Sobredispersão em Modelos de Contagem

Modelagem Estatística

Pedro Henrique Coterli

May 20, 2025

1. Introdução

O presente trabalho tem como objetivo investigar a sobredispersão em dados de contagem e seus efeitos no ajuste de 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.

2. Métodos

2.1. Os dados

Os dados utilizados para este estudo foram retirados da biblioteca Applied Econometrics with R (AER)[1]. Será utilizada a base RecreationalDemand[2], 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;
- **quality**: assume valores de 1 a 5 e classifica subjetivamente a qualidade da instalação;
- **ski**: assume valores "yes" ou "no" e indica se o indivíduo estava praticando esqui aquático no lago;
- **income**: assume valores naturais e representa a renda familiar anual do entrevistado (milhares de dólares);
- **userfee**: assume valores "yes" ou "no" e indica se o indivíduo pagou uma taxa anual de uso no 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.

2.2. 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.

[1] https://rdrr.io/cran/AER/
[2] https://rdrr.io/cran/AER/man/RecreationDemand.html

Serão considerados nesse estudo 3 modelos estatísticos da família dos modelos lineares generalizados (GLMs), descritos a seguir.

2.2.1. Modelo Poisson

Frequentemente utilizada para modelar dados de contagem, a regressão de 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 (no nosso 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 simples que a necessária na abordagem Bayesiana. Enquanto a primeira utiliza de métodos simples como Fisher Scoring e Iteractive 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:

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

  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.

2.2.2. 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 seguinte[3]:

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

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:

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

Com isso, quanto menor o valor de ϕ, maior é a variância dos dados e mais distante da média ela está. Analogamente, quando $\phi \to \infty$, 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.
2. Com β fixo, ϕ é encontrado numericamente maximizando a verossimilhança atual.

Esse processo é repetido até os parâmetros convergirem.

2.2.3. 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 forma[4]:

[3] https://mc-stan.org/docs/functions-reference/unbounded_discrete_distributions.html#nbalt
[4] https://mc-stan.org/docs/stan-users-guide/finite-mixtures.html#zero-inflation

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

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:

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

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.

2.3. 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.

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

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

Assim, a hipótese nula diz que $\alpha = 0$ e, 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:

$$\mathbb{E}[Z_i] = \frac{\mathbb{E}[(Y_i - \hat{\mu_i})^2] - \mathbb{E}[Y_i]}{\hat{\mu_i}} = \frac{Var(Y_i) - \hat{\mu_i}}{\hat{\mu_i}} = \frac{\hat{\mu_i} - \hat{\mu_i}}{\hat{\mu_i}} = 0$$

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.

2.4. Critérios de avaliação

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

2.4.1. Intervalo de confiança aproximado dos parâmetros

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

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

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.

2.4.2. 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 (2001). 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.

2.4.3. 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:

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

onde $l(\dot{\beta}; y)$ 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 (2001), 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.

2.4.4. 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:

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

com os mesmos significados já definidos anteriormente.

2.4.5. 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 (2001), 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.

2.5. 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.

[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

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 2.3.

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:

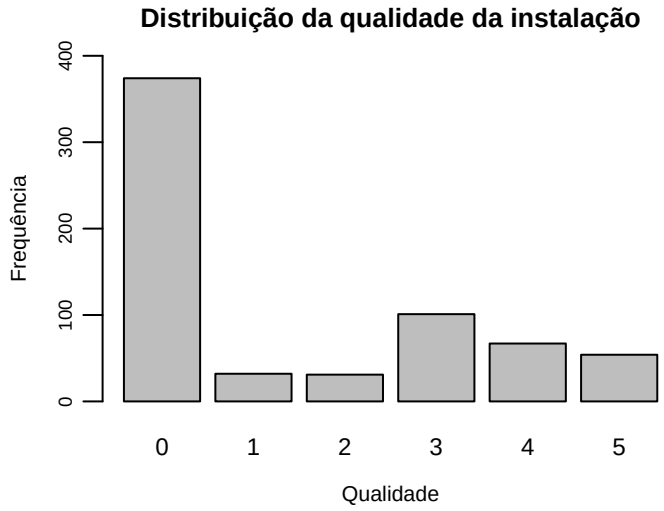
```
## Média:     2.24431
```

```
## 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.

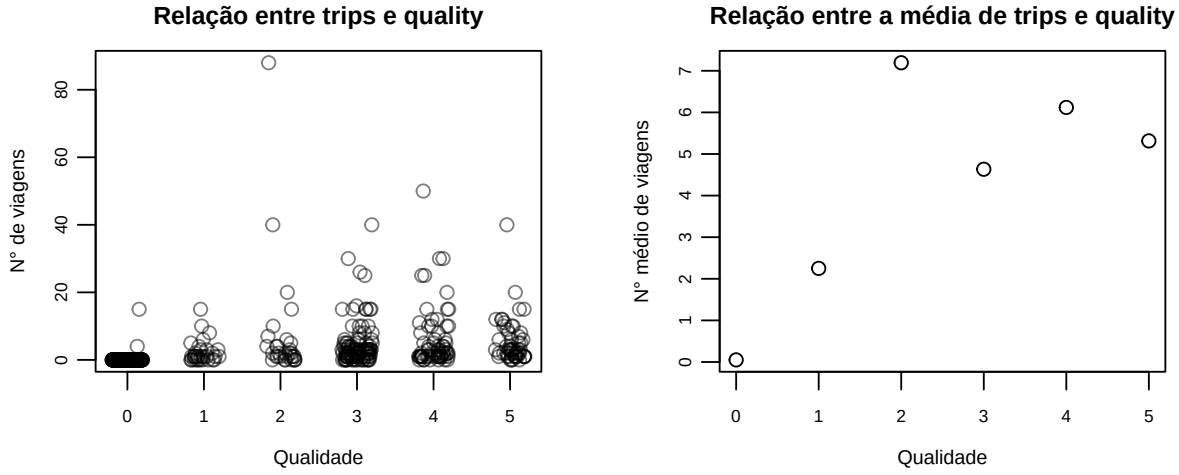
3.1.2. Covariável *quality*

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 à esquerda (com jitter para facilitar a visualização) e apenas as médias dentro de cada valor de **quality** à direita.



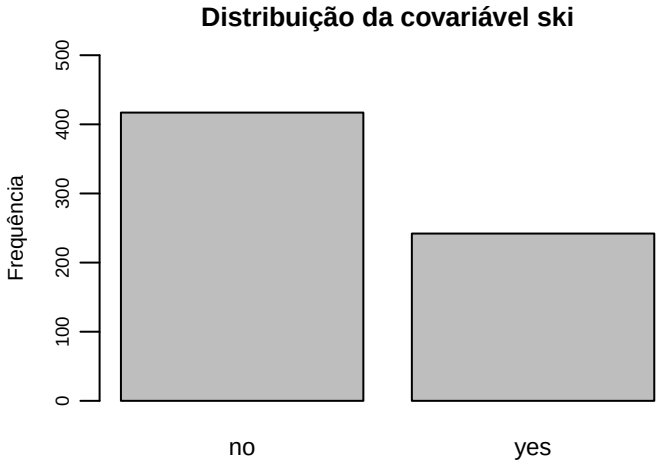
Aparentemente, existe uma correlação entre essa covariável e a variável resposta, tornando-a potencialmente relevante para a modelagem.

Além disso, apesar do problema de dados "desconhecidos" citado anteriormente, o uso dessa covariável mostra-se razoável, pois, como visto no gráfico da esquerda, a grande maioria dos registros com qualidade 0 apresenta valor 0 também para **trips**. Isso é semanticamente correto, dado que, como explicado na seção 2.1, **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.

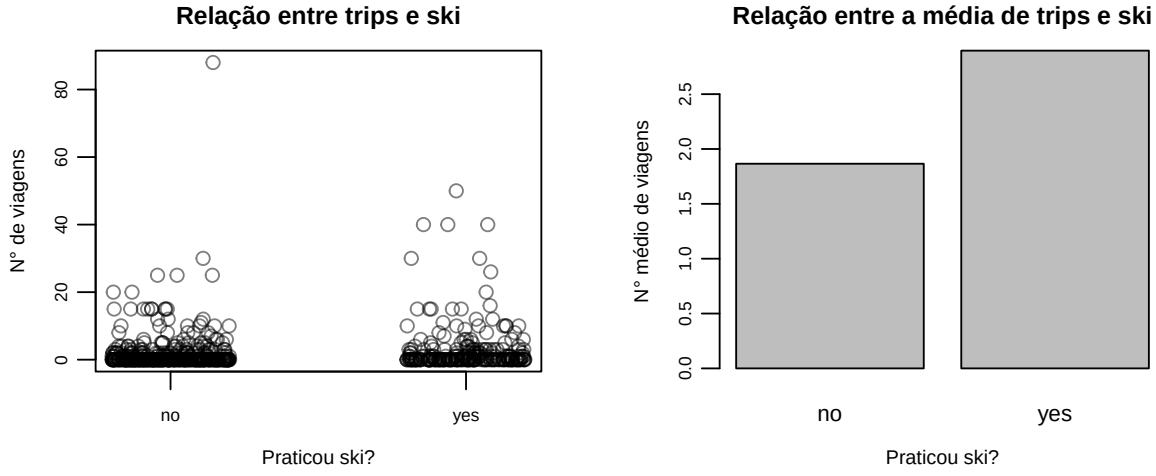
3.1.3. Covariável *ski*

Abaixo está exibida a distribuição dessa variável:



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

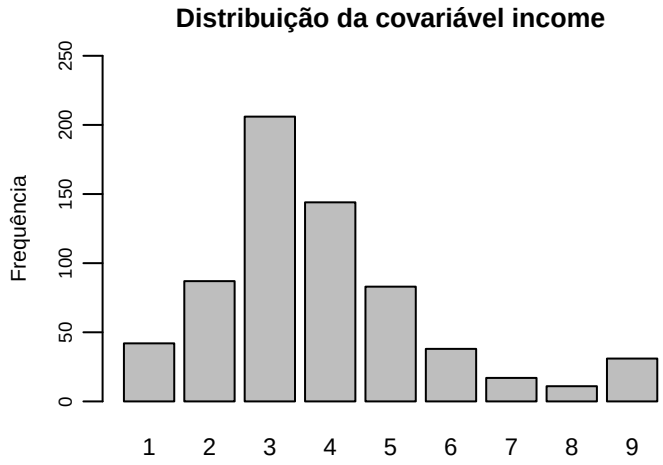
A seguir, está apresentada a distribuição da variável resposta em função dessa covariável, tanto de forma bruta quanto agregada por meio da média.



Analisando o gráfico à direita, é possível inferir que há uma certa influência da prática de ski aquático na quantidade de viagens de barco recreativas realizadas. Portanto, a covariável **ski** será considerada para os ajustes dos modelos de interesse.

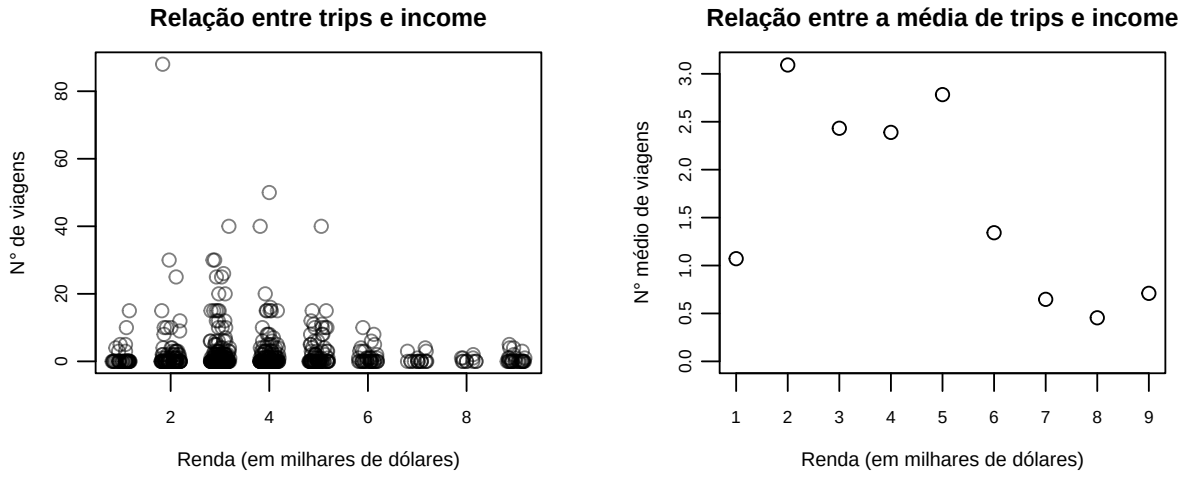
3.1.4. Covariável *income*

Abaixo está apresentada a distribuição dessa variável:



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

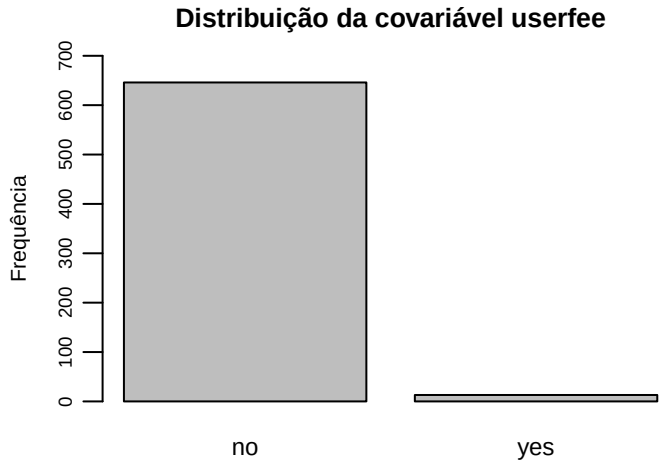
A seguir, está plotada a distribuição de **trips** em relação a essa covariável, da mesma forma que com as anteriores.



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.

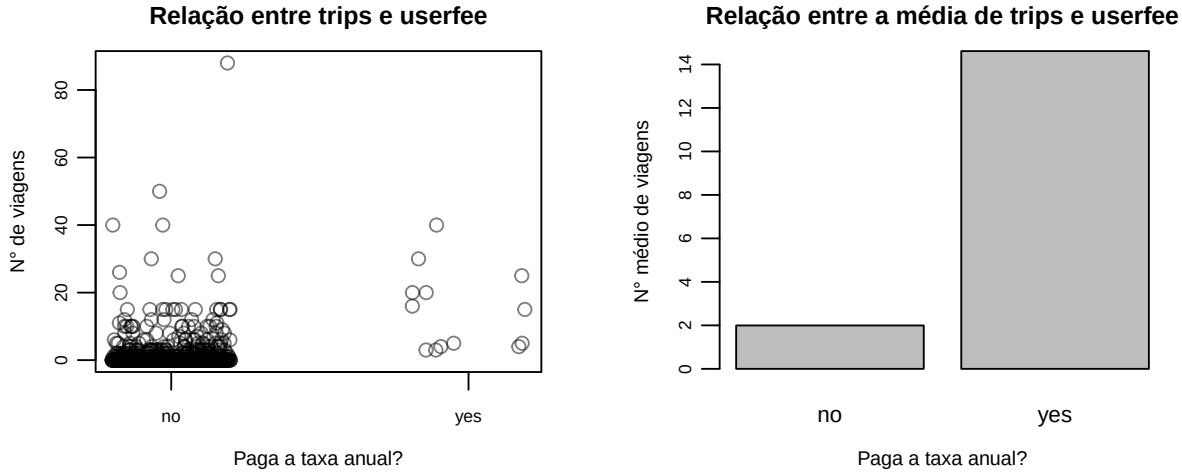
3.1.5. Covariável *userfee*

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.

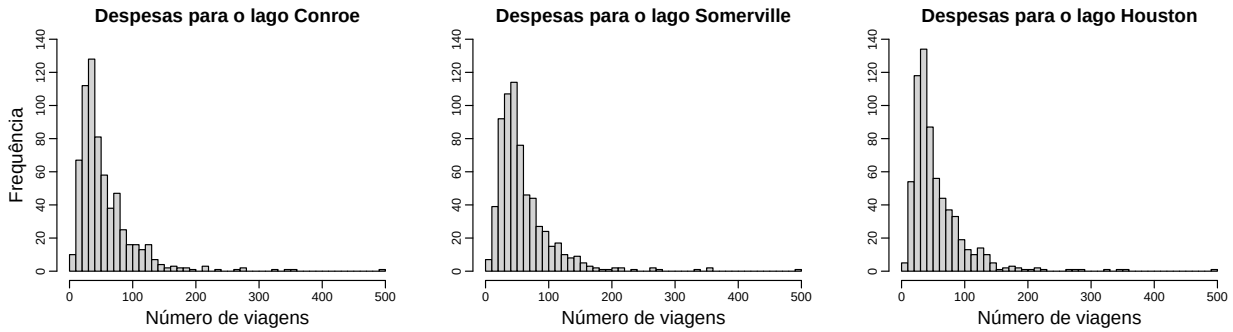
Ainda assim, será verificada a relação entre essa covariável e **trips**. Abaixo está o resultado:



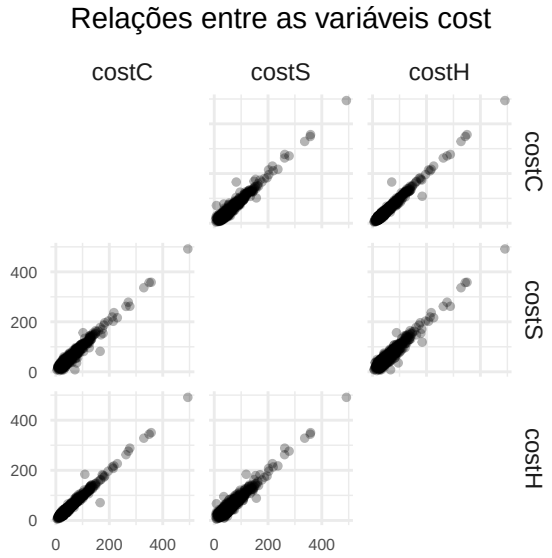
Como visto, a covariável **userfee** parece influenciar de forma considerável o valor da variável resposta. No entanto, devido ao problema de desequilíbrio mostrado anteriormente, os dados da categoria rara podem introduzir ruído excessivo, não sendo adequados para o ajuste de modelos. Portanto, essa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



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 2.2.1, 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 possuia 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:

```
##
## Call:
## glm(formula = trips ~ quality + ski + income_c + costS_100_c,
##     family = poisson(), data = RecreationDemand)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.97360    0.06797 -14.323   <2e-16 ***
## quality      0.55895    0.01574  35.511   <2e-16 ***
## skiyes       0.53674    0.05563   9.649   <2e-16 ***
## income_c    -0.16754    0.01922  -8.719   <2e-16 ***
## costS_100_c -1.66734    0.10259 -16.253   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4849.7  on 658  degrees of freedom
## Residual deviance: 2835.9  on 654  degrees of freedom
## AIC: 3599
##
## Number of Fisher Scoring iterations: 7
```

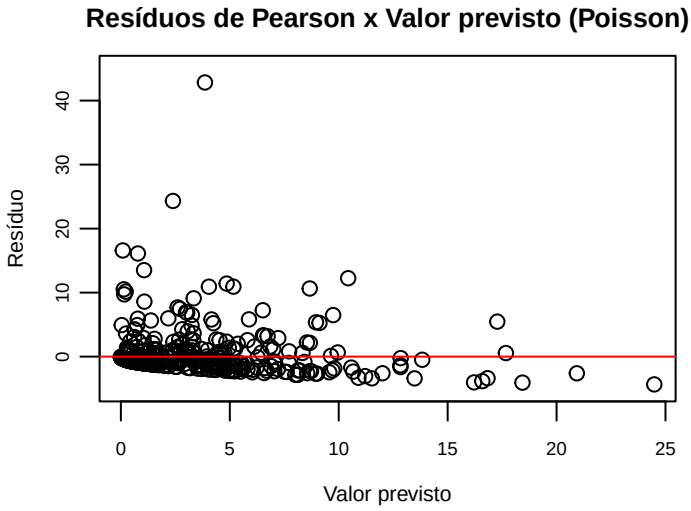
Como pode ser observado por meio dos intervalos de confiança de cada parâmetro e de seus respectivos p-valores, todos eles se mostraram estatisticamente significativos sob esse modelo.

3.2.1. Avaliação do modelo

A seguir, serão aplicadas algumas das estatísticas e testes descritos na seção 2.4 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:



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.

Desempenho da modelagem dos valores zero

Por último, será analisado o desempenho desse modelo na estimação dos valores zero do dado. Para isso, será gerada, para cada dado, uma previsão para **trips** por meio de uma amostragem da distribuição Poisson com a média sendo a média estimada pelo modelo para esse ponto de dado. Em seguida, serão contabilizadas quantas dessas previsões receberam o valor 0. Esse processo será realizado 1000 vezes e a média dos resultados será considerada, a fim de obter uma estimativa estável.

Abaixo estão os resultados obtidos:

```
## -> Número de 0s:
```

```
## No dado:   417
```

```
## Previstos: 260.458
```

Como observado, o modelo Poisson subestima a ocorrência de zeros em comparação à quantidade observada nos dados, o que é confirmado pelo gráfico dos resíduos mostrado acima. Nele, há uma grande quantidade de resíduos pequenos negativos, sugerindo que o modelo tenha atribuído para esses pontos valores pequenos da variável resposta, mas ainda maiores que 0. Isso levará, mais adiante, ao ajuste de outros modelos de modo a tentar mitigar esse efeito.

3.3. Teste para sobredispersão

Nesse momento, será realizado o teste de sobredispersão definido na seção 2.3, aplicado sobre os dados de interesse. Após o ajuste da estatística de teste às previsões, o resultado obtido foi o seguinte:

```
##
## Call:
## glm(formula = z ~ pred + 0, data = overdispersion_data)
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## pred   21.028      7.961   2.641  0.00846 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 74336.33)
##
##     Null deviance: 49431888  on 659  degrees of freedom
## Residual deviance: 48913308  on 658  degrees of freedom
## AIC: 9264.7
##
## Number of Fisher Scoring iterations: 2
```

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 2.2.2 e com as mesmas covariáveis e respectivas transformações que o modelo de Poisson. Será utilizado o método `glm.nb` da biblioteca **MASS** do R.

```
##
## Call:
## glm.nb(formula = trips ~ quality + ski + income_c + costS_100_c,
##     data = RecreationDemand, init.theta = 0.4459740411, link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.82879    0.14713 -12.430  < 2e-16 ***
## quality      0.91343    0.04257  21.456  < 2e-16 ***
## skiyes       0.55795    0.17019   3.278  0.00104 **
## income_c    -0.06561    0.04817  -1.362  0.17314
## costS_100_c -1.29493    0.24695  -5.244 1.57e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.446) family taken to be 1)
##
##     Null deviance: 923.49  on 658  degrees of freedom
## Residual deviance: 455.70  on 654  degrees of freedom
## AIC: 1817.5
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.4460
##           Std. Err.:  0.0418
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -1805.4710
```

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 (ou θ pela parametrização do método `glm.nb`), 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.

```
##
## Call:
## glm.nb(formula = trips ~ quality + ski + costS_100_c, data = RecreationDemand,
##     init.theta = 0.4398523177, link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.82952    0.14719 -12.429  < 2e-16 ***
## quality      0.92045    0.04283  21.493  < 2e-16 ***
## skiyes       0.52564    0.16509   3.184  0.00145 **
## costS_100_c -1.40482    0.24875  -5.647 1.63e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.4399) family taken to be 1)
##
##     Null deviance: 915.43  on 658  degrees of freedom
## Residual deviance: 453.57  on 655  degrees of freedom
## AIC: 1817.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.4399
##           Std. Err.:  0.0410
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -1807.1110
```

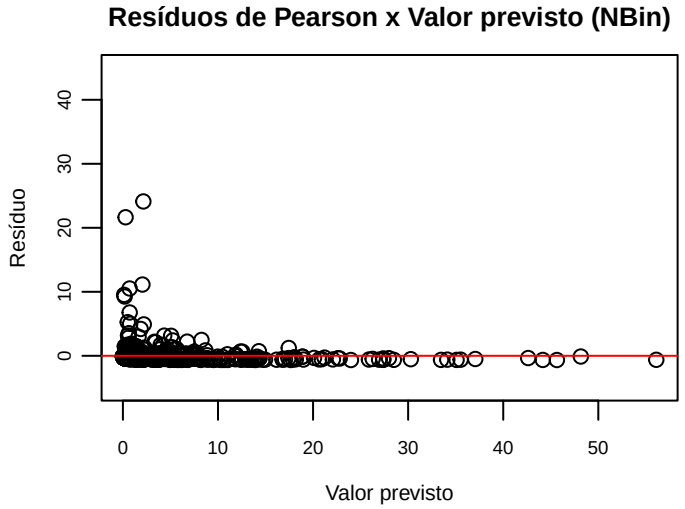
Os coeficientes estimados para as três covariáveis restantes apresentam significância estatística.

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 2.4.5, 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 alguns testes, 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:

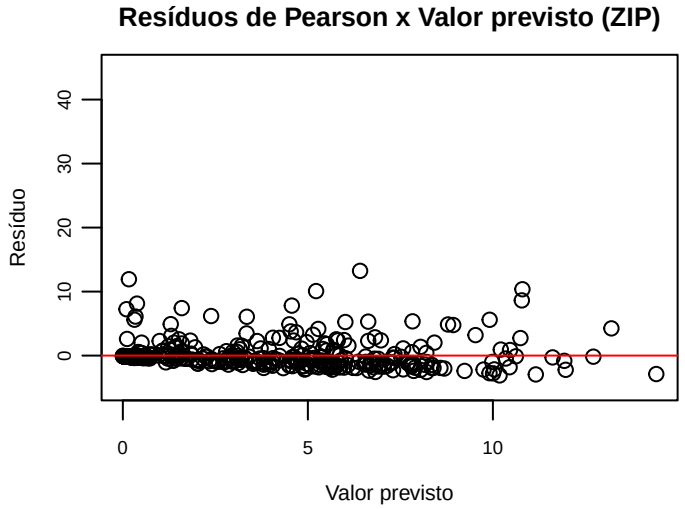
```
##  Family: poisson  ( log )
## Formula:          trips ~ ski + income_c + costS_100_c
## Zero inflation:         ~quality
## Data: RecreationDemand
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    2631.8    2658.7   -1309.9    2619.8       653
##
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.38474    0.04610  30.037  < 2e-16 ***
## skiyes       0.52836    0.05752   9.185  < 2e-16 ***
## income_c    -0.12716    0.01981  -6.418 1.38e-10 ***
## costS_100_c -1.50350    0.10260 -14.654  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.9731     0.2283   13.02   <2e-16 ***
## quality      -1.6184     0.1213  -13.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dada a seleção prévia realizada, todos os coeficientes mostram-se relevantes.

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

3.5.1. Avaliação do ajuste

Resíduos de Pearson



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 iguais. Na próxima seção, será verificado se ele conseguiu mitigar o problema do excesso de zeros.

Estatística qui-quadrado de razão de verossimilhança

```r
# 1. Obtenha as probabilidades da parte de zero inflado
p_zero <- predict(model_zip, type = "zprob")

# 2. Obtenha os valores esperados da parte Poisson
mu <- predict(model_zip, type = "conditional")

# 3. Calcule a probabilidade total de zero
p_zero_total <- p_zero + (1 - p_zero) * exp(-mu)

# 4. Simulação
n_sim <- 1000
zero_counts <- numeric(n_sim)

for (i in 1:n_sim) {
  sampled <- rbinom(length(p_zero_total), size = 1, prob = p_zero_total)
  zero_counts[i] <- sum(sampled)
}

# 5. Resultados
mean_predicted_zeros <- mean(zero_counts)
ci <- quantile(zero_counts, c(0.025, 0.975))

cat("Média total de zeros previstos em", n_sim, "simulações:", mean_predicted_zeros, "\n")
```

```
## Média total de zeros previstos em 1000 simulações: 420.449
```

Discussão e Conclusão

...

Referências

Ros Dobson Cameron & Trivedi