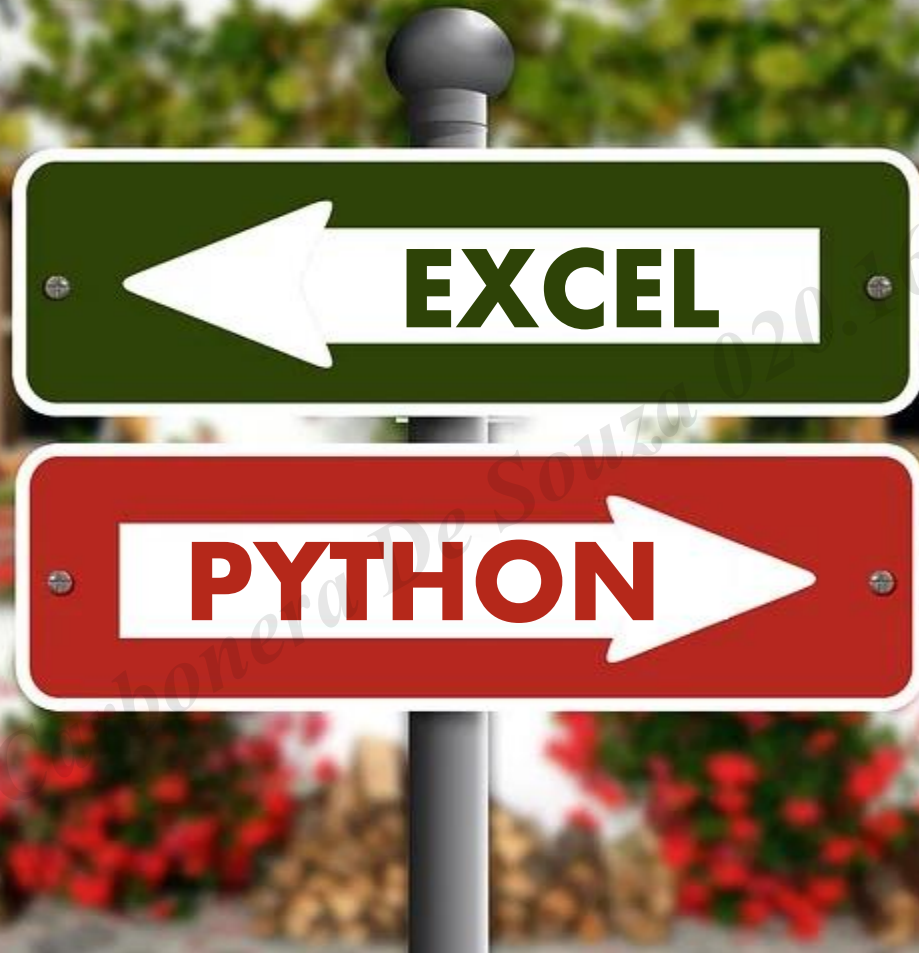
Média:
$$E(Y) = \sum_{m=0}^{\infty} m \cdot \frac{e^{-\lambda} \cdot \lambda^m}{m!} = \lambda \cdot \sum_{m=1}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{m-1}}{(m-1)!} = \lambda \cdot 1 = \lambda$$

Variância:
$$\begin{aligned} Var(Y) &= \sum_{m=0}^{\infty} m \cdot \frac{e^{-\lambda} \cdot \lambda^m}{m!} \cdot (m - \lambda)^2 = \sum_{m=0}^{\infty} m \cdot \frac{e^{-\lambda} \cdot \lambda^m}{m!} \cdot (m^2 - 2 \cdot m \cdot \lambda + \\ &\lambda^2) = \lambda^2 \cdot \sum_{m=2}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{m-2}}{(m-2)!} + \lambda \cdot \sum_{m=1}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{m-1}}{(m-1)!} - \lambda^2 = \lambda \end{aligned}$$

O Modelo Poisson:

$$\ln(\hat{Y}_i) = \ln(\lambda_{poisson_i}) = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki}$$





Teste de Superdispersão

$$Y_i^* = \frac{\left[\left(Y_i - \lambda_{poisson_i} \right)^2 - Y_i \right]}{\lambda_{poisson_i}}$$

$$Y_i^* = \beta \cdot \lambda_{poisson_i} \quad (\text{Estimação OLS sem intercepto})$$

Cameron e Trivedi (1990) salientam que, se ocorrer o fenômeno da superdispersão nos dados, o parâmetro estimado β deste **modelo auxiliar sem intercepto** será estatisticamente diferente de zero, a determinado nível de significância (5%, usualmente).

CAMERON, A. C.; TRIVEDI, P. K. Regression-based tests for overdispersion in the Poisson model. **Journal of Econometrics**, v. 46, n. 3, p. 347-364, 1990.

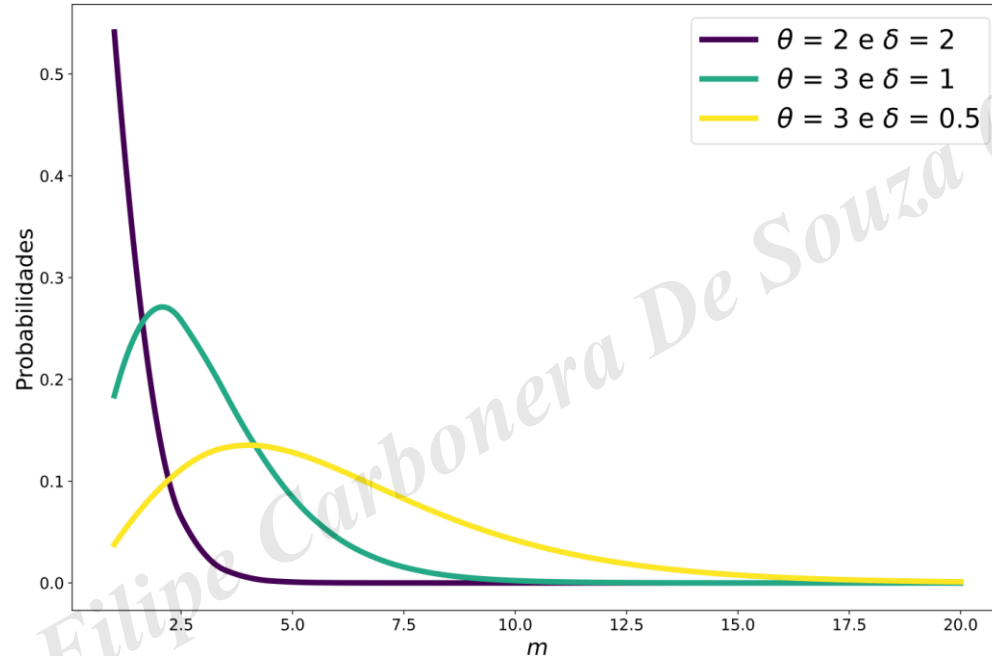
A Distribuição Poisson-Gama ou Binomial Negativa

Para determinada observação i ($i = 1, 2, \dots, n$, em que n é o tamanho da amostra), a função da distribuição de probabilidade da variável dependente Y será dada por:

$$p(Y_i = m) = \frac{\delta^\theta \cdot m_i^{\theta-1} \cdot e^{-m_i \cdot \delta}}{(\theta-1)!}$$

em que θ é chamado de parâmetro de forma ($\theta > 0$) e δ é chamado de parâmetro de taxa de decaimento ($\delta > 0$).

A Distribuição Poisson-Gama ou Binomial Negativa



- **Média:**

$$E(Y) = \lambda_{bneg}$$

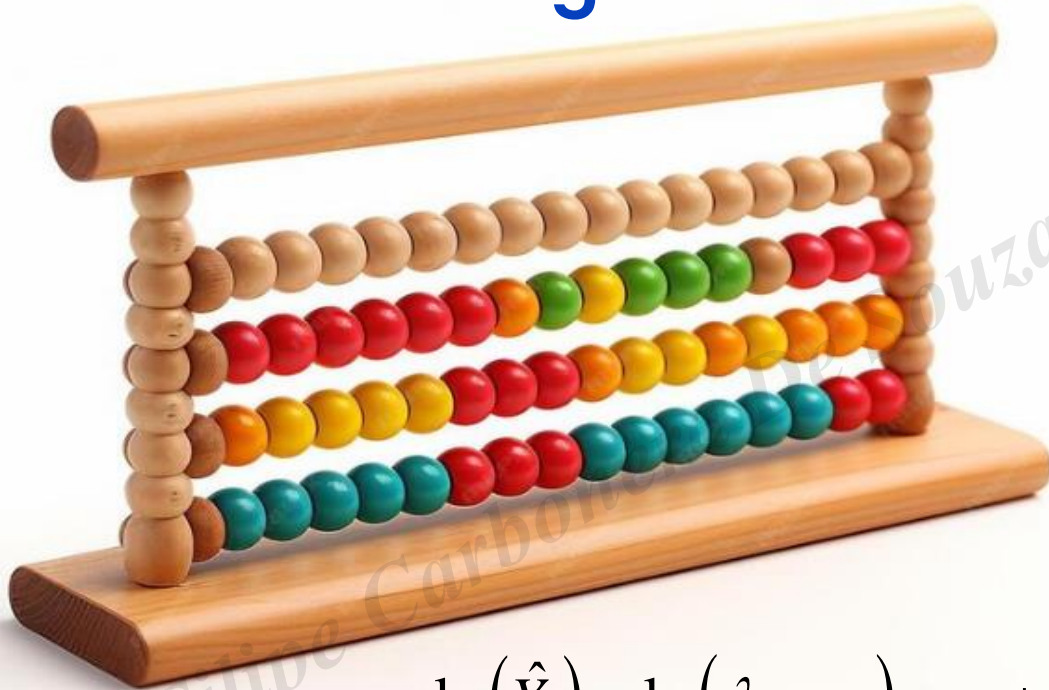
- **Variância:**

$$Var(Y) = \lambda_{bneg} + \phi \cdot (\lambda_{bneg})^2$$

$$\phi = \frac{1}{\theta}$$

Modelos NB2

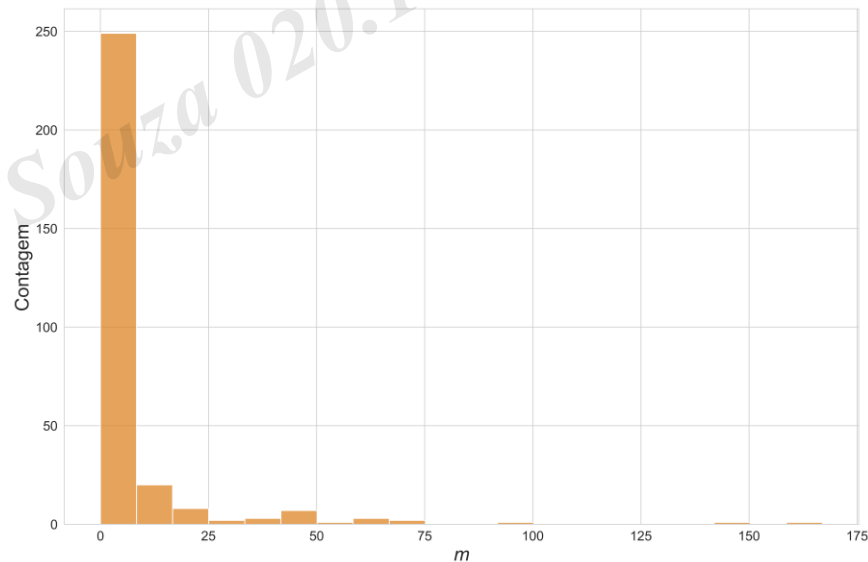
O Modelo Poisson-Gama ou Binomial Negativo



$$\ln(\hat{Y}_i) = \ln(\lambda_{bneg_i}) = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \dots + \beta_k.X_{ki}$$



Modelos Inflacionados de Zeros



Modelos Inflacionados de Zeros

São considerados uma combinação entre um modelo para dados de contagem e um modelo para dados binários, já que são utilizados para investigar as razões que levam a determinada quantidade de ocorrências (contagens) de um fenômeno, bem como as razões que levam (ou não) à ocorrência propriamente dita deste fenômeno, independentemente da quantidade de contagens observadas.

Enquanto um modelo Poisson inflacionado de zeros é estimado a partir da **combinação de uma distribuição Bernoulli com uma distribuição Poisson**, um modelo binomial negativo inflacionado de zeros é estimado por meio da **combinação de uma distribuição Bernoulli com uma distribuição Poisson-Gama**.

LAMBERT, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. **Technometrics**, v. 34, n. 1, p. 1-14, 1992.

Modelos Inflacionados de Zeros

A definição sobre a existência ou não de uma quantidade excessiva de zeros na variável dependente Y é elaborada por meio de um teste específico, conhecido por **teste de Vuong** (1989), que representará um importante *output* a ser analisado na estimação de modelos de regressão para dados de contagem, quando houver a suspeita de existência de inflação de zeros.

VUONG, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. **Econometrica**, v. 57, n. 2, p. 307-333, 1989.

Modelos Inflacionados de Zeros do Tipo Poisson (ZIP)

Em relação especificamente aos **modelos de regressão Poisson inflacionados de zeros**, podemos definir que, enquanto a **probabilidade p de ocorrência de nenhuma contagem** para dada observação i ($i = 1, 2, \dots, n$, em que n é o tamanho da amostra), ou seja, $p(Y_i = 0)$, é calculada levando-se em consideração a soma de um componente dicotômico com um componente de contagem e, portanto, deve-se definir a probabilidade p_{logit} de não ocorrer nenhuma contagem devido exclusivamente ao componente dicotômico, a **probabilidade p de ocorrência de determinada contagem m** ($m = 1, 2, \dots$), ou seja, $p(Y_i = m)$, segue a própria expressão da probabilidade da distribuição Poisson, multiplicada por $(1 - p_{logit})$.

Modelos Inflacionados de Zeros do Tipo Poisson (ZIP)

$$\begin{cases} p(Y_i = 0) = p_{logit_i} + (1 - p_{logit_i}) \cdot e^{-\lambda_i} \\ p(Y_i = m) = (1 - p_{logit_i}) \cdot \frac{e^{-\lambda_i} \cdot \lambda_i^m}{m!}, \quad m = 1, 2, \dots \end{cases}$$

$$p_{logit_i} = \frac{1}{1 + e^{-(\gamma + \delta_1 \cdot W_{1i} + \delta_2 \cdot W_{2i} + \dots + \delta_q \cdot W_{qi})}}$$

$$\lambda_{poisson_i} = e^{(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki})}$$

Os modelos de regressão Poisson inflacionados de zeros apresentam dois processos geradores de zeros, sendo um devido à distribuição binária (neste caso, são gerados os chamados zeros estruturais) e outro devido à distribuição Poisson (nesta situação, são gerados dados de contagem, entre os quais os chamados zeros amostrais).



Modelos Inflacionados de Zeros do Tipo Binomial Negativo (ZINB)

Já em relação aos **modelos de regressão do tipo binomial negativo inflacionados de zeros**, podemos definir que, enquanto a **probabilidade p de ocorrência de nenhuma contagem** para dada observação i , ou seja, $p(Y_i = 0)$, é também calculada levando-se em consideração a soma de um componente dicotômico com um componente de contagem, a **probabilidade p de ocorrência de determinada contagem m** ($m = 1, 2, \dots$), ou seja, $p(Y_i = m)$, segue agora a expressão da probabilidade da distribuição Poisson-Gama.

Modelos Inflacionados de Zeros do Tipo Binomial Negativo (ZINB)

$$\begin{cases} p(Y_i = 0) = p_{logit_i} + (1 - p_{logit_i}) \cdot \left(\frac{1}{1 + \theta^{-1} \cdot \lambda_{bneg_i}} \right)^\theta \\ p(Y_i = m) = (1 - p_{logit_i}) \cdot \left[\frac{\delta^\theta \cdot m_i^{\theta-1} \cdot e^{-m_i \cdot \delta}}{(\theta - 1)!} \right], \quad m = 1, 2, \dots \end{cases}$$

$$p_{logit_i} = \frac{1}{1 + e^{-(\gamma + \delta_1 \cdot W_{1i} + \delta_2 \cdot W_{2i} + \dots + \delta_q \cdot W_{qi})}}$$

$$\lambda_{bneg_i} = e^{(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki})}$$





Escolha do Modelo

Verificação	Modelo de Regressão para Dados de Contagem			
	Poisson	Binomial Negativo	Poisson Inflacionado de Zeros (ZIP)	Binomial Negativo Inflacionado de Zeros (ZINB)
Superdispersão nos Dados da Variável Dependente	Não	Sim	Não	Sim
Quantidade Excessiva de Zeros na Variável Dependente	Não	Não	Sim	Sim





MUITO OBRIGADO!

Prof. Dr. Luiz Paulo Fávero

