UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Modelagem de Dados de Contagem com Inflação de Zeros

Teoria, Diagnóstico e Aplicação

Márcio Antônio Vieira Batista

Lourdes C. C. Montenegro

Belo Horizonte
2026

Sumário

1 Introdução **4**

2 Fundamentos de Dados de Contagem **6**

2.1 Distribuição de Poisson 6

2.2 Modelagem Generalizada 7

2.2.1 Modelo de Regressão Poisson 7

2.2.2 Modelo de Regressão Geométrico 7

2.2.3 Modelo de Regressão Binomial Negativo 7

2.3 Limitações dos Modelos Clássicos 8

2.4 Inflação de Zeros 8

2.4.1 Tipos de Zeros 8

2.4.2 Implicações para a Modelagem 8

3 Modelos Inflacionados de Zeros **10**

3.1 Estrutura Geral dos Modelos Zero-Inflados 10

3.2 Modelo Poisson Zero-Inflado (ZIP) 10

3.2.1 Definição do Modelo 11

3.2.2 Função de Probabilidade 11

3.2.3 Probabilidade de zero 12

3.2.4 Esperança e Variância 12

3.2.5 Estimação 12

3.3 Modelo Geométrico Zero-Inflado (ZIG) 15

3.3.1 Definição do Modelo 15

3.3.2 Função de Probabilidade 15

3.3.3 Probabilidade de zero 16

3.3.4 Esperança e variância 16

3.3.5 Estimação 17

3.4 Modelo Binomial Negativo Zero-Inflado (ZINB) 18

3.4.1 Definição do Modelo 19

3.4.2 Função de probabilidade 19

3.4.3 Probabilidade de zero 20

3.4.4 Esperança e variância 20
3.4.5 Estimação 20
3.5 Modelo Poisson Generalizada Zero-Inflado (ZIGP) 22
3.6 Definição do Modelo 22
3.6.1 Função de probabilidade 23
3.6.2 Probabilidade de zero 23
3.6.3 Esperança e variância 24
3.6.4 Estimação 24
3.7 Comparação Teórica entre os Modelos 26
3.7.1 ZIP: O Modelo Básico 26
3.7.2 ZIG: Variância Quadrática Fixa 26
3.7.3 ZINB: Variância Flexível 26
3.7.4 ZIGP: Máxima Flexibilidade 27

4 Interpretação dos Parâmetros e Inferência 28
4.1 Componente de Contagem (μ) 28
4.1.1 Razão de Taxas (Rate Ratio - RR) 28
4.2 Componente de Inflação (π) 28
4.2.1 Razão de Chances (Odds Ratio - OR) 29
4.3 Interpretação Conjunta 29
4.4 Testes de Significância e Intervalos de Significância 29
4.4.1 Matriz de Informação: Observada vs. Esperada 29
4.4.2 Intervalos de Confiança: Wald vs. LRT 30

5 Diagnóstico de Modelos (`DHARMa`) 31
5.1 Limitações dos resíduos tradicionais 31
5.2 Resíduos Quantílicos Randomizados (RQR) 32
5.3 Interpretação dos os Gráficos do `DHARMa` 32
5.3.1 QQ Plot Residuals 32
5.3.2 Resíduos vs. Valores Ajustados 32
5.4 Vantagens do Diagnóstico via Simulação 33

6 Seleção e Comparação de Modelos 34
6.1 Comparação de Modelos 34
6.1.1 Teste de Razão de Verossimilhança (LRT) 34
6.1.2 Teste de Vuong 35
6.2 Critérios de Informação 36
6.2.1 Critério de Informação de Akaike (AIC) 36
6.2.2 Critério de Informação Bayesiano (BIC) 36
6.3 Comparação entre Modelos Inflacionados (ZIP vs ZINB) 36

6.4 Avaliação de Resíduos com DHARMa 37

7 Aplicação I — Modelo ZIP (bioChemists) 38
7.1 Descrição dos dados 38
7.2 Análise descritiva 39
7.3 Ajuste 39
7.4 Diagnóstico 40
7.5 Resultados 41
7.6 Conclusão 42

8 Aplicação II — Modelo ZIG (fishing) 43
8.1 Descrição dos dados 43
8.2 Análise descritiva 43
8.3 Ajuste 44
8.4 Diagnóstico 45
8.5 Resultados 45
8.6 Conclusão 46

9 Aplicação III — Modelo ZINB (Student Performance) 48
9.1 Descrição dos dados 48
9.2 Análise descritiva 48
9.3 Ajuste 49
9.4 Diagnóstico 50
9.5 Resultados 51
9.6 Conclusão 52

10 Aplicação IV — Modelo ZIGP (Salamanders) 53
10.1 Descrição dos dados 53
10.2 Análise descritiva 53
10.3 Ajuste 54
10.4 Diagnóstico 54
10.5 Resultados 55
10.6 Conclusão 56

11 Conclusão Geral 58

Agradecimentos 59

Capítulo 1

Introdução

A análise de dados de contagem constitui um pilar fundamental da estatística aplicada, sendo amplamente utilizada para descrever a frequência de ocorrência de eventos em um determinado intervalo de tempo, espaço ou unidade de observação. Desde a modelagem do número de sinistros em apólices de seguro até a contagem de espécies em ecologia, a correta especificação do processo gerador desses dados é essencial para a obtenção de inferências válidas e precisas.

Historicamente, os Modelos Lineares Generalizados (MLG), propostos por Nelder e Wedderburn (1972), estabeleceram o arcabouço metodológico padrão para lidar com variáveis respostas discretas. Nesse contexto, a distribuição de Poisson desponta como a principal escolha teórica. Contudo, a aplicação prática do modelo de Poisson frequentemente esbarra em uma de suas propriedades mais restritivas: a equidispersão, que impõe a igualdade entre a média e a variância teóricas. Em dados reais, especialmente nas ciências biológicas e sociais, é comum observar que a variabilidade dos dados supera significativamente a média estipulada pelo modelo, fenômeno conhecido como sobredispersão.

Embora distribuições alternativas, como a Binomial Negativa, ofereçam maior flexibilidade para acomodar a sobredispersão geral, elas frequentemente se mostram insuficientes perante uma anomalia específica e muito recorrente: o excesso de zeros. Em muitos fenômenos reais, a frequência de observações nulas é substancialmente maior do que a massa de probabilidade atribuída por modelos clássicos. Esse fenômeno sugere que os dados não são provenientes de um processo homogêneo, mas sim da coexistência de múltiplos mecanismos geradores. Alguns zeros podem representar a mera ausência momentânea do evento (zeros amostrais), enquanto outros refletem uma impossibilidade estrutural de sua ocorrência (zeros estruturais).

A negligência dessa distinção na modelagem estatística resulta em estimativas enviesadas, erros padrão distorcidos e conclusões equivocadas. Para contornar essa limitação, a literatura estatística desenvolveu os modelos inflacionados de zeros, popularizados por trabalhos seminais como os de Lambert (1992) e Greene (1994). Esses modelos de mistura combinam uma distribuição degenerada no zero com uma distribuição de contagem

padrão, permitindo modelar simultaneamente a probabilidade de o evento não ocorrer de forma absoluta e a intensidade de sua ocorrência entre os indivíduos suscetíveis.

Neste cenário, o presente trabalho tem como objetivo explorar, de forma integrada, a teoria, o diagnóstico e a aplicação de modelos de contagem com inflação de zeros. A análise abrange desde a extensão do modelo Poisson (**ZIP**), passando por formulações mais robustas para caudas pesadas, como o Geométrico (**ZIG**) e o Binomial Negativo (**ZINB**), até a elevada flexibilidade do modelo Poisson Generalizado (**ZIGP**).

Além da formulação matemática, este estudo confere especial atenção aos desafios computacionais e ao diagnóstico de ajuste. Dada a limitação dos resíduos clássicos para dados discretos inflacionados, destaca-se o uso de resíduos quantílicos randomizados Dunn e Smyth (1996), operacionalizados via simulação no ambiente R R Core Team (2023), o que permite uma validação mais rigorosa das premissas assumidas.

Para tanto, além desta introdução, o trabalho está estruturado da seguinte forma: o segundo capítulo revisa os fundamentos teóricos dos dados de contagem e detalha as limitações estruturais dos modelos clássicos. No terceiro capítulo, os modelos de mistura com inflação de zeros (**ZIP**, **ZIG**, **ZINB** e **ZIGP**) são formalmente introduzidos, dissecando suas propriedades estatísticas e o processo de estimação via algoritmo EM. Posteriormente, os capítulos quatro, cinco e seis abordam, respectivamente, a interpretação paramétrica e inferência, as técnicas de diagnóstico baseadas em simulação (`DHARMa`), e os critérios para seleção e comparação metodológica. Por fim, toda a teoria desenvolvida é consolidada por meio de quatro aplicações práticas a conjuntos de dados reais, demonstrando a adequação dos modelos frente a diferentes arranjos de dispersão, finalizando o estudo com as conclusões gerais.

Capítulo 2

Fundamentos de Dados de Contagem

A análise de dados de contagem desempenha um papel fundamental em diversas áreas do conhecimento. Sob essa ótica, ao trabalharmos com amostras independentes e identicamente distribuídas (i.i.d.), torna-se possível realizar inferências assumindo que a população de origem segue uma distribuição de probabilidade específica.

2.1 Distribuição de Poisson

O modelo de referência primordial é a Distribuição de Poisson, definida pela seguinte função de probabilidade:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

Uma característica central dessa distribuição é sua função geradora de momentos (FGM), dada por:

$$M(t) \coloneqq E[e^{tY}] = \exp\{\lambda(e^{t} - 1)\}$$

Para a obtenção do k-ésimo momento ordinário, derivamos $M(t)$ em relação a t por k vezes, avaliando o resultado em $t = 0$. A partir desse processo, estabelecem-se os quatro primeiros momentos:

$$\begin{aligned}
\mu_1' &= \lambda \\
\mu_2' &= \lambda + \lambda^2 \\
\mu_3' &= \lambda + 3\lambda^2 + \lambda^3 \\
\mu_4' &= \lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4
\end{aligned}$$

Um ponto crítico na aplicação deste modelo é a propriedade de equidispersão, na qual

o valor esperado e a variância são idênticos ($E[Y] = Var[Y] = \lambda$). Contudo, essa premissa é frequentemente violada em dados biológicos e sociais, onde a variabilidade dos dados supera a média.

2.2 Modelagem Generalizada

É comum observar que os parâmetros das distribuições apresentam dependência em relação a covariáveis observadas. Quando essa dependência pode ser modelada de forma paramétrica, os Modelos Lineares Generalizados (MLG) Nelder e Wedderburn (1972) constituem a principal ferramenta para relacionar a média da variável resposta às covariáveis explicativas.

2.2.1 Modelo de Regressão Poisson

Seja, para $i = 1, \ldots, n$, a variável resposta y_i associada ao vetor de covariáveis $\mathbf{x}_i$. O modelo de Poisson assume que $y_i \mid \mathbf{x}_i$

$$f(y_i \mid \mathbf{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

A média μ_i é relacionada ao preditor linear por meio da função de ligação logarítmica:

$$\log(\mu_i) = \mathbf{x}_i^\top \beta$$

2.2.2 Modelo de Regressão Geométrico

No modelo de regressão geométrico, a distribuição é parametrizada em termos de sua média μ_i. A função de densidade é dada por:

$$f(y_i \mid \mathbf{x}_i) = \frac{\mu_i^{y_i}}{(1+\mu_i)^{y_i+1}}, \quad y_i = 0, 1, 2, \ldots$$

A estrutura para a média segue a forma log-linear:

$$\log(\mu_i) = \mathbf{x}_i^\top \beta$$

2.2.3 Modelo de Regressão Binomial Negativo

O modelo Binomial Negativo introduz um parâmetro adicional de dispersão k. A função de densidade é expressa por:

$$f(y_i \mid \mathbf{x}_i) = \frac{\Gamma(y_i + k)}{\Gamma(k) y_i!} \left(\frac{k}{\mu_i + k}\right)^k \left(\frac{\mu_i}{\mu_i + k}\right)^{y_i}, \quad y_i = 0, 1, 2, \ldots$$

A média μ_i é modelada via função de ligação log:

$$\log(\mu_i) = \mathbf{x}_i^\top \beta$$

2.3 Limitações dos Modelos Clássicos

Embora os modelos clássicos de contagem, como Poisson e Binomial Negativo, sejam amplamente utilizados, sua aplicação em dados reais frequentemente revela limitações importantes. Uma das mais recorrentes é a presença de excesso de zeros, isto é, uma frequência de observações iguais a zero significativamente maior do que a prevista por esses modelos. Esse fenômeno ocorre em diversas áreas, como saúde pública, economia e ciências ambientais, e indica que a estrutura dos dados pode ser mais complexa do que aquela assumida pelos modelos tradicionais.

2.4 Inflação de Zeros

A inflação de zeros pode ser compreendida a partir da própria natureza do que está sendo medido.

Um exemplo prático ocorre em pesquisas sobre o consumo de produtos específicos, como o cigarro. Em uma amostra, a presença de zeros pode decorrer de dois grupos distintos: há quem fume, mas não tenha fumado no dia da pesquisa, e há o grupo dos não-fumantes. Estes últimos representam os chamados zeros estruturais, pois sua chance de consumo é nula por uma questão de hábito ou escolha de vida. Quando o modelo ignora essa distinção e trata todos os zeros como se fossem a mesma coisa, a estimativa acaba distorcida, o que justifica o uso de modelos que tratam essa massa de dados excedente separadamente.

2.4.1 Tipos de Zeros

Para compreender esse problema, é fundamental distinguir dois tipos de zeros:

- **Zeros Amostrais:** correspondem a situações em que o evento poderia ter ocorrido, mas não foi observado durante o período de análise.
- **Zeros Estruturais:** correspondem a situações em que a unidade experimental não possui possibilidade de gerar o evento.

2.4.2 Implicações para a Modelagem

Quando modelos tradicionais ignoram que os zeros podem ter origens distintas e tratam todas essas observações como resultado de um único processo, ocorre uma distorção

na estimação dos parâmetros, o que pode comprometer a qualidade do ajuste e a interpretação dos resultados.

Essa limitação motiva o desenvolvimento de modelos mais flexíveis, capazes de capturar explicitamente a presença de múltiplos processos geradores de dados, como os modelos com inflação de zeros, que serão apresentados na próxima parte.

Capítulo 3

Modelos Inflacionados de Zeros

3.1 Estrutura Geral dos Modelos Zero-Inflados

Os modelos de contagem inflacionados de zeros constituem uma classe de modelos de mistura que combinam dois processos geradores distintos: um responsável pela geração de zeros estruturais e outro associado a uma distribuição de contagem tradicional.

A ideia central é que a variável resposta pode ser gerada por dois mecanismos:

- um processo degenerado, que produz apenas zeros; um processo de contagem, que pode gerar zeros e valores positivos.
- um processo de contagem, que pode gerar zeros e valores positivos.

A função de probabilidade geral pode ser expressa como:

$$P(Y = y) = \begin{cases} \pi + (1 - \pi)f(0), & \text{se } y = 0 \\ (1 - \pi)f(y), & \text{se } y > 0 \end{cases}$$

onde:

- $\pi \in (0, 1)$ representa a probabilidade de ocorrência de um zero estrutural;
- $f(\cdot)$ é a função de probabilidade de uma distribuição de contagem com suporte em $\mathbb{N} \cup \{0\}$.

3.2 Modelo Poisson Zero-Inflado (ZIP)

Nesta seçâo, apresenta-se o modelo **Poisson Zero-Inflado (ZIP)**, proposto originalmente por Lambert (1992) em seu trabalho *"Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing"*. O modelo surge como uma extensão natural para dados de contagem que apresentam uma frequência de zeros superior à esperada por

um processo de Poisson simples, permitindo a modelagem de fenômenos onde a população é composta por dois subgrupos distintos: um que gera exclusivamente zeros e outro cujas contagens seguem uma distribuição de Poisson.

3.2.1 Definição do Modelo

Seja Y_i a variável resposta associada à i-ésima observação, com $i = 1, \ldots, n$.
O modelo assume que:

$$Y_i|\mathbf{X_i}, \mathbf{Z_i} \sim ZIP(\mu_i, \pi_i)$$

onde:

- $\mathbf{X_i}$ são as covariáveis associadas ao processo de contagem;
- $\mathbf{Z_i}$ são as covariáveis associadas ao processo de inflação de zeros;

3.2.2 Função de Probabilidade

Componente de Contagem

A componente de contagem é descrita por uma distribuição de Poisson::

$$f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots,$$

onde μ_i é a média da contagem, geralmente modelada por uma função de ligação logarítmica tal que $\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$. Nesse modelo, $\mathbf{x}_i$ representa o vetor de variáveis explicativas e $\boldsymbol{\beta}$ é o vetor de parâmetros (ou coeficientes) de regressão a serem estimados, que indicam a magnitude e a direção do efeito de cada variável sobre o logaritmo da média.

Estrutura de Mistura

A função de probabilidade do modelo **ZIP** é expressa como uma combinação linear de uma massa pontual em zero e a distribuição Poisson definida anteriormente. A função de probabilidade conjunta é dada por:

$$f(y_i \mid \mathbf{x}_i, \mathbf{z}_i) = \pi(\mathbf{z}_i, \boldsymbol{\psi}) \cdot \mathbf{I}(y_i = 0) + \big[1 - \pi(\mathbf{z}_i, \boldsymbol{\psi})\big] \cdot f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}).$$

onde:

- $\mathbf{I}(y_i = 0)$ é a função indicadora, que assume valor 1 se $y_i = 0$ e 0 caso contrário;

- $\pi_i = \pi(\mathbf{z}_i, \boldsymbol{\psi})$ representa a probabilidade de a observação pertencer ao grupo dos zeros estruturais. Para garantir que esteja contida no intervalo $(0, 1)$, utiliza-se usualmente a função de ligação *logit*, definida por:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i^\top \boldsymbol{\psi}$$

- $\boldsymbol{\beta}$ e $\boldsymbol{\psi}$ são os vetores de parâmetros a serem estimados.

3.2.3 Probabilidade de zero

Vale destacar algumas particularidades da distribuição **Poisson Zero-Inflada**. Em particular, a probabilidade de observar o valor zero pode ser decomposta da seguinte forma:

$$P(Y_i = 0) = \underbrace{\pi_i}_{\substack{\text{Zero} \\ \text{Estrutural}}} + \underbrace{(1-\pi_i) \cdot \exp(-\mu_i)}_{\substack{\text{Zero} \\ \text{Amostral}}}$$

A probabilidade pode ser decomposta em duas componentes: os zeros estruturais, provenientes de um processo degenerado que gera apenas valores nulos, e os zeros amostrais, oriundos do próprio processo de contagem.

3.2.4 Esperança e Variância

Pode-se verificar que a esperança e a variância condicionais da variável Y_i são dadas por:

$$E[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1-\pi_i)\mu_i$$

$$Var[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1-\pi_i)\mu_i \cdot (1+\mu_i\pi_i)$$

Note que a variância é maior que a média, evidenciando a capacidade do modelo em capturar sobredispersão.

3.2.5 Estimação

A estimação dos parâmetros do modelo de **Poisson Zero-Inflado (ZIP)** é realizada, predominantemente, por meio do método de **Máxima Verossimilhança (MLE)**. Considerando os vetores de parâmetros $\boldsymbol{\beta}$ (associado à média Poisson μ) e $\boldsymbol{\psi}$ (associado à probabilidade de inflação de zeros, π), assume-se que a variável Y_i pode ser gerada por dois processos: um estado de zero estrutural, com probabilidade π_i, ou um processo de contagem Poisson (μ_i), com probabilidade $1 - \pi_i$.

Log-verossimilhança

A função de log-verossimilhança para o modelo **ZIP**, assumindo que os parâmetros $\boldsymbol{\psi}$ e $\boldsymbol{\beta}$ são distintos e não estão funcionalmente relacionados, pode ser expressa como:

$$\ell(\boldsymbol{\psi}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{y_i=0} \log\left(e^{\mathbf{z}_i^\top \boldsymbol{\psi}} + \exp(-e^{\mathbf{x}_i^\top \boldsymbol{\beta}})\right) + \sum_{y_i>0} (y_i \mathbf{x}_i^\top \boldsymbol{\beta} - e^{\mathbf{x}_i^\top \boldsymbol{\beta}}) - \sum_{i=1}^{n} \log(1 + e^{\mathbf{z}_i^\top \boldsymbol{\psi}}) - \sum_{y_i>0} \log(y_i!).$$

onde $\mathbf{X}$ e $\mathbf{Z}$ denotam as matrizes de covariáveis associadas, respectivamente, aos processos de contagem e de inflação de zeros. A maximização direta desta função apresenta elevada complexidade numérica devido ao termo logarítmico nas observações iguais a zero. Essa dificuldade surge porque não se conhece, *a priori*, a origem de cada valor zero: se provém do estado estrutural ou da distribuição de contagem (zero amostral).

Variável Latente

Para contornar essa limitação, introduz-se uma variável latente W_i, tal que:

- $W_i = 1$: a observação pertence ao estado de inflação (zero estrutural);
- $W_i = 0$: a observação é gerada pelo processo de contagem.

Caso essa variável fosse observável, a log-verossimilhança dos dados completos, denotada por (ℓ_c), poderia ser decomposta como:

$$\ell_c(\boldsymbol{\psi}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{w}) = \ell_c(\boldsymbol{\psi}; \mathbf{y}, \mathbf{w}) + \ell_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{w}) - \mathrm{C}$$

onde C é uma constante que não depende dos parâmetros.

Essa decomposição torna o problema de estimação significativamente mais simples, permitindo a estimação separada dos parâmetros ψ e β.

Algoritmo EM

Dada a natureza não observável da variável W_i, Lambert (1992) propôs o uso do algoritmo EM. Trata-se de um procedimento iterativo que busca obter os estimadores de máxima verossimilhança por meio da alternância entre dois passos:

1. **Passo E (Expectation):** Neste passo, calcula-se o valor esperado da variável latente W_i, isto é, a probabilidade a *posteriori* de que a observação pertença ao estado de inflação, dados os estimadores atuais $\boldsymbol{\psi}^{(k)}$ e $\boldsymbol{\beta}^{(k)}$.
    - Se $y_i > 0$, então $W_i^{(k)} = 0$.

- Se $y_i = 0$, calcula-se a probabilidade de o zero ser proveniente do estado de inflação:

$$W_i^{(k)} = \left[1 + \exp(-\mathbf{z}_i^\top \boldsymbol{\psi}^{(k)} - \exp(\mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}))\right]^{-1}$$

2. **Passo M (Maximization):** Com os pesos $W_i^{(k)}$ fixados, maximiza-se a log-verossimilhança esperada:

   - Para $\boldsymbol{\beta}^{(k+1)}$: realiza-se uma **regressão Poisson log-linear ponderada**, utilizando como pesos $(1 - W_i^{(k)})$.
   - Para $\boldsymbol{\psi}^{(k+1)}$: realiza-se uma **regressão logística ponderada**, considerando $W_i^{(k)}$ como a variável resposta esperada.

Convergência e Desafios Computacionais

O algoritmo EM apresenta propriedades robustas para a estimação do modelo **ZIP**:

- **Estabilidade Numérica:** o algoritmo é monótono, garantindo que a função de verossimilhança não decresça a cada iteração, o que assegura convergência para um ponto estacionário.
- **Escolha de valores iniciais:** A convergência tende a ser eficiente quando se utilizam bons valores iniciais; em particular, estimativas provenientes de uma regressão de Poisson truncada em zero fornecem um ponto de partida adequado para o vetor $\boldsymbol{\beta}$.
- **Comparação com métodos baseados em gradiente:** Comparação com métodos baseados em gradiente: embora algoritmos como Newton-Raphson apresentem convergência quadrática nas proximidades do ótimo, eles podem ser sensíveis à escolha dos valores iniciais. O algoritmo EM, por sua vez, tende a ser mais estável e de implementação mais direta.

Apesar dessas vantagens, alguns desafios práticos podem surgir:

- Separação na componente logística: em amostras finitas, pode ocorrer separação completa ou quase completa. Nesses casos, quando os zeros estruturais são perfeitamente explicados por uma combinação de covariáveis, as estimativas dos parâmetros $\boldsymbol{\psi}$ podem divergir.
- Questões numéricas: a implementação computacional requer cuidado no cálculo das probabilidades, sendo recomendável o uso de técnicas numericamente estáveis, como a função log-sum-exp, para evitar problemas de overflow e underflow.

3.3 Modelo Geométrico Zero-Inflado (ZIG)

Neste capítulo, tratamos do modelo **Geométrico Zero-Inflado (ZIG)**. Embora a distribuição geométrica seja um caso particular da Binomial Negativa (onde o parâmetro de dispersão é igual a 1).

O uso de distribuições com caudas mais pesadas como alternativa para lidar com excesso de zeros remonta a trabalhos como os de Heilbron (1994) e Long (1997). Esses estudos evidenciam que a estrutura geométrica oferece uma abordagem parsimoniosa e adequada para modelar processos caracterizados por "tentativas até o sucesso", especialmente quando há a possibilidade de interrupção por um estado de inatividade estrutural.

3.3.1 Definição do Modelo

Seja Y_i a variável resposta para a i-ésima observação, para $i = 1, \ldots, n$. Assume-se que Y_i segue uma distribuição **Geométrica Zero-Inflada** condicionada aos vetores de covariáveis $\mathbf{x}_i$ (associado ao processo de contagem) e $\mathbf{z}_i$ (associados ao processo de inflação).

3.3.2 Função de Probabilidade

Componente de Contagem

Considera-se, inicialmente, a função de probabilidade da distribuição geométrica padrão, denotada por $f_2(\cdot)$, a qual descreve o processo de contagem para a subpopulação não associada ao estado de zero estrutural:

$$f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}) = \frac{\mu_i^{y_i}}{(1+\mu_i)^{y_i+1}}, \quad y_i = 0, 1, 2, \ldots,$$

onde μ_i representa a média da contagem, modelada por uma função de ligação logarítmica.

$$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Estrutura de Mistura

A distribuição **ZIG** pode ser representada como uma combinação linear entre uma massa pontual em zero e a distribuição Geométrica definida anteriormente. Sua função de probabilidade é dada por:

$$f(y_i \mid \mathbf{x}_i, \mathbf{z}_i) = \pi(\mathbf{z}_i, \boldsymbol{\psi}) \cdot \mathbf{I}(y_i = 0) + \big[1 - \pi(\mathbf{z}_i, \boldsymbol{\psi})\big] \cdot f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta})$$

onde:

- $\mathbf{I}(y_i = 0)$ é a função indicadora, que assume valor 1 se $y_i = 0$ e 0 caso contrário;

- $\pi_i = \pi(\mathbf{z}_i, \boldsymbol{\psi})$ representa a probabilidade de a observação pertencer ao grupo de zeros estruturais. Para garantir que essa probabilidade π_i esteja contida no intervalo $(0, 1)$, utiliza-se a função de ligação *logit*, definida por:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i^\top \boldsymbol{\psi}$$

- $f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta})$ corresponde à função de probabilidade da distribuição geométrica associada ao processo de contagem;

- $\boldsymbol{\beta}$ e $\boldsymbol{\psi}$ são os vetores de parâmetros a serem estimados.

3.3.3 Probabilidade de zero

Diferente do modelo Poisson, a distribuição geométrica já atribui uma probabilidade maior ao valor zero. No contexto do modelo **ZIG**, a probabilidade de observar uma contagem nula é dada por:

$$P(Y_i = 0) = \underbrace{\pi_i}_{\substack{\text{Zero} \\ \text{Estrutural}}} + \underbrace{(1-\pi_i) \cdot \left(\frac{1}{1+\mu_i}\right)}_{\substack{\text{Zero} \\ \text{Amostral}}}$$

Essa expressão evidencia que a probabilidade de zero é composta por duas contribuições:

- zeros estruturais, representados por π_i;

- zeros amostrais, provenientes da distribuição geométrica, representados por $(1-\pi_i) \cdot (1/1+\mu_i)$.

3.3.4 Esperança e variância

Além disso, a esperança e a variância condicionais da variável Y_i são dadas por:

$$E[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1-\pi_i)\mu_i$$

$$Var[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1-\pi_i)\mu_i \cdot (1+\mu_i+\pi_i\mu_i)$$

Observa-se que a variância excede a média, refletindo a capacidade do modelo em acomodar sobredispersão, tanto pela estrutura geométrica quanto pela presença de inflação de zeros.

3.3.5 Estimação

A estimação dos parâmetros do modelo **Geométrico Zero-Inflado (ZIG)** é realizada, em geral, por meio do método de **Máxima Verossimilhança (MLE)**. O modelo assume que cada observação Y_i pode ser gerada por um dos dois processos: um estado de zero estrutural, com probabilidade π_i, ou um processo de contagem que segue uma distribuição geométrica com média μ_i, com probabilidade $1 - \pi_i$.

Log-verossimilhança

A função de log-verossimilhança do modelo **ZIG** pode ser expressa como:

$$\ell(\boldsymbol{\psi}, \boldsymbol{\beta}; \mathbf{y}) = \sum_{y_i=0} \log\left(e^{\mathbf{z}_i^\top \boldsymbol{\psi}} + \frac{1}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}\right) + \sum_{y_i>0} \left[y_i \mathbf{x}_i^\top \boldsymbol{\beta} - (y_i + 1)\log(1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}})\right] - \sum_{i=1}^{n} \log(1 + e^{\mathbf{z}_i^\top \boldsymbol{\psi}})$$

A maximização direta dessa função apresenta desafios numéricos, principalmente devido à dificuldade de identificar a origem das observações iguais a zero, isto é, se são provenientes do estado estrutural ou do processo de contagem.

Variável Latente

Para contornar essa limitação, introduz-se uma variável latente W_i, tal que:

- $W_i = 1$: a observação pertence ao estado de inflação (zero estrutural);
- $W_i = 0$: a observação é gerada pelo processo de contagem.

Caso essa variável fosse observável, a log-verossimilhança dos dados completos, denotada por (ℓ_c), poderia ser decomposta como:

$$\ell_c(\boldsymbol{\psi}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{w}) = \ell_c(\boldsymbol{\psi}; \mathbf{y}, \mathbf{w}) + \ell_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{w}) - \mathrm{C}$$

onde C é uma constante que não depende dos parâmetros.

Essa decomposição torna o problema de estimação significativamente mais simples, permitindo a estimação separada dos parâmetros ψ e β.

Algoritmo EM

O processo de estimação via algoritmo EM para o modelo **ZIG** consiste na alternância entre dois passos iterativos:

1. **Passo E (Expectation):** Neste passo, estima-se a probabilidade *a posteriori* de que a observação pertença ao estado de inflação, isto é, $W_i = 1$ dados os estimadores atuais:

   - Se $y_i > 0$, então $W_i^{(k)} = 0$.
   - Se $y_i = 0$, calcula-se a probabilidade de o zero observado pertencer ao estado de inflação:

$$W_i^{(k)} = \left[1 + \exp(-\mathbf{z}_i^\top \boldsymbol{\psi}^{(k)} - \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}^{(k)})))\right]^{-1}$$

2. **Passo M (Maximization):** Maximização da ℓ_c com pesos $W_i^{(k)}$: Com os valores esperados $W_i^{(k)}$ fixados, procede-se à maximização da log-verossimilhança dos dados completos.

   - Para $\boldsymbol{\beta}^{(k+1)}$: realiza-se uma **regressão Geométrica log-linear ponderada** (ou Binomial Negativa com parâmetro de dispersão unitário), utilizando pesos $(1 - W_i^{(k)})$.
   - Para $\boldsymbol{\psi}^{(k+1)}$: realiza-se uma **regressão logística ponderada** sobre $W_i^{(k)}$.

Convergência e Desafios Computacionais

O modelo **ZIG** compartilha propriedades de convergência semelhantes às do modelo **ZIP**, especialmente no que se refere ao uso do algoritmo EM. No entanto, apresenta algumas características que o tornam mais adequado em cenários com maior variabilidade nos dados.

- **Flexibilidade:** diferentemente do modelo **ZIP**, a variância no modelo **ZIG** é uma função quadrática da média, o que permite maior capacidade de ajuste em situações de sobredispersão residual.
- **Estabilidade Numérica:** o algoritmo EM mantém sua propriedade de monotonicidade, garantindo que a função de verossimilhança não diminua ao longo das iterações. Contudo, no caso do modelo geométrico, a função de verossimilhança pode ser mais sensível a valores de contagem y_i elevados quando comparada ao modelo de Poisson.

3.4 Modelo Binomial Negativo Zero-Inflado (ZINB)

Neste capítulo, apresenta-se o modelo **Binomial Negativo Zero-Inflado (ZINB)**, amplamente difundido na literatura econométrica por Greene (1994) em seu trabalho *"Accounting for Antecedent Conditions in Models of Count Data"*. Esse modelo surge

como uma extensão natural do **ZIP** para lidar com cenários onde, além do excesso de zeros, os dados de contagem apresentam sobredispersão residual (variância superior à média) que não pode ser explicada apenas pela estrutura de inflação.

Ao introduzir o parâmetro de dispersão da distribuição Binomial Negativa, o modelo **ZINB** permite capturar de forma mais adequada a variabilidade não explicada pela estrutura de inflação de zeros, proporcionando maior flexibilidade no ajuste.

3.4.1 Definição do Modelo

Seja Y_i a variável resposta associada à a i-ésima observação, onde $i = 1, \ldots, n$. Assume-se que Y_i segue uma distribuição **Binomial Negativa Zero-Inflada**, condicionada aos vetores de covariáveis $\mathbf{x}_i$ (associado ao processo de contagem) e $\mathbf{z}_i$ (associado ao processo de inflação).

3.4.2 Função de probabilidade

Componente de Contagem

Inicialmente, define-se a função de probabilidade de uma distribuição Binomial Negativa padrão, denotada por $f_2(\cdot)$, que descreve o processo de contagem para a parcela da população não sujeita ao estado de "zero estrutural":

$$f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}, \theta) = \frac{\Gamma(y_i + \theta)}{y_i! \Gamma(\theta)} \left(\frac{\theta}{\theta + \mu_i}\right)^{\theta} \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i}, \quad y_i = 0, 1, 2, \ldots,$$

onde μ_i representa a média da contagem, por meio da função de ligação logarítmica, e θ é o parâmetro de dispersão, que permite que a variância seja modelada de forma independente da média.

Estrutura de Mistura

A distribuição **ZINB** é edefinida como uma combinação linear entre uma massa pontual em zero e a distribuição Binomial Negativa apresentada anteriormente. A função de probabilidade é dada por:

$$f(y_i \mid \mathbf{x}_i, \mathbf{z}_i, \theta) = \pi(\mathbf{z}_i, \boldsymbol{\psi}) \cdot \mathbf{I}(y_i = 0) + \left[1 - \pi(\mathbf{z}_i, \boldsymbol{\psi})\right] \cdot f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}, \theta),$$

onde:

- $\mathbf{I}(y_i = 0)$ é a função indicadora, que assume valor 1 se $y_i = 0$ e 0 caso contrário;
- $= \pi_i = \pi(\mathbf{z}_i, \boldsymbol{\psi})$ é a probabilidade de a observação pertencer ao grupo dos zeros

estruturais. Para garantir que $\pi_i \in (0,1)$, utiliza-se:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i^\top \boldsymbol{\psi}$$

- $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ e θ são os parâmetros a serem estimados.

3.4.3 Probabilidade de zero

A função de probabilidade do modelo **ZINB** apresenta características particulares decorrentes da combinação entre a inflação de zeros e a flexibilidade da distribuição Binomial Negativa.

Em particular, a probabilidade de observar uma contagem nula é dada por:

$$P(Y_i = 0) = \underbrace{\pi_i}_{\substack{\text{Zero} \\ \text{Estrutural}}} + \underbrace{(1-\pi_i) \cdot \left(\frac{\theta}{\theta+\mu_i}\right)^\theta}_{\substack{\text{Zero} \\ \text{Amostral}}}$$

Essa expressão evidencia que a probabilidade de zero é composta por duas parcelas:

- zeros estruturais, representados por π_i;
- zeros amostrais, provenientes da distribuição Binomial Negativa, representados por $(1-\pi_i) \cdot (\theta/\theta+\mu_i)^\theta$.

3.4.4 Esperança e variância

A esperança e a variância condicionais da variável Y_i são dadas por:

$$E[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1-\pi_i)\mu_i$$

$$Var[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1-\pi_i)\mu_i \cdot \left(1 + \frac{\mu_i}{\theta} + \pi_i\mu_i\right)$$

Observa-se que a variância incorpora dois mecanismos de amplificação: o termo μ_i/θ, associado à dispersão da Binomial Negativa, e o termo $\pi_i\mu_i$, decorrente da inflação de zeros, o que confere ao modelo elevada flexibilidade para acomodar sobredispersão.

3.4.5 Estimação

A estimação dos parâmetros do modelo de **Regressão Binomial Negativa Zero-Inflada (ZINB)** é realizada, em geral, por meio do método de **Máxima Verossimilhança (MLE)**.O modelo assume que cada observação Y_i pode ser gerada por um dos dois processos: um estado de zero estrutural, com probabilidade π_i, ou um processo de

contagem que segue uma distribuição Binomial Negativa com média μ_i e parâmetro de dispersão θ, com probabilidade $1 - \pi_i$.

Log-verossimilhança

A função de log-verossimilhança para o modelo **ZINB** pode ser expressa como:

$$\begin{aligned}\ell(\boldsymbol{\psi}, \boldsymbol{\beta}, \theta; \mathbf{y}) = &\sum_{y_i=0} \log\left(e^{\mathbf{z}_i^\top \boldsymbol{\psi}} + \left(\frac{\theta}{\theta + e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}\right)^\theta\right) \\ &+ \sum_{y_i>0}\left[\log\frac{\Gamma(y_i+\theta)}{y_i!\Gamma(\theta)} + y_i\mathbf{x}_i^\top\boldsymbol{\beta} - (y_i+\theta)\log(e^{\mathbf{x}_i^\top\boldsymbol{\beta}}+\theta) + \theta\log\theta\right] \\ &- \sum_{i=1}^{n}\log(1+e^{\mathbf{z}_i^\top\boldsymbol{\psi}})\end{aligned}$$

A maximização direta dessa função apresenta desafios numéricos, principalmente devido à impossibilidade de identificar, a priori, a origem das observações iguais a zero, isto é, se são provenientes do estado estrutural ou do processo de contagem.

Variável Latente

De forma análoga aos modelos anteriores, a introdução de uma variável latente W_i, que indica o processo gerador de cada observação, permite reescrever a log-verossimilhança dos dados completos (ℓ_c) em componentes separáveis: uma associada ao processo de inflação (logístico) e outra ao processo de contagem (Binomial Negativa), simplificando o procedimento de estimação.

Algoritmo EM

O algoritmo EM para o modelo **ZINB** baseia-se na alternância entre dois passos iterativos:

1. **Passo E (Expectation):** Neste passo, estima-se a probabilidade *a posteriori* de a observação pertencer ao estado de inflação:
   - Se $y_i > 0$, então $W_i^{(k)} = 0$.
   - Se $y_i = 0$, calcula-se:

$$W_i^{(k)} = \left[1 + \exp(-\mathbf{z}_i^\top\boldsymbol{\psi}^{(k)}) \cdot \left(\frac{\theta^{(k)}}{\theta^{(k)} + \exp(\mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})}\right)^{\theta^{(k)}}\right]^{-1}$$

2. **Passo M (Maximization):** Com os pesos $W_i^{(k)}$ fixados, procede-se à maximização da log-verossimilhança dos dados completos:

- Para $\boldsymbol{\beta}^{(k+1)}$ e $\theta^{(k+1)}$: realiza-se uma **regressão Binomial Negativa ponderada** pelos pesos $(1 - W_i^{(k)})$.
- Para $\boldsymbol{\psi}^{(k+1)}$: realiza-se uma **regressão logística ponderada**, tomando $W_i^{(k)}$ como resposta.

Convergência e Desafios Computacionais

O modelo **ZINB** compartilha características de convergência com o modelo **ZIP**, mas apresenta desafios adicionais decorrentes da maior complexidade de sua estrutura.

- **Estimação de θ:** a convergência pode ser mais lenta, uma vez que o parâmetro de dispersão θ deve ser estimado simultaneamente aos coeficientes de regressão.
- **Robustez:** o modelo **ZINB** é mais flexível que o **ZIP**, pois, na ausência de sobredispersão ($\theta \to \infty$), reduz-se ao modelo Poisson Zero-inflado. No entanto, em termos numéricos, essa situação pode gerar instabilidade quando os dados não apresentam sobredispersão significativa.
- **Valores Iniciais:** é comum utilizar estimativas provenientes de uma regressão Binomial Negativa padrão ou de um modelo **ZIP** como valores iniciais para os vetores $\boldsymbol{\beta}$ e $\boldsymbol{\psi}$, favorecendo uma convergência mais rápida e estável do algoritmo EM.

3.5 Modelo Poisson Generalizada Zero-Inflado (ZIGP)

Neste capítulo, apresenta-se o modelo **Poisson Generalizada Zero-Inflado (ZIGP)**. A distribuição Poisson Generalizada foi originalmente proposta por Consul e Jain (1973) e posteriormente estendida para modelos de regressão inflacionados por Famoye e Singh (2006).

O principal diferencial dessa abordagem reside em sua elevada flexibilidade. Diferentemente dos modelos **ZIP** e **ZINB**, o modelo baseado na distribuição Poisson Generalizada é capaz de acomodar tanto sobredispersão quanto subdispersão, *isto é, situações em que a variância pode ser maior ou menor que a média. Essa característica torna o modelo **ZIGP** particularmente adequado para contextos em que o comportamento dos dados de contagem não segue o padrão de variância quadrática associado à distribuição Binomial Negativa, ampliando o escopo de modelagem em cenários mais gerais.

3.6 Definição do Modelo

Seja Y_i a variável resposta associada à i-ésima observação, para $i = 1, \ldots, n$. Assume-se que Y_i segue uma distribuição **Poisson Generalizada Zero-Inflada** condicionada

aos vetores de covariáveis $\mathbf{x}_i$ (associados ao processo de contagem) e $\mathbf{z}_i$ (associados ao processo de inflação).

3.6.1 Função de probabilidade

Componente de Contagem

Inicialmente, define-se a função de probabilidade da distribuição Poisson Generalizada padrão, denotada por $f_2(\cdot)$, conforme a parametrização de Consul e Famoye (1992):

$$f_2(y_i; \mu_i, \varphi) = \left(\frac{\mu_i}{1+\varphi\mu_i}\right)^{y_i} \frac{(1+\varphi y_i)^{y_i-1}}{y_i!} \exp\left(\frac{-\mu_i(1+\varphi y_i)}{1+\varphi\mu_i}\right)$$

em que $\mu_i > 0$ representa média da contagem, usualmente modelada por meio da função de ligação logarítmica, isto é, $\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, e φ é o parâmetro de dispersão.

- Se $\varphi > 0$, indica sobredispersão (variância maior que a média);
- Se $\varphi < 0$, indica subdispersão (variância menor que a média);
- Se $\varphi = 0$, reduz o modelo à distribuição Poisson padrão.

Estrutura de Mistura

A distribuição **ZIGP** é definida como uma combinação linear entre uma massa pontual no zero e a distribuição Poisson Generalizada apresentada anteriormente. Sua função de probabilidade é dada por:

$$f(y_i \mid \mathbf{x}_i, \mathbf{z}_i, \varphi) = \pi(\mathbf{z}_i, \boldsymbol{\psi}) \cdot \mathbf{I}(y_i = 0) + \left[1 - \pi(\mathbf{z}_i, \boldsymbol{\psi})\right] \cdot f_2(y_i; \mathbf{x}_i, \boldsymbol{\beta}, \varphi),$$

onde:

- $\mathbf{I}(y_i = 0)$ é a função indicadora;
- $\pi_=\pi(\mathbf{z}_i, \boldsymbol{\psi})i$ representa a probabilidade de a observação pertencer ao estado de zero estrutural, . para garantir que $\pi_i \in (0, 1)$, utiliza-se usualmente a função de ligação *logit*:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i^\top \boldsymbol{\psi}$$

3.6.2 Probabilidade de zero

A probabilidade de observar uma contagem nula no modelo **ZIGP** é dada por:

$$P(Y_i = 0) = \underbrace{\pi_i}_{\text{Zero Estrutural}} + \underbrace{(1-\pi_i) \cdot \exp\left(\frac{-\mu_i}{1+\varphi\mu_i}\right)}_{\text{Zero Amostral}}$$

Essa expressão evidencia que a probabilidade de zero é composta por duas parcelas:

- zeros estruturais, representados por π_i;
- zeros amostrais, provenientes da distribuição Poisson Generalizada, representados por$(1 - \pi_i) \cdot \exp(-\mu_i/1 + \varphi\mu_i)$.

3.6.3 Esperança e variância

A esperança e a variância condicionais da variável Y_i no modelo **ZIGP** são:

$$E[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1 - \pi_i)\mu_i$$

$$Var[Y_i|\mathbf{X}_i, \mathbf{Z}_i] = (1 - \pi_i)\mu_i(1 + \varphi\mu_i)^2 + \pi_i(1 - \pi_i)\mu_i^2$$

3.6.4 Estimação

A estimação dos parâmetros do modelo **ZIGP** $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ e φ é realizada por meio do método de Máxima Verossimilhança.

Log-verossimilhança

A função de log-verossimilhança do modelo **ZIGP** pode ser escrita como:

$$\begin{aligned}\ell(\boldsymbol{\psi}, \boldsymbol{\beta}, \varphi; \mathbf{y}) = &\sum_{y_i=0} \log\left(e^{\mathbf{z}_i^\top \boldsymbol{\psi}} + \exp\left(\frac{-e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + \varphi e^{\mathbf{x}_i^\top \boldsymbol{\beta}}}\right)\right) \\ &+ \sum_{y_i>0}\left[y_i(\mathbf{x}_i^\top \boldsymbol{\beta} - \log(1 + \varphi e^{\mathbf{x}_i^\top \boldsymbol{\beta}})) + (y_i - 1)\log(1 + \varphi y_i) - \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}(1 + \varphi y_i)}{1 + \varphi e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} - \log(y_i!)\right] \\ &- \sum_{i=1}^{n} \log(1 + e^{\mathbf{z}_i^\top \boldsymbol{\psi}})\end{aligned}$$

Variável Latente

A maximização direta dessa função apresenta desafios numéricos relevantes, especialmente devido à presença simultânea do parâmetro de dispersão φ e à impossibilidade de identificar, a priori, a origem das observações nulas. Assim como nos modelos anteriores, a introdução de uma variável latente permite reescrever a log-verossimilhança dos dados completos em componentes separáveis, facilitando o processo de estimação via algoritmos iterativos.

Algoritmo EM

O algoritmo EM para o modelo **ZIGP** trata a incerteza sobre a origem dos zeros por meio de uma variável latente W_i, que indica o estado de inflação. O procedimento iterativo alterna entre os seguintes passos:

1. **Passo E:** Estima-se o peso da variável latente W_i:

$$W_i^{(k)} = \begin{cases} \left[1 + \exp(-\mathbf{z}_i^\top \boldsymbol{\psi}^{(k)}) \cdot \exp\left(\frac{-\mu_i^{(k)}}{1+\varphi^{(k)}\mu_i^{(k)}}\right)\right]^{-1} & \text{se } y_i = 0 \\ 0 & \text{se } y_i > 0 \end{cases}$$

2. **Passo M: Com os valores esperados W_i^k fixados, procede-se à maximização da log-verossimilhança dos dados completos.**

   - Atualiza-se $\boldsymbol{\psi}$ via **regressão logística ponderada** por $W_i^{(k)}$.
   - Atualizam-se $\boldsymbol{\beta}$ e φ via regressão Poisson Generalizada ponderada por $(1 - W_i^{(k)})$.

O algoritmo é iterado até a convergência dos parâmetros

Convergência e Desafios Computacionais

O modelo **ZIGP** apresenta particularidades matemáticas que impactam diretamente o processo de estimação e a estabilidade numérica.:

- **Restrições de Parâmetros:** para que a probabilidade seja válida, o parâmetro φ deve satisfazer $1+\varphi\mu_i > 0$ e $1+\varphi y_i > 0$. Em casos de subdispersão ($\varphi < 0$), o essas restrições podem implicar um suporte finito para a distribuição, exigindo atenção na implementação.

- **Flexibilidade de Dispersão:** o modelo **ZIGP** mostra-se particularmente adequado em situações nas quais a variância não segue um crescimento estritamente quadrático em relação à média, ou quando há evidências de subdispersão. Nesses contextos, pode apresentar desempenho superior ao modelo **ZINB**.

- **Complexidade da Função Gamma:** Embora a distribuição Poisson Generalizada não envolva diretamente a função gama em sua forma básica, o que simplifica alguns cálculos, sua estrutura exponencial mais complexa pode demandar métodos de otimização numérica robustos para garantir estabilidade e convergência.

3.7 Comparação Teórica entre os Modelos

Ao longo deste capítulo, foram apresentadas quatro formulações distintas para a modelagem de dados de contagem com excesso de zeros. Embora todas compartilhem a mesma estrutura de mistura combinando um processo degenerado para os zeros estruturais e uma distribuição clássica para o processo de contagem, elas diferem substancialmente na flexibilidade com que acomodam a relação entre a média e a variância teóricas. A escolha do modelo reflete, portanto, um compromisso entre parcimônia e capacidade de ajuste à dispersão dos dados observados.

3.7.1 ZIP: O Modelo Básico

O modelo **Poisson Zero-Inflado (ZIP)** constitui a abordagem fundamental para lidar com a inflação de zeros. Sua premissa é que, uma vez superado o estado de inatividade absoluta, o processo de contagem segue uma distribuição de Poisson, caracterizada pela rigorosa propriedade de equidispersão (onde a média é igual à variância). Dessa forma, o modelo **ZIP** consegue acomodar a sobredispersão global dos dados *exclusivamente* através da probabilidade de mistura π. Contudo, mostra-se restritivo e insuficiente em cenários onde a própria subpopulação suscetível ao evento já apresenta variabilidade superior à média estipulada pelo modelo.

3.7.2 ZIG: Variância Quadrática Fixa

Como uma evolução natural para cenários de caudas mais pesadas, o modelo **Geométrico Zero-Inflado (ZIG)** substitui a componente de Poisson pela distribuição Geométrica. Esta formulação introduz uma estrutura de variância quadrática fixa ($Var(Y) = \mu + \mu^2$) para a componente de contagem, garantindo uma sobredispersão inerente ao processo gerador. Apesar de oferecer maior capacidade de acomodação de variabilidade do que o **ZIP**, o modelo **ZIG** ainda é restritivo por assumir uma relação engessada entre a média e a variância, correspondendo a um caso particular da distribuição Binomial Negativa com o parâmetro de dispersão travado em um ($\theta = 1$).

3.7.3 ZINB: Variância Flexível

Para superar a rigidez funcional do modelo anterior, o **Binomial Negativo Zero-Inflado (ZINB)** introduz um parâmetro de dispersão adicional (θ), conferindo variância flexível ao processo de contagem. Com a relação da variância dada por $Var(Y) = \mu + \mu^2/\theta$, o modelo **ZINB** generaliza elegantemente tanto o **ZIP** (quando $\theta \to \infty$) quanto o **ZIG** (quando $\theta = 1$). Essa parametrização permite que a sobredispersão residual seja estimada diretamente a partir dos dados, consolidando o **ZINB** como a escolha padrão e altamente versátil na literatura para dados discretos superdispersos.

3.7.4 ZIGP: Máxima Flexibilidade

Por fim, o modelo **Poisson Generalizada Zero-Inflado (ZIGP)** eleva a modelagem ao seu grau de máxima flexibilidade estrutural. Enquanto o **ZINB** restringe-se a lidar com cenários de sobredispersão estrita, o modelo **ZIGP**, por meio de seu parâmetro de dispersão φ, é capaz de acomodar de forma unificada tanto a sobredispersão ($\varphi > 0$) quanto a subdispersão ($\varphi < 0$), além de reduzir-se ao caso Poisson clássico ($\varphi = 0$). Com a estrutura de variância expressa por $Var(Y) = \mu(1 + \varphi\mu)^2$, esta abordagem ajusta-se a arranjos de dispersão atípicos e complexos, tornando-se a alternativa definitiva quando a relação quadrática convencional da **ZINB** se mostra insuficiente ou mal especificada.

Capítulo 4

Interpretação dos Parâmetros e Inferência

Neste capítulo, a estrutura matemática dos modelos zero-inflados é traduzida em termos práticos, com foco na interpretação dos parâmetros estimados. A interpretação adequada dos resultados depende diretamente das funções de ligação utilizadas para relacionar os preditores lineares aos parâmetros do modelo.

4.1 Componente de Contagem (μ)

A média do processo de contagem é, em geral, modelada por meio da função de ligação logarítmica, tal que:

4.1.1 Razão de Taxas (Rate Ratio - RR)

Para interpretar o efeito de uma covariável, utiliza-se a exponencial do coeficiente estimado, e^{β_j}, conhecida como **Razão de Taxas (RR)**.

- **Interpretação:** um aumento unitário na covariável x_j, mantendo as demais constantes, multiplica a média esperada da contagem por e^{β_j}.
- **Exemplo:** Se $\hat{\beta} = 0,42$ para um "relacionamento romântico", então $e^{0,42} \approx 1,52$. Isso indica que estudantes nessa condição apresentam uma taxa esperada de ocorrência aproximadamente 52% maior em comparação ao grupo de referência.

4.2 Componente de Inflação (π)

A probabilidade de ocorrência de um zero estrutural (π_i) é modelada por meio da função de ligação *logit*, que relaciona as covariáveis à log-chance (log-odds) de a observação pertencer ao grupo dos "zeros absolutos".

4.2.1 Razão de Chances (Odds Ratio - OR)

A exponencial do coeficiente da parte logística, e^{ψ_j}, corresponde à **Razão de Chances (OR)** de pertencer ao estado de inflação de zeros.

- **Interpretação:** um aumento unitário na covariável z_j, mantendo as demais constantes, multiplica a chance de a observação pertencer ao grupo de zeros estruturais por e^{ψ_j}.
- **Sinal do Coeficiente:** Se $\psi_j > 0$ ($e^{\psi_j} > 1$), o aumento da covariável aumenta a chance de o indivíduo ser um "zero estrutural"(ou seja, ele tem maior chance de ser "imune"ao evento de contagem).
- **Exemplo:** Se a **Razão de Chances (OR)** associada à produtividade do orientador é $0,88$, então cada artigo adicional reduz a chance de o aluno pertencer ao grupo de "zero estrutural"em aproximadamente 12% ($1 - 0,88$).

4.3 Interpretação Conjunta

A análise final deve considerar simultaneamente as duas componentes do modelo. Uma mesma covariável pode influenciar tanto a intensidade do processo de contagem quanto a probabilidade de pertencimento ao estado de zero estrutural.

Assim, um preditor pode ser significativo na componente de contagem (afetando a média entre indivíduos suscetíveis ao evento) e, ao mesmo tempo, na componente de inflação (alterando a probabilidade de ausência estrutural). A interpretação integrada desses efeitos é essencial para uma compreensão completa do fenômeno em estudo.

4.4 Testes de Significância e Intervalos de Significância

A inferência estatística nos modelos zero-inflados baseia-se em resultados assintóticos, segundo os quais os estimadores de máxima verossimilhança apresentam, sob condições regulares, distribuição aproximadamente normal multivariada em grandes amostras.

4.4.1 Matriz de Informação: Observada vs. Esperada

Para o cálculo dos erros padrão, recomenda-se fortemente o uso da **matriz de informação observada**. Em modelos zero-inflados, a utilização da informação esperada pode levar à subestimação da variabilidade dos estimadores, resultando em erros padrão artificialmente reduzidos e, consequentemente, inferências potencialmente enganosas.

4.4.2 Intervalos de Confiança: Wald vs. LRT

A avaliação da significância de parâmetros pode ser conduzida por diferentes abordagens:

1. **Teste de Wald:** baseia-se na aproximação normal dos estimadores, utilizando intervalos do tipo $(\hat{\beta} \pm Z \cdot SE)$. Embora computacionalmente simples, pode apresentar desempenho insatisfatório em amostras pequenas ($n \leq 50$) ou quando a função de verossimilhança apresenta forte assimetria.

2. **Teste de Razão de Verossimilhança (LRT):** trata-se da abordagem recomendada. Os intervalos são obtidos a partir da região de valores dos parâmetros para os quais a log-verossimilhança não se reduz além de um limite crítico. Esse método tende a apresentar melhor desempenho, especialmente na componente logística, garantindo níveis de cobertura mais confiáveis.

Capítulo 5

Diagnóstico de Modelos (DHARMa)

A avaliação da qualidade de ajuste em modelos de contagem, especialmente na presença de inflação de zeros ou estruturas complexas, como nos modelos **ZIP**, **ZINB** e **ZIGP**, requer ferramentas que superem as limitações dos resíduos tradicionais. Nesse contexto, o pacote `DHARMa` (Diagnostic Residuals for Hierarchical Regression Models) Hartig (2022) tornou-se o como uma referência na literatura, ao empregar resíduos quantílicos randomizados Dunn e Smyth (1996) obtidos via simulação.

O pacote `DHARMa` baseia-se na abordagem de resíduos quantílicos randomizados. A ideia central consiste em transformar a função de distribuição acumulada (FDA) discreta do modelo em uma variável contínua com distribuição uniforme, opcionalmente, ser transformada para a na escala normal.

5.1 Limitações dos resíduos tradicionais

Em modelos lineares clássicos, os resíduos de Pearson e Deviance são amplamente utilizados. No entanto, para modelos de contagem discretos, essas abordagens apresentam limitações importantes:

- **Artefatos Visuais:** devido à natureza discreta dos dados ($Y \in \{0, 1, 2, ...\}$), gráficos de resíduos versus valores preditos frequentemente exibem padrões estruturais (como faixas paralelas ou formatos em "leque") que não necessariamente indicam mau ajuste, mas sim refletem a discretização dos dados.
- **Ausência de Normalidade:** mesmo sob especificação correta do modelo, os resíduos tradicionais raramente seguem uma distribuição normal, o que compromete a interpretação de gráficos QQ-Plot convencionais.
- **Sensibilidade à Inflação:** em modelos com excesso de zeros, a variância utilizada na padronização dos resíduos de Pearson depende fortemente da correta es-

pecificação do parâmetro de inflação (π). Estimativas imprecisas podem distorcer significativamente o diagnóstico.

5.2 Resíduos Quantílicos Randomizados (RQR)

Seja $F(y;\hat{\theta})$ a FDA estimada do modelo. O resíduo quantílico associado à observação i é obtido através de:

$$r_{Q,i} = \Phi^{-1}(u_i)$$

onde Φ^{-1} denota a função quantílica da normal padrão e u_i é uma variável aleatória uniforme no intervalo:

$$u_i \sim U[F(y_i - 1;\hat{\theta}_i), F(y_i;\hat{\theta}_i)]$$

Na prática, o `DHARMa` implementa esse procedimento por meio de simulação: novos conjuntos de dados são gerados a partir do modelo ajustado, e as observações reais são comparadas com a distribuição simulada.

5.3 Interpretação dos os Gráficos do `DHARMa`

O comando `plot(simulateResiduals(model))` produz gráficos de diagnósticos essenciais para a avaliação do ajuste.

5.3.1 QQ Plot Residuals

Este gráfico avalia se a distribuição dos resíduos está de acordo com o esperado sob o modelo ajustado:

- **Teste KS (Kolmogorov-Smirnov)** Verifica a uniformidade global dos resíduos. Valores de $p < 0,05$ indicam possível inadequação do modelo.
- **Teste de dispersão:** avalia a presença de sobredispersão ou subdispersão residual.
- **Teste de outliers:** investiga se a frequência de valores extremos excede o esperado teoricamente.

5.3.2 Resíduos vs. Valores Ajustados

Este gráfico permite identificar desvios sistemáticos ao longo da faixa de predição:

- Os resíduos devem estar aproximadamente uniformemente distribuídos no intervalo $[0, 1]$.

- As curvas de quantis (tipicamente nos níveis 0,25, 0,50 e 0,75) devem ser aproximadamente horizontais.
- curvaturas ou inclinações, sugerem falhas na especificação do modelo, por exemplo, na forma funcional ou na estrutura de variância.

5.4 Vantagens do Diagnóstico via Simulação

1. **Distribuição teórica bem definida:** sob especificação correta, os resíduos seguem exatamente uma distribuição uniforme (ou normal, após transformação), independentemente da natureza discreta dos dados.
2. **Aplicabilidade geral:** a abordagem é válida para uma ampla classe de modelos, incluindo Poisson, Binomial Negativa, modelos inflacionados e modelos com efeitos mistos.
3. **Interpretação mais clara:** ao eliminar artefatos decorrentes da discretização, os gráficos tornam-se mais informativos, permitindo avaliar diretamente se os dados observados são compatíveis com a distribuição gerada pelo modelo.

Capítulo 6

Seleção e Comparação de Modelos

A escolha do modelo mais adequado constitui uma etapa fundamental na análise de dados de contagem. Não é suficiente que o modelo produza estimativas estatisticamente significativas; é necessário avaliar se o aumento de complexidade, decorrente, por exemplo, da inclusão de componentes de inflação de zeros ou parâmetros de dispersão, resulta em melhoria efetiva na qualidade do ajuste, sem comprometer a parcimônia.

6.1 Comparação de Modelos

A comparação entre modelos é essencial para verificar a necessidade de estruturas mais complexas. Em particular, para avaliar a presença de inflação de zeros, é comum comparar o modelo **Poisson Zero-Inflado (ZIP)** com o modelo de Poisson padrão correspondente.

6.1.1 Teste de Razão de Verossimilhança (LRT)

O principal método para comparação entre modelos aninhados é o Teste de Razão de Verossimilhança. No contexto de modelos de contagem, o modelo Poisson pode ser visto como um caso especial do modelo **ZIP**, correspondente à situação em que a probabilidade de inflação de zeros (π) é nula.

Sejam L_{ZIP} e L_{Pois} as log-verossimilhanças dos modelos **ZIP** e Poisson, respectivamente. A estatística do teste é definida por:

- **Estatística de Teste:** $-2\ln\Lambda = 2(L_{ZIP} - L_{Pois})$.
- **Distribuição assintótica:** Sob condições regulares, essa estatística segue aproximadamente uma distribuição qui-quadrado χ^2, com graus de liberdade iguais à diferença no número de parâmetros entre os modelos (geralmente o número de parâmetros associadas a à componente de inflação (ou 1, no caso de π constante).).

- **Critério de decisão:** Valores de p-valor inferiores ao nível de significância (por exemplo, $p < 0,05$) indicam evidências contra o modelo Poisson simples, favorecendo o modelo inflacionado.

6.1.2 Teste de Vuong

Diferente do Teste de Razão de Verossimilhança (LRT), que requer modelos aninhados, o teste proposto por Vuong (1989) é apropriado para comparar modelos não aninhados, sobrepostos ou estritamente distintos. O teste baseia-se na comparação das log-verossimilhanças individuais de cada observação sob os modelos concorrentes.

Fundamentação Teórica e Hipóteses

O teste avalia a hipótese nula de que os dois modelos, denotados por $f(Y|X)$ e $g(Y|X)$, são estatisticamente equivalentes em termos de sua proximidade com a distribuição real dos dados:

- $H_0 : E[\ln f(Y|X)] = E[\ln g(Y|X)]$ (os modelos são equivalentes)
- $H_1 : E[\ln f(Y|X)] \neq E[\ln g(Y|X)]$ (um dos modelos é superior ao outro)

Estatística de Teste

A estatística de teste V é calculada com base na média ($\bar{m}$) e no desvio padrão (s_m) das diferenças de log-verossimilhança entre os modelos para as N observações:

$$V = \frac{\sqrt{N}}{s_m}\bar{m}$$

onde ($\bar{m}$) representa a média das diferenças individuais de log-verossimilhança entre os modelos, e (s_m) é o desvio padrão dessas diferenças.

Interpretação e Critérios de Decisão

Seja $z_{1-\alpha/2}$ o valor crítico da distribuição Normal Padrão para um nível de significância α (geralmente $1,96$ para $\alpha = 5\%$). A decisão entre os modelos segue:

- **Se** $V > z_{1-\alpha/2}$**:** há evidências estatísticas significativas que favorecem o primeiro modelo (f).
- **Se** $V < -z_{1-\alpha/2}$**:** há evidências estatísticas significativas que favorecem o segundo modelo (g).
- **Se** $|V| \leq z_{1-\alpha/2}$**:** não se rejeita H_0, indicando que os modelos são estatisticamente equivalentes quanto à qualidade de ajuste.

6.2 Critérios de Informação

A seleção de modelos pode ser conduzida por meio da maximização da log-verossimilhança penalizada pela complexidade do modelo, de modo a evitar o sobreajuste (*overfitting*) e preservar a parcimônia.

6.2.1 Critério de Informação de Akaike (AIC)

O AIC é amplamente utilizado para comparar modelos ajustados a um mesmo conjunto de dados. Sua definição é dada por:

$$AIC = 2k - 2\ln(\hat{L}),$$

onde k representa o número total de parâmetros estimados (incluindo $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ e, quando aplicável, parâmetros de dispersão como θ ou φ) e $\ln(\hat{L})$ é a log-verossimilhança maximizada. Modelos com menores valores de AIC são preferidos, pois oferecem um melhor compromisso entre qualidade de ajuste e simplicidade.

6.2.2 Critério de Informação Bayesiano (BIC)

O BIC introduz uma penalidade mais rigorosa sobre para a complexidade do modelo, especialmente em amostras de grande tamanho:

$$BIC = k\ln(n) - 2\ln(\hat{L}),$$

Enquanto o AIC foca na capacidade preditiva, o BIC adota uma postura mais conservadora, favorecendo modelos mais parcimoniosos à medida que n aumenta.

6.3 Comparação entre Modelos Inflacionados (ZIP vs ZINB)

Quando há evidências de sobredispersão mesmo após a inclusão da componente de inflação de zeros, torna-se necessário comparar o modelo **Poisson Zero-Inflado (ZIP)** com o modelo **Binomial Negativo Zero-Inflado (ZINB)**.

Do ponto de vista teórico, o modelo **ZIP** pode ser interpretado como um caso limite do **ZINB** quando o parâmetro de dispersão $\theta \to \infty$. No entanto, essa relação de fronteira impõe cautela na aplicação direta de testes clássicos de razão de verossimilhança para o parâmetro de dispersão. Na prática, No entanto, essa relação de fronteira impõe cautela na aplicação direta de testes clássicos de razão de verossimilhança para o parâmetro de dispersão.

6.4 Avaliação de Resíduos com DHARMa

De forma complementar aos testes formais, a avaliação do ajuste do modelo deve incluir a análise visual de resíduos quantílicos simulados. O uso de ferramentas como o histograma baseado na Transformada Integral de Probabilidade (Probability Integral Transform) e o gráfico QQ possibilita a identificação de padrões sistemáticos não capturados por medidas globais de ajuste.

Capítulo 7

Aplicação I — Modelo ZIP (bioChemists)

Este capítulo apresenta a aplicação dos modelos discutidos ao conjunto de dados *bioChemists*, extraído do estudo seminal de Long (1997), intitulado *"Regression Models for Categorical and Limited Dependent Variables"*. O banco de dados investiga a produtividade científica de doutorandos em bioquímica e tornou-se um padrão na literatura estatística para o estudo de fenômenos com excesso de zeros. O objetivo central é modelar o número de artigos publicados nos últimos três anos do doutorado, enfrentando o desafio de lidar com a massa de observações nulas que a distribuição Poisson simples não consegue acomodar.

7.1 Descrição dos dados

O estudo baseia-se em uma amostra de 915 observações. As covariáveis selecionadas para explicar tanto a média de contagem quanto a probabilidade de inflação de zeros são apresentadas na Tabela 7.1.

Tabela 7.1: Descrição das variáveis do conjunto de dados.

Variável	**Descrição**	**Tipo**
`art` (Y)	Número de artigos do aluno	Contagem
`fem`	Gênero (Feminino/Masculino)	Categórica
`mar`	Estado Civil (Casado/Solteiro)	Categórica
`kid5`	Filhos menores de 5 anos	Contagem
`ment`	Artigos do orientador (Mentor)	Contagem

Para a realização da análise, o banco de dados foi incorporado ao ambiente estatístico R R Core Team (2023) por meio do pacote `pscl` Zeileis, Kleiber e Jackman (2008), conforme demonstrado no código a seguir:

Script R: Carregamento do Banco de Dados bioChemists

```r
# 1. Carregar o pacote que contem o dataset
library(pscl)

# 2. Carregar o banco de dados bioChemists
data("bioChemists", package = "pscl")

# 3. Inspecionar a estrutura das variaveis
str(bioChemists)
```

7.2 Análise descritiva

A variável resposta, `art`, apresenta forte assimetria à direita, além de uma elevada frequência de zeros. Esse padrão sugere a presença de dois mecanismos geradores distintos: alunos que não possuem condições estruturais de publicação e alunos ativos que, por fatores aleatórios, não obtiveram aceites no período analisado.

A análise descritiva inicial dos dados permitiu identificar padrões relevantes que justificam a utilização de modelos com inflação de zeros. Destacam-se os seguintes pontos:

- **Excesso de Zeros:** a distribuição empírica da variável resposta evidencia uma frequência de zeros significativamente superior àquela esperada sob um modelo de Poisson.
- **Fatores familiares e Gênero:** observa-se que em contextos de maior estabilidade familiar tendem a apresentar níveis ligeiramente mais elevados de produtividade. Por outro lado, a presença de filhos pequenos (`kid5`) está associada a uma redução na produção científica.
- **Influência do Orientador:** identifica-se uma associação positiva entre a produtividade do orientador (`ment`) e a do aluno, indicando que o ambiente acadêmico exerce papel relevante na geração de publicações.

7.3 Ajuste

Para a modelagem da produção científica, utilizou-se o pacote `glmmTMB` Brooks et al. (2017), que permite a especificação conjunta das componentes de contagem e de inflação de zeros.

Script R: Modelagem ZIP e Diagnóstico RQR

```r
# 4. Carregar pacotes necessarios
library(glmmTMB)
library(DHARMa)

# 5. Ajuste do modelo ZIP
# ziformula define a probabilidade de zeros estruturais (logit)
modelo_zip <- glmmTMB(art ~ fem + mar + ment,
                      ziformula = ~ ment,
                      data = bioChemists,
                      family = poisson)

# 6. Gerar Residuos Quantilicos Randomizados
# O DHARMa simula dados para criar residuos continuos uniformes
res_sim <- simulateResiduals(fittedModel = modelo_zip, n = 1000)

# 7. Diagnostico Visual
# Gera o QQ-plot e o grafico de Residuos vs. Preditos
plot(res_sim)
```

7.4 Diagnóstico

A validação do modelo **ZIP** baseada na análise de resíduos quantílicos randomizados (RQR) permitiu identificar limitações relevantes que não seriam detectadas por resíduos tradicionais. O diagnóstico foi realizado por meio do pacote `DHARMa`. Em particular, a análise (Figura 7.1) indicou que o modelo **Poisson Zero-Inflado** não foi plenamente capaz de capturar a estrutura de variabilidade dos dados.

- **Desvio de Uniformidade (KS Test):** o teste de Kolmogorov-Smirnov apresentou significância ($p = 0,0084$), indicando que os resíduos não seguem a distribuição uniforme teórica.
- **Sobredispersão Residual:** o teste de dispersão resultou em um p-valor de 0, evidenciando que a variabilidade dos dados excede aquela prevista pelo modelo, mesmo após o ajuste para inflação de zeros.
- **Presença de Outliers:** o teste de *outliers* ($p = 0,00012$) e o comportamento da

cauda superior no gráfico QQ indicam que o modelo subestima as observações com alta produtividade.

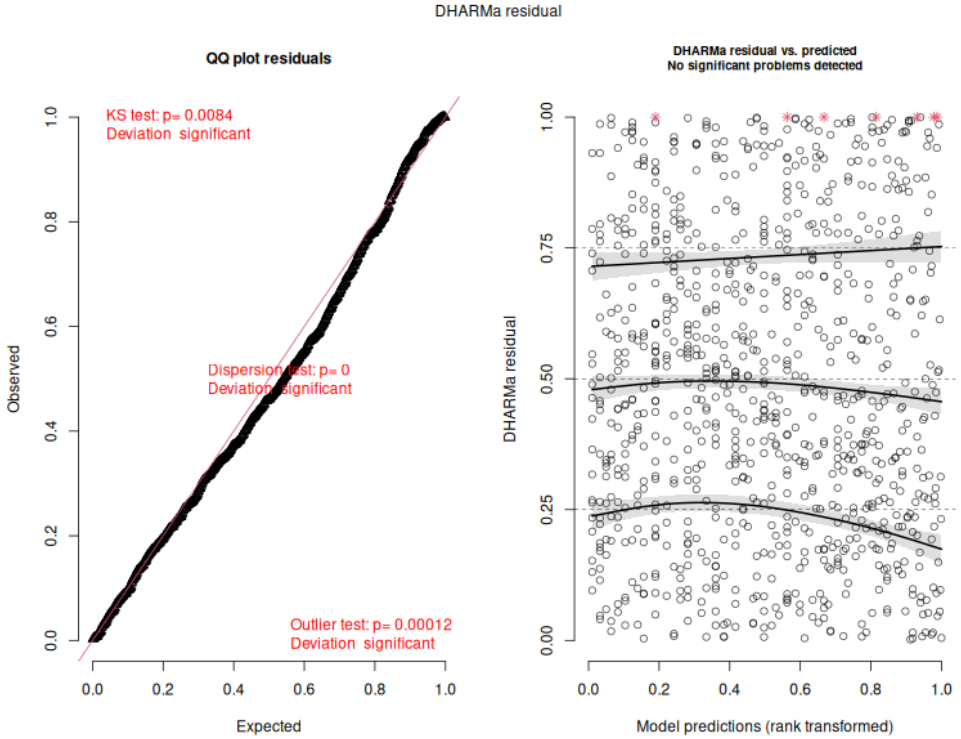

Figura 7.1: Diagnóstico `DHARMa` evidenciando sobredispersão e falha de ajuste nas caudas.

7.5 Resultados

Para facilitar a interpretação, os coeficientes estimados foram exponencializados, resultando em Razões de Taxas (RR) para a componente de contagem e Razões de Chances (OR) para a componente de inflação. A Tabela 7.2 resume as estimativas do modelo.

Tabela 7.2: Estimativas do modelo ZIP (Coeficientes Exponenciados).

Covariável	**Razão de Taxas (RR)**	**Razão de Chances (OR)**
Componente de Contagem (Poisson)		
Gênero (Feminino)	0,84	—
Estado Civil (Casado)^{ns}	1,03	—
Artigos do Orientador	1,02	—
Componente de Inflação (Logit)		
Artigos do Orientador	—	0,88

^{ns} *Não significativo (p-valor = 0,604).*

A análise dos parâmetros estimados indica:

- **Gênero:** indivíduos do gênero feminino apresentam uma taxa de publicação aproximadamente 16% menor em relação aos homens ($RR \approx 0,84$), mantendo-se constantes as demais variáveis.

- **Produtividade do Orientador:** o orientador atua como um fator de proteção contra a inatividade acadêmica. Cada artigo adicional publicado pelo orientador está associado a uma redução de aproximadamente 12% na chance de o aluno pertencer ao grupo de zeros estruturais ($OR \approx 0,88$).

- **Estado Civil:** não foi identificada evidência estatística de associação significativa entre estado civil e produtividade acadêmica ($p = 0,604$).

7.6 Conclusão

A aplicação empírica demonstrou que o modelo **ZIP** constitui uma alternativa mais adequada do que o modelo de Poisson simples para lidar com dados de contagem com excesso de zeros.

Entretanto, a análise diagnóstica evidenciou que a distribuição de Poisson ainda se mostra restritiva para este conjunto de dados, não sendo capaz de capturar adequadamente a sobredispersão residual e a presença de valores extremos.

Os resultados indicam que a produtividade do orientador desempenha papel central na produção científica dos alunos, atuando tanto na redução da probabilidade de inatividade quanto no aumento da taxa de publicações.

Por fim, os indícios de falta de ajuste sugerem que modelos mais flexíveis, como o modelo **Binomial Negativo Zero-Inflado (ZINB)**, constituem uma extensão natural e recomendada para análises futuras.

Capítulo 8

Aplicação II — Modelo ZIG (fishing)

Neste capítulo, apresenta-se a aplicação do modelo **Geométrico Zero-Inflado (ZIG)** ao conjunto de dados *fishing*. Este exemplo ilustra a imposição de uma restrição teórica ao parâmetro de dispersão, de modo que o processo de contagem siga estritamente a distribuição Geométrica, cuja variância é definida por $Var(Y) = \mu + \mu^2$.

8.1 Descrição dos dados

Esta aplicação utiliza o conjunto de dados *fishing*, disponibilizado originalmente pelo portal de consultoria estatística da **UCLA** UCLA Statistical Consulting (2021). O banco registra a atividade de 250 grupos em um parque estadual e é amplamente utilizado na literatura para ilustrar a distinção entre zeros estruturais e amostrais.

Para a realização da análise, o conjunto de dados pode ser importado no ambiente R diretamente a partir de sua fonte original, conforme ilustrado no código a seguir:

Script R: Carregamento dos dados Fishing

```r
# 1. Carregar os dados diretamente do portal da UCLA
url_fish <- "https://stats.idre.ucla.edu/stat/data/fish.csv"
fishing_data <- read.csv(url_fish)

# 2. Inspecionar as primeiras linhas
head(fishing_data)
```

8.2 Análise descritiva

A variável resposta corresponde ao número de peixes capturados (`count`), sendo influenciada por diversas covariáveis, entre as quais se destacam: número de crianças (`child`),

uso de acampamento (`camper`), total de pessoas (`persons`) e uso de isca viva (`livebait`).

Uma análise preliminar revela que a variável `count` apresenta uma massa considerável de valores nulos (grupos que não pescaram nenhum peixe). Esse excesso de zeros justifica a necessidade de incorporar uma componente logística para a inflação, distinguindo os zeros provenientes de falha na captura daqueles originados pela ausência da prática de pesca.

8.3 Ajuste

Para garantir o rigor da distribuição Geométrica no pacote `glmmTMB` Brooks et al. (2017), o parâmetro de dispersão (ϕ) foi fixado em 1 através do argumento `map`. Essa restrição impede sua estimação e assegura que a estrutura de variância do modelo seja compatível com a distribuição geométrica.

Script R: Ajuste Geométrico Estrito via `glmmTMB`

```r
# 3. Carregar pacotes necessarios
library(glmmTMB)
library(DHARMa)

# 4. Ajuste do modelo ZIG (phi fixado em 1)
# O parametro betadisp e travado em 0 para garantir phi = exp(0) = 1
modelo_fishing <- glmmTMB(count ~ child + camper + persons + livebait,
                          ziformula = ~ child,
                          data = fishing_data,
                          family = nbinom2,
                          start = list(betadisp = 0),
                          map = list(betadisp = factor(NA))
                             )

# 5. Geracao de Residuos Quantilicos simulados
res_sim <- simulateResiduals(fittedModel = modelo_fishing, n = 1000)

# 6. Diagnostico visual
plot(res_sim)
```

8.4 Diagnóstico

O diagnóstico visual via pacote `DHARMa` (Figura 8.1) e os testes formais baseados em resíduos quantílicos randomizados (RQR) confirmam que a distribuição Geométrica, mesmo sob a restrição $\phi = 1$, é adequada para este fenômeno:

- **Uniformidade (KS Test):** o p-valor de $0,240$ não fornece evidências de desvios em relação à distribuição uniforme dos resíduos, conforme esperado sob boa especificação do modelo.
- **Dispersão:** o teste de dispersão ($p = 0,44$) sugere que a estrutura de variância da distribuição geométrica ($\mu + \mu^2$) foi suficiente para acomodar a variabilidade observada, não havendo indícios de sobredispersão residual.
- **Outliers:** o teste de *outliers* ($p = 1$) não indica presença de observações extremas incompatíveis com o modelo ajustado.

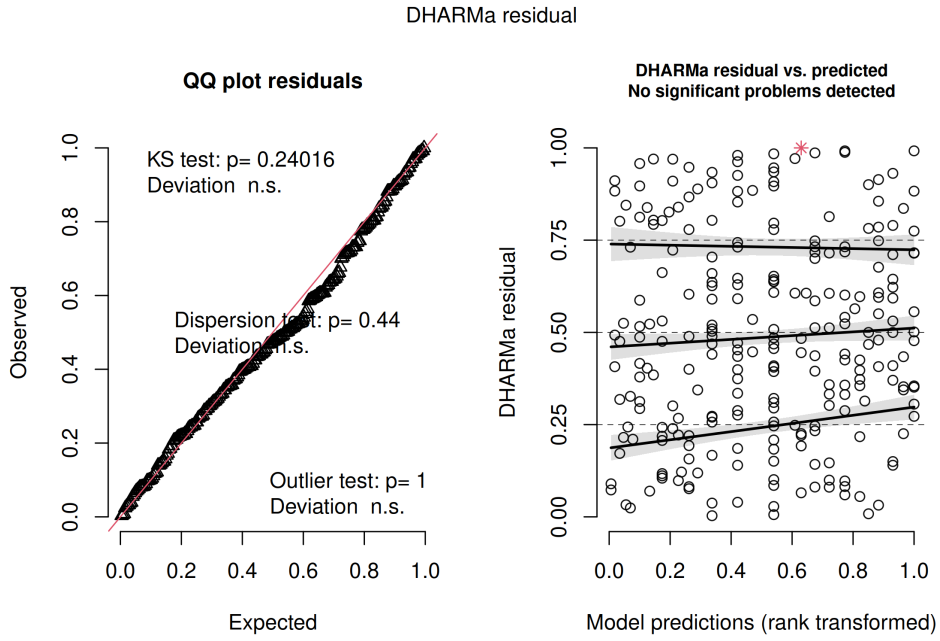

Figura 8.1: Resíduos quantílicos simulados evidenciando ajuste adequado do modelo **ZIG**.

8.5 Resultados

Os resultados da estimação, com o parâmetro de dispersão fixado em 1, são apresentados na Tabela 8.1. Para facilitar a interpretação, os coeficientes foram exponencializados, sendo expressos em termos de Razões de Taxas (RR), para a componente de contagem, e Razões de Chances (OR), para a componente de inflação de zeros.

Tabela 8.1: Estimativas do modelo **ZIG** (Coeficientes Exponenciados).

Covariável	**Estimativa**	**RR / OR**	**p-valor**
Componente de Contagem (Geométrica)			
Intercepto	-2,9038	0,055	$< 0{,}001$
Crianças (`child`)	-1,1079	0,330	$< 0{,}001$
Acampamento (`camper`)	0,4310	1,539	0,039
Pessoas (`persons`)	1,0362	2,819	$< 0{,}001$
Isca Viva (`livebait`)	1,5458	4,691	$< 0{,}001$
Componente de Inflação (Logit)			
Intercepto	-3,1007	0,045	$< 0{,}001$
Crianças (`child`)	2,3947	10,965	$< 0{,}001$

Interpretação dos Efeitos:

- **Efeito das Crianças:** A variável `child` apresenta um impacto relevante em ambas as componentes do modelo. Na componente de contagem, cada criança adicional reduz a taxa esperada de captura em aproximadamente 67% ($RR = 0,330$), mantendo-se as demais variáveis constantes. Na componente de inflação, observa-se um efeito oposto: cada criança adicional aumenta em cerca de **11 vezes** a chance de o grupo pertencer ao estado de zero estrutural ($OR = 10,965$), isto é, de não realizar capturas.

- **Sucesso na Captura:** O uso de isca viva (`livebait`) destaca-se como o principal fator associado à intensidade da pesca, elevando a taxa esperada de captura em aproximadamente **4,69 vezes**. Esse resultado sugere que a escolha da estratégia de pesca exerce papel determinante no sucesso da atividade.

8.6 Conclusão

A aplicação do modelo **Geométrico Zero Inflado (ZIG)** mostrou-se adequada para descrever os dados fishing, sobretudo por sua capacidade de acomodar variabilidade superior à média.

Em comparação ao modelo Poisson inflacionado, a estrutura geométrica permitiu capturar melhor a dispersão dos dados, sem a necessidade de um parâmetro adicional. Os resultados evidenciaram que o número de crianças (`kids`) atua tanto na redução da taxa de captura quanto no aumento da probabilidade de zeros estruturais, reforçando a importância de modelar separadamente os dois processos.

O diagnóstico baseado em resíduos quantílicos simulados confirmou a adequação do modelo, sem indícios de desvios de uniformidade, problemas de dispersão ou outliers.

Em síntese, o modelo **ZIG** apresentou maior flexibilidade que o **ZIP** neste contexto, sendo uma alternativa eficaz em cenários com variância crescente em relação à média. Como extensão, modelos mais gerais, como o **ZINB**, podem ser considerados em situações de variabilidade ainda mais elevada.

Capítulo 9

Aplicação III — Modelo ZINB (Student Performance)

Nesta seção, apresenta-se a aplicação do modelo **Binomial Negativo Zero-Inflado (ZINB)** ao problema de absenteísmo escolar.

9.1 Descrição dos dados

A análise baseia-se no conjunto de dados *Student Performance*, proveniente do estudo de Cortez e Silva (2008), amplamente difundido na plataforma **Kaggle**. O conjunto de dados reúne informações socioeconômicas e comportamentais de estudantes de duas escolas portuguesas.

Para a realização da análise, o conjunto de dados pode ser carregado no ambiente R a partir de um arquivo CSV, conforme ilustrado no código a seguir:

Script R: Carregamento dos dados Student Absences

```r
# 1. Carregar o arquivo CSV (especifico para estudantes de
   matematica)
student_data <- read.csv("student-mat.csv", sep = ";")

# 2. Visualizar o sumario da variavel de contagem
summary(student_data$absences)
```

9.2 Análise descritiva

Após o pré-processamento dos dados (exclusão de registros com dados faltantes na educação dos pais), a análise foi conduzida com uma amostra de 389 observações. A

variável resposta corresponde ao número de faltas escolares (`absences`), caracterizada por apresentar sobredispersão e um aglomerado significativo no valor zero.

A variável de contagem é modelada em função de covariáveis comportamentais e demográficas, destacando-se a idade (`age`), o envolvimento em relacionamentos românticos (`romantic_relationship`) e a percepção de tempo livre após a escola (`free_time`).

9.3 Ajuste

Diferentemente do modelo Poisson, o ajuste **ZINB** por meio do pacote `glmmTMB` Brooks et al. (2017) permite a estimação simultânea da componente de contagem, do parâmetro de dispersão e da probabilidade de inflação. A utilização da família `nbinom2` garante a parametrização quadrática da variância, em conformidade com a formulação teórica adotada.

Script R: Ajuste ZINB via `glmmTMB`

```r
# 3. Carregar pacotes necessarios
library(glmmTMB)
library(DHARMa)

# 4. Limpeza previa
student_data <- student_data[student_data$mother_education
   !='none' &
                             student_data$father_education
                                != 'none', ]

# 5. Ajuste do modelo ZINB (Parametrizacao NB2)
modelo_estudantes <- glmmTMB(absences ~ age + romantic_
   relationship + as.factor(free_time),
                             ziformula = ~ 1,
                             data = student_data,
                             family = nbinom2)

# 6. Geracao de Residuos Quantilicos simulados (DHARMa)
res_sim <- simulateResiduals(fittedModel = modelo_
   estudantes, n = 1000)

# 7. Diagnostico visual
plot(res_sim)
```

9.4 Diagnóstico

O diagnóstico visual (Figura 9.1) e os testes formais, realizados via resíduos quantílicos randomizados (RQR) no pacote `DHARMa`, confirmam que a distribuição **Binomial Negativa Zero-Inflada** é altamente adequada para este fenômeno, corrigindo as patologias observadas no modelo Poisson simples:

- **Uniformidade (KS Test):** O p-valor de $0,520$ não indica evidências de desvio da distribuição uniforme esperada, sugerindo boa adequação global do modelo.
- **Dispersão:** O teste de dispersão ($p = 0,096$) indica que a variabilidade dos dados foi adequadamente capturada pelo modelo, especialmente por meio do parâmetro de dispersão estimado ($\theta = 1,67$).
- **Outliers:** O teste de *outliers* ($p = 0,24$) não aponta evidências de valores extremos mal ajustados, indicando que a cauda da distribuição Binomial Negativa foi suficiente para acomodar observações mais elevadas.

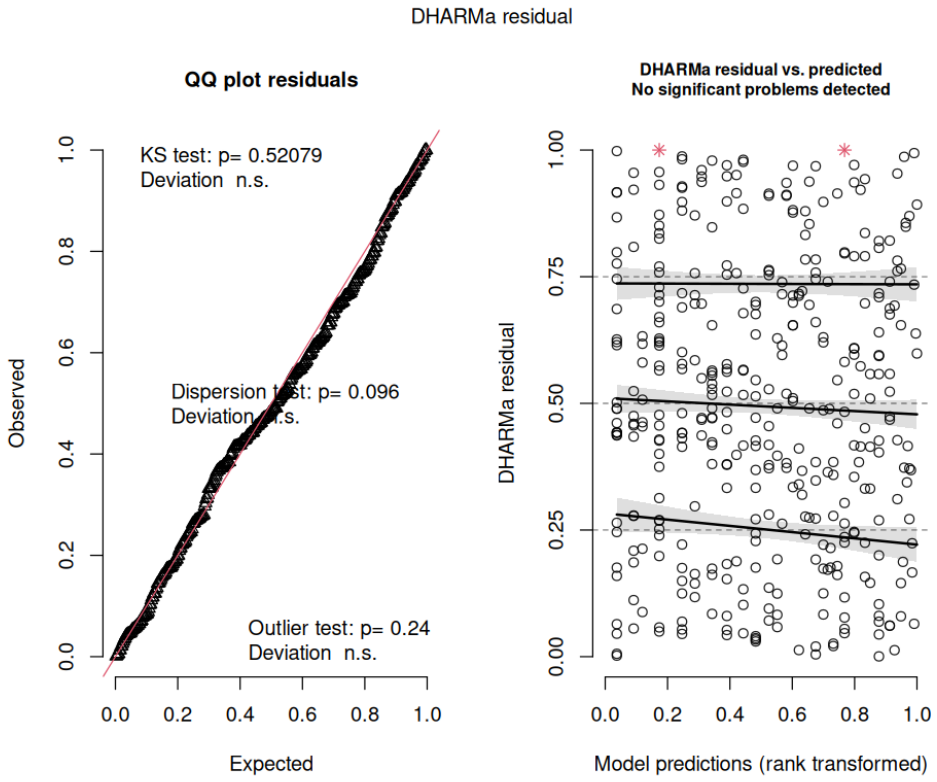

Figura 9.1: Resíduos quantílicos simulados indicando bom ajuste do modelo **ZINB** para os dados de absenteísmo escolar.

9.5 Resultados

Os resultados da estimação são apresentados na Tabela 9.1. O parâmetro de dispersão foi estimado em $\hat{\theta} = 1,67$, evidenciando a presença de sobredispersão significativa nos dados, aspecto que não seria adequadamente capturado por um modelo de Poisson.

Tabela 9.1: Estimativa do modelo ZINB (Coeficientes Exponenciados).

Variável	**Estimativa**	**RR / OR**	**p-valor**
Componente de Contagem (Binomial Negativa)			
Intercepto	0,0247	1,025	0,975
Idade (`age`)	0,1364	1,146	0,002
Relacionamento (`romantic_yes`)	0,4226	1,526	$<$ 0,001
Tempo Livre (`2`)	-0,4555	0,634	0,074
Tempo Livre (`3`)	-0,5267	0,591	0,029
Tempo Livre (`4`)	-0,5212	0,594	0,031
Tempo Livre (`5`)	-0,3273	0,721	0,233
Componente de Inflação (Logit)			
Intercepto (ψ_0)	-1,1444	0,318	$<$ 0,001

Interpretação dos Efeitos:

- **Efeito da maturidade e socialização:** as variáveis idade (`age`) e envolvimento em relacionamento (`romantic`) apresentam impacto positivo na frequência de faltas. Em particular, estudantes em relacionamentos possuem uma taxa de faltas aproximadamente 52,6% maior ($RR = 1,526$) em comparação aos demais. Além disso, cada ano adicional de idade está associado a um aumento de 14,6% na taxa esperada de faltas ($RR = 1,146$).

- **Gestão de Tempo Livre:** níveis intermediários de tempo livre (categorias 3 e 4) estão associados a reduções de aproximadamente 40% na taxa de faltas em relação ao nível de referência, sugerindo que um equilíbrio nas atividades fora do ambiente escolar pode contribuir para menor absenteísmo.

- **Inflação de Zeros:** o intercepto da componente logística indica uma **Razão de Chances (OR)** de 0,318, sugerindo que uma parcela não negligenciável dos estudantes apresenta propensão estrutural a não registrar faltas, independentemente das covariáveis incluídas no modelo de contagem.

9.6 Conclusão

A aplicação empírica demonstrou que o modelo **ZINB** apresenta um ajuste altamente satisfatório para os dados de absenteísmo escolar, superando as limitações de distribuições menos flexíveis.

O uso da distribuição Binomial Negativa permitiu incorporar adequadamente a sobredispersão gerada pela elevada frequência de faltas de alguns estudantes (parâmetro estimado em $1,67$), enquanto a componente logística acomodou a parcela de alunos estruturalmente assíduos.

Os achados indicam que fatores como a idade e o envolvimento em relacionamentos românticos são impulsionadores diretos do absenteísmo, e que o equilíbrio no tempo livre fora da escola pode ser um fator de atenuação. O modelo ZINB confirmou-se como a ferramenta robusta e metodologicamente rigorosa para a análise deste fenômeno educacional.

Capítulo 10

Aplicação IV — Modelo ZIGP (Salamanders)

Neste capítulo, apresenta-se a aplicação do modelo **Poisson Generalizada Zero-Inflado (ZIGP)** ao conjunto de dados *Salamanders*. Este exemplo ilustra a flexibilidade da distribuição Poisson Generalizada na modelagem de dados de contagem em contextos ecológicos, nos quais fatores como a presença de mineração e diferenças entre espécies influenciam tanto a abundância quanto a probabilidade de ausência.

10.1 Descrição dos dados

Para a análise de impactos ambientais, utiliza-se o conjunto de dados *Salamanders*, disponibilizado no pacote `glmmTMB` Brooks et al. (2017) e baseado no estudo ecológico de Price et al. (2016).

O carregamento dos dados pode ser realizado diretamente no ambiente R, conforme ilustrado a seguir:

Script R: Carregamento dos dados Salamanders

```r
# 1. Carregar o banco de dados nativo
data("Salamanders", package = "glmmTMB")

# 2. Verificar a frequencia de zeros por local minerado
table(Salamanders$count == 0, Salamanders$mined)
```

10.2 Análise descritiva

O banco de dados registra a contagem de indivíduos observados em diferentes locais, com o objetivo de avaliar como a degradação ambiental afeta a densidade populacional.

A variável resposta corresponde ao número de indivíduos encontrados (`count`).

A observação preliminar dos dados evidencia uma grande massa de valores nulos estruturais, oriundos possivelmente da inabilidade da espécie sobreviver em determinadas condições ou por ausência natural no local amostrado. A contagem é modelada em função de covariáveis cruciais como a espécie analisada (`spp`) e a condição do local quanto à atividade prévia de mineração (`mined`).

10.3 Ajuste

O ajuste do modelo foi realizado utilizando a família `genpois` no `glmmTMB`. Diferentemente de especificações mais restritivas, o parâmetro de dispersão φ foi estimado livremente, permitindo que a variância se ajuste de forma flexível à média. Sob essa parametrização, a estrutura de variância é dada por $Var(Y) = \mu(1 + \varphi\mu)^2$.

Script R: Ajuste ZIGP via `glmmTMB`

```r
# 3. Carregar pacotes necessarios
library(glmmTMB)
library(DHARMa)

# 4. Ajuste do modelo ZIGP (Poisson Generalizada)
# A inflacao de zeros e modelada apenas pelo intercepto
   para simplificacao
modelo_salamanders <- glmmTMB(count ~ spp + mined,
                              ziformula = ~ 1,
                              data = Salamanders,
                              family = genpois)

# 5. Geracao de Residuos Quantilicos simulados (DHARMa)
res_sim <- simulateResiduals(fittedModel = modelo_
   salamanders, n = 1000)

# 6. Diagnostico visual
plot(res_sim)
```

10.4 Diagnóstico

O diagnóstico visual (Figura 10.1), via pacote `DHARMa`, em conjunto com os testes formais de resíduos quantílicos randomizados (RQR), indica que a distribuição Poisson Generalizada ajusta-se adequadamente aos dados analisados:

- **Uniformidade (KS Test):** O elevado p-valor ($p = 0,844$) não fornece evidências de desvio da uniformidade, sugerindo que os resíduos seguem o comportamento esperado sob um modelo bem especificado.
- **Dispersão:** O teste de dispersão ($p = 0,244$) demonstra que o parâmetro $\varphi = 2,97$ capturou com precisão a variabilidade dos dados, eliminando problemas de super-dispersão e indicando ausência de problemas de ajuste relacionados à variabilidade.
- **Outliers:** O teste de outliers ($p = 0,368$) não aponta excesso de valores extremos, indicando que a cauda da distribuição está bem representada pelo modelo.

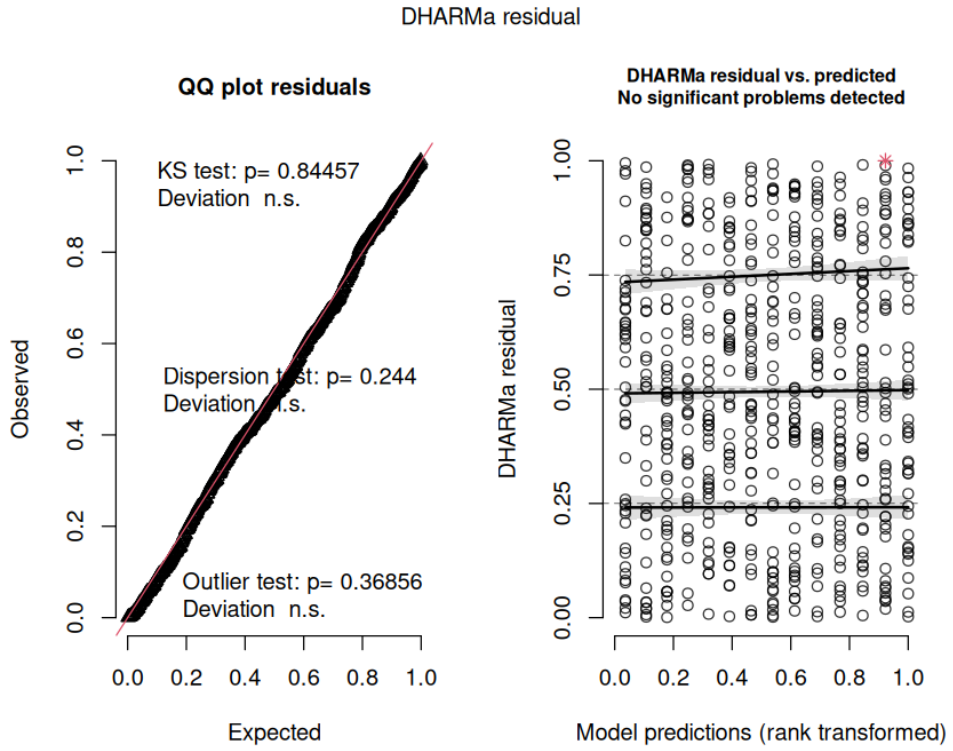

Figura 10.1: Resíduos quantílicos simulados evidenciando o ajuste superior do modelo **ZIGP**.

10.5 Resultados

As estimativas do modelo, com o parâmetro de dispersão estimado em 2,97, são apresentadas na Tabela 10.1. Os coeficientes foram exponencializados para interpretação em termos de **Razão de Taxas (RR)** e **Razão de Chances (OR)**.

Tabela 10.1: Estimativas do modelo Poisson Generalizada Zero-Inflado (ZIGP).

Variável	**Estimativa**	**RR / OR**	**p-valor**
Componente de Contagem (Poisson Generalizada)			
Intercepto	-1,1328	0,322	$< 0{,}001$
Espécie PR (`sppPR`)	-1,3596	0,256	$< 0{,}001$
Espécie DM (`sppDM`)	0,2631	1,301	0,188
Espécie EC-A (`sppEC-A`)	-0,7594	0,467	0,003
Espécie EC-L (`sppEC-L`)	0,4685	1,597	0,020
Espécie DES-L (`sppDES-L`)	0,6725	1,959	$< 0{,}001$
Espécie DF (`sppDF`)	0,1263	1,134	0,542
Mineração (Não) (`minedno`)	1,9785	7,232	$< 0{,}001$
Componente de Inflação (Logit)			
Intercepto	-1,9723	0,139	$< 0{,}001$

Interpretação dos Efeitos:

- **Impacto da Mineração:** a ausência de atividade de mineração (`minedno`) destaca-se como o principal fator associado à abundância de salamandras. Locais não minerados apresentam uma taxa de ocorrência aproximadamente **7,23 vezes maior** em comparação aos locais minerados, mantendo-se constantes as demais variáveis.

- **Variabilidade entre Espécies:** observa-se heterogeneidade relevante entre as espécies. A espécie *sppDES-L* apresenta uma taxa de contagem cerca de **duas vezes superior** ($RR = 1,959$) em relação à categoria de referência. Em contraste, a espécie sppPR apresenta a menor densidade relativa, com redução aproximada de 74% na taxa esperada ($RR = 0,256$).

- **Inflação de Zeros:** o intercepto da componente logística resultou em uma **Razão de Chances (OR)** de $0,139$, indicando baixa probabilidade de pertencimento ao estado de zero estrutural. Isso sugere que a maior parte das ausências observadas é explicada pelo próprio processo de contagem da Poisson Generalizada, e não por uma impossibilidade estrutural de ocorrência.

10.6 Conclusão

O modelo **ZIGP** provou ser uma ferramenta robusta e flexível para analisar dados ecológicos caracterizados por excesso de zeros e comportamentos complexos de sobredispersão.

A utilização da família Poisson Generalizada permitiu que o modelo acomodasse a variabilidade inerente às contagens populacionais por meio do parâmetro estimado ($\varphi = 2,97$), eliminando as falhas de ajuste comumente observadas sob a suposição restritiva da distribuição de Poisson simples.

A análise evidenciou empiricamente o drástico impacto ecológico causado pela mineração, sendo as áreas preservadas um fator de altíssima influência ($RR = 7{,}23$) para a ocorrência e abundância de indivíduos. Adicionalmente, verificou-se uma distinção clara entre as diferentes espécies quanto às suas frequências bases. A abordagem metodológica adotada validou-se como sendo adequada e necessária para a avaliação precisa desse cenário ambiental.

Capítulo 11

Conclusão Geral

O presente trabalho percorreu, de forma integrada, os aspectos teóricos e aplicados da modelagem de dados de contagem na presença de inflação de zeros. Partindo das limitações da distribuição de Poisson, especialmente quanto à suposição de equidispersão, evidenciou-se como a estrutura de modelos de mistura permite distinguir zeros estruturais de zeros amostrais, proporcionando inferências mais realistas e precisas..

As aplicações desenvolvidas com os conjuntos de dados *bioChemists*, *fishing*, *Student Performance* e *Salamanders* demonstraram que a escolha entre os modelos **ZIP**, **ZIG**, **ZINB** ou **ZIGP** não é meramente arbitrária, mas dependente da estrutura de de dispersão dos dados e pelas características do fenômeno em análise. Em particular, observou-se que modelos mais flexíveis, como o **ZINB** e o **ZIGP**, são capazes de acomodar padrões de variabilidade que vão além da simples inflação de zeros.

Adicionalmente, a utilização do pacote `DHARMa`, por meio de resíduos quantílicos simulados, mostrou-se fundamental para a validação dos modelos ajustados. Essa abordagem permitiu identificar desvios de especificação, sobredispersão residual e presença de outliers, superando limitações inerentes aos métodos tradicionais de diagnóstico.

Dessa forma, conclui-se que a modelagem com inflação de zeros constitui uma ferramenta indispensável na análise contemporânea de dados de contagem. Ao incorporar maior flexibilidade estrutural e métodos robustos de diagnóstico, esses modelos possibilitam representar de maneira mais fiel a complexidade de fenômenos reais, contribuindo para inferências estatísticas mais confiáveis e interpretáveis.

Agradecimentos

O presente trabalho foi realizado com o apoio da Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais (PRPq - UFMG) por meio do Programa de Iniciação Científica (PROBIC) e do Edital 11/2024, com financiamento da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Referências Bibliográficas

BROOKS, M. E. et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, v. 9, n. 2, p. 378–400, 2017.

CONSUL, P. C.; FAMOYE, F. Generalized poisson regression model. *Communications in Statistics - Theory and Methods*, v. 21, n. 1, p. 89–109, 1992.

CONSUL, P. C.; JAIN, G. C. A generalization of the poisson distribution. *Technometrics*, v. 15, n. 4, p. 791–799, 1973.

CORTEZ, P.; SILVA, A. M. G. Using data mining to predict secondary school student performance. In: *Proceedings of 5th Future Business Technology Conference*. Porto, Portugal: [s.n.], 2008. p. 5–12.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, v. 5, n. 3, p. 236–244, 1996.

FAMOYE, F.; SINGH, K. P. Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science*, v. 4, p. 117–130, 2006.

GREENE, W. H. Accounting for antecedent conditions in models of count data. *Journal of Econometrics*, v. 62, n. 1-2, p. 141–163, 1994.

HARTIG, F. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. [S.l.], 2022. R package version 0.4.6. Disponível em: <https://CRAN.R-project.org/package=DHARMa>.

HEILBRON, D. C. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, v. 36, n. 5, p. 531–547, 1994.

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

LONG, J. S. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage Publications, 1997.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, v. 135, n. 3, p. 370–384, 1972.

PRICE, S. J. et al. Effects of mountaintop removal mining and valley filling on the occupancy and abundance of stream salamanders. *Biological Conservation*, v. 202, p. 25–33, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>.

UCLA Statistical Consulting. *Zero-Inflated Data Analysis Examples*. 2021. Institute for Digital Research and Education. Disponível em: <https://stats.idre.ucla.edu/>.

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

ZEILEIS, A.; KLEIBER, C.; JACKMAN, S. Regression models for count data in r. *Journal of Statistical Software*, v. 27, n. 8, p. 1–25, 2008.