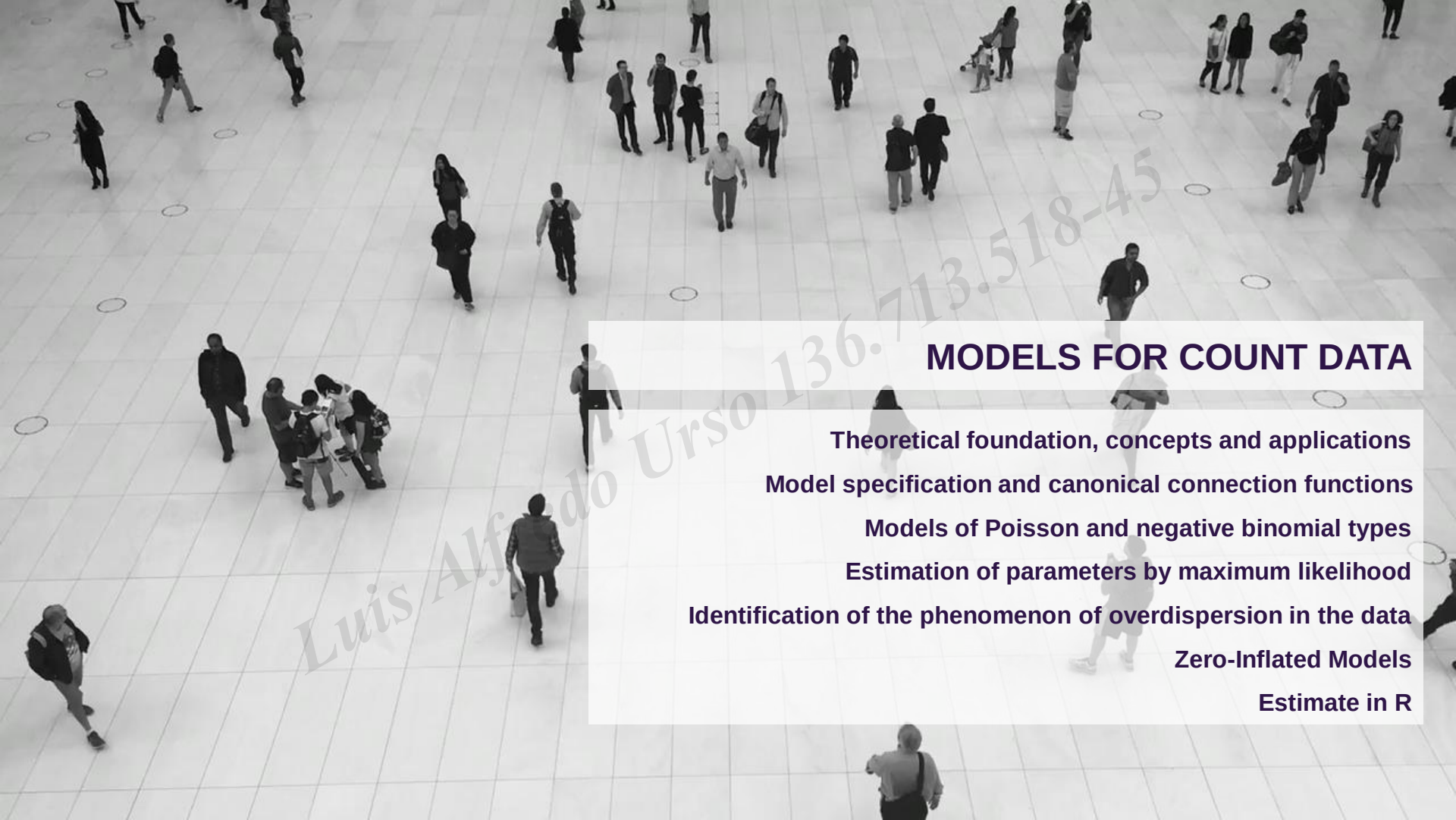
MBA USP ESALQ

SUPERVISED MACHINE LEARNING: REGRESSION MODELS FOR COUNT DATA

Prof. Dr. Luiz Paulo Fávero

*The responsibility for trustworthiness, originality, and legality of the didactic content presented is responsibility of the professor.



MODELS FOR COUNT DATA

Theoretical foundation, concepts and applications

Model specification and canonical connection functions

Models of Poisson and negative binomial types

Estimation of parameters by maximum likelihood

Identification of the phenomenon of overdispersion in the data

Zero-Inflated Models

Estimate in R

Generalized Linear Model (GLM)

$$\eta_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \ldots \beta_k.X_{ki}$$

Modelos lineares generalizados, características da variável dependente e funções de ligação canônica.

Modelo de Regressão	Característica da Variável Dependente	Distribuição	Função de Ligação Canônica (η)
Linear	Quantitativa	Normal	$\hat{Y}$
Com Transformação de Box-Cox	Quantitativa	Normal Após a Transformação	$\frac{\hat{Y}^{\lambda}-1}{\lambda}$
Logística Binária	Qualitativa com 2 Categorias (*Dummy*)	Bernoulli	$\ln\left(\frac{p}{1-p}\right)$
Logística Multinomial	Qualitativa M ($M > 2$) Categorias	Binomial	$\ln\left(\frac{p_m}{1-p_m}\right)$
Poisson	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson	$\ln\left(\lambda_{poisson}\right)$
Binomial Negativo	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson-Gama	$\ln\left(\lambda_{bneg}\right)$

Siméon Denis Poisson

(1781-1840)

Models for Count Data

Poisson and negative binomial regression models are part of what is known as regression models for count data, and aim to analyze the behavior, according to predictor variables, of a determined dependent variable that is presented in the quantitative form with discrete and non-negative values. Exposure must also be defined (temporal, spatial, social unit, among others).

Models for Count Data: Examples and Applications

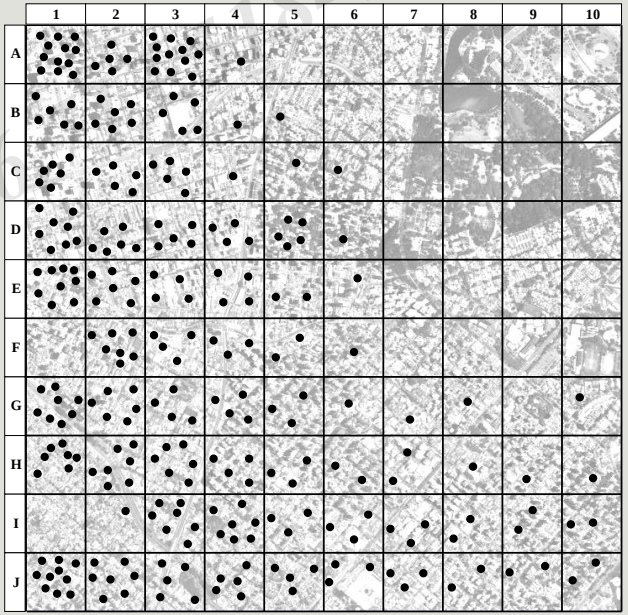
- Evaluation of the number of times that a group of elderly patients goes to the doctor per year, according to the age of each of them, the gender and the characteristics of their health plans.

- A study on the number of public offers of actions that are carried out in a sample of developed and emerging countries in a given year, based on their economic performance, such as inflation, interest rate, gross domestic product, and foreign investment rate.

  Notice that the number of going to the doctor or the number of public offers of shares are the dependent variables in both cases, being represented by quantitative data that assume discrete, non-negative values, and with annual exposure. That is, they offer count data.

Models for Count Data: Examples and Applications

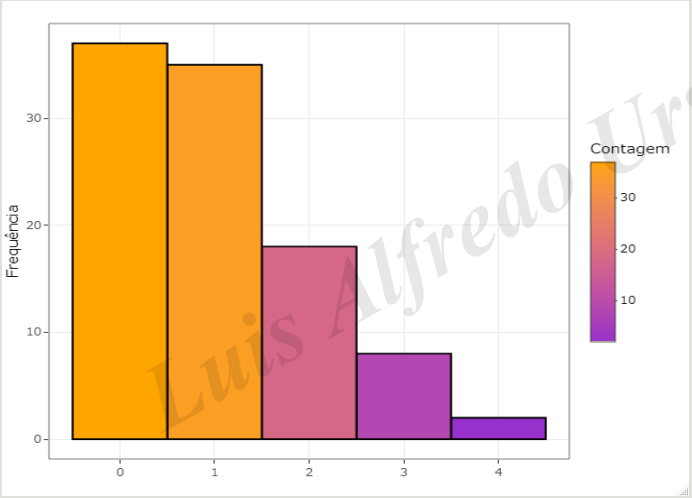
Environment

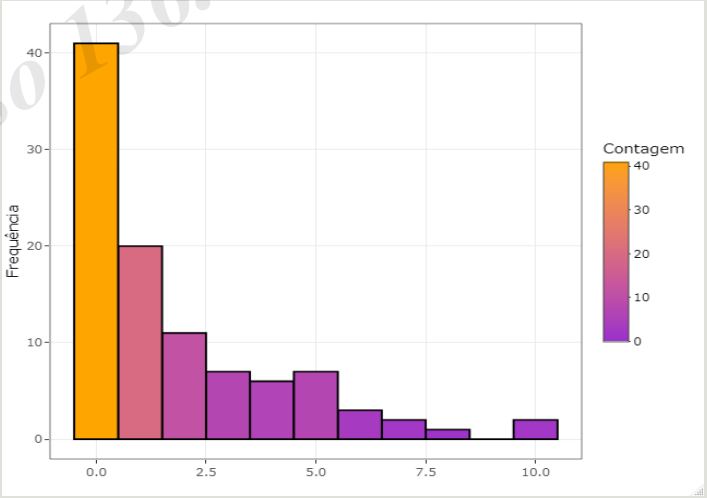


Property Market

Poisson and Binomial Negative Distributions

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



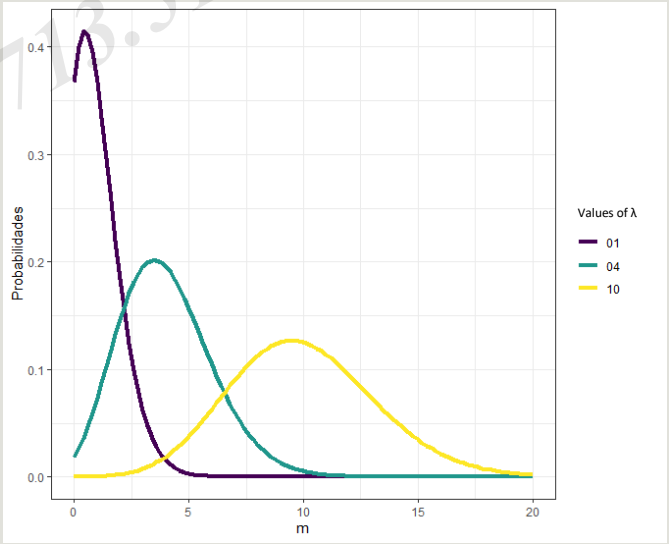


Poisson Distribution

Certain observation *i* (*i* = 1, 2, ..., *n*, in which *n* is the sample size) has the following probability of occurrence of a count *m* in a certain exposure (period, area, region, among other examples):

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

in which λ is the expected number of occurrences or the estimated average incidence rate of the phenomenon under study for a given exposure.



The Poisson Distribution and the Poisson Model

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

Variance: $$Var(Y) = \sum_{m=0}^{\infty} m.\frac{e^{-\lambda}.\lambda^{m}}{m!}.(m-\lambda)^2 = \sum_{m=0}^{\infty} m.\frac{e^{-\lambda}.\lambda^{m}}{m!}.(m^2 - 2.m.\lambda + \lambda^2)$$ $$\lambda^2.\sum_{m=2}^{\infty}\frac{e^{-\lambda}.\lambda^{m-2}}{(m-2!)} + \lambda.\sum_{m=1}^{\infty}\frac{e^{-\lambda}.\lambda^{m-1}}{(m-1!)} - \lambda^2 = \lambda$$

General Model:

$$\ln\left(\hat{Y}_i\right) = \ln\left(\lambda_{poisson_i}\right) = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \ldots + \beta_k.X_{ki}$$

Overdispersion Test

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

$$Y_i^* = \beta \cdot \lambda_{poisson_i}$$

Cameron and Trivedi (1990) state that if the phenomenon of overdispersion in the data occurs, the estimated parameter β of this **auxiliary model without intercept** will be statistically different from zero, at a given level of significance (5%, usually).

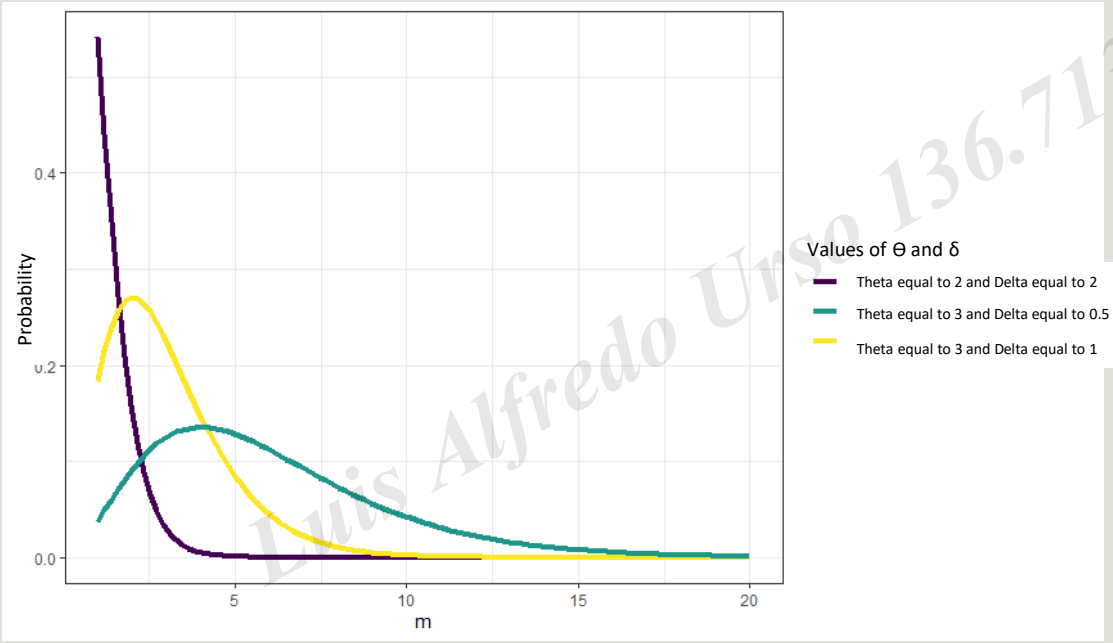
Gamma-Poisson Distribution or Negative Binomial

For a certain observation *i* (*i* = 1, 2, ..., *n*, in which *n* is the sample size), the function of the probability distribution of the dependent variable *Y* will be given by:

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

in which θ is called a form parameter ($\theta > 0$) and δ is called a decay rate parameter ($\delta > 0$).

Gamma-Poisson Distribution or Negative Binomial



- Mean:

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

- Variance:

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

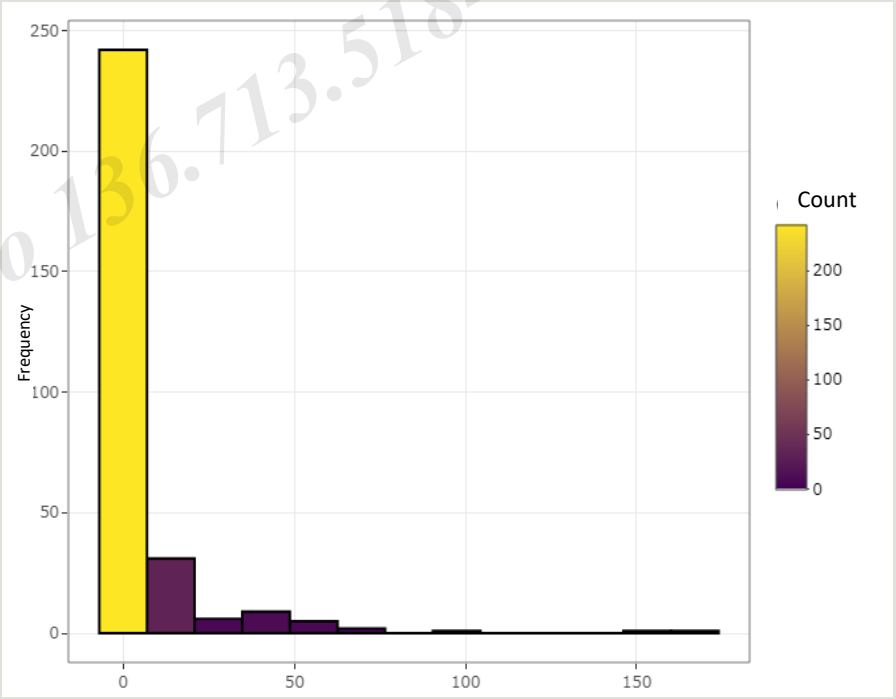
$$\phi = \frac{1}{\theta}$$

NB2 Models

The Gamma-Poisson Model or Negative Binomial

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

Zero-Inflated Models



Model Choice

Verification	Modelo de Regressão para Dados de Contagem			
	Poisson	Negative Binomial	Zero-Inflated Poisson (ZIP)	Zero-Inflated Negative Binomial (ZINB)
Overdispersion in Data of the Dependent Variable	No	Yes	No	Yes
Excessive Number Of Zeros In The Dependent Variable	No	No	Yes	Yes

Zero-Inflated Models

They are considered a combination of a model for count data and a model for binary data, since they are used to investigate the reasons that conduct to a certain number of occurrences (counts) of a phenomenon, as well as the reasons (or not) that conduct to the occurrence of this phenomenon, regardless of the amount of counts observed.

While a Poisson model inflated with zeros is estimated from the **combination of a Bernoulli distribution with a Poisson distribution**, a negative binomial model inflated with zeros is estimated by combining **a Bernoulli distribution with a Gamma-Poisson** distribution.

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

Zero-Inflated Models

The definition of the existence of an excessive number of zeros in the dependent variable *Y* is elaborated through a specific test, known as **the Vuong** (1989) test, which will represent an important *output* to be analyzed in the estimation of regression models for count data, when there is a suspicion of existence of zero-inflation.

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

Zero-Inflated Models of the Poisson Type (ZIP)

Regarding specifically the **Poisson regression models inflated with zeros**, we can define that, while the **probability *p* of occurrence of no count** for a given observation *i* (*i* = 1, 2, ..., *n*, in which *n* is the sample size), that is, $\boldsymbol{p(Y_i = 0)}$, is calculated taking into account the sum of a dichotomous component with a counting component, and therefore, the $_{plogit}$ probability of not occurring any count due to the dichotomous component, the **probability *p* of occurrence of a certain count *m*** (*m* = 1, 2, ...), i.e., $\boldsymbol{p(Y_i = m)}$, follows the expression of the probability of the Poisson distribution, multiplied by $(1 - {}_{plogit})$.

Zero-Inflated Models of the Poisson Type (ZIP)

$$\begin{cases} p\left(Y_i = 0\right) = p_{logit_i} + \left(1 - p_{logit_i}\right).e^{-\lambda_i} \\ p\left(Y_i = m\right) = \left(1 - p_{logit_i}\right).\dfrac{e^{-\lambda_i}.\lambda_i^m}{m!}, \qquad m = 1,\ 2,\ \ldots \end{cases}$$

$$p_{logit_i} = \frac{1}{1 + e^{-\left(\gamma + \delta_1.W_{1i} + \delta_2.W_{2i} + \ldots + \delta_q.W_{qi}\right)}}$$

$$\lambda_{poisson_i} = e^{\left(\alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \ldots + \beta_k.X_{ki}\right)}$$

The zero-inflated Poisson regression models have two processes that generate zeros, one of them due to the binary distribution (in this case, the so-called structural zeros are generated) and the other one due to the Poisson distribution (in this situation, count data are generated, including the so-called sample zeros).

Zero-Inflated Models of the Negative Binomial Type (ZINB)

Regarding the zero-**inflated negative binomial regression models**, we can define that, while the **probability *p* of occurrence of no count** for a given observation *i*, that is, $\boldsymbol{p(Y_i = 0)}$, is also calculated taking into account the sum of a dichotomous component with a count component, the **probability *p* of occurrence of a certain count *m*** (*m* = 1, 2, ...), that is, $\boldsymbol{p(Y_i = m)}$, follows now the expression of the probability of the Gamma-Poisson distribution.

Zero-Inflated Models of the Negative Binomial Type (ZINB)

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

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

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