Sobredispersão em Modelos de Contagem: Análise da Demanda por Viagens Recreativas ao Lago Somerville

Paloma Vieira Borges*

2025

Resumo

Este trabalho investiga a presença de sobredispersão e excesso de zeros em dados de contagem por meio da análise da demanda por viagens recreativas ao Lago Somerville, no Texas. Utilizando o conjunto de dados `RecreationDemand`, foram ajustados três modelos: regressão de Poisson, regressão binomial negativa e modelo de Poisson inflado de zeros (ZIP). A análise exploratória revelou variância significativamente maior que a média e acentuada concentração de zeros, indicando potencial inadequação da Poisson. O teste de sobredispersão confirmou estatisticamente essa hipótese. Comparações com base em AIC, BIC e número de zeros estimados indicaram que a regressão binomial negativa apresenta melhor ajuste aos dados, ao mesmo tempo em que o modelo ZIP contribui para capturar a estrutura de zeros estruturais. Conclui-se que, em contextos com variância inflada e/ou excesso de zeros, é fundamental avaliar modelos alternativos para garantir estimativas e inferências consistentes.

Palavras-chaves: modelos de contagem; sobredispersão; modelo de regressão poisson; modelo binomial negativa; Poisson inflado de zeros.

*FGV EMAp, paloma.borges@fgv.edu.br

Introdução

Modelos estatísticos de contagem são amplamente utilizados para modelar eventos com um número inteiro e não negativo de ocorrências, como o número de acidentes, visitas ou viagens realizadas, por exemplo. Um dos modelos mais comumente utilizados para a análise de dados de contagem é a distribuição de Poisson, que representa a probabilidade de que um evento ocorra um certo número de vezes em um intervalo de tempo ou espaço, assumindo que os eventos ocorrem independentemente uns dos outros.

No entanto, uma limitação de utilizar este modelo é a premissa de *equidispersão*, ou seja, de que a média e a variância da variável resposta são iguais, o que raramente ocorre na prática com dados reais. Quando essa suposição não é satisfeita e a variância excede a média, ocorre o fênomeno de *sobredispersão*, o que pode levar à estimação inconsistente das variâncias para estimativas de parâmetros e, consequentemente, a inferências inválidas (CAMERON; TRIVEDI, 1990).

Neste contexto, este trabalho tem como objetivo investigar a presença de sobredispersão em um conjunto de dados sobre a demanda por viagens recreativas ao Lago Somerville, no Texas, no ano de 1980. O conjunto `RecreationDemand`, originalmente apresentado por SELLER; STOLL; CHAVAS, 1985, baseia-se em uma pesquisa com proprietários de barcos de lazer, registrando características econômicas e comportamentais dos entrevistados, bem como os custos associados a destinos de lazer similares ao Lago Somerville.

A análise será conduzida por meio da comparação de diferentes modelos de contagem — Poisson e modelos alternativos — com o intuito de identificar qual abordagem melhor representa os dados observados. Além da modelagem, também será realizado o teste de hipóteses proposto por CAMERON; TRIVEDI, 1990 para detectar formalmente a presença de sobredispersão.

Esta investigação busca contribuir para a escolha adequada de modelos estatísticos em contextos com a presença de dados com variância inflada (sobredispersão) e excesso de zeros, duas situações frequentemente observadas em aplicações envolvendo dados de contagem.

1 Métodos

Para esta análise, utilizamos o conjunto de dados `RecreationDemand`, disponível no pacote `AER` da linguagem R. A base consiste em uma amostra com 659 observações sobre o número de viagens recreativas ao Lago Somerville, Texas, baseada em uma pesquisa realizada com proprietários de barcos de lazer registrados em 23 condados do leste do estado.

A variável de interesse principal é o número de viagens ao lago (`trips`), sendo esse um dado de contagem. As demais variáveis envolvem características individuais (como renda e prática de atividades no local), custos associados a diferentes destinos alternativos e a avaliação da qualidade da infraestrutura do lago.

A Tabela 1 descreve as variáveis presentes no conjunto de dados, incluindo tipo e domínio de cada uma:

Tabela 1 – Descrição das variáveis do conjunto de dados *RecreationDemand*

Variável	**Descrição**	**Tipo**	**Domínio**
trips	Número de viagens recreativas ao Lago	Numérico	Inteiros não negativos
quality	Avaliação subjetiva da qualidade das instalações	Numérico	0 (não visitou) a 5
ski	Prática de esqui aquático no lago	Categórico	Sim, Não
income	Renda anual (mil dólares)	Numérico	$[0, \infty)$
userfee	Pagamento de taxa anual de uso do lago	Categórico	Sim, Não
costC	Custo da visita ao Lago Conroe (USD)	Numérico	$[0, \infty)$
costS	Custo da visita ao Lago Somerville (USD)	Numérico	$[0, \infty)$
costH	Custo da visita ao Lago Houston (USD)	Numérico	$[0, \infty)$

1.1 Modelo de Regressão de Poisson

O modelo de regressão de Poisson é uma generalização da distribuição de Poisson para casos em que a média da variável resposta depende de covariáveis. Seja Y_i o número de viagens do indivíduo i, assume-se que:

$$Y_i \sim \text{Poisson}(\mu_i), \quad \text{com} \quad \mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

onde μ_i é a média condicional de Y_i dado o vetor de covariáveis $\mathbf{x}_i$, e $\boldsymbol{\beta}$ é o vetor de parâmetros a ser estimado. A função de ligação utilizada é a função logarítmica, o que garante que $\mu_i > 0$ para todos os indivíduos.

Uma característica importante desse modelo é a *equidispersão*, isto é, $\mathbb{E}[Y_i] = \mathbb{V}[Y_i] = \mu_i$. No entanto, quando essa condição não é satisfeita, ou seja, quando a variância empírica excede a média, o modelo de Poisson tende a subestimar a variabilidade dos dados, tornando-se inadequado.

1.2 Teste de Sobredispersão

Para verificar a presença de sobredispersão nos dados, foi utilizado o teste de hipóteses proposto por CAMERON; TRIVEDI, 1990, que considera a seguinte parametrização para a variância:

$$\mathbb{V}[Y] = \mu + \alpha \cdot \mu^2,$$

em que α é um parâmetro que mede a intensidade da sobredispersão. Assim definimos a hipótese nula e alternativa como:

$$H_0 : \alpha = 0 \quad \text{(sem sobredispersão)}$$
$$H_1 : \alpha \neq 0 \quad \text{(com sobredispersão)}$$

O teste se baseia no fato de que sobre o modelo Poisson

$$E[(Y - \mu)^2 - Y] = 0$$

Assim, o teste é realizado em 3 etapas:

Primeiro, ajustamos um GLM Poisson e obtemos os valores de média ajustados $\hat{\mu}_i$. Então, o teste utiliza a seguinte estatística:

$$Z_i = \frac{(Y_i - \hat{\mu}_i)^2 - Y_i}{\hat{\mu}},$$

e o próximo passo consiste em regredir Z_i sobre $\hat{\mu}_i$ (sem intercepto).

O teste-t sobre o coeficiente obtido na regressão permite verificar a sua significância, que indica a presença (ou ausência) de sobredispersão. Um coeficiente significativamente diferente de zero sugere que a variância cresce mais rapidamente do que a média, violando a suposição de equidispersão do modelo de Poisson e motivando o uso de modelos alternativos.

1.3 Modelo de Regressão Binomial Negativa

Diante da evidência de sobredispersão, uma alternativa ao modelo de Poisson é a regressão binomial negativa, uma generalização da Poisson que inclui um parâmetro adicional para a dispersão. Essa extensão permite que a variância da variável resposta exceda sua média.

$$Y_i \mid X_i \sim NB(\mu_i, \phi),$$
$$\text{com } \mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \text{ e função de ligação } g(\mu_i) = log(\mu_i)$$

A forma da variância condicional é dada por:

$$\text{Var}(Y_i \mid X_i) = \mu_i + \phi\mu_i^2$$

em que ϕ representa o parâmetro de sobredispersão.

A função de ligação utilizada foi a função logarítmica, pois garante que os valores de μ_i sejam positivos, além de possuir interpretação direta dos coeficientes como efeitos multiplicativos, assim como no modelo de Poisson.

O modelo foi ajustado por meio da função `glm.nb()` do pacote `MASS` na linguagem R, que realiza a estimação dos parâmetros por máxima verossimilhança. O algoritmo utiliza um processo iterativo que alterna entre a estimação dos coeficientes do modelo e a atualização do valor de ϕ, maximizando a verossimilhança marginal. Essa estratégia fornece estimativas mais precisas e confiáveis, especialmente em contextos com forte sobredispersão.

1.4 Modelo de Poisson Inflado de Zeros (ZIP)

A elevada proporção de observações com valor zero na variável resposta sugere que esses valores não sejam explicados apenas pelo processo de contagem, mas também por um processo adicional de geração de dados, o que pode ser uma outra possível causa de sobredispersão.

Diante disso, um modelo frequentemente usado neste cenário é o modelo de Poisson Inflado de Zeros (ZIP), que combina dois componentes: um modelo de contagem de Poisson e um modelo logit para prever o excesso de zeros.

Ou seja, assume-se que os zeros vêm de duas fontes distintas: os "zeros verdadeiros" e os "zeros em excesso". O modelo inflacionado com zeros estima duas equações em simultâneo, uma para o modelo de contagem e outra para prever a probabilidade de uma observação ser um zero em excesso.

No ajuste realizado neste estudo, a parte inflacionada foi especificada apenas com um intercepto — isto é, assumiu-se que a probabilidade de um zero estrutural é constante entre os indivíduos.

1.5 Métricas de Avaliação de Ajuste

Para avaliar o desempenho dos modelos ajustados, foram utilizadas as métricas de bondade do ajuste AIC (*Akaike Information Criterion*) e BIC (*Bayesian Information Criterion*), que consideram tanto a qualidade do ajuste quanto a complexidade do modelo.

O AIC é definido como $\text{AIC} = 2k - 2\ln(\hat{L})$, onde k é o número de parâmetros e $\hat{L}$ é o valor máximo da função de verossimilhança.

Já o BIC é calculado por $\text{BIC} = k\log(n) - 2\ln(\hat{L})$, sendo n o número de observações. O BIC penaliza mais severamente a complexidade do modelo.

Em ambos os casos, menores valores indicam melhor equilíbrio entre ajuste e parcimônia.

Além dessas métricas, a comparação entre o número de zeros observados e o número de zeros estimados por cada modelo também é utilizada como medida de comparação, especialmente relevante em dados com inflação de zeros.

2 Resultados

2.1 Análise Exploratória

O estudo teve início com a realização de uma análise exploratória dos dados, a fim de investigar a relação entre as covariáveis e a variável resposta `trips`, que representa o número de viagens recreativas ao Lago Somerville. Nesta etapa, foi utilizada a linguagem Python.

A Tabela 2 apresenta estatísticas descritivas da variável `trips`. A média de viagens é de aproximadamente 2,24, enquanto a variância é de 39,60, valor que excede amplamente a média e indica forte indício sobredispersão.

Tabela 2 – Estatísticas descritivas da variável `trips`

Medida	**Valor**
Número total de observações	659
Média	2,24
Variância	39,60
Valor mínimo	0
Valor máximo	88
Número de observações 0	417

A Figura 1 apresenta o histograma de `trips`. Nota-se uma distribuição fortemente assimétrica à direita, com uma concentração expressiva de zeros (cerca de 63% das observações).

A matriz de correlação (Figura 2) mostra que as variáveis `quality` e `userfee` apresentam as maiores correlações com `trips` (0,39 e 0,28, respectivamente). Esse resultado está de acordo com a interpretação intuitiva das variáveis: indivíduos que avaliam melhor

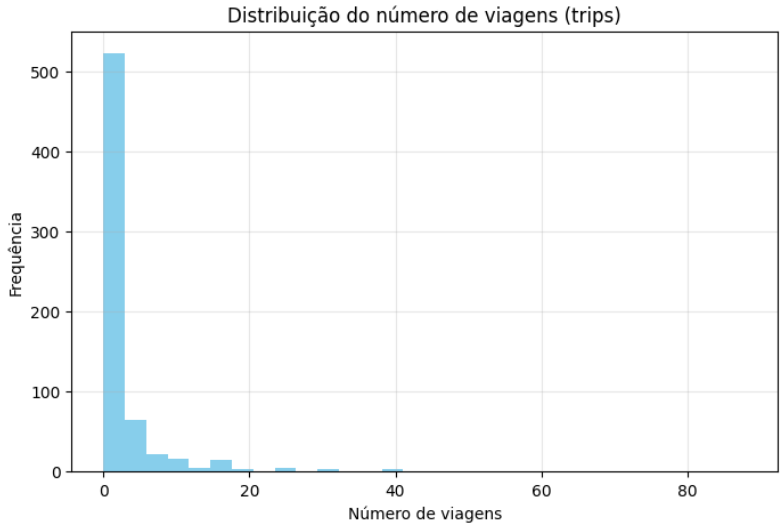

Figura 1 – Distribuição da variável ‘trips’

a qualidade das instalações do lago ou que pagam a taxa de uso tendem a realizar mais visitas ao local.

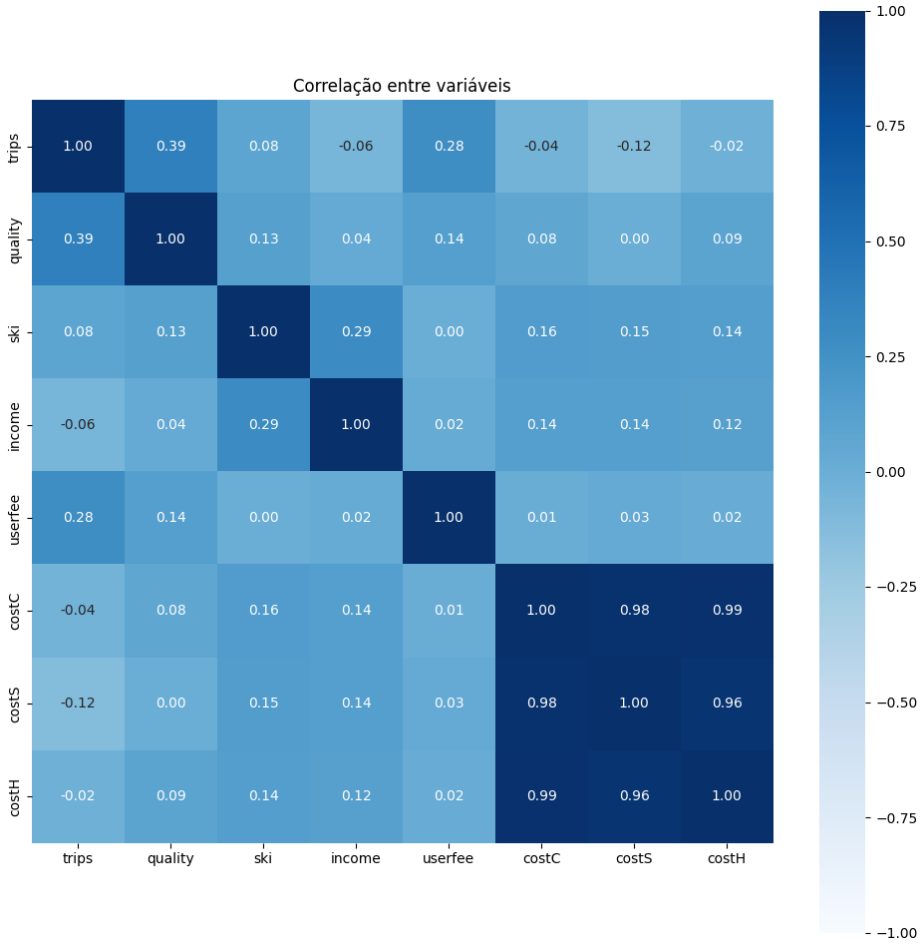

Figura 2 – Matriz de correlação entre as variáveis

O boxplot na Figura 3 reforça essa tendência, confirmando que indivíduos que pagaram a taxa de uso apresentaram média de aproximadamente 15 viagens, enquanto aqueles que não pagaram, apenas 2. A variabilidade no número de viagens também é consideravelmente maior entre os pagantes.

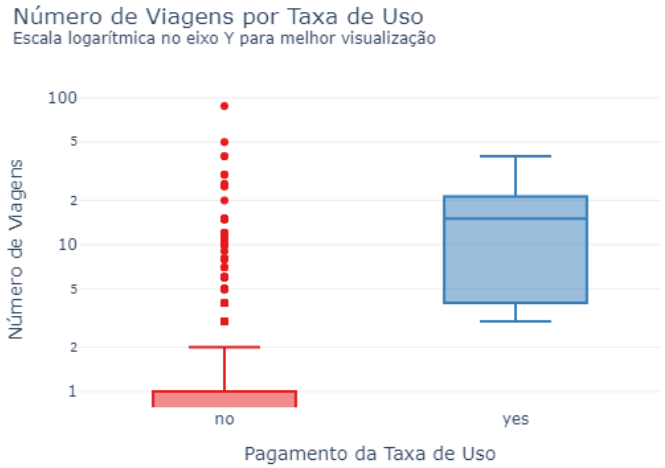

Figura 3 – Boxplot do número de viagens por pagamento da taxa

A Figura 4 evidencia uma relação positiva entre a avaliação da qualidade do lago e o número médio de viagens.

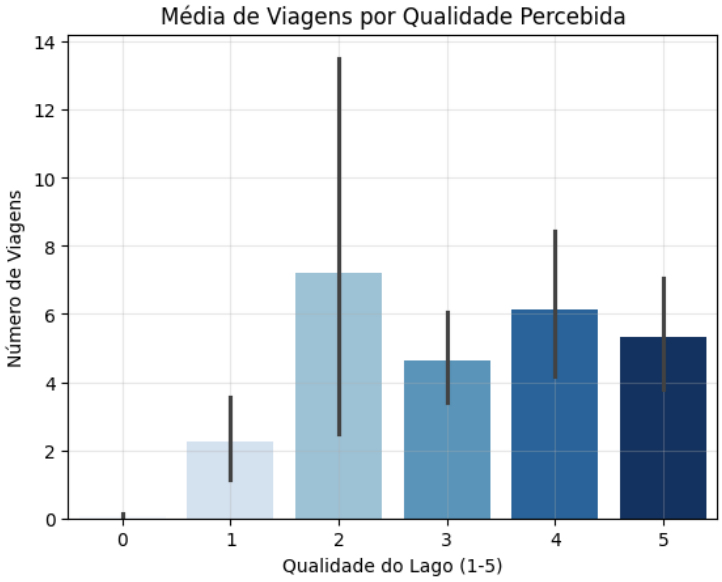

Figura 4 – Média de viagens por avaliação da qualidade das instalações

As variáveis de custo de visita (`costC`, `costS`, `costH`) apresentam correlação negativa, ainda que fraca, com `trips`, indicando que maiores custos tendem a estar associados a um menor número de viagens. Isso é visualizado na Figura 5, que mostra uma tendência decrescente no número de viagens conforme os custos aumentam, embora com

grande dispersão e predominância de zeros. A alta correlação entre os custos de visita aos diferentes lagos é explicada pelos valores semelhantes dessas variáveis.

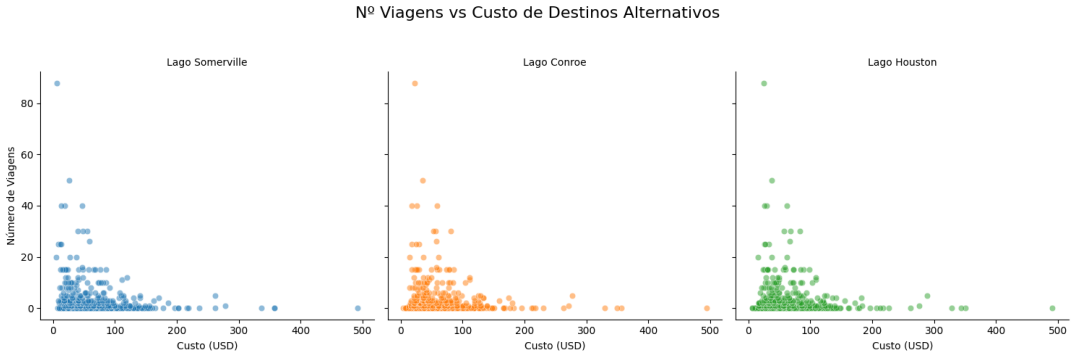

Figura 5 – Relação entre número de viagens e custos alternativos

A variável `ski` também apresentou correlação fraca com `trips`, mas a Figura 6 mostra que indivíduos que praticaram esqui aquático tendem a realizar, em média, mais viagens ao lago do que os demais.

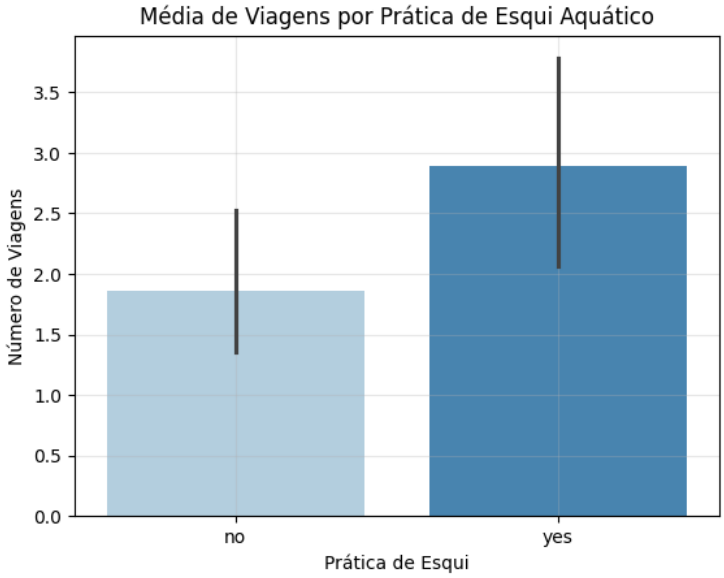

Figura 6 – Média de viagens por prática de esqui aquático

Por fim, ao analisar a relação de *income* com *trips*, vemos que existe uma baixa correlação entre as duas e a dispersão observada na Figura 7 não indica uma relação clara entre renda e número de viagens. Diante disso, optou-se por não incluir `income` nos modelos ajustados.

2.2 Modelagem dos Dados

Com base nas evidências de sobredispersão identificadas na análise exploratória, foram ajustados três modelos aos dados: primeiramente um modelo de regressão de Poisson; em seguida, um modelo de regressão binomial negativa, que incorpora um parâmetro adicional de dispersão; e, por fim, um modelo de Poisson Inflado de Zeros (ZIP). O

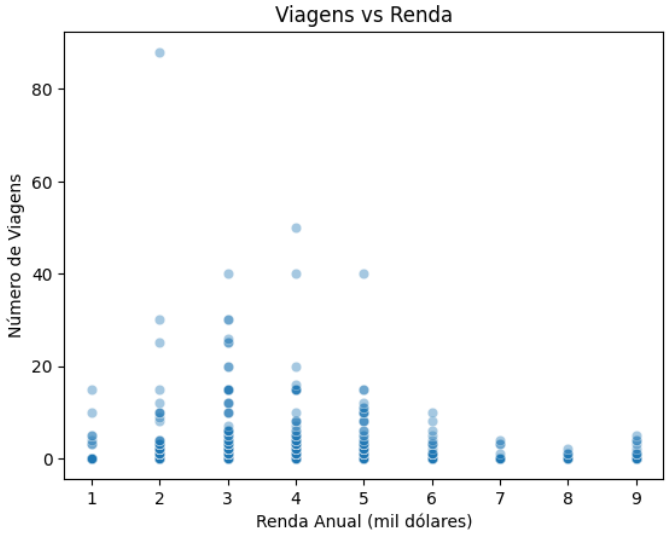

Figura 7 – Relação da renda e número de viagens

objetivo foi comparar o desempenho desses modelos em representar adequadamente os dados observados.

Os modelos foram ajustados na linguagem R, por meio das funções `glm()` para o modelo de Poisson, `glm.nb()` do pacote `MASS` para o modelo Binomial Negativo, e `zeroinfl()` do pacote `pscl` para o modelo ZIP. A escolha das covariáveis foi orientada pela análise exploratória, que indicou relação significativa entre `trips` e as covariáveis `quality`, `userfee`, `ski`, `costS`, `costC` e `costH`.

A Tabela 3 apresenta os coeficientes estimados para os modelos de Poisson e Binomial Negativa:

Tabela 3 – Coeficientes estimados nos modelos de Poisson e Binomial Negativa

	Poisson	**Binomial Negativa**
Intercepto	-0.077	-1.207
quality	0.467	0.723
ski (yes)	0.331	0.599
userfee (yes)	0.921	0.668
costS	-0.044	-0.093
costH	0.037	0.039
costC	-0.004	0.048

Os sinais e significâncias das variáveis permanecem consistentes entre os modelos (com excessão de `costC`). Em ambos, destaca-se o efeito positivo de `quality`, `ski` e `userfee`, e negativo para `costS`, o que é coerente com a expectativa de que maiores custos reduzem o número esperado de viagens ao lago. Observa-se, também, que os efeitos de `costH` e `costC` foram mais discretos e com menor impacto.

Após o ajuste do modelo de Poisson, foi aplicado o teste proposto por CAMERON; TRIVEDI, 1990, para verificar a hipótese de sobredispersão formalmente. O teste resultou em um p-valor muito pequeno (< 0.05), além de um coeficiente estimado significativamente diferente de zero ($\hat{\alpha} = 39.09$), confirmando a presença de sobredispersão e a inadequação

do modelo de Poisson para os dados em questão.

Outro aspecto relevante é a capacidade do modelo de representar os zeros. O conjunto de dados apresenta 417 observações com valor zero ($\approx 63\%$), enquanto o modelo de Poisson estimou apenas 272 zeros, subestimando expressivamente essa quantidade. Isso indica que o modelo falha também em capturar o excesso de zeros.

Assim, adicionalmente aos dois modelos iniciais, foi ajustado um modelo de Poisson Inflado de Zeros (ZIP), que incorpora uma estrutura para zeros adicionais. A parte inflacionada do modelo (probabilidade de uma observação ser um zero estrutural) foi especificada com apenas um intercepto, assumindo uma probabilidade constante de inflação de zeros entre os indivíduos. Os coeficientes estimados estão resumidos na Tabela 4:

Tabela 4 – Coeficientes estimados no modelo ZIP

	Estimativa	**Erro Padrão**
Parte de contagem		
Intercepto	1.390	0.140
quality	0.155	0.034
ski (yes)	0.403	0.058
userfee (yes)	0.667	0.080
costS	-0.039	0.002
costH	0.029	0.003
costC	-0.002	0.004
Parte inflacionada		
Intercepto	0.254	0.109

Observa-se que os coeficientes da parte de contagem mantêm sinais similares aos modelos anteriores. Além disso, a parte inflacionada teve intercepto significativamente positivo, sugerindo que há, de fato, a possibilidade de um processo gerador de zeros estruturais presente nos dados.

Para avaliar o desempenho dos modelos, foram calculados o AIC, o BIC e a quantidade de zeros estimada por cada modelo (Tabela 5):

Tabela 5 – Comparação de AIC, BIC e Estimativa de Zeros dos modelos ajustados

Modelo	**AIC**	**BIC**	**Zeros Estimados**
Poisson	3107.1	3138.5	272
Poisson Inflado de Zeros (ZIP)	2902.8	2938.7	218
Binomial Negativa	1667.4	1703.4	370

Em resumo, o modelo de Poisson mostrou-se inadequado, tanto por subestimar a variância quanto por não representar adequadamente a presença de zeros nos dados, além do teste formal ter confirmado a presença de sobredispersão.

O modelo ZIP melhora substancialmente os critérios de AIC e BIC em relação ao Poisson, mas ainda assim apresenta desempenho inferior ao da Binomial Negativa.

A regressão binomial negativa apresentou os menores valores de AIC e BIC, além de prever um número de zeros mais próximo do observado. Esses resultados, aliados à

confirmação da sobredispersão, justificam sua escolha como o modelo mais adequado para representar os dados.

3 Conclusão

Este trabalho teve como objetivo investigar a presença de sobredispersão e excesso de zeros em um conjunto de dados de contagem e, a partir disso, avaliar a adequação de diferentes modelos estatísticos. Utilizando o conjunto de dados `RecreationDemand`, observou-se desde a análise exploratória que a variável resposta apresentava uma variância muito superior à média, além de uma concentração elevada de observações com valor zero.

A aplicação do teste de sobredispersão confirmou estatisticamente a inadequação do modelo de Poisson, justificando a necessidade de modelos alternativos. A regressão binomial negativa, que incorpora um parâmetro adicional de dispersão, apresentou o melhor ajuste, com menores valores de AIC e BIC e maior capacidade de capturar a variabilidade presente nos dados.

Em paralelo, também foi ajustado um modelo Poisson Inflado de Zeros (ZIP), motivado pela alta proporção de zeros observados. O modelo apresentou desempenho superior ao de Poisson, mas inferior ao modelo binomial negativa. Ainda assim, o intercepto significativo na parte inflacionada do modelo ZIP indica que parte dos zeros pode, de fato, ser estrutural, revelando que ambos os mecanismos podem estar presentes, o que justifica a consideração em contextos semelhantes.

Esses resultados reforçam a importância de diagnosticar adequadamente a estrutura dos dados antes da escolha do modelo estatístico. Em contextos onde a variância excede a média ou há acúmulo de zeros, modelos como a binomial negativa ou ZIP devem ser considerados, de modo a obter estimativas mais consistentes e inferências válidas, minimizando o risco de se obter estimativas imprecisas com o uso da Poisson padrão.

Uma limitação deste estudo é a suposição de independência entre observações e a simplicidade da especificação dos modelos, especialmente na parte inflacionada do ZIP. Além disso, não foram explorados modelos ainda mais flexíveis, como o modelo binomial negativa inflado de zeros (ZINB).

Para superar essas limitações, estudos futuros podem considerar a inclusão de variáveis explicativas na parte inflacionada do ZIP, bem como avaliar o desempenho de modelos mais complexos, como o ZINB. Também seria interessante explorar abordagens bayesianas para modelar melhor a heterogeneidade da amostra.

Em conclusão, este estudo fornece uma base para a escolha de modelos mais robustos em contextos com variância inflada e excesso de zeros, características comuns em aplicações reais envolvendo dados de contagem.

Referências

CAMERON, A.; TRIVEDI, P. K. Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, Elsevier, v. 46, n. 3, p. 347–364, 1990. ISSN 0304-4076. Disponível em: <https://www.sciencedirect.com/science/article/pii/030440769090014K>. Citado 3 vezes nas páginas 2, 3 e 9.

SELLER, C.; STOLL, J. R.; CHAVAS, J.-P. Validation of empirical measures of welfare change: A comparison of nonmarket techniques. *Land Economics*, University of Wisconsin Press, v. 61, n. 2, p. 156–175, 1985. ISSN 00237639. Disponível em: <https://www.jstor.org/stable/3145808>. Citado na página 2.

1

[1] Os códigos utilizados nesta análise estão disponíveis em: <https://github.com/palomavb/stats_modelling/>