Comparação de Modelos de Regressão Para Dados de Contagem Inflacionados de Zeros Por Meio de Simulações

Maicon Michael Fridrich Gottselig [1]

Juliana Sena de Souza[2]

Silvana Schneider[3]

Resumo: Modelos inflacionados de zeros são ferramentas importantes no desarme de dados não identicamente distribuídos provenientes da mistura de duas populações com processos distintos. Esta classe de modelos é evidenciada por Diane Lambert (1992) que postula uma família de modelos de mistura que permite a modelagem de dados com excesso de zeros, lidando com a sobredispersão decorrente desta característica. Posto isso, este trabalho tem como foco executar por meio de simulações computacionais uma comparação de modelos de contagem sob ótica de excesso de zeros. Os seguintes modelos: ZIP (Zero-Inflated Poisson), ZIG (Zero-Inflated Geometric), ZIB (Zero-Inflated Binomial), ZINB (Zero-Inflated Negative Binomial), ZIPIG (Zero-Inflated Poisson Inverse Gaussian), ZIBB (Zero-Inflated Beta Binomial), ZIBNB (Zero-Inflated Beta Negative Binomial), ZICMP (Zero-Inflated Conway-Maxwell Poisson) e ZIDelaporte (Zero-Inflared Delaporte); São utilizados como base para simulações e ajustes cruzados afim de avaliar e testar adaptilidade de cada modelo a diferentes cenários de sobredispersão e inflação de zeros. Notou-se que modelos os modelos relativamente novos ZID e ZICMP performam muito bem e se posicionam paralelamente aos modelos ZIPIG e ZINB. Negativamente destacam-se os modelos ZIBNB, ZIB, ZIBB e ZIG que não obtiveram estimativas satisfatórias.

Palavras-chave:*Modelos de contagem*, *Inflação de zeros*, *sobresipersão*, *Comparação*, *Simulação*

1 Introdução

Frank A. Haight (1967) explica que dados de contagem são definidos como o número de sucessos de experimentos realizado num período finito. Quando existe o intuito de se modelar variáveis de contagem, afim de se inferir acerca da relação desta esperança condicionada à variáveis explicativas, é necessária a suposição de distribuições discretas sobre a variável dependende, como exemplos bastante explorados menciona-se Poisson, Binomial e Geométrica. Como tais distribuição pertencem a família exponencial de distribuições toda a construção teórica exposta em Nelder e Wedderburn (1972) estende-se de forma natural.

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *maiconmfg@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *julianass.estatistica@gmail.com*
[3]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *sschneider@ufrgs.br*

Na maioria dos estudos entretanto, surge o fenômeno da sobredispersão, que é caracterizada como uma variabilidade superior a qual o modelo de contagem empregado é capaz de incorporar. No caso da distribuição de Poisson que impõe equidispersão, quando é registrado $\mathbb{E}(\bar{Y}) \neq \mathbb{V}A\bar{R}(Y)$ há indícios que colocam em cheque a Regressão de Poisson. A direção desde desbalanço entre esperança e variância caracteriza sub ou sobredispersão e tem como justificativa uma grande gama de justificativas: caudas pesadas, assimetria, excesso de zeros, entre outros.

O eixo principal deste trabalho é verificar a adaptilidade de modelos de contagem à sobredispersão e excesso de zeros. Segundo proposição de Lambert (1992) que sugere mistura de distribuições de contagem com distribuição de Bernoulli afim de captação de efeitos associados ao processo de zeros. É importante ressaltar a existência de outras alternativas para ajuste de dados inflacionados de zeros, como modelos Hurdle de Ridout (1998) e modelos de zeros alterados de Heilbron (1989).

2 Modelos Inflacionados de Zero

Em seu artigo, D. Lambert (1992) discorre acerca de dados provenientes da amostragem de um conjunto de duas populações com processos distintos. Uma população contendo apenas indivíduos com valor zero e outra população cujos indivíduos se adequam a alguma distribução de contagem.

Desta maneira assumindo $Y = (y_1, y_2, ..., y_n)$ como uma amostra aleatória independente do processo acima descrito tem-se: $P(y_i \in \text{Sempre Zero}) = \pi$ e $P(y_i \notin \text{Sempre Zero}) = 1 - \pi$, o que compila em:

$$P(Y_{ZI} = y|\theta, \pi) = \begin{cases} \pi + (1-\pi)f_y(y=0|\theta), & \text{y=0.,} \\ (1-\pi)f_y(y|\theta), & \text{y > 0 .} \\ 0, & \text{caso contrário} \end{cases} \tag{1}$$

onde f_y denota a distribuição de probabilidade indexada pelo parâmetro θ do processo de contagem e π assume posição de parâmetro que define a probabilidade da contagem de zero decorrente dos indivídios da população que apenas fornece contagem zero. Lambert percebeu que sob condições ideais a contagem de falhas de soldas eram sempre zero, e quando fora de controle, o processo observada falhas que se adequavam a distribuição de Poisson. Assim, propôs assumir que os processos sob controle e fora de controle eram na verdade populações distintas.

A formulação (1) caracteriza a família de distribuições infladas de zeros e, frente à diferentes f_y, novas propriedades são observadas e diferentes fontes de sobredispersão são captadas conforme mostrado por Paula (2004). Inicialmente é necessário verificar que

$$\mathbb{E}(Y_{ZI}) = (1-\pi)\mathbb{E}(Y) \qquad \text{e} \qquad \mathbb{V}AR(Y_{ZI}) = (1-\pi)(\mathbb{V}AR(Y) + \pi\mathbb{E}(Y)^2).$$

Por meio do índice de sobredispersão proposto por Cox e Lewis (1966), e denotado como $OI(Y)$ *(overdispersion index)*, tem a fórmula denotada por $OI(Y) = \mathbb{V}AR(Y)/\mathbb{E}(Y)$. É possível verificar que a proposição de Lambert com a insersão do parâmetro π de fato há absorção de sobredispersão de ordem $\pi\mathbb{E}(Y)$ uma vez que $OI(Y_{ZI}) = OI(Y) + \pi\mathbb{E}(Y)$.

Lambert (1992) ainda propõe a modelagem via covariáveis da proporção π de zeros estruturais e da média μ do processo. Explica também que ambos os parâmetros podem ou não ser modelados pelo mesmo conjunto de covariáveis, o que os torna ou não relacionados, conforme examinado por Daniel B. Hall (2000). Como marginalmente $\mathbb{E}(Y_i) = (1 - p_i)\mu_i$ pode haver confundimento nas estimativas dos coeficientes dos dois processos, o que atribui maior variabilidade aos coeficientes associados ao processo logístico bem como acrésimo de erros padrões.

As estimativas dos coeficientes e demais parâmetros são obtidas pela maximização da verossimilhança por meio do emprego de um método computacional recursivo. Lambert (1992) demonstra como se dá a construção do algorítmo EM, o que requer o cálculo de esperanças condicionais que podem ser complexas, por isso geralmente utiliza-se o método de Ficher Scoring que é um algortimo de *hill climbing*, conforme explicitado por Sampson (1976).

A proposição de Lambert é flexível e se estende a diversas $f_y(y)$, que acaba por incorporar ao modelo inflado de zeros seus momentos e permite melhor adequamento à diferentes perfis de dados, captando sobredisperão e modelando excesso de zeros. Outra alternativa para incorporar maior sobredispersão ao modelo por meio da inclusão de novos parâmetros é via suposição de variáveis latentes $Y|W \sim P(\lambda)$ que frente a diferentes W, confere novos parâmetros e complexidade à Y . Há também a posibilidade de se assumir Y como sendo resultado de alguma função de variáveis aleatórias do tipo $Y = W + Z$, com W e Z variáveis aleatórias. Estas táticas corroboram com a construção de modelos mais fléxiveis. A Tabela abaixo traz os modelos selecionados e expõe suas construções, bem como a paralela distribuição inflacionada de zeros e o índice de sobredispersão.

Tabela 1: Tabela resumo das distribuições infladas de zeros

Descrição	**Distribuição ($\underline{\theta}$)**	**Distribuição ZI ($\underline{\theta}$)**	**OI(Y_{ZI})**
-	P(λ)	ZIP (λ, π)	$1 + \pi\lambda$
$Y\|W \sim P(\lambda), W \sim G(\alpha, \beta))$	NB(α, β)	ZINB (α, β, π)	$\frac{(\alpha+\beta+\alpha\beta\pi)}{\beta}$
-	G(p)	ZIG(p, π)	$\frac{1+p\pi}{p}$
-	Bin(k, p)	ZIB(k, p, π)	$1 - p + kp\pi$
$Y\|W \sim P(\lambda), W \sim IG(\mu, \sigma))$	PIG(μ, σ)	ZIPIG(μ, σ, π)	$e^{\mu+\sigma^2/2}[e^{\sigma^2} - 1 + \pi]$
$Y\|p \sim Bin(n, p)$,$p \sim Beta(\alpha, \beta))$	BB(n, α, β)	ZIBB(n, α, β, π)	$\frac{n\beta}{\alpha+\beta} + \pi\frac{n\alpha}{\alpha+\beta}$
$Y\|p \sim NB(r, p)$, $p \sim Beta(\alpha, \beta))$	BNB(r, α, β)	ZIBNB(r, α, β, π)	$\frac{r\beta}{\alpha-1}\left[\frac{(r+\alpha-1)(\alpha+\beta-1)}{r\beta(\alpha-2)} - \pi^2\right]$
-	CMP(λ, v)	ZICMP(λ, v, π)	$\frac{\frac{\lambda^{1/v}}{v} + \pi(\lambda^{1/v} + \frac{1-v}{2v})^2}{\lambda^{1/v} + \frac{1-v}{2v}}$
Convolução entre NB(α, β) e P(λ)	Delaporte(λ, α, β)	ZIDelaporte($\lambda, \alpha, \beta, \pi$)	$\frac{\lambda+\alpha\beta(1+\beta)+\pi(\lambda+\alpha\beta)^2}{\lambda+\alpha\beta}$

3 Metodologia e Simulações

Com a premissa de comparar a capacidade de absorção de sobredispersão dos modelos apresentados na Tabela 1 e verificar o ajuste destes frente a dados com excesso de zeros foram realizadas simulações computacionais de dados de regressão com as distribuições alvo via software R (versão 3.4.1) com auxílio dos pacotes VGAM, gamlss.dist COMPoissonReg, pscl, Delaporte e gamlss.

Foram gerados 1000 bancos de dados de cada um dos k=9 modelos abordados, cada qual com n=500. Tomando $\beta = [1, 0.5, -0.5]'$ e $\gamma = [-2, 1, -2]'$ como coeficientes regressores, além da relação $log(\mu_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$ e $log(\frac{\pi}{1-\pi}) = \gamma_0 + \gamma_1 X_{i,2} + \gamma_2 X_{i,3}$ sendo que $x_{i,1} \sim N(3.5, 0.6)$, $x_{i,2} \sim Gamma(20, 100)$ e $x_{i,3} \sim Gamma(1, 1)$.

Com isso gerou-se $y_i \sim D_k(\mu = e^{X'_{1:2}\beta}; \pi = \dfrac{e^{X'_{2:3}\gamma}}{1 + e^{X'_{2:3}\gamma}})$, D_k expressando o k-ésimo modelo, portanto D_k é modelo de origem de Y condicionado em X. Desta forma foram observados para μ valores que se estendem de 2.95 e 70.9 e para π foram observados valores no intervalo de 0.02 a 0.17. Os demais parâmetros de sobredispersão foram setados de forma a se obter grande variedade de índices de sobredispersão, cujos valores observados se extendem de 1.87 a 29.15.

Gerados os dados, procedeu-se o ajuste dos modelos. Para cada banco foram ajustado os nove modelos inflacionados de zero abordados neste estudo, além da regressão de Poisson tradicional, o que confere a cada banco dez ajustes. Como métricas para avaliar a adaptabilidade dos modelos aos dados foram coletadas as estimativas dos coeficientes e seus erros padrões.

Já para a verificação da qualidade do ajuste foram utilizados o logarítmo da função de verossimilhança maximizada, que consta nas tabelas como *LogLik*; o critério de informação de Akaike (AIC), de Hilbe (2014) já bastante utilizado, o critério de informação de Hannan-Quinn (HQC), que é frequentemente usado como um critério para a seleção de modelos entre um conjunto finito de modelos e o critério de informação bayesiano (BIC), uma medida de ajuste que possui um termo que penaliza o número de parâmetros do modelo de uma forma mais grave que o AIC.

4 Resultados

O ajuste dos modelos e obtenção das estimativas dos coeficientes regressores foi realizada via maximização de verossimilhaça que se deu pelo método iterativo de Fisher Scoring, um algoritmo Hill Climbing com critério de convergência definido por uma diferença absoluta mínima entre as verossimilhanças de duas iterações sucessivas. Essa classe de algorítmos apesar de amplamente versátil, apresenta problemas de convergência frente a alguns cenários dentro de um número limitado de iterações. Este trabalho conforme esperado encontrou problemas de convergência em alguns bancos e modelos, conforme já exposto por Silva (2017) em sua dissertação. Globalmente obtivemos convergência em 91.18%

dos ajustes. A regressão de Poisson, ZIP e ZIB convergiram em 100% do ajustes. ZIBN, ZIG e ZIBB apresentaram convergência na casa dos 97%, já ZIBNB, ZICMP, ZIDelaporte e ZIPIG retornaram 80% de convergência.

Nota-se uma proporcionalidade entre percentual de convergência e complexidade do modelo ajustado. Já a convergência segundo o modelo do qual o dados foram gerados apresentou percentual homogêneo na casa dos 91%. Dados simulados de ZIP, ZIB e ZIG foram os com menor índice (88%), justamente os modelos mais simplistas. Ou seja, evidenciamos em nossos dados que frente uma sobreparametrização há maiores chances de se registrar uma falha na convergência do modelo.

A Tabela 2 apresentada abaixo apresenta as estimativas médias dos coeficientes de regressão de μ e de seus erros padrões relativos ao modelos convergentes.

Tabela 2: Estimativas para β_0, β_1 e β_2 e seus respectivos erros padrões dos modelos de regressão aplicados à simulações de diferentes tipos de dados inflacionados de zeros

		POIS	**ZIP**	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.00 (0.09)	1.00 (0.09)	0.95 (0.35)	1.00 (0.09)	1.00 (0.10)	0.98 (0.12)	0.98 (0.01)	1.00 (0.09)	0.97 (0.08)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.08)	0.50 (0.02)	0.50 (0.02)	0.51 (0.03)	0.53 (0.00)	0.50 (0.02)	0.51 (0.02)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.56 (0.27)	-0.51 (0.27)	-0.51 (0.28)	-0.54 (1.06)	-0.51 (0.27)	-0.51 (0.28)	-0.53 (0.36)	-0.55 (0.04)	-0.51 (0.27)	-0.47 (0.17)
ZINB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.01 (0.09)	1.00 (0.16)	0.95 (0.35)	1.01 (0.09)	1.00 (0.16)	1.02 (0.16)	1.06 (0.14)	0.97 (0.16)	1.04 (0.16)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.04)	0.50 (0.08)	0.50 (0.02)	0.50 (0.04)	0.48 (0.04)	0.49 (0.03)	0.51 (0.04)	0.49 (0.03)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.50 (0.27)	-0.50 (0.48)	-0.52 (1.06)	-0.50 (0.27)	-0.50 (0.48)	-0.49 (0.48)	-0.50 (0.40)	-0.51 (0.48)	-0.46 (0.46)
ZIG	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.94 (0.09)	1.16 (0.09)	0.99 (0.36)	0.99 (0.36)	1.17 (0.09)	1.08 (0.38)	1.43 (0.32)	1.46 (0.33)	1.42 (0.41)	1.48 (0.33)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.47 (0.02)	0.50 (0.08)	0.50 (0.08)	0.47 (0.02)	0.49 (0.08)	0.36 (0.07)	0.37 (0.07)	0.14 (0.00)	0.37 (0.07)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.53 (0.27)	-0.47 (0.27)	-0.49 (1.09)	-0.49 (1.10)	-0.47 (0.27)	-0.47 (1.15)	-0.35 (0.98)	-0.36 (0.99)	-0.50 (0.03)	-0.44 (1.04)
ZIB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.00 (0.09)	1.00 (0.13)	0.95 (0.35)	1.00 (0.09)	1.00 (0.12)	0.98 (0.14)	0.98 (0.01)	1.00 (0.07)	0.99 (0.08)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.03)	0.50 (0.08)	0.50 (0.02)	0.50 (0.03)	0.51 (0.03)	0.52 (0.00)	0.50 (0.02)	0.50 (0.02)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.57 (0.27)	-0.51 (0.27)	-0.51 (0.37)	-0.53 (1.05)	-0.51 (0.27)	-0.51 (0.35)	-0.52 (0.42)	-0.55 (0.04)	-0.51 (0.21)	-0.50 (0.19)
ZIPIG	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.95 (0.09)	1.00 (0.09)	0.99 (0.18)	0.95 (0.35)	1.00 (0.09)	0.99 (0.18)	1.03 (0.18)	1.01 (0.17)	0.93 (0.18)	1.05 (0.17)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.04)	0.51 (0.08)	0.50 (0.02)	0.50 (0.04)	0.48 (0.04)	0.50 (0.04)	0.52 (0.04)	0.49 (0.04)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.01 (0.27)	-0.50 (0.52)	-0.52 (1.06)	-0.50 (0.27)	-0.49 (0.53)	-0.48 (0.52)	-0.49 (0.51)	-0.50 (0.53)	-0.47 (0.50)
ZIBB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.09 (0.09)	1.07 (0.20)	0.96 (0.35)	1.09 (0.09)	0.88 (0.21)	1.00 (0.20)	0.92 (0.19)	0.93 (0.18)	0.92 (0.19)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.48 (0.02)	0.48 (0.04)	0.50 (0.08)	0.48 (0.02)	0.54 (0.05)	0.53 (0.04)	0.52 (0.04)	0.99 (0.19)	0.52 (0.04)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.46 (0.27)	-0.45 (0.59)	-0.50 (1.06)	-0.46 (0.27)	-0.50 (0.61)	-0.53 (0.57)	-0.50 (0.56)	-0.48 (0.56)	-0.47 (0.56)
ZIBNB	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.89 (0.09)	1.02 (0.09)	0.99 (0.22)	0.90 (0.36)	1.02 (0.09)	0.99 (0.22)	1.09 (0.22)	1.01 (0.22)	0.90 (0.23)	1.09 (0.21)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.05)	0.51 (0.08)	0.50 (0.02)	0.50 (0.05)	0.46 (0.05)	0.50 (0.05)	0.52 (0.05)	0.47 (0.05)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.50 (0.28)	-0.50 (0.28)	-0.51 (0.65)	-0.53 (1.09)	-0.50 (0.27)	-0.51 (0.67)	-0.45 (0.65)	-0.50 (0.66)	-0.51 (0.67)	-0.46 (0.63)
ZICMP	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.96 (0.09)	1.00 (0.09)	1.00 (0.11)	0.96 (0.35)	1.00 (0.09)	1.00 (0.11)	0.93 (0.11)	1.01 (0.01)	1.00 (0.11)	0.99 (0.09)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.02)	0.50 (0.08)	0.50 (0.02)	0.50 (0.02)	0.51 (0.02)	0.52 (0.00)	0.50 (0.02)	0.50 (0.02)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.55 (0.27)	-0.49 (0.27)	-0.49 (0.32)	-0.53 (1.06)	-0.49 (0.27)	-0.49 (0.32)	-0.51 (0.33)	-0.49 (0.04)	-0.50 (0.32)	-0.48 (0.24)
ZID	$\hat{\beta}_0(\hat{\sigma}_{\beta 0})$	0.97 (0.09)	1.01 (0.09)	1.00 (0.20)	0.96 (0.35)	1.01 (0.09)	0.93 (0.20)	0.99 (0.20)	0.97 (0.20)	0.92 (0.20)	1.00 (0.19)
	$\hat{\beta}_1(\hat{\sigma}_{\beta 1})$	0.50 (0.02)	0.50 (0.02)	0.50 (0.04)	0.50 (0.08)	0.50 (0.02)	0.52 (0.05)	0.49 (0.04)	0.51 (0.04)	0.52 (0.04)	0.50 (0.04)
	$\hat{\beta}_2(\hat{\sigma}_{\beta 2})$	-0.57 (0.27)	-0.51 (0.27)	-0.52 (0.59)	-0.55 (1.06)	-0.51 (0.27)	-0.54 (0.60)	-0.52 (0.57)	-0.53 (0.58)	-0.53 (0.60)	-0.51 (0.57)

Verifica-se que dentro de um limiar, em média as estimativas são satisfatórias e parecem pouco viesadas. Silva (2017) mostra via simulação que EM em comparação a Hill Climbing é superior e preferível, pois apresenta menor viés e melhor índice de convergência. Fica evidente ainda que vício e convergência são afetados conjuntamente pelo π e n. Este projeto por atribuir um grau baixo a moderado de zeros e um n amistoso não lida com problema de grandes viéses e raras convergências. Referente aos erros padrões percebe-se que modelos mais complexos tendem a apresentar erros padrões maiores, com excessão do ZIG, que retorna erros bastante superiores aos outros modelos.

Na Tabela 3 apresentada abaixo são expostas as estimativas médias dos coeficientes regressores associados ao processo logístico que modela a probabilidade de pertencer ao grupo sempre zero, bem como

seus erros padrões.

Tabela 3: Estimativas para γ_0, γ_1 e γ_2 e seus respectivos erros padrões dos modelos de regressão aplicados à simulações de diferentes tipos de dados inflacionados de zeros

		ZIP	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,00 (1,00)	-0,99 (1,04)	-1,75 (3,25)	-2,00 (1,00)	-1,98 (1,01)	-2,49 (1,05)	-2,04 (1,06)	-2,02 (1,00)	-2,45 (0,99)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,94 (4,75)	0,89 (4,93)	0,64 (14,07)	0,94 (4,75)	0,88 (4,77)	0,73 (4,98)	0,83 (5,03)	1,10 (4,73)	1,12 (4,62)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,13 (0,64)	-2,13 (0,68)	-2,23 (6,23)	-2,13 (0,64)	-2,12 (0,64)	-2,02 (0,70)	-2,14 (0,7)	-2,12 (0,64)	-2,05 (0,62)
ZINB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,02 (1,00)	-2,01 (1,02)	-1,59 (3,28)	-2,02 (1,00)	-2,01 (1,01)	-2,17 (1,02)	-2,03 (1,03)	-2,11 (1,07)	-2,06 (1,01)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,92 (4,75)	0,91 (4,84)	0,56 (13,92)	0,93 (4,75)	0,91 (4,81)	0,84 (4,83)	0,76 (4,87)	1,48 (5,06)	1,27 (4,76)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,04 (0,62)	-2,13 (0,67)	-1,62 (6,33)	-2,04 (0,62)	-2,10 (0,65)	-1,93 (0,70)	-2,12 (0,67)	-2,32 (0,77)	-2,12 (0,65)
ZIG	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-1,75 (0,67)	-1,92 (1,51)	-1,87 (1,49)	-1,75 (0,67)	-1,82 (0,84)	-2,35 (1,59)	-1,91 (1,38)	-1,90 (1,09)	-1,99 (1,30)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,37 (3,18)	1,41 (7,21)	1,31 (7,12)	1,37 (3,18)	1,25 (3,98)	1,09 (7,59)	1,10 (6,66)	1,14 (4,98)	1,21 (6,01)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,44 (0,20)	-1,56 (1,52)	-2,49 (1,42)	-2,44 (0,2)	-1,78 (0,38)	-2,00 (1,85)	-2,05 (1,19)	-2,44 (0,85)	-1,67 (1,11)
ZIB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-1,99 (1,01)	1,99 (1,07)	-1,93 (3,22)	-1,99 (1,01)	-1,98 (1,06)	-1,99 (1,07)	-2,03 (1,08)	-1,94 (1,00)	-1,64 (1,00)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,88 (4,77)	0,98 (5,06)	1,38 (13,39)	0,88 (4,77)	0,83 (5,00)	0,85 (5,07)	0,84 (5,11)	0,73 (4,75)	0,96 (4,63)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,14 (0,64)	-2,12 (0,70)	-1,60 (6,68)	-2,14 (0,64)	-2,13 (0,69)	-2,12 (0,71)	-2,14 (0,71)	-2,14 (0,64)	-2,02 (0,61)
ZIPIG	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,00 (1,00)	-1,98 (1,03)	-1,55 (3,26)	-2,00 (1,00)	-1,99 (1,02)	-1,93 (1,04)	-1,99 (1,03)	-2,14 (1,13)	-1,98 (1,01)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,89 (4,74)	0,89 (4,88)	0,88 (13,78)	0,89 (4,73)	0,88 (4,82)	0,80 (4,90)	0,89 (4,86)	0,65 (5,30)	0,90 (4,76)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,06 (0,62)	-2,20 (0,69)	-2,33 (6,40)	-2,06 (0,62)	-2,15 (0,67)	-1,98 (0,73)	-2,18 (0,69)	-1,51 (0,88)	-2,11 (0,65)
ZIBB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,17 (0,87)	-2,09 (0,94)	-1,63 (2,82)	-2,17 (0,87)	-2,10 (0,92)	-1,76 (0,98)	-2,05 (0,99)	-2,10 (1,05)	-2,39 (0,98)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,30 (4,11)	1,31 (4,44)	0,58 (12,36)	1,30 (4,11)	1,30 (4,32)	0,97 (4,63)	1,20 (4,67)	0,59 (4,91)	0,83 (4,58)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,14 (0,38)	-1,52 (0,56)	-2,38 (5,15)	-2,14 (0,38)	-2,40 (0,50)	-1,61 (0,66)	-1,72 (0,62)	-1,94 (0,75)	-1,66 (0,62)
ZIBNB	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,13 (0,70)	-2,17 (0,73)	-1,98 (2,12)	-2,14 (0,70)	-2,15 (0,71)	-2,07 (0,75)	-2,16 (0,72)	-1,54 (0,84)	-2,26 (0,72)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	0,87 (3,33)	0,78 (3,52)	0,57 (10,17)	0,87 (3,33)	0,82 (3,41)	0,83 (3,60)	0,80 (3,47)	0,97 (3,96)	1,38 (3,41)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,02 (0,15)	-2,03 (0,16)	-1,80 (0,96)	-2,02 (0,15)	-2,02 (0,16)	-2,01 (0,17)	-2,02 (0,16)	-2,04 (0,19)	-2,03 (0,16)
ZICMP	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,04 (1,01)	-2,04 (1,01)	-1,87 (3,24)	-2,04 (1,01)	-2,04 (1,01)	-2,15 (1,00)	-2,11 (1,08)	-2,04 (1,01)	-2,23 (1,00)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,08 (4,78)	1,08 (4,79)	1,00 (13,84)	1,08 (4,78)	1,09 (4,79)	0,98 (4,71)	1,15 (5,12)	1,11 (4,78)	1,01 (4,71)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-2,12 (0,64)	-2,13 (0,64)	-1,67 (6,59)	-2,12 (0,64)	-2,13 (0,64)	-1,96 (0,65)	-2,12 (0,72)	-2,14 (0,65)	-2,05 (0,62)
ZID	$\hat{\gamma}_0(\hat{\sigma}_{\gamma 0})$	-2,03 (0,99)	-2,00 (1,05)	-2,24 (3,22)	-2,03 (0,99)	-2,00 (1,02)	-1,52 (1,06)	-2,00 (1,04)	-2,21(1,20)	-2,00 (1,02)
	$\hat{\gamma}_1(\hat{\sigma}_{\gamma 1})$	1,00 (4,66)	0,97 (4,97)	0,68 (13,74)	1,00 (4,66)	0,96 (4,83)	0,82 (5,03)	0,94 (4,95)	0,90 (5,63)	0,99 (4,81)
	$\hat{\gamma}_2(\hat{\sigma}_{\gamma 2})$	-1,95 (0,59)	-2,22 (0,73)	-1,71 (6,38)	-1,95 (0,59)	-2,11 (0,67)	-1,99 (0,79)	-2,19 (0,71)	-1,65 (1,00)	-2,12 (0,68)

As estimativas médias de γ deixam de ser tão satisfatórias quanto as estimativas dos coeficientes associados à μ, o que deixa claro o vício decorrente do método de otimização e possível confundimento, uma vez que há uma covariável ($X_{i,2}$) que estabele uma interseção entre conjunto de covariáveis de μ e π, ou seja, está associada ao β_2 e γ_1. O erro padrão médio assume também novas escalas (em comparação com os erros padrões do vetor β) justamente pela interseção de covariáveis anteriormente mencionado, sendo observado valor médio máximo de 14.07 no modelo ZIG (que já apresentou erros padrões altos para o vetor $\hat{\beta}$). Da mesma maneira que nas estimativas do vetor β, o ZIB apresenta o menor erro padrão, seguido pelo modelo ZINB e ZIP, ZIPIG e ZID.

A Tabela 4 traz as médias das medidas de ajuste selecionadas para verificar qualidade do ajuste e comparar modelos. São apresentadas as medidas: o logaritmo da verossimilhança Maximizada, AIC, HQC e BIC. Estas últimas três medidas buscam por meio da verossimilhança tornar diferentes modelos comparáveis.

Os modelos que obtiveram as menores médias nos critérios de qualidade de ajuste estão evidenciados na tabela abaixo em *negrito*. Se mais de uma estimativa possui médias muito similares entre os mesmos dados, então ambas estão destacadas.

Tabela 4: Qualidade de ajuste do modelo frente aos dados inflacionado de zeros

		POIS	**ZIP**	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	LogLik	-1659,96	**-1384,91**	**-1384,98**	-1814,10	**-1384,91**	-1386,08	-2612,01	-1434,20	**-1384,83**	-1387,31
	AIC	3325,92	**2781,83**	**2783,96**	3640,20	**2781,81**	2786,16	5238,02	2884,36	**2783,66**	2790,63
	HQC	3330,88	**2791,75**	**2795,53**	3650,12	**2791,73**	2797,74	5249,60	2897,55	**2795,24**	2800,20
	BIC	3338,56	**2807,11**	**2813,46**	3665,49	**2807,10**	2815,66	5267,52	2917,98	**2813,17**	2818,13
ZINB	LogLik	-2125,73	-1851,37	**-1626,48**	-1814,55	-1859,57	**-1627,87**	-1640,95	-1630,41	-1632,34	**-1628,69**
	AIC	4257,46	3714,74	**3266,97**	3641,10	3731,15	**3269,74**	3295,91	3276,82	3278,68	**3273,39**
	HQC	4262,42	3724,66	**3278,54**	3651,02	3741,07	**3281,32**	3307,48	3290,06	3290,26	**3282,96**
	BIC	4270,11	3740,03	**3296,47**	3666,39	3756,43	**3299,24**	3325,41	3310,54	3308,19	**3300,89**
ZIG	LogLik	-4476,17	-3826,92	**-1803,02**	**-1803,60**	-3882,01	-1816,54	-1809,42	-1807,83	-1828,78	-1808,24
	AIC	8958,35	7665,83	**3620,05**	**3619,20**	7776,02	3647,08	3632,84	3631,66	3671,57	3632,48
	HQC	8963,31	7675,75	**3631,62**	**3629,12**	7785,94	3658,65	3644,41	3644,89	3683,14	3642,06
	BIC	8970,99	7691,12	**3649,55**	**3644,49**	7801,31	3676,58	3662,34	3665,38	3701,07	3659,98
ZIB	LogLik	-1565,17	-1291,84	-1292,31	-1814,55	-1290,25	-1293,25	-2517,81	-1388,65	**-1262,28**	-1293,52
	AIC	3136,34	2595,67	2598,62	3641,11	2592,50	2600,49	5049,63	2793,30	**2538,56**	2603,04
	HQC	3141,30	2605,59	2610,20	3651,03	2602,42	2612,07	5061,20	2806,53	**2550,14**	2612,62
	BIC	3148,99	2620,96	2628,13	3666,39	2617,79	2629,99	5079,13	2827,02	**2568,06**	2630,54
ZIPIG	LogLik	-2274,02	-1998,31	-1659,07	-1814,22	-2009,82	**-1656,95**	-1673,09	-1666,11	-1671,27	**-1658,77**
	AIC	4554,03	4008,61	3332,13	3640,44	4031,65	**3327,90**	3360,18	3348,21	3356,55	**3333,54**
	HQC	4559,00	4018,54	3343,71	3650,37	4041,57	**3339,48**	3371,75	3361,44	3368,12	**3343,11**
	BIC	4566,68	4033,90	3361,63	3665,73	4056,93	**3357,40**	3389,68	3381,93	3386,05	**3361,04**
ZIBB	LogLik	-2416,31	-2079,16	-1708,62	-1812,52	-2088,75	-1721,87	-1699,84	**-1695,36**	-1689,54	**-1694,78**
	AIC	4838,61	4170,32	3431,24	3637,03	4189,49	3457,75	3413,69	**3406,72**	3393,08	**3405,56**
	HQC	4843,58	4180,24	3442,82	3646,96	4199,42	3469,33	3425,26	**3419,95**	3404,66	**3415,14**
	BIC	4851,26	4195,61	3460,74	3662,32	4214,78	3487,25	3443,19	**3440,44**	3422,58	**3433,07**
ZIBNB	LogLik	-2870,55	-2314,99	-1710,29	-1785,20	-2332,88	**-1709,35**	-1725,21	**-1708,09**	-1720,44	-1711,33
	AIC	5747,11	4641,97	3434,58	3582,40	4677,75	**3432,70**	3464,43	**3432,18**	3454,88	3438,67
	HQC	5752,07	4651,90	3446,16	3592,33	4687,67	**3444,28**	3476,01	**3445,41**	3466,45	3448,24
	BIC	5759,75	4667,26	3464,08	3607,69	4703,04	**3462,21**	3493,93	**3465,90**	3484,38	3466,17
ZICMP	LogLik	-1750,53	-1479,49	-1464,98	-1815,12	-1480,93	-1465,39	-1478,03	-1504,38	**-1463,99**	**-1464,12**
	AIC	3507,07	2970,98	2943,96	3642,23	2973,87	2944,77	2970,05	3024,76	**2941,97**	**2944,24**
	HQC	3512,03	2980,91	2955,54	3652,16	2983,79	2956,35	2981,63	3037,99	**2953,55**	**2953,82**
	BIC	3519,71	2996,27	2973,46	3667,52	2999,16	2974,28	2999,56	3058,47	**2971,48**	**2971,75**
ZID	LogLik	-2464,62	-2182,17	-1701,07	-1813,56	-2196,32	-1698,73	-1707,18	**-1698,44**	-1708,60	**-1696,34**
	AIC	4935,23	4376,33	3416,13	3639,11	4404,63	3411,46	3428,35	**3412,88**	3431,19	**3408,68**
	HQC	4940,20	4386,25	3427,71	3649,04	4414,56	3423,03	3439,93	**3426,11**	3442,77	**3418,26**
	BIC	4947,88	4401,62	3445,63	3664,40	4429,92	3440,96	3457,85	**3446,60**	3460,70	**3436,19**

Espera-se que a diagonal apresente sempre indicativos de um bom ajuste, uma vez que representa a situação onde o modelo correto foi ajustado. Ou seja, o modelo originário é o mesmo que o ajustado. Com isso em mente percebe-se que únicas situações onde a diagonal não pertence ao grupo dos bons ajustes são os modelos ZIB e ZIBB, ambos provenientes da distribuição binomial.

Na contramão, o modelo que mais recebeu indicação de melhor ajuste foi o ZID, que além de ter tido a menor média dos critérios de qualidade de ajuste para os dados provenientes dessa mesma distribuição, também obteve a menor média com os dados simulados pelas distribuições ZINB,ZIPIG ,ZIBB e ZICMP.

Em seguida com três indicações de melhor ajuste surgem os modelos ZINB, ZIPIG, ZIBNB e ZICMP, todos modelos flexíveis com grande cobertura de índice de sobredispersão.

Na Tabela 5 abaixo estão expostas as proporções de vezes que dentre os 1000 ajustes o modelo empregado é dono da menor medida de HQC segundo a origem dos dados. Desta maneira espera-se similarmente que a diagonal *(em negrito)* contenha as maiores proporções, uma vez que é aguardado que o ajuste do modelo correto forneça uma alta taxa de melhor adequamento.

Casos onde a diagonal não representa a maior proporção de HQC mínimo estão indicados em vermelho. Importante notar que ZICMP apresenta três indicações de maior proporção de menor HQC. Outros modelos que conseguem apresentar uma proporção superior ao modelo de origem são ZINB, ZIB e ZIPIG.

Verifica-se que mais uma vez os modelos mais flexíveis conseguem se adequar bem a dados que são oriundos de distribuições simples, por exemplo ZINB, ZIPIG e ZICMP representam mais de 50% dos ajustes de menor HQC, ao passo que ZIP forneceu o melhor ajuste apenas 4% das vezes e Poisson 0%.

Tabela 5: Porcentagem de modelos com HQC mínimo

	POIS	**ZIP**	**ZINB**	**ZIG**	**ZIB**	**ZIPIG**	**ZIBB**	**ZIBNB**	**ZICMP**	**ZID**
ZIP	0,00	**0,41**	0,00	0,00	**0,49**	0,00	0,00	0,04	0,02	0,02
ZINB	0,00	0,00	**0,64**	0,00	0,00	0,14	0,03	0,08	0,02	0,09
ZIG	0,00	0,00	0,07	**0,90**	0,00	0,00	0,01	0,00	0,00	0,02
ZIB	0,00	0,00	0,00	0,00	**0,11**	0,00	0,00	0,00	**0,88**	0,00
ZIPIG	0,00	0,00	0,12	0,00	0,00	**0,65**	0,01	0,01	0,00	0,21
ZIBB	0,00	0,00	0,00	0,00	0,00	0,00	**0,14**	0,01	**0,84**	0,01
ZIBNB	0,00	0,00	**0,34**	0,00	0,00	**0,51**	0,01	**0,05**	0,00	0,09
ZICMP	0,00	0,00	0,10	0,00	0,00	0,03	0,23	0,28	**0,31**	0,05
ZID	0,00	0,00	0,05	0,00	0,00	0,15	0,08	0,06	0,00	**0,67**

5 Conclusões

As simulações evidenciam a flexibilidade de cada modelo, tanto na simulação de dados como no ajuste destes. Obviamente espera-se melhores resultados quando se ajusta aos dados o modelo que de fato gere o processo de contagem da população, mas esse modelo é tão desconhecido quanto os próprios parâmetros, logo o processo de ajuste, antes de passar pela estimação de parâmentros, requer a definição de modelo apropriado.

Dos modelos aqui abordados e brevemente testados se fortificam evidências da sobrepujança de alguns sobre outros. É possível dizer que a inclusão da inflação de zeros no modelo de fato é crucial quando se está lidando com dados desta natureza, bem como atenção a sobredispersão é fundamental. É recomendado testes mais exaustivos, trabalhando com diferentes graus de sobredispersão em cada modelo e buscar táticas mais adequedas para obtenção de estimativas, pois conforme mostrado por Silva (2017), para certas configurações de parâmetros e tamanho amostral, a maximização sem emprego do EM representa um grande risco.

Chama-se a atenção para modelos tradicionais que incorporam sobredispersão como PIG e NB, que ao serem inflados de zero passam não somente a absorver a sobredispersão comum, mas também a modelar os zeros estruturais que reduzem a média e elevam a variância dos dados, o que os torna ainda mais versáteis.

Houveram boas surpresas com as distribuições discretas Delaporte e sobretudo CMP, que neste estudo apresentaram desempenho tão bom quanto, ou melhor, que NB e PIG frente a casos de sobredispersão. Suas derivações infladas de zeros ZID e ZICMP performaram muito bem mais uma vez e se posicionaram paralelamente ou a frente dos modelos ZINB e ZIPIG. Mais estudos e simulações são requeridos, pois a utilização destas distribuições para análise de dados de contagem é escassa.

Negativamente menciona-se a distribuição ZIBNB, que além de superparametrizada, apresentou resultados similares ao ZIB e ZIBB. Além disso menciona-se peculiaridade do modelo ZIG cujos dados não foram bem ajustados por nenhum modelo além do próprio ZIG, sendo estes dados um dos poucos que o modelo ajusta de maneira satisfatória. À cerca do ZIP, conforme esperado ele apenas desempenhou bem com dados que além da inflação de zeros eram equidispersos, já o modelo Poisson tradicional desempenhou mal em todos os cenários, o que corrobora com a atenção necessária que devemos ter com a sobredispersão e inflação de zeros, uma vez que regressão de Poisson costuma ser tomada como procedimento padrão frente a dados de contagem.

6 Referências

CONWAY, R.W. and MAXWELL, W.L.*A queuing model with state dependent service rates*. J. Ind. Eng. 12, 132–136.1962.

COX, D. R.,LEWIS, P. A. W.*The Statistical Analysis of Series of Events*. 1966

DELAPORTE, P.J. *Quelques problèmes de statistiques mathématiques poses par l'Assurance Automobile et le Bonus pour non sinistre [Some problems of mathematical statistics as related to automobile insurance and no-claims bonus]*. Bulletin Trimestriel de l'Institut des Actuaires Français (in French). 1960. 87–102 p.

HAIGHT, F.A. *Handbook of the Poisson Distribution*. New York: John Wiley & Sons, 1967.

HAL, D.B. , *Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study*, Department of Statistics, University of Georgia. 2000

HEILBRON, D.C. *Generalized linear models for altered zero probabilities and overdispersion in count data*. SIMS Technical Report 9, Department of Epidemiology and Biostatistics, University of California, San Francisco. 1989.

JENNRICH, R. I., and SAMPSON, P.F. *Newton-Raphson and related algorithms for maximum likelihood variance component estimation*. Technometrics, 18. 1976. 11-17 p.

LAMBERT, D. *Zero-Inflated Poisson Regression, With An Application to Defects in Manufacturing*, 1992.

NELDER, J.A. and WEDDERBURN, R.W.M. *Generalized Linear Models*. Journal of the Royal Statistical Society. 1972.

PAULA, G.P. *Modelos de Regressão com apoio Computacional*

RIDOUT, M. S., DEMÉTRIO, C.G.B. and HINDE, J.P. *Models for count data with many zeros*. 1998.

SELLERS, K.F. and RAIM, A. *A flexible zero-inflated model to address data dispersion*, Computational Statistics and Data Analysis. 2016.

SELLERS, K.F., SHMUELI, G. and BORLE, S. *The COM-Poisson model for count data: a survey of methods and applications*. Appl. Stoch. Models Bus. Ind. 28. 2011. 104–116 p.

SILVA, J.G. *Zero-Inflated Mixed Poisson Regression Models*. 2017

SIN, C. Y., WHITE, H. *Information criteria for selecting possibly misspecified parametric models*. Journal of Econometrics, 71(1), 1996. 207-225.

WANG, Z. *One mixed negative binomial distribution with application*. Journal of Statistical Planning and Inference. 2011.

WILLMOT, G.E. *The Poisson-Inverse Gaussian distribution as an alternative to the negative binomial*, Scandinavian Actuarial Journal, DOI: 10.1080/03461238.1987.10413823. 1987.

VIEIRA, A. M. C., HINDE, J. P., DEMETRIO, C. G. B. *Zero-inlated proportion data models applied to a biological control assay*. Journal of Applied Statistics 27(3), 2000. 373-389.

RIDOUT, M., HINDE, J., DEMETRIO, C. G. B. *A score test for testing a zero-inflated Poisson regression model against zero-inlated negative binomial alternatives*. Biometrics 57(1), 2001. 219-223.