

Sobredispersão em Modelos de Contagem

Modelagem Estatística

Pedro Henrique Coterli

24 de maio de 2025

1 Introdução

O presente trabalho tem como objetivo investigar a sobredispersão em dados de contagem e seus efeitos no ajuste de modelos estatísticos, como modelos lineares generalizados (GLMs). Serão considerados modelos de regressão Poisson, Binomial Negativo e Poisson inflado de zeros, aplicados a dados de turismo da região do Texas, nos Estados Unidos. Busca-se avaliar como a sobredispersão afeta o desempenho desses modelos, bem como discutir estratégias para reduzir seus impactos e melhorar a qualidade dos ajustes.

1.1 Os dados

Os dados utilizados para este estudo foram retirados da biblioteca Applied Econometrics with R (AER)¹. Será utilizada a base RecreationDemand², que contém dados sobre o número de viagens de barco recreativas ao Lago Somerville, no Texas, em 1980, com base em uma pesquisa administrada a 2000 proprietários de barcos de lazer registrados em 23 condados do leste do estado.

Estão presentes 659 observações com 8 variáveis, que estão descritas a seguir:

- **trips**: assume valores naturais (incluindo 0) e representa o número de viagens de barco recreativas ao lago Somerville em 1980;
- **quality**: assume valores de 1 a 5 e representa a classificação fornecida pelo indivíduo sobre a qualidade das instalações do lago;
- **ski**: assume valores “yes” ou “no” e indica se o indivíduo praticava (ou se praticaria, caso nunca tivesse ido) esqui aquático no lago;
- **income**: assume valores naturais e representa a renda familiar anual do entrevistado (em milhares de dólares);
- **userfee**: assume valores “yes” ou “no” e indica se o indivíduo pagou uma taxa anual de uso do Lago Somerville;
- **costC**: assume valores positivos e representa as despesas estimadas para visitar o Lago Conroe (em dólares);
- **costS**: assume valores positivos e representa as despesas estimadas para visitar o Lago Somerville (em dólares);

- **costH**: assume valores positivos e representa as despesas estimadas para visitar o Lago Houston (em dólares).

Uma limitação apresentada por esses dados diz respeito à variável **quality**: apesar de possuir uma escala de 1 a 5, ela recebe o valor 0 para indivíduos que não haviam visitado o lago Somerville, que, como veremos mais adiante, compõem a maioria dos registros.

2 Métodos

2.1 Os modelos

No presente trabalho, o objetivo no que se refere à modelagem será regredir a variável de contagem **trips** sobre outras das variáveis descritas anteriormente. A seleção dessas variáveis será realizada por meio de uma análise exploratória de cada uma delas, em que serão exploradas tanto suas distribuições univariadas quanto suas relações com a variável resposta.

Serão considerados nesse estudo 3 modelos estatísticos, descritos a seguir.

2.1.1 Modelo Poisson

Frequentemente utilizada para modelar dados de contagem, a regressão Poisson possui a seguinte forma:

$$Y_i \sim \text{Poisson}(\theta_i) \\ \mathbb{E}[Y_i] = \mu_i = e^{X_i^T \beta}$$

onde Y_i é a variável resposta (nesse caso, **trips**) e X_i é o vetor das covariáveis de interesse, ambos indexados pelo i -ésimo registro. Aqui, a função de ligação utilizada é a canônica para a família Poisson: $g(\mu_i) = \log(X_i^T \beta)$.

Para o processo de estimação do vetor de parâmetros β , será utilizado o método de máxima verossimilhança (MV). Algumas das razões para tal escolha são:

- Não há informação a priori para ser fornecida ao modelo. Assim, a distribuição a priori de β seria uma distribuição pouco informativa e sua posteriori seria praticamente a verossimilhança, sendo, portanto, equivalente ao método de MV.
- A otimização numérica para a busca da estimativa de MV é matemática e computacionalmente mais

¹<https://rdrr.io/cran/AER/>

²<https://rdrr.io/cran/AER/man/RecreationDemand.html>

simples que a necessária na abordagem Bayesiana. Enquanto a primeira utiliza de métodos simples como Fisher Scoring e Iterative Weighted Least Squares (IWLS), a segunda necessita de algoritmos mais complexos como Markov Chain Monte Carlo (MCMC) e Variational Inference.

- Dado que a base utilizada possui uma quantidade razoável de amostras (mais de 600), as estimativas de MV provavelmente apresentarão um bom desempenho, graças a sua normalidade assintótica.

Para a aproximação numérica da estimativa de máxima verossimilhança, será utilizado o método de Fisher Scoring, que atualiza o vetor de parâmetros da seguinte forma:

$$\beta^{(t+1)} = \beta^{(t)} + [\mathbb{E}[-\nabla^2 l(\beta^{(t)})]]^{-1} \nabla l(\beta^{(t)})$$

onde $l(\beta)$ é a função de log-verossimilhança, $\nabla l(\beta)$ é o gradiente (score) e $\nabla^2 l(\beta)$ é a Hessiana, todos avaliados em β . Esse método difere do Newton-Raphson clássico ao substituir a matriz Hessiana pela sua esperança, que, com alguns ajustes de sinais, torna-se a informação de Fisher. Algumas razões por essa preferência são:

- Em GLMs, a informação de Fisher tem uma forma conhecida, podendo ser calculada como $\mathbb{I}(\beta) = X^T W X$, onde X é a matriz de desenho (dos dados) e W é uma matriz diagonal com a i -ésima entrada sendo:

$$w_i = \left(\frac{d\mu_i}{d\eta_i} \right)^2 \cdot \frac{1}{\text{Var}(Y_i)},$$

com $\eta_i = X_i^T \beta$. Graças a isso, esse algoritmo pode ser implementado como um simples Iterative Weighted Least Squares (IWLS), sendo resolvido de forma computacionalmente eficiente ao solucionar um problema de mínimos quadrados ponderados a cada iteração.

- Se a função de ligação utilizada for a canônica, derivada da distribuição da família exponencial, então os métodos são equivalentes, coincidindo assintoticamente. No entanto, o Fisher Scoring geralmente funciona de forma mais robusta e é mais eficiente computacionalmente, como citado no item anterior.

2.1.2 Modelo Binomial Negativo

Utilizado principalmente como um substituto do modelo Poisson, o modelo Binomial Negativo possui a seguinte forma:

$$Y_i \sim \text{NBin}(\mu_i, \phi) \\ \mathbb{E}[Y_i] = \mu_i = e^{X_i^T \beta}$$

onde a função de massa de probabilidade dessa parametrização da distribuição Binomial Negativa é a seguinte³:

$$\text{NBin}(y|\mu, \phi) = \binom{y + \phi - 1}{y} \left(\frac{\mu}{\mu + \phi} \right)^y \left(\frac{\phi}{\mu + \phi} \right)^\phi$$

Esse modelo é adequado para ajustar dados de contagem, assim como o Poisson. No entanto, ao contrário deste, ele fornece a possibilidade de modelar separadamente os valores da média e da variância (fixos como iguais no modelo Poisson). Com isso, ele possui a capacidade de ajustar-se melhor a dados de contagem que apresentam sobredispersão.

Nessa parametrização, o parâmetro da média permanece com o mesmo significado. Entretanto, agora há um parâmetro a mais: ϕ , que modela o inverso da sobredispersão. Isso é possível pois a variância passa a ser definida como:

$$\text{Var}(Y) = \mu + \frac{\mu^2}{\phi}$$

Com isso, quanto menor o valor de ϕ , maior é a variância dos dados e mais distante da média ela está. Analogamente, quando $\phi \rightarrow \infty$, essa variância se aproxima da própria média e o modelo volta a ser o de Poisson. Dessa forma, passa a ser possível modelar dados de contagem de modo que a variância não seja idêntica à média.

Mais uma vez, a função de ligação desse modelo é o logaritmo: $\log(\mu_i) = X_i^T \beta$, devido a ela ser a função de ligação canônica dessa parametrização da distribuição Binomial Negativa na forma da família exponencial. Sua utilização proporciona as vantagens descritas na seção anterior com relação ao uso do método de Fisher Scoring para a aproximação numérica da estimativa de MV.

Assim como com o modelo Poisson, será aplicado o método de máxima verossimilhança para estimativa dos parâmetros, pelas mesmas justificativas.

Além disso, será novamente aplicado o método de otimização por Fisher Scoring para o cálculo da estimativa do vetor de parâmetros β , e a explicação para tal escolha é a mesma do modelo Poisson. A única diferença é que, como agora há “dois” parâmetros a serem estimados (o vetor β e o valor ϕ), será adotada uma estratégia de otimização em duas etapas:

1. Com ϕ fixo, β é encontrado por MV com Fisher Scoring.

³https://mc-stan.org/docs/functions-reference/unbounded_discrete_distributions.html#nbalt

2. Com β fixo, ϕ é encontrado numericamente maximizando a verossimilhança atual.

Esse processo é repetido até os parâmetros convergirem.

2.1.3 Modelo Poisson inflado de zeros

O modelo Poisson inflado de zeros (ZIP da sigla em inglês) é uma adaptação do modelo Poisson para cenários com excesso de zeros na variável resposta. É um modelo misto definido da seguinte forma⁴:

$$\begin{aligned} y_i &= 0 && \text{com probabilidade } \theta_i, \text{ e} \\ y_i &\sim \text{Poisson}(\mu_i) && \text{com probabilidade } 1 - \theta_i \end{aligned}$$

onde β é estimado por um modelo Poisson para gerar $\mu_i = e^{Z_i^T \beta}$ e γ é estimado por um modelo logístico para gerar $\theta_i = \text{logit}^{-1}(W_i^T \gamma)$. Aqui, Z_i e W_i são vetores com covariáveis do i -ésimo ponto de dado, podendo conter variáveis diferentes ou comuns e até dimensões diferentes.

Assim, para o i -ésimo ponto de dado, existe probabilidade θ_i de observar um 0 e probabilidade $1 - \theta_i$ de observar uma amostra de uma distribuição Poisson(μ_i). Com isso, esse modelo permite separar zeros estruturais, ou seja, que não são explicados pelas covariáveis explicativas da contagem, de zeros ocasionais, gerados pela combinação de valores dessas covariáveis. Dessa forma, ele é capaz de modelar todos esses zeros, que prejudicariam o desempenho de um modelo Poisson convencional.

Sua função de verossimilhança pode ser escrita como a seguir:

$$p(y_i | \theta_i, \mu_i) = \begin{cases} \theta_i + (1 - \theta_i) \cdot e^{-\mu_i}, & \text{se } y_i = 0 \\ (1 - \theta_i) \cdot \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, & \text{se } y_i > 0 \end{cases}$$

Para estimar os coeficientes β e γ , será aplicada a estratégia de máxima verossimilhança, utilizando as mesmas justificativas dos modelos anteriores. Além disso, para realizar a aproximação numérica, serão utilizados métodos numéricos que não serão discutidos aqui, dada sua complexidade. O método de Fisher Scoring não pode ser aplicado diretamente devido à estrutura mista do problema.

2.2 Teste de sobredispersão sob modelo Poisson

Após o ajuste do modelo Poisson aos dados de interesse, será realizado um teste estatístico para identificar a possibilidade de existência de sobredispersão nos dados, ou seja, para verificar se sua variância difere de sua média.

⁴<https://mc-stan.org/docs/stan-users-guide/finite-mixtures.html#zero-inflation>

O teste utilizado foi descrito por Cameron & Trivedi (1990) e baseia-se na seguinte modelagem da variância da variável resposta:

$$\text{Var}(Y_i) = \mu_i + \alpha \mu_i^2$$

Assim, a hipótese nula diz que $\alpha = 0$ e que, portanto, o modelo de Poisson é adequado. Por outro lado, a hipótese alternativa afirma que $\alpha > 0$, tornando esse modelo inadequado.

Primeiramente, deve-se ajustar um modelo de Poisson aos dados. Em seguida, é calculada a seguinte estatística de teste:

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

onde $\hat{\mu}_i$ é o valor estimado para a média da i -ésima observação obtido pelo modelo de Poisson ajustado. Essa estatística mede a diferença entre a variância observada e a variância esperada sob a hipótese nula. Nota-se que, sob essa hipótese:

$$\begin{aligned} \mathbb{E}[Z_i] &= \frac{\mathbb{E}[(Y_i - \hat{\mu}_i)^2] - \mathbb{E}[Y_i]}{\hat{\mu}_i} = \frac{\text{Var}(Y_i) - \hat{\mu}_i}{\hat{\mu}_i} = \\ &= \frac{\hat{\mu}_i - \hat{\mu}_i}{\hat{\mu}_i} = 0 \end{aligned}$$

Com base nisso, a etapa final desse teste consiste em ajustar uma regressão linear sem intercepto de Z sobre $\hat{\mu}$, de forma que $\mathbb{E}[Z_i] = \beta \hat{\mu}_i$. Assim, caso o modelo Poisson seja adequado, a expectativa é que β seja próximo de 0. Em outras palavras, β aproxima o parâmetro α da parametrização da variância descrita acima. Portanto, caso β seja estatisticamente significativo e positivo, há forte evidência da presença de sobredispersão.

2.3 Critérios de avaliação

A seguir, estão descritos os principais critérios utilizados para a avaliação do desempenho dos modelos ajustados.

2.3.1 Intervalo de confiança aproximado dos parâmetros

Para a avaliação da significância dos parâmetros do ajuste, será utilizada a seguinte aproximação assintótica:

$$z_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \approx \mathcal{N}(0, 1)$$

onde $se(\hat{\beta}_j) \approx \sqrt{[\mathbb{I}(\hat{\beta})^{-1}]_{jj}}$.

Assim, é possível calcular um intervalo de confiança aproximado para cada parâmetro e verificar se ele contém o valor 0. Caso não contenha, é altamente provável que ele seja estatisticamente significativo.

2.3.2 Resíduos de Pearson

Os resíduos de Pearson são uma forma de quantificar a distância entre os valores observados e os valores preditos por um modelo. O resíduo para o i -ésimo ponto de dado é calculado da seguinte forma:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}$$

Dessa forma, quanto mais esses valores se aproximarem de 0, melhor é o ajuste do modelo. Esses resíduos serão plotados em função do valor da variável de resposta observada y_i (**trips**) para melhor análise tanto interior ao modelo quanto entre modelos.

Uma possibilidade considerada para avaliação do ajuste foi a estatística qui-quadrado de Pearson:

$$X^2 = \sum r_i^2$$

citada por Dobson (2018). De acordo com a autora, sob a hipótese de que o modelo está correto, essa estatística teria aproximadamente a distribuição $\chi^2(N - p)$, onde N é o número de pontos de dados e p é o número de parâmetros do modelo. No entanto, essa aproximação é pobre se as frequências esperadas são muito pequenas, que é o caso desse problema, como será mostrado mais adiante. Portanto, a utilização dessa estatística não é viável.

2.3.3 Estatística qui-quadrado de razão de verossimilhança

Seja $l(\hat{\beta}; y)$ a log-verossimilhança do modelo de interesse avaliada no estimador de máxima verossimilhança (EMV) e $l(\tilde{\beta}; y)$ a log-verossimilhança do modelo minimal (ou seja, ajustado apenas com o intercepto) também avaliada em seu EMV. Assim, a estatística qui-quadrado de razão de verossimilhança (C) pode ser calculada como:

$$C = 2[l(\hat{\beta}; y) - l(\tilde{\beta}; y)]$$

Isso é equivalente a dizer que:

$$C = D_{min} - D_{model}$$

onde D_{min} e D_{model} são as deviances dos modelos minimal e de interesse, respectivamente. Essas deviances são calculadas como:

$$D_{\mathcal{M}} = 2[l(\hat{\beta}; y) - l(\tilde{\beta}; y)]$$

onde $l(\hat{\beta}; y)$ é a log-verossimilhança do modelo saturado (ou seja, com o número máximo de parâmetros que podem ser estimados) avaliada em seu EMV.

Segundo Dobson (2018), a distribuição amostral aproximada para C é $\chi^2(p - 1)$ sob a hipótese de que todos os p parâmetros, exceto o termo do intercepto, são zero. Assim, C é uma estatística de teste para a hipótese de que nenhuma das covariáveis é necessária para um modelo parcimonioso. Dessa forma, se o valor de C for estatisticamente significativo comparado com essa distribuição qui-quadrado, então é altamente provável que as covariáveis utilizadas sejam relevantes.

2.3.4 AIC

O Akaike Information Criterion (AIC) é uma medida utilizada para avaliar e comparar modelos ajustados a um conjunto de dados. Ele pode ser interpretado como uma estimativa do risco preditivo de um modelo, ou seja, do erro esperado ao utilizá-lo para prever novos dados. Assim, teoricamente, quanto menor seu valor, melhor o modelo.

Seu cálculo dá-se pela seguinte fórmula:

$$AIC = 2p - 2l(\hat{\beta}; y)$$

com os mesmos significados já definidos anteriormente.

2.3.5 Pseudo- R^2

O pseudo- R^2 é um análogo do R^2 da regressão linear múltipla para outros modelos, como regressão logística, Poisson e Binomial Negativo. Ele é calculado da seguinte forma:

$$\text{pseudo-}R^2 = \frac{l(\tilde{\beta}; y) - l(\hat{\beta}; y)}{l(\tilde{\beta}; y)}$$

Segundo Dobson (2018), ele pode representar a melhora proporcional na função de log-verossimilhança ocasionada pelos termos no modelo de interesse, comparada ao modelo minimal. Assim, assume valores entre 0 e 1, com valores próximos de 0 indicando ajuste ruim e valores próximos de 1 indicando ótimo ajuste.

No entanto, vale destacar que essa estatística não é adequada para a comparação de modelos de famílias diferentes, dado que as funções de log-verossimilhança podem possuir escalas distintas. Assim, ela será utilizada apenas para a avaliação interna de modelos.

2.4 Ferramentas

Para a realização de todas as análises, ajustes de modelos e gerações de visualizações, foi utilizado o software estatístico R. A biblioteca **AER**⁵ foi utilizada para a obtenção dos dados, e a **ggplot2**⁶ e a **GGally**⁷, para a geração dos gráficos. Além disso, foi utilizado um método da biblioteca **MASS**⁸ para o ajuste do modelo Binomial Negativo e um da biblioteca **glmmTMB**⁹ para o ajuste do modelo de inflação de zeros.

3 Resultados

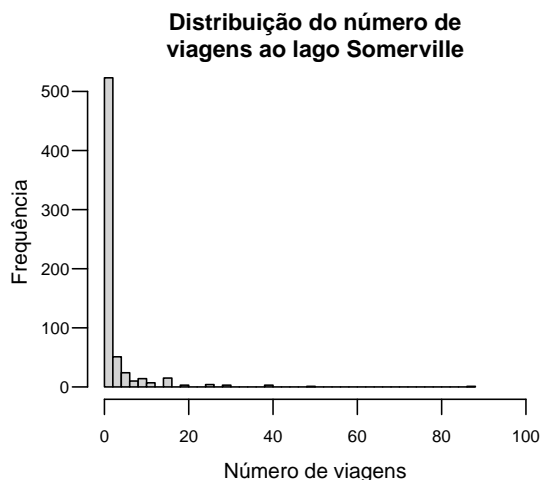
A seguir, serão exibidos e analisados os resultados obtidos a partir da análise exploratória dos dados, seguidos dos ajustes e interpretações dos modelos de interesse, além da aplicação do teste de sobredispersão descrito na seção 2.2.

3.1 Análise exploratória e seleção de covariáveis

Inicialmente, será apresentada uma análise a respeito da variável de interesse a ser modelada, **trips**. Em seguida, as demais covariáveis da base em estudo serão analisadas com o objetivo de selecionar previamente variáveis potencialmente mais adequadas para o ajuste dos modelos.

3.1.1 Variável resposta *trips*

Abaixo está a distribuição dessa variável:



É fácil notar a forte presença de dados com valor 0, indicando a realização de nenhuma viagem recreativa de barco ao lago Somerville em 1980. Isso impacta diretamente nos valores da média e da variância desses dados, como mostrado a seguir:

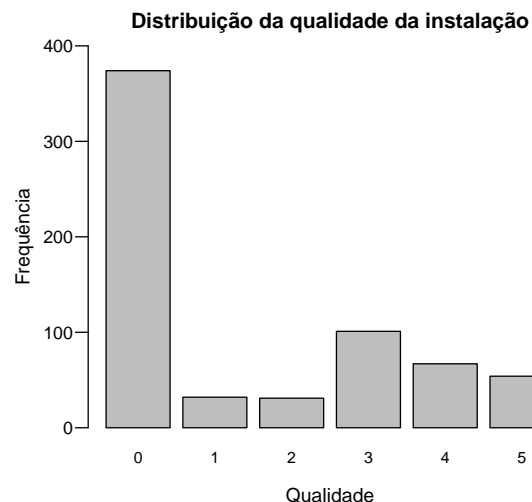
Média: 2.24431

Variância: 39.59524

A variância é consideravelmente maior que a média, o que é causado justamente pelo excesso de zeros nessa variável, que reduz a média e aumenta o efeito na variância dos valores não nulos. Assim, há um forte indício de sobredispersão que possivelmente afetará o desempenho do modelo Poisson, como será discutido nas próximas seções.

3.1.2 Covariável *quality*

Abaixo está a distribuição dessa variável:



Mais uma vez, há um grande número de valores iguais a zero. No entanto, a razão é diferente da fornecida para a variável **trips**: segundo a documentação dos dados, esses valores correspondem a indivíduos que não haviam visitado o lago e que, com isso, não avaliaram a qualidade de suas instalações. Portanto, são equivalentes a valores desconhecidos.

A seguir, é apresentada a distribuição dos valores de **trips** em função da covariável **quality**, sendo plotados todos os pontos acima (com jitter para facilitar a visualização) e apenas as médias dentro de cada valor de **quality** abaixo.

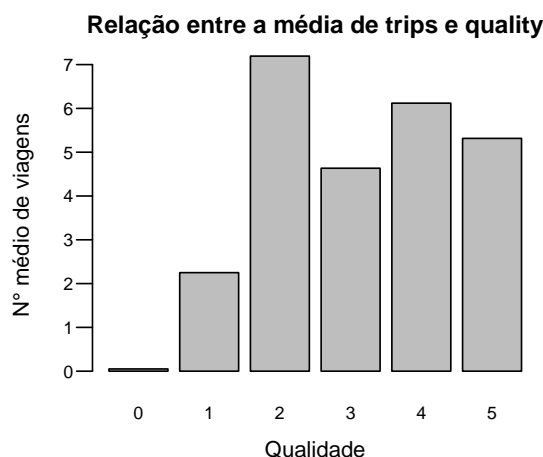
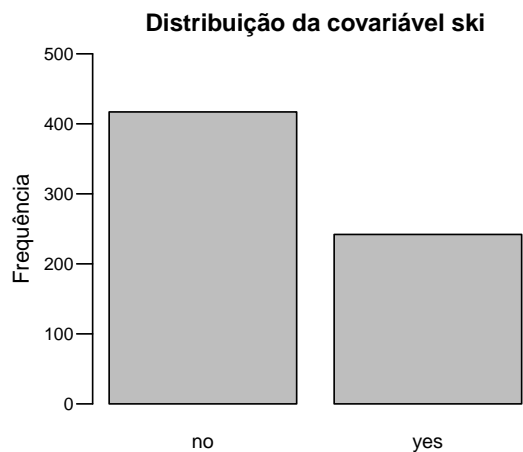
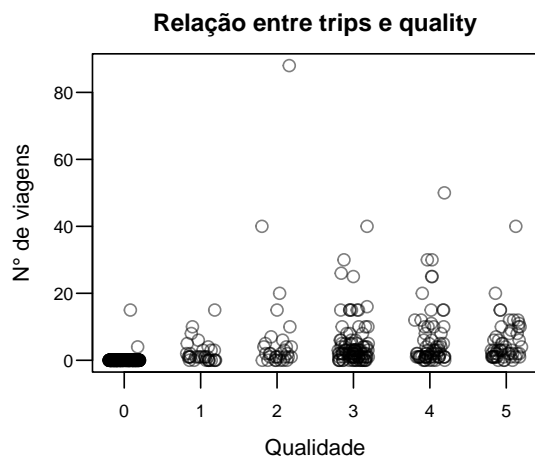
⁵<https://rdrr.io/cran/AER/>

⁶<https://ggplot2.tidyverse.org>

⁷<https://cran.r-project.org/web/packages/GGally/index.html>

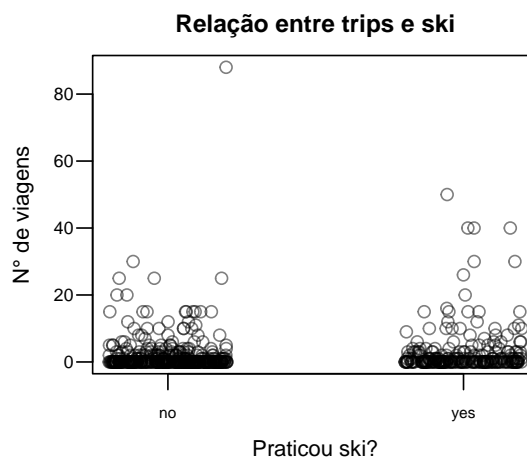
⁸<https://cran.r-project.org/web/packages/MASS/index.html>

⁹<https://cran.r-project.org/web/packages/glmmTMB/index.html>



Existe um relativo equilíbrio entre ambos os valores de **ski**, tornando-a adequada para o uso nos modelos em estudo.

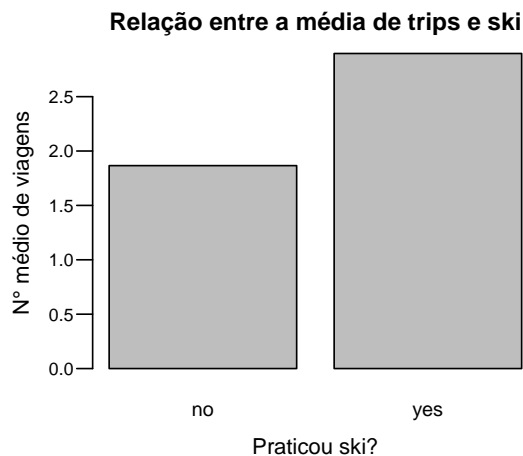
Abaixo, está apresentada a distribuição da variável resposta em função dessa covariável, tanto de forma bruta quanto agregada por meio da média.



Aparentemente, existe uma correlação entre essa covariável e a variável resposta, tornando-a potencialmente relevante para a modelagem.

Além disso, apesar do problema de dados “desconhecidos” citado anteriormente, o uso dessa covariável mostra-se razoável, pois, como visto no gráfico superior, a grande maioria dos registros com qualidade 0 apresenta valor 0 também para **trips**. Isso é semanticamente correto, dado que, como explicado na seção 1.2, **quality** é 0 para registros de indivíduos que não viajaram ao lago Somerville, ou seja, cujo **trips** é 0. Os dois registros que divergem dessa interpretação podem ser dados incorretos, perdidos ou cujos indivíduos apenas não avaliaram a qualidade das instalações do lago.

Portanto, a covariável **quality** será incluída nos modelos que serão ajustados.



3.1.3 Covariável *ski*

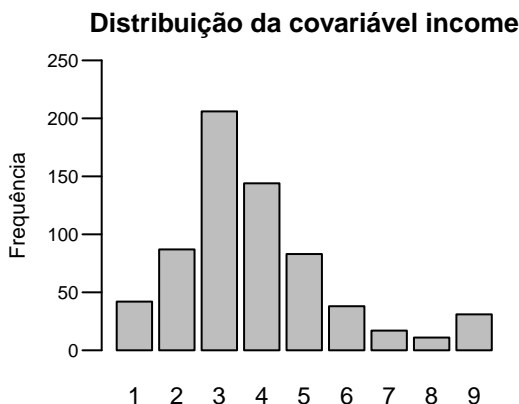
A seguir está exibida a distribuição dessa variável:

Analisando o gráfico inferior, é possível inferir que há uma certa influência da prática de ski aquático na quantidade de viagens de barco recreativas realizadas. Por-

tanto, a covariável **ski** será considerada para os ajustes dos modelos de interesse.

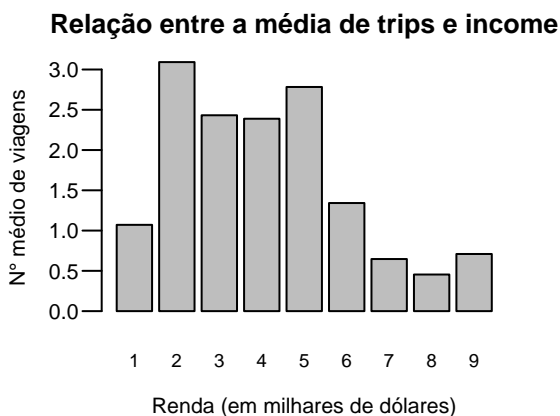
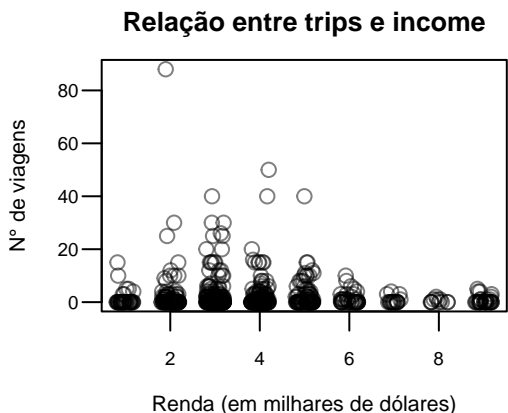
3.1.4 Covariável *income*

Abaixo está apresentada a distribuição dessa variável:



A distribuição dos dados sobre essa variável não apresenta problemas visíveis.

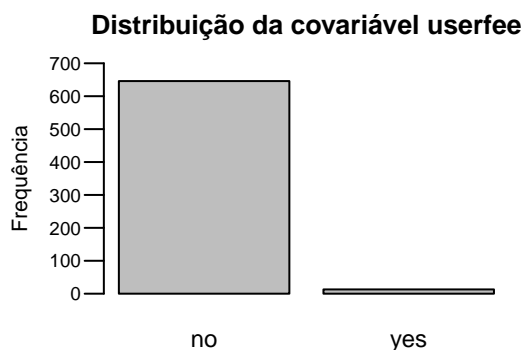
A seguir, está plotada a distribuição de **trips** em relação a essa covariável, da mesma forma que com as anteriores.



Parece existir uma relação entre essa covariável e a variável resposta **trips**, com valores maiores de renda implicando números menores de viagens. Dessa forma, a covariável **income** também será incluída nos ajustes dos modelos considerados.

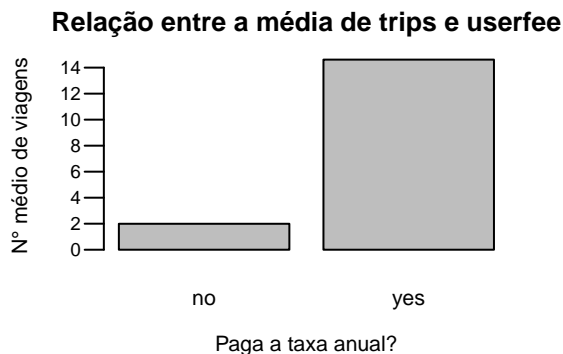
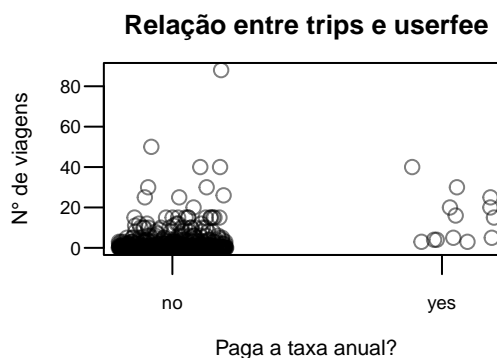
3.1.5 Covariável *userfee*

Visualizando a distribuição dessa covariável, obtém-se:



Nota-se um forte desbalanceamento nos valores dessa variável binária, o que pode afetar o desempenho dos modelos. Isso pode ocorrer devido à dificuldade de estimar precisamente o parâmetro referente à categoria mais rara (“yes”) e à possibilidade de o modelo não detectar o efeito dessa covariável por causa da pequena quantidade de registros em uma das categorias.

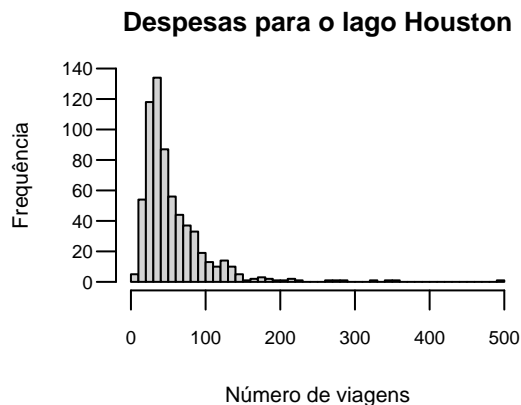
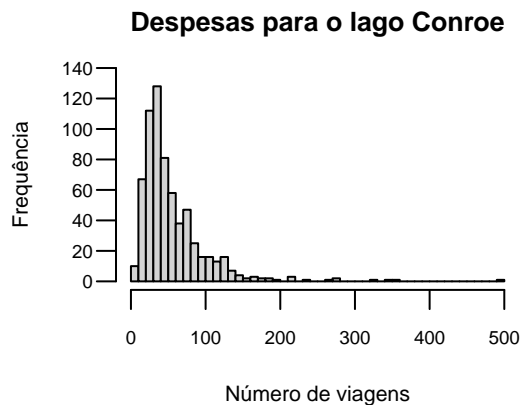
Ainda assim, será verificada a relação entre essa covariável e **trips**. Abaixo está o resultado:



Como visto, a covariável **userfee** parece influenciar de forma considerável o valor da variável resposta. No entanto, devido ao problema de desequilíbrio mostrado anteriormente, os dados da categoria rara podem introduzir ruído excessivo, não sendo adequados para o ajuste de modelos. Portanto, essa covariável não será considerada.

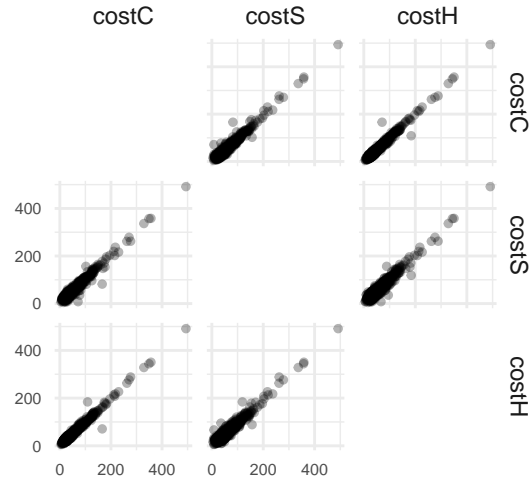
3.1.6 Covariáveis *costC*, *costS* e *costH*

Por último, serão analisadas conjuntamente as 3 variáveis de despesas estimadas. A seguir, estão as distribuições dessas covariáveis.



Suas distribuições são bem similares, o que sugere uma relação de colinearidade entre elas. Para investigar isso, será verificada a correlação entre essas três variáveis aos pares:

Relações entre as variáveis *cost*



É evidente a colinearidade entre essas variáveis. Assim, para manter o princípio da independência das covariáveis e evitar problemas como falta de identificabilidade, será utilizada apenas uma delas na modelagem. Para realizar essa escolha, serão analisadas as correlações entre cada uma dessas variáveis e a variável resposta **trips**. Abaixo estão os resultados:

-> Correlação com **trips**:

costC: -0.04221274

costS: -0.1237036

costH: -0.02051193

Dessa forma, conclui-se que, do ponto de vista de correlação, a variável **costS** é a mais apropriada. Além disso, sob uma perspectiva semântica, essa mesma variável também seria a mais adequada, dado que a variável **trips** corresponde ao número de viagens ao lago Somerville, e **costS** representa o custo estimado de uma viagem a esse mesmo lago, melhorando a interpretação do modelo.

Portanto, as variáveis **costC** e **costH** serão descartadas, mantendo-se apenas **costS**.

3.1.7 Resumo

Finalmente, com base em todas as análises realizadas previamente, as variáveis que serão utilizadas no ajuste dos modelos são:

- **quality**
- **ski**
- **income**
- **costS**

3.2 Ajuste do modelo Poisson

A seguir, será realizado o ajuste do modelo Poisson, como definido na seção 2.1.1, aos dados de interesse, modelando a variável resposta **trips** em função das covariáveis listadas na seção anterior.

Vale ressaltar que a covariável **costS** foi transformada por meio de uma divisão por 100. Isso mostrou-se adequado após alguns testes que mostraram que seu coeficiente possuía magnitude inferior à dos demais, o que é causado pelo fato de sua unidade ser *dólares*, enquanto **income**, por exemplo, é medida em *milhares de dólares*. Assim, com sua nova unidade de medida sendo *centenas de dólares*, seu coeficiente torna-se mais comparável e mais interpretável.

Além disso, tanto a covariável **income** quanto a versão transformada de **costS** foram centradas para facilitar sua interpretação.

Abaixo está o resultado do ajuste segundo `glm()` do R:

```
## Resultados do GLM Poisson (AIC = 3599)

## Deviances: nulo ----- 4849.69

##          residual -- 2835.88

## Coef.      Est.      SE      p-valor
## -----
## (Intercept) -0.974    0.068    <2e-16
## quality      0.559    0.016    <2e-16
## skiyes       0.537    0.056    <2e-16
## income_c    -0.168    0.019    <2e-16
## costS_100_c -1.667    0.103    <2e-16
```

Como pode ser observado por meio dos intervalos de confiança de cada parâmetro e de seus respectivos p-valores, todos eles se mostraram estatisticamente significativos sob esse modelo.

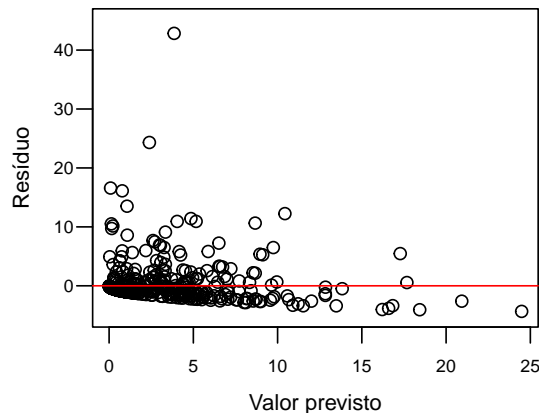
3.2.1 Avaliação do modelo

A seguir, serão aplicadas algumas das estatísticas e testes descritos na seção 2.3 para avaliar esse modelo.

Resíduos de Pearson

Primeiramente, será avaliada a distribuição dos resíduos de Pearson em função dos valores preditos para a variável resposta:

Resíduos de Pearson x Valor previsto (Poisson)



Ao analisar o gráfico, é possível notar que a maioria dos resíduos consideráveis, ou seja, com valores absolutos maiores, é positiva. Isso indica que, para diversos registros, o modelo subestimou o valor de **trips**, isto é, o valor verdadeiro era muito maior que o previsto. Esse resultado demonstra que o modelo de Poisson não está conseguindo modelar adequadamente a alta variância dos dados, não sendo capaz de prever corretamente valores mais extremos.

Estatística qui-quadrado de razão de verossimilhança

A seguir, será calculada a estatística qui-quadrado de razão de verossimilhança (C) para esse modelo.

```
## Estatística C: 2013.809
```

```
## Graus de liberdade: 4
```

```
## p-valor: 0
```

O valor de C é altamente significativo quando comparado à distribuição qui-quadrado de 4 graus de liberdade, tanto que seu p-valor é praticamente nulo. Assim, isso sugere que os parâmetros do modelo não são nulos e que, portanto, suas covariáveis associadas são relevantes para a modelagem de **trips**.

Pseudo- R^2

Agora, será calculado o valor do pseudo- R^2 desse modelo:

```
## Pseudo-R2: 0.3594313
```

Com base nessa estatística, é possível inferir que o modelo de Poisson ajustado obteve uma melhoria de aproximadamente 36% em seu desempenho quando comparado com o modelo minimal da mesma família, o que indica um resultado razoável, mas que pode ser melhorado com o uso de outros modelos, como será apresentado a seguir.

3.3 Teste para sobredispersão

Nesse momento, será realizado o teste de sobredispersão definido na seção 2.2, aplicado sobre os dados de interesse. Após o ajuste da estatística de teste às previsões, o resultado obtido foi o seguinte:

```
## Resultados da Sobredispersão

## Coef.      Est.      SE      p-valor
## -----
## pred      21.028    7.961    0.008
```

Como é possível observar, o parâmetro do modelo possui considerável significância estatística, o que sugere a existência de sobredispersão nesses dados. Isso justifica o desempenho mediano do modelo Poisson sobre eles. Na próxima seção, será ajustado um modelo Binomial Negativo com o objetivo de tentar contornar esse problema.

3.4 Ajuste do modelo Binomial Negativo

A seguir, será ajustado um modelo Binomial Negativo aos dados em estudo conforme definido na seção 2.1.2 e com as mesmas covariáveis e respectivas transformações utilizadas no modelo Poisson. Será utilizado o método `glm.nb` da biblioteca **MASS** do R.

```
## Resultados do GLM NegBin (AIC = 1817.5)

## Phi = 0.4460 (SE = 0.0418)

## Deviances: nulo ----- 923.49

##          residual -- 455.70

## Coef.      Est.      SE      p-valor
## -----
## (Intercept) -1.829    0.147    <2e-16
## quality      0.913    0.043    <2e-16
## skiyes       0.558    0.170    0.001
## income_c     -0.066    0.048    0.173
## costS_100_c  -1.295    0.247    0.000
```

Ao contrário do modelo de Poisson, nesse modelo, nem todos os coeficientes se mostraram significativos: o coeficiente da covariável **income** perdeu sua significância. Uma possível razão para isso é a inclusão do parâmetro de sobredispersão ϕ , que absorveu a variância que o modelo Poisson estava tentando ajustar por meio da covariável **income**. Dessa forma, será ajustado um novo modelo sem essa covariável de forma a respeitar o princípio da parcimônia.

```
## Resultados do GLM NegBin (AIC = 1817.1)

## Phi = 0.4399 (SE = 0.0410)

## Deviances: nulo ----- 915.43

##          residual -- 453.57

## Coef.      Est.      SE      p-valor
## -----
## (Intercept) -1.830    0.147    <2e-16
## quality      0.920    0.043    <2e-16
## skiyes       0.526    0.165    0.001
## costS_100_c  -1.405    0.249    0.000
```

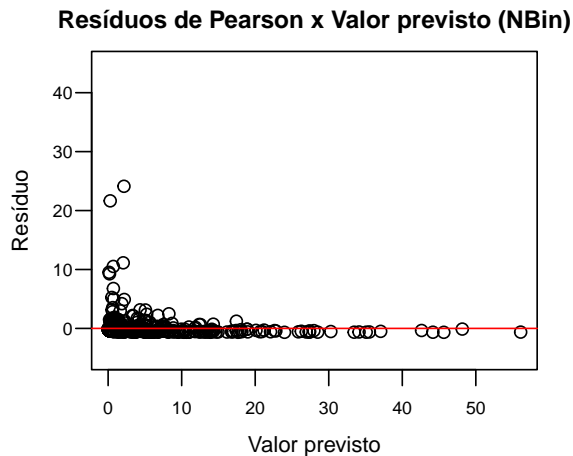
Os coeficientes estimados para as três covariáveis restantes apresentam significância estatística.

A seguir, serão realizadas as avaliações desse ajuste semelhantemente a como foi feito com o modelo Poisson. Comparações entre os resultados dos modelos serão efetuadas após o ajuste dos 3 modelos em estudo.

3.4.1 Avaliação do modelo

Resíduos de Pearson

A primeira análise realizada será sobre a distribuição dos resíduos de Pearson:



É notável a concentração dos resíduos em valores próximos de zero, o que sugere um bom ajuste. Além disso, nota-se que os poucos resíduos com maior magnitude são positivos, indicando que o modelo ainda subestima alguns pontos de dados, ou seja, o valor verdadeiro da variável resposta é bem maior que o previsto. Isso mostra que, apesar de modelar a variância dos dados mais eficientemente que o modelo Poisson, o modelo Binomial Negativo ainda comete alguns erros semelhantes aos cometidos por esse último.

Estatística qui-quadrado de razão de verossimilhança

Abaixo, estão os resultados do teste sobre a estatística C .

```
## Estatística C: 461.859
```

```
## Graus de liberdade: 3
```

```
## p-valor: 8.784565e-100
```

Novamente, o valor da estatística C mostra-se altamente significativo quando comparado com a distribuição $\chi^2(3)$. Portanto, há fortes indícios de que os parâmetros referentes às covariáveis utilizadas são estatisticamente significativos, o que representa uma melhoria no desempenho gerada pelo uso dessas covariáveis.

Pseudo- R^2

Por fim, abaixo está o valor calculado para o pseudo- R^2 desse modelo:

```
## Pseudo-R2: 0.1513699
```

Dessa forma, segundo essa estatística, o modelo Binomial Negativo ajustado gerou uma melhoria de aproximadamente 15% no desempenho quando comparado com o modelo minimal da mesma família. Isso parece fraco, mas pode ser causado pelo fato de a verossimilhança do modelo nulo já ser considerável, o que diminui a relevância de ajustes melhores.

Além disso, esse valor é menor que o obtido pelo modelo Poisson, mas, como explicado na seção 2.3.5, esses valores não são comparáveis por virem de modelos de famílias diferentes.

3.5 Ajuste do modelo Poisson inflado de zeros

Para o ajuste do modelo Poisson inflado de zeros, foram escolhidas, com base em testes preliminares, a covariável **quality** para a modelagem da etapa logística e as demais para a modelagem da etapa de contagem. Assim, feito o ajuste, o resultado obtido foi o seguinte:

```
## Resultados do Modelo ZIP (AIC = 2631.8)
```

```
## Modelo Condicional:
```

## Coef.	Est.	SE	p-valor
## (Intercept)	1.385	0.046	<2e-16
## skiyes	0.528	0.058	<2e-16
## income_c	-0.127	0.020	0.000
## costS_100_c	-1.504	0.103	<2e-16

```
## Modelo de Inflação Zero:
```

## Coef.	Est.	SE	p-valor
## (Intercept)	2.973	0.228	<2e-16
## quality	-1.618	0.121	<2e-16

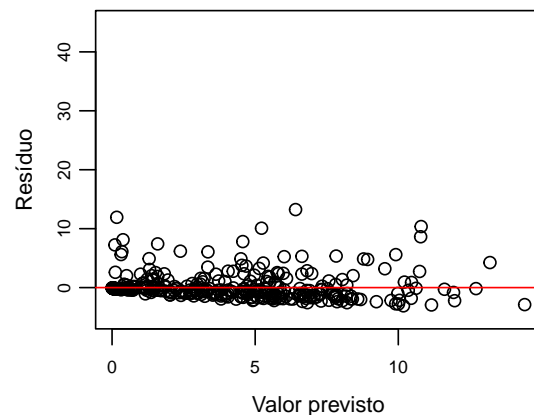
Dada a seleção de covariáveis realizada, todos os coeficientes mostraram-se estatisticamente significativos.

A seguir, esse modelo será avaliado segundo as mesmas métricas aplicadas sobre os modelos anteriores.

3.5.1 Avaliação do ajuste

Resíduos de Pearson

Resíduos de Pearson x Valor previsto (ZIP)



Nota-se uma concentração dos resíduos mais proeminentes em valores positivos, indicando que, assim como o modelo Poisson, o modelo ZIP não foi tão eficiente ao absorver a variância gerada pela sobredispersão. Isso é perfeitamente válido, já que, apesar de ser um modelo mais elaborado, a base de sua modelagem de contagem ainda é uma distribuição Poisson convencional, que considera a média e a variância como iguais. Na próxima seção, será verificado se ele conseguiu mitigar o problema do excesso de zeros.

Estatística qui-quadrado de razão de verossimilhança

Aqui, será considerado como modelo nulo a versão do modelo ZIP sem o uso de covariáveis, ou seja, ainda será mantida sua estrutura mista, mas tanto a etapa Binomial quanto a etapa Poisson serão regredidas apenas sobre interceptos.

```
## Estatística C: 2500.429
```

```
## Graus de liberdade: 4
```

```
## p-valor: 0
```

Novamente, conclui-se que os coeficientes referentes às covariáveis utilizadas provavelmente são estatisticamente significativos.

Pseudo- R^2

Por fim, a seguir está o valor calculado para o pseudo- R^2 desse modelo:

```
## Pseudo-R2: 0.252546
```

Portanto, esse modelo apresenta uma melhoria de aproximadamente 25% na verossimilhança em comparação com o modelo minimal, ajustado sem o uso de covariáveis explicativas. Assim, ele apresenta uma performance considerável, embora ainda permita a exploração de modelos mais elaborados ou a inserção de covariáveis adicionais.

3.6 Comparação dos modelos

Finalmente, nesta seção, será efetuada a comparação dos desempenhos dos três modelos ajustados, analisando-se, para esse fim, seus coeficientes estimados, as distribuições de seus resíduos de Pearson, suas estatísticas AIC (definido e justificado na seção 2.3.4) e seus resultados na modelagem dos valores zero.

3.6.1 Coeficientes estimados

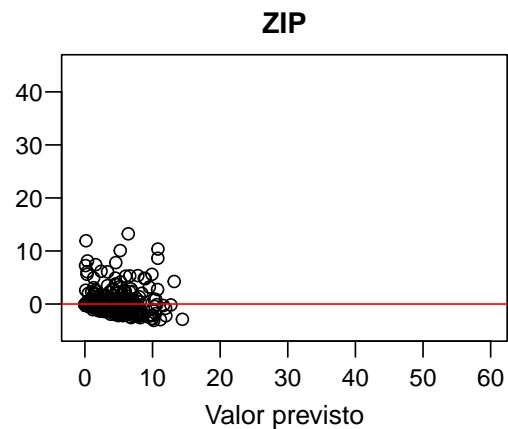
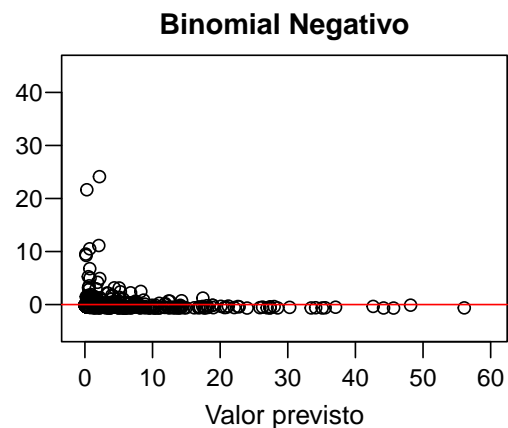
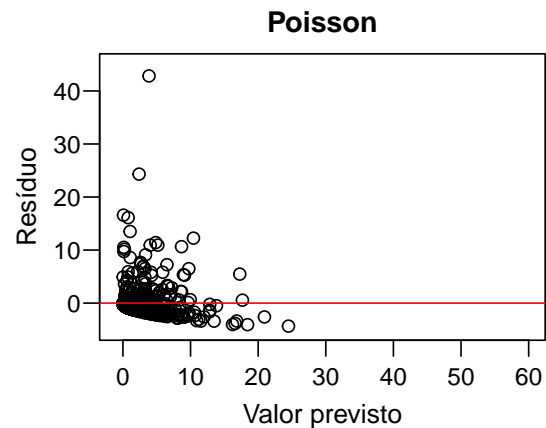
A seguir, apresenta-se uma tabela com os valores estimados por cada modelo para os coeficientes de cada covariável utilizada. Vale lembrar que a interpretação desses parâmetros varia entre os modelos, principalmente no caso do modelo misto ZIP.

Termo	Poisson	NegBin	ZIP.Pois	ZIP.Bin
(Intercept)	-0.97	-1.83	1.38	2.97
quality	0.56	0.92		-1.62
skiyes	0.54	0.53	0.53	
income_c	-0.17		-0.13	
costS_100_c	-1.67	-1.40	-1.50	

Em resumo, todos os modelos apontam direções semelhantes nos efeitos das covariáveis utilizadas sobre a variável resposta **trips**, o que pode ser observado pelos sinais dos coeficientes estimados em cada modelo. Os únicos casos que não obedecem a esse padrão são os interceptos e o coeficiente de **quality**. Com relação a esse último, enquanto os modelos de Poisson e Binomial Negativo apresentam valores positivos para esse parâmetro, a etapa Binomial do modelo ZIP atribui a ele um valor negativo. Embora pareça contraintuitivo, esse resultado é coerente, já que, nesse modelo ZIP, quanto maior o valor de **quality**, menor a probabilidade de a variável resposta ser zerada e, com isso, maior seu valor potencial, indicando uma relação positiva entre essas variáveis, consistente com os demais modelos. Portanto, eles estão de acordo com respeito às direções dessas correlações.

3.6.2 Resíduos de Pearson

A seguir, estão dispostos em conjunto os gráficos das distribuições dos resíduos de Pearson de cada modelo para que seja realizada uma comparação entre seus desempenhos.



Como pode ser observado, tanto o modelo Poisson quanto o ZIP (que também utiliza a distribuição Poisson) estimam valores muito menores que o modelo Binomial Negativo, o que é explicado pela falta de modelagem do fator de sobredispersão por parte desses modelos. Essa subestimação também é confirmada pela grande quantidade de resíduos altamente positivos em ambos os ajustes, que sugere sua falta de capacidade para modelar valores extremos.

Por outro lado, o modelo Binomial Negativo apresenta uma distribuição mais simétrica e próxima do zero, indicando um ajuste mais apropriado aos dados. Isso pode ser justificado pela inclusão de um parâmetro de sobre-dispersão, que permite a ele absorver de forma mais eficiente a variância gerada pelos registros extremos.

3.6.3 AIC

Em seguida, serão comparadas as estatísticas AIC dos três modelos em estudo.

Modelo	AIC
Poisson	3598.95
NegBin	1817.11
ZIP	2631.78

Notavelmente, o modelo Binomial Negativo obteve um desempenho superior ao dos modelos baseados na distribuição Poisson, devido à incorporação do coeficiente de sobre-dispersão, que permite a ele separar a média da variância e, assim, modelar de maneira mais adequada os valores extremos.

Enquanto isso, o modelo ZIP apresenta um risco estimado menor que o Poisson tradicional, o que se justifica pela sua capacidade de modelar os zeros estruturais da variável resposta (o que será demonstrado a seguir). Entretanto, ainda não apresenta um ajuste tão apropriado quanto o do modelo Binomial Negativo devido à falta da modelagem da sobre-dispersão, presente nesse último.

3.6.4 Desempenho na modelagem dos valores zero

Por fim, será avaliado como cada um desses modelos desempenha com relação à modelagem e previsão dos valores zero da variável resposta **trips**. Para isso, em cada modelo:

1. Será feito seu ajuste.
2. Serão geradas amostras de previsões para os dados a partir da distribuição assumida por cada modelo com os parâmetros estimados.
3. Serão contabilizados os valores zero previstos.

Esse processo será realizado 1000 vezes e a média dos resultados será considerada para garantir uma estimativa estável dessa quantidade. Os resultados estão exibidos a seguir:

Modelo	MediaZeros
REAL	417.00
Poisson	261.22
NegBin	423.51
ZIP	437.54

Analisando a tabela, é possível concluir que, de fato, o modelo Poisson não estava ajustando de forma adequada o excesso de zeros dos dados da variável **trips**,

prevendo muito menos valores zero que a quantidade verdadeira. Por outro lado, tanto o modelo Binomial Negativo quanto o modelo ZIP apresentaram desempenho aceitável nesse quesito. Nota-se, então, que a maior elaboração do modelo Poisson ao construir um modelo misto, o ZIP, foi capaz de corrigir a deficiência do modelo clássico com relação à modelagem desses valores. Enquanto isso, a própria inclusão de um coeficiente de sobre-dispersão já foi suficiente para o modelo Binomial Negativo obter uma boa capacidade de representar os zeros observados.

4 Discussão e Conclusão

Portanto, com base nos experimentos e análises conduzidos, pode-se concluir que a presença de sobre-dispersão na variável resposta a ser modelada pode comprometer significativamente o desempenho de modelos estatísticos quando não levada em consideração. Isso é exemplificado pela aplicação de um modelo Poisson aos dados em estudo, o que gerou resultados insatisfatórios na modelagem tanto de valores extremos quanto de valores zero excessivos.

Ao refinar essa estratégia por meio da adoção de um modelo Poisson inflado de zeros (ZIP), o desempenho relacionado ao excesso de zeros obteve uma melhora significativa, com o modelo sendo capaz de absorver e modelar os zeros estruturais presentes na base. Entretanto, ele ainda apresentou desempenho insatisfatório na modelagem de valores extremos, resultantes da alta variância dos dados.

Por fim, a inclusão de um coeficiente de sobre-dispersão por meio da utilização de um modelo Binomial Negativo permitiu que as consequências geradas pela alta variância fossem mitigadas, como mostrado pelos resíduos relativamente pequenos quando comparados aos resíduos dos modelos Poisson e ZIP. Ao mesmo tempo, essa nova característica também forneceu ao modelo a capacidade de modelar os zeros estruturais, proporcionando um desempenho geral superior aos demais modelos, sendo, portanto, preferível dentre os modelos analisados.

No entanto, apesar dos desempenhos variados, todos os modelos estudados apresentaram resultados medianos, não sendo capazes de alcançar um ajuste plenamente satisfatório. Uma possível causa para isso é a falta de dados mais relevantes, dado que o estudo foi limitado aos dados presentes na base da biblioteca AER. Com isso, é provável que a incorporação de novas covariáveis resultasse em modelos melhor ajustados.

Além disso, com o objetivo de obter modelos mais adequados a esses dados, esse trabalho poderia ser expandido por meio da aplicação e da experimentação de novos modelos estatísticos, sejam variantes dos apresenta-

dos aqui ou baseados em abordagens distintas. Uma possibilidade seria o modelo Binomial Negativo inflado de zeros (ZINB da sigla em inglês), que agrega a capacidade de absorção da sobredispersão do modelo Binomial Negativo com a boa modelagem dos zeros estruturais fornecida pelo modelo ZIP. Esse e outros modelos poderiam gerar resultados mais robustos que os obtidos pelos estudados nesse projeto.

5 Referências

[Cameron & Trivedi (1990)] Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46(3), 347–364. [https://doi.org/10.1016/0304-4076\(90\)90014-K](https://doi.org/10.1016/0304-4076(90)90014-K)

[Dobson (2018)] Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models* (4th ed.). CRC Press. <https://doi.org/10.1201/9781315182780>

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. <https://doi.org/10.1017/9781139161879>

Stan Development Team. (n.d.). *Finite mixtures*. Stan User’s Guide. <https://mc-stan.org/docs/stan-users-guide/finite-mixtures.html>