Sobredispersão em Modelos de Contagem

Modelagem Estatística

Paula Eduarda de Lima

Ciência de Dados e Inteligência Artificial

5º Período

Maio, 2025

1 Introdução

Modelos de regressão para dados de contagem são amplamente utilizados em diversas áreas aplicadas, como economia, saúde, ciências ambientais e sociais. O modelo de regressão de Poisson é, tradicionalmente, o ponto de partida para a classe de dados de contagem, sendo caracterizado pela suposição de que a variável resposta $Y \in \mathbb{N} \cup \{0\}$ segue uma distribuição de Poisson com parâmetro $\lambda > 0$, de modo que: $Y \sim \text{Poisson}(\lambda) \Rightarrow \mathbb{E}[Y] = \text{Var}(Y) = \lambda$.

Esta propriedade de equidispersão — igualdade entre a média e a variância — é central na formulação do modelo de Poisson. No entanto, em aplicações reais é comum observar situações em que a variância dos dados excede significativamente a média amostral, fenômeno conhecido como *sobredispersão*.

A sobredispersão pode comprometer seriamente a validade inferencial do modelo Poisson, tornando os erros padrão subestimados e inflando a significância estatística dos coeficientes. Isso motiva a investigação cuidadosa da dispersão dos dados antes da adoção de um modelo de Poisson.

Neste trabalho, analisamos o conjunto de dados `RecreationDemand`, proveniente do pacote `AER` do `R`, que registra o número de viagens recreativas feitas por proprietários de barcos ao Lago Somerville, Texas, em 1980. A partir de uma análise exploratória dos dados, ajustamos um modelo de regressão de Poisson, avaliamos a presença de sobredispersão utilizando o teste proposto por Cameron e Trivedi (1990), e comparamos diferentes alternativas de modelos, como a regressão Binomial Negativa e o modelo de Poisson Inflado de Zeros (ZIP).

Nosso objetivo é discutir, de forma fundamentada, os efeitos da sobredispersão em modelos de contagem e propor estratégias adequadas para seu tratamento, analisando métricas de ajuste e inferência estatística em cada cenário.

Objetivo do Estudo

O principal objetivo deste estudo é investigar a adequação do modelo de regressão de Poisson para dados de contagem observados no conjunto `RecreationDemand`, analisando em particular a presença de sobredispersão — isto é, situações em que a

variância excede a média. Para isso, buscamos: (i) realizar uma análise exploratória dos dados, com foco na estrutura da variável resposta e suas covariáveis; (ii) ajustar e interpretar um modelo de regressão de Poisson; (iii) aplicar testes formais para detecção de sobredispersão; e (iv) explorar modelos alternativos, como a regressão Binomial Negativa e o modelo de Poisson Inflado de Zeros, comparando o desempenho entre eles com base em métricas de ajuste. Este estudo visa compreender melhor os limites da Poisson clássica e apresentar soluções estatísticas mais robustas quando suas suposições não são atendidas.

2 Análise exploratória

Descrição do dataset

O conjunto de dados analisado é descrito na Tabela 2

Variável	Tipo	Descrição
`trips`	double	Número de viagens recreativas de barco realizadas pelo indivíduo.
`quality`	double	Avaliação subjetiva da qualidade da instalação (de 1 a 5), com 0 para quem não visitou o lago.
`ski`	factor	O indivíduo praticou ski aquático no lago? (`"yes"` ou `"no"`)
`income`	double	Renda anual da família do entrevistado (em milhares de dólares).
`userfee`	factor	O indivíduo pagou uma taxa anual de uso no Lago Somerville? (`"yes"` ou `"no"`)
`costC`	double	Custo ao visitar o Lago Conroe (em dólares).
`costS`	double	Custo ao visitar o Lago Somerville (em dólares).
`costH`	double	Custo ao visitar o Lago Houston (em dólares).

Tabela 1: Descrição das variáveis do conjunto `RecreationDemand`.

Detalhes

De acordo com a fonte original (Seller, Stoll e Chavas, 1985, p. 168), a avaliação de qualidade é feita em uma escala de 1 a 5, sendo atribuída nota 0 para os indivíduos que não visitaram o lago. Isso explica a média notavelmente baixa dessa variável, mas também sugere que seu tratamento em algumas publicações mais recentes está longe do ideal. Para manter a consistência com outras fontes, trataremos a variável como numérica, incluindo os valores zero.

Focaremos na variável "trips", analisando sua média, variância e investigando sua relação com as outras variáveis presentes no banco.

Variáveis categóricas

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
ski	0	1	FALSE	2	no: 417, yes: 242
userfee	0	1	FALSE	2	no: 646, yes: 13

Variáveis quantitativas

Variável	n_miss	Comp.	Média	DP	p0	p25	p50	p75	p100
trips	0	1	2.244	6.292	0.000	0.000	0.00	2.000	88.000
quality	0	1	1.419	1.812	0.000	0.000	0.00	3.000	5.000
income	0	1	3.853	1.852	1.000	3.000	3.00	5.000	9.000
costC	0	1	55.424	46.683	4.340	28.240	41.19	69.675	493.770
costS	0	1	59.928	46.377	4.767	33.312	47.00	72.573	491.547
costH	0	1	55.990	46.133	5.700	28.964	42.38	68.560	491.049

Não há dado faltante nas variáveis categóricas, nem nas quantitativas.

Resumo estatístico

```
summary(RecreationDemand)
```

```
##      trips           quality       ski          income       userfee
##  Min.   : 0.000   Min.   :0.000   no :417   Min.   :1.000   no :646
##  1st Qu.: 0.000   1st Qu.:0.000   yes:242   1st Qu.:3.000   yes: 13
##  Median : 0.000   Median :0.000             Median :3.000
##  Mean   : 2.244   Mean   :1.419             Mean   :3.853
##  3rd Qu.: 2.000   3rd Qu.:3.000             3rd Qu.:5.000
```

```
##  Max.   :88.000   Max.   :5.000            Max.   :9.000
##      costC            costS             costH
##  Min.   :  4.34   Min.   :  4.767   Min.   :  5.70
##  1st Qu.: 28.24   1st Qu.: 33.312   1st Qu.: 28.96
##  Median : 41.19   Median : 47.000   Median : 42.38
##  Mean   : 55.42   Mean   : 59.928   Mean   : 55.99
##  3rd Qu.: 69.67   3rd Qu.: 72.573   3rd Qu.: 68.56
##  Max.   :493.77   Max.   :491.547   Max.   :491.05
```

A variável de contagem "trips"apresenta mediana 0, o que nos aponta grande quantidade de zeros, além de assimetria, por ter média 2.244. Isso indica que a distribuição é fortemente concentrada em valores baixos, com muitos indivíduos realizando poucas viagens — o que é evidenciado pelo valor máximo de 88.

Análise univariada de "trips

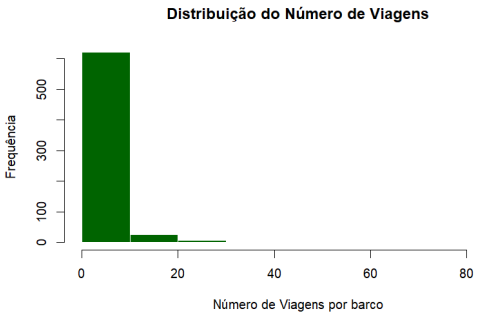


Figura 1: Visualização - Distribuição Número de Viagens

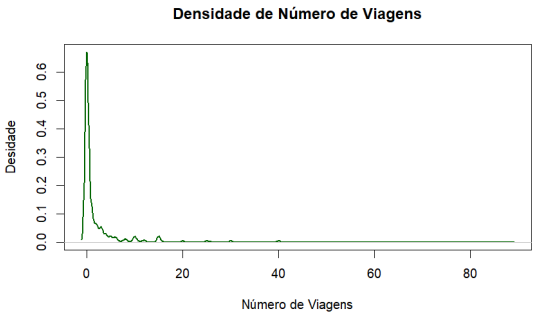


Figura 2: Visualização - Densidade do Número de Viagens

Confirmando o que foi visto no resumo estatístico, os gráficos de distribuição e densidade (Figuras 1 e 2) evidenciam ainda mais que a maioria dos valores da variável *trips* é bastante pequeno ou igual a zero.

Média e variância de "trips"

```{r}
```

```
summary_stats <- c(
  media = mean(RecreationDemand$trips),
  variancia = var(RecreationDemand$trips)
)
summary_stats
```

```
  media variancia
2.24431  39.59524
```

A variável trips apresenta média baixa e variância elevada, o que indica:

Distribuição assimétrica à direita (maioria fez 0 ou poucas viagens). Pode haver sobredispersão, o que influencia na escolha do modelo estatístico posteriormente (ex: Poisson vs NegBin).

Análise bivariada de "trips"e as demais

Covariâncias

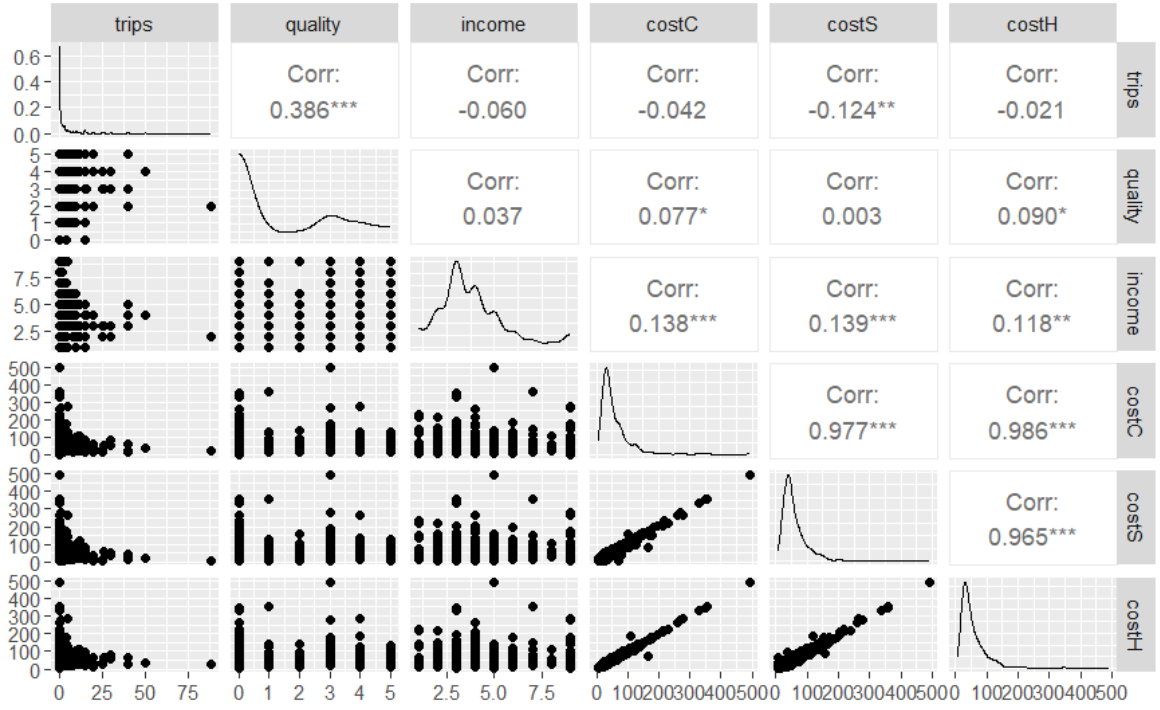


Figura 3: Tabela de covariância

Na parte superior direita da matriz, onde aparecem os coeficientes de correlação de Pearson, os asteriscos indicam significância estatística do valor da correlação, conforme a convenção comum de p-valor:

Símbolo	**p-valor**	**Interpretação**
***	$p < 0.001$	Correlação muito significativa
**	$p < 0.01$	Correlação bastante significativa
*	$p < 0.05$	Correlação significativa
.	$p < 0.1$	Tendência de significância
	$p \geq 0.1$	Não significativo

Na matriz de correlação, observamos que as três variáveis relacionadas a custo — `costC`, `costS` e `costH` — apresentam **alta correlação entre si**:

- `costC` $\sim$ `costS`: 0,977***
- `costC` $\sim$ `costH`: 0,986***
- `costS` $\sim$ `costH`: 0,965***

Esse padrão é um indicativo claro de **multicolinearidade**, o que também faz sentido do ponto de vista prático: o custo total de uma viagem para um lago é diretamente ligado ao custo de aos demais lagos.

Para evitar problemas de multicolinearidade no modelo, optamos por **manter apenas uma dessas variáveis**. A escolha será feita com base no nível **significância estatística** na relação com a variável resposta `trips`. Entre as três, a variável `costS` apresenta a **correlação mais forte e significativa com `trips`**:

- `costS` $\sim$ `trips`: -0,124**

As demais variáveis de custo têm correlações muito menores e estatisticamente menos significativas com a variável resposta.

Boxplots de "trips"por cada variável categorica

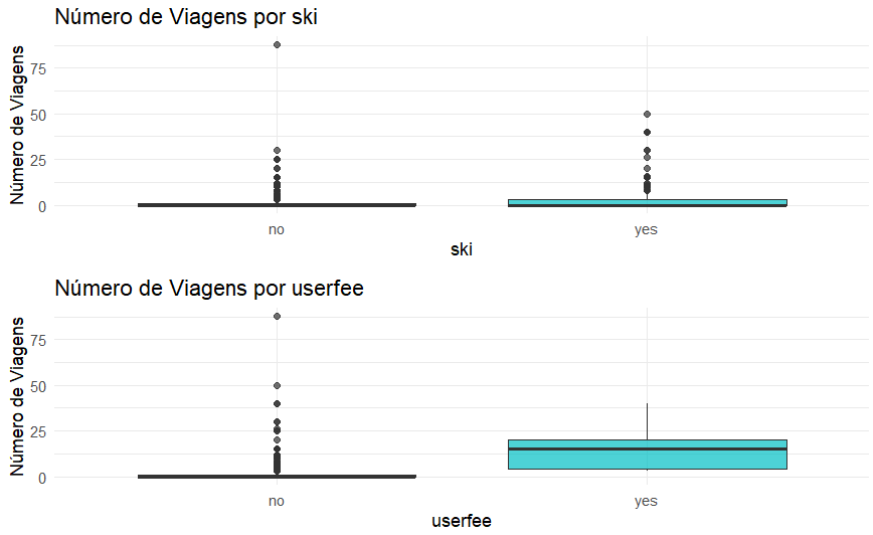
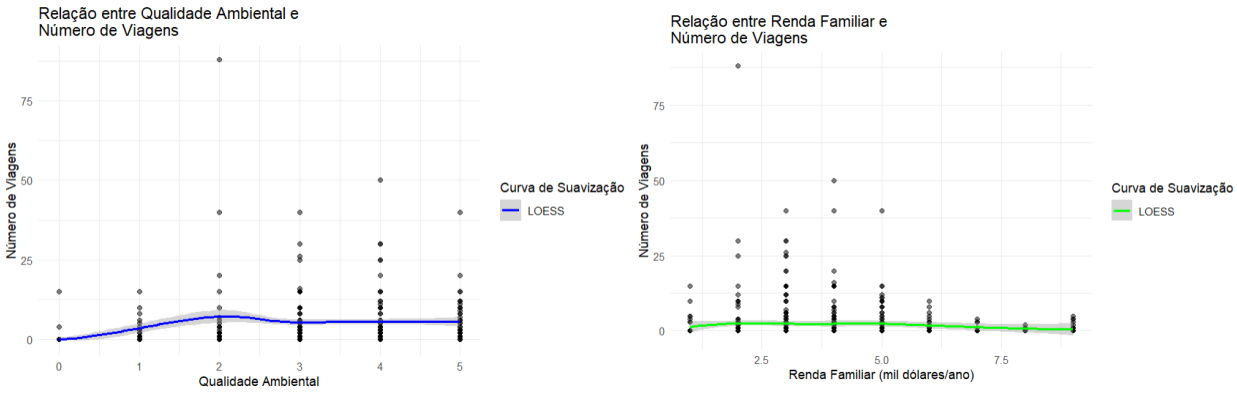


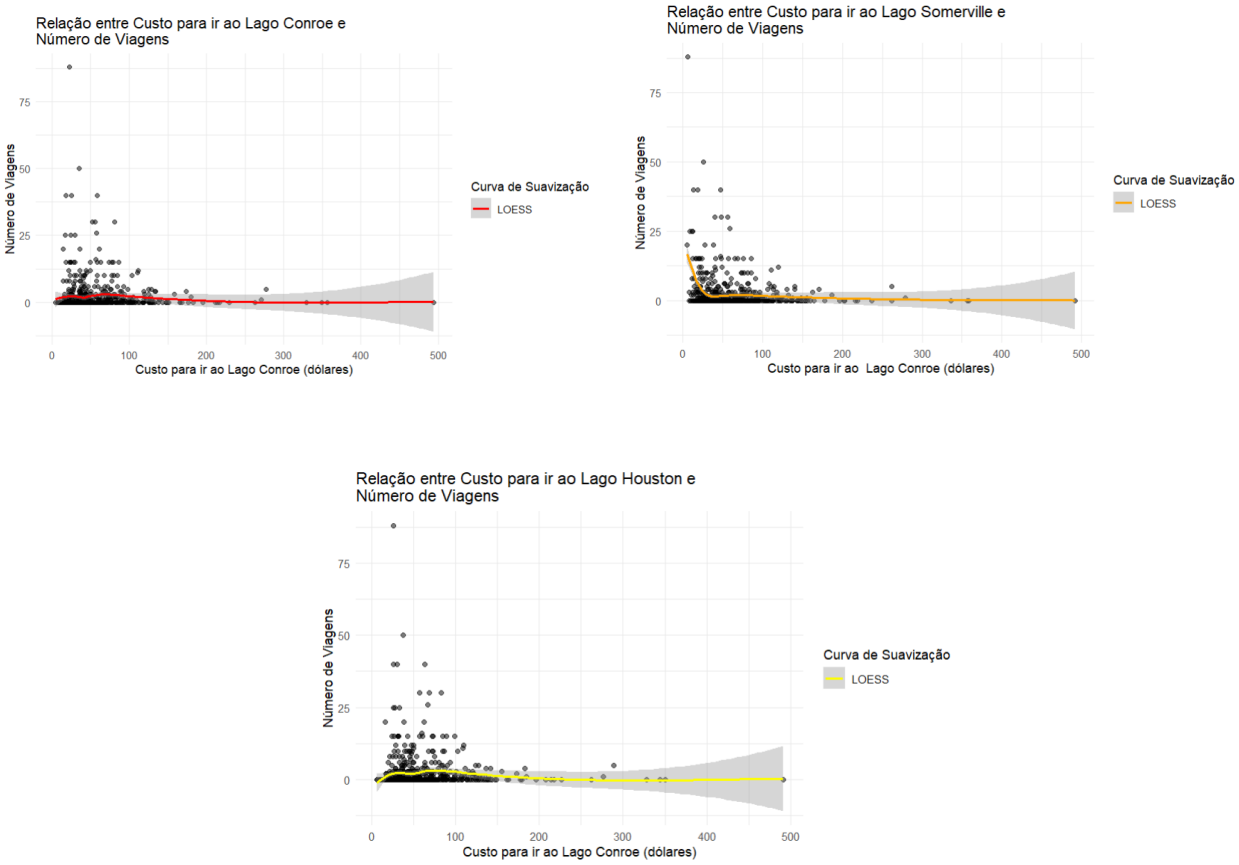
Figura 4: Visualização - Número de viagens pela variável “ski” e “userfree”

O número de viagens com barcos de ski tendem a ser um pouco maior, mas a diferença não é tão significativa.

Há maior número de viagens quando há taxa de uso.

Scatterplot de "trips"por cada variável quantitativa





Há tendência de aumento de viagens de barco com a renda, mas com bastante disperssão, o mesmo com a qualidade da instalação.

Em geral, quanto maior o custo, menor o número de viagens (relação negativa), no entanto os dados são bem variados e não aparentam tão grande correlação. Quanto maior o custo, maior a incerteza pela diminuição da quantidade de dados presente no dataset.

Covariáveis escolhidas

Como visto na análise de covariância, escolhemos apenas uma covariável para representar o custo, “costH “ e “costC“ serão desonciderадas por isso, para evitar multicolinearidade no modelo. Todas as demais serão incluidas, resultando nas covariáveis: income, costS , userfee , ski e quality

3 Métodos

Ajuste do Modelo Poisson

A escolha inicial por um modelo de regressão de Poisson se justifica pelas características da variável resposta *trips*, que representa contagens não-negativas inteiras. O modelo de Poisson é amplamente utilizado para modelar esse tipo de dado, a suposição inicial é de que os dados de contagem observados poderiam ser adequadamente explicados por esse modelo. Além disso, utilizou-se a formulação de um modelo linear generalizado (GLM), com função de ligação logarítmica, que permite relacionar a média condicional λ_i com uma combinação linear dos preditores disponíveis no conjunto de dados *RecreationDemand*. Esse procedimento fornece uma estrutura flexível para estimar os efeitos das covariáveis sobre a taxa média de viagens realizadas pelos indivíduos da amostra.

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{com} \quad \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} \quad (1)$$

- Y_i: número de viagens (*trips*) para o indivíduo i;
- λ_i: média da distribuição de Poisson (esperança de Y_i);
- $\log(\lambda_i)$: função de ligação logarítmica usada no modelo de Poisson;
- $x_{i1}, x_{i2}, \ldots, x_{i6}$: variáveis explicativas (*quality*, *income*, *costS*, *userfee* , *ski* e *quality.*);
- $\beta_0, \beta_1, \ldots, \beta_p$: coeficientes do modelo.

Formulação Matemática

A função de massa de probabilidade da distribuição de Poisson é:

$$\mathbb{P}(Y_i = y_i \mid \mathbf{x}_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

Com a função de ligação logarítmica, temos:

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

A função de verossimilhança para n observações é dada por:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \left(e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\right)^{y_i}}{y_i!}$$

E a log-verossimilhança correspondente é:

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^\top \boldsymbol{\beta} - e^{\mathbf{x}_i^\top \boldsymbol{\beta}} - \log(y_i!) \right]$$

A estimação dos parâmetros $\boldsymbol{\beta}$ é realizada pela maximização dessa log-verossimilhança. Para modelar o número de viagens (*trips*), que é uma variável de contagem, foi utilizado um modelo de regressão de Poisson ajustado via `glm()` em R

```
Call:
glm(formula = trips ~ income + costS + userfee + ski + quality,
    family = poisson, data = RecreationDemand)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.586097   0.091906   6.377 1.80e-10 ***
income      -0.157829   0.019502  -8.093 5.82e-16 ***
costS       -0.015315   0.001014 -15.098  < 2e-16 ***
userfeeyes   1.101518   0.079901  13.786  < 2e-16 ***
skiyes       0.454188   0.056463   8.044 8.69e-16 ***
quality      0.540831   0.015942  33.924  < 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: 2687.5  on 653  degrees of freedom
```

```
AIC: 3452.6

Number of Fisher Scoring iterations: 7
```

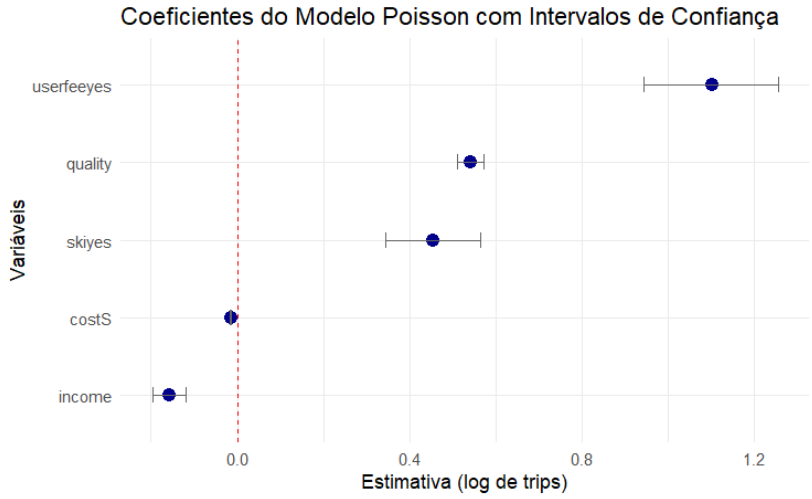


Figura 5: Coeficiente previstos e Intervalo de confiança

- As variáveis explicativas apresenta efeito estatisticamente significativo sobre o número de viagens realizadas, considerando o p-valor e o intervalo de confiança, apesar do coeficiente de *costS*, ser próximo de 0.
- **quality**, **skiyes** e **userfeeyes** estão positivamente associadas ao número de viagens.
- A variável **income** possui coeficiente negativo ($\beta = -0{,}157829$), indicando que maiores rendas estão associadas a uma menor frequência de viagens.
- O **custo ao visitar o Lago Somerville (costS)** reduz o número esperado de viagens e seu coeficiente é bem próximo de zero, mas o intervalo de confiança não inclui o zero.
- A redução da deviance (de 4849,7 para 2687.5) e o AIC de 3452,6 indicam um bom ajuste do modelo.

Ajuste de um Modelo de Regressão de Poisson

- **Modelo de regressão de Poisson:** apropriado para modelar variáveis de contagem com valores inteiros não-negativos.
- **Função de ligação logarítmica:** padrão no modelo de Poisson, garante que as médias previstas sejam positivas.
- **Estimação por máxima verossimilhança:** método usado automaticamente pela função `glm()` para GLMs.

Métricas de Avaliação do Modelo

- **Deviance:** a *null deviance* (4849,7) foi reduzida para *residual deviance* (2687,5), o que indica que o modelo ajustado explica parte substancial da variabilidade dos dados em relação ao modelo nulo. A deviance é dada por:

$$D = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

  no caso da distribuição Poisson, sendo usada para comparar o modelo ajustado com o modelo saturado.

- **AIC (Akaike Information Criterion):** mede a qualidade do modelo em relação à sua complexidade. É definido por:

$$\text{AIC} = -2\log(L) + 2k$$

  onde L é a verossimilhança do modelo e k é o número de parâmetros. O valor de 3452,6 sugere bom equilíbrio entre ajuste e parcimônia. Quanto menor o AIC, melhor o modelo (ao comparar modelos aninhados ou com os mesmos dados).

- **Significância dos coeficientes:** testada via estatística z, definida por:

$$z = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

  e o p-valor é obtido com base na distribuição normal padrão.

- **Erro padrão dos coeficientes:** fornece uma medida da variabilidade das estimativas dos coeficientes. É calculado a partir da matriz de variância-covariância dos estimadores:

$$\text{SE}(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}$$

onde $\widehat{\text{Var}}(\hat{\beta}_j)$ é a variância estimada do coeficiente $\hat{\beta}_j$, normalmente obtida da diagonal da inversa da matriz de informação de Fisher.

Teste para Sobredispersão

Vamos seguir o procedimento descrito por Cameron e Trivedi (1990) para testar a presença de sobredispersão nos dados do conjunto `RecreationDemand`. O objetivo é verificar se a variância da variável resposta `trips` excede o valor esperado sob o modelo de Poisson. Formalmente, testamos a hipótese:

- $H_0 : \alpha = 0$ (dispersão adequada, o modelo de Poisson é apropriado)
- $H_1 : \alpha \neq 0$ (há sobredispersão, o modelo de Poisson é inadequado)

Passo 1: Calcular os valores ajustados $\hat{\mu}_i$ do modelo de Poisson:

```{r}
mu_hat <- fitted(poisson_model)
y <- RecreationDemand$trips
```

Passo 2: Calcular a estatística $Z_i = \frac{(Y_i - \hat{\mu}_i)^2 - Y_i}{\hat{\mu}_i}$:

```{r}
Z <- ((y - mu_hat)^2 - y) / mu_hat
```

Passo 3: Regressar Z_i sobre $\hat{\mu}_i$, sem intercepto, e verificar o coeficiente:

```{r}
sobredisp_test <- lm(Z ~ 0 + mu_hat)
summary(sobredisp_test)
```

Resultado:

```
Call:
lm(formula = Z ~ 0 + mu_hat)

Residuals:
    Min      1Q  Median      3Q     Max
 -71.65   -3.23   -0.46   -0.27 1999.93

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
mu_hat   1.6406     0.7664   2.141   0.0327 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.13 on 658 degrees of freedom
Multiple R-squared:  0.006916, Adjusted R-squared:  0.005406
F-statistic: 4.582 on 1 and 658 DF,  p-value: 0.03267
```

Interpretação:

O p-valor do coeficiente de `mu_hat` foi inferior a 0,05, o que nos leva a rejeitar a hipótese nula H_0. Isso indica a presença de sobredispersão nos dados, ou seja, a variância da variável resposta é maior do que a média, violando uma das principais suposições do modelo de Poisson.

Conclusão:

Como a hipótese de sobredispersão adequada foi rejeitada, o modelo de Poisson não é apropriado. Dessa forma, recomenda-se o uso da regressão com distribuição Binomial Negativa como alternativa mais robusta e adequada para os dados.

Modelos Alternativos: Binomial Negativo

A distribuição Binomial Negativa é uma extensão da Poisson que permite acomodar sobredispersão, introduzindo um parâmetro de dispersão adicional θ. Esta distribuição é definida por:

$$Y_i \sim \text{NB}(\mu_i, \theta)$$
$$\mathbb{E}[Y_i] = \mu_i, \quad \text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\theta}$$

$$\log\left(\mathbb{E}[Y]\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Quando $\theta \to \infty$, a variância se aproxima da média, recuperando o modelo de Poisson como caso particular.

Formulação Matemática

A função de massa de probabilidade da distribuição Binomial Negativa, parametrizada pela média μ_i e pelo parâmetro de dispersão θ, é dada por:

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

Com a função de ligação logarítmica, temos:

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

A log-verossimilhança correspondente (omitindo termos constantes) é:

$$\log \mathcal{L}(\boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \left[\log \Gamma(y_i + \theta) - \log \Gamma(\theta) - \log y_i! + \theta \log\left(\frac{\theta}{\theta + \mu_i}\right) + y_i \log\left(\frac{\mu_i}{\theta + \mu_i}\right)\right]$$

A estimação dos parâmetros $\boldsymbol{\beta}$ e θ é realizada por máxima verossimilhança, normalmente por métodos iterativos como o algoritmo de Fisher scoring.

1. Por que a Binomial Negativa é mais adequada?

- O modelo de Poisson assume $\text{Var}(Y) = \mu$, o que no caso analisado é irrealista, como visto na análise da variável "trips".
- A Binomial Negativa adiciona um parâmetro de dispersão θ, permitindo que a variância cresça mais rapidamente que a média.

- Garante inferência mais precisa, com erros-padrão corretos e testes mais confiáveis.
- Melhora o ajuste geral do modelo, como indicado pela redução do AIC.

Ajuste do Modelo Binomial Negativo

O modelo é ajustado com a função `glm.nb()` do pacote `MASS`, utilizando a função de ligação logarítmica padrão:

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

Essa ligação:

- Garante que $\mu_i > 0$, compatível com dados de contagem;
- Permite interpretar os coeficientes como efeitos multiplicativos;
- É a escolha padrão para modelos de contagem em `glm()` e `glm.nb()`.

Exemplo: Um coeficiente de 0,05 para `income` indica que um aumento de uma unidade em `income` está associado a um aumento de $e^{0.05} \approx 1.05$ vezes no número esperado de viagens.

```
Call:
glm.nb(formula = trips ~ income + costS + userfee + ski + quality,
    data = RecreationDemand, init.theta = 0.4713992214, link = log)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.836942   0.227014  -3.687 0.000227 ***
income      -0.066335   0.047669  -1.392 0.164051
costS       -0.012606   0.002433  -5.182  2.2e-07 ***
userfeeyes   1.522432   0.425474   3.578 0.000346 ***
skiyes       0.553319   0.167553   3.302 0.000959 ***
quality      0.886736   0.042212  21.007  < 2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.4714) family taken to be 1)

    Null deviance: 956.34  on 658  degrees of freedom
Residual deviance: 453.89  on 653  degrees of freedom
AIC: 1802.7

Number of Fisher Scoring iterations: 1


              Theta:  0.4714
          Std. Err.:  0.0448
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -1788.7030
```

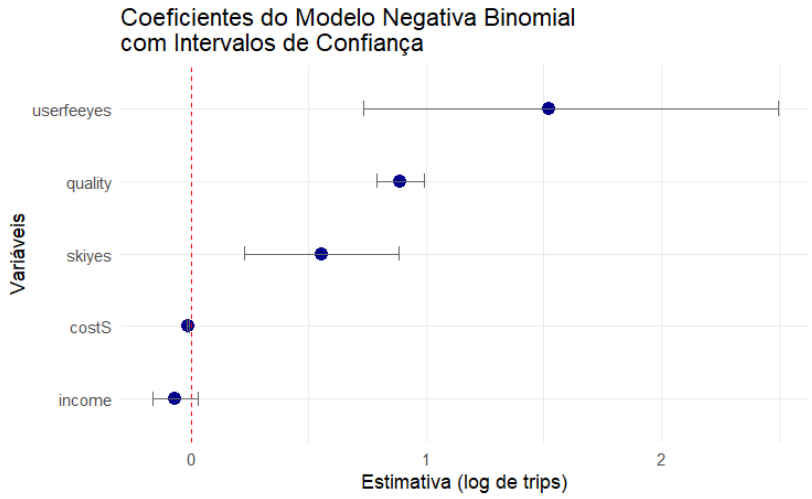


Figura 6: Coeficiente previstos e Intervalo de confiança

- As variáveis explicativas apresenta efeito estatisticamente significativo sobre o número de viagens realizadas, considerando o p-valor e o intervalo de confiança, apesar do coeficiente de *costS*, ser próximo de 0. A única exceção foi a variável

"Income"que se mostrou pouco siginificativa de acordo com o p-valor do teste e o intervalo de confiança que inclui o 0.

- **quality**, **skiyes** e **userfeeyes** estão positivamente associadas ao número de viagens, como no Modelo Poisson.
- O **custo ao visitar o Lago Somerville (costS)** reduz o número esperado de viagens e seu coeficiente é bem próximo de zero, mas o intervalo de confiança não inclui o zero.
- **Income** (renda) não apresentou significância estatística (p = 0,164051), sugerindo ausência de efeito relevante sobre as viagens.
- A redução da deviance (de 956.34 para 453.89) e o AIC de 1802.7 indicam um bom ajuste do modelo.

Comparação com o modelo de Poisson

AIC:

```{r}
AIC(poisson_model)
AIC(nb_model)
```

```
[1] 3074.057
[1] 1677.56
```

O modelo com menor AIC é preferível: negativo binomial.

Teste de razão de verossimilhança:

Podemos comparar os dois modelos usando o teste de razão de verossimilhança:

```{r}
lrtest(poisson_model, nb_model)
```

```
Likelihood ratio test

Model 1: trips ~ income + costS + userfee + ski + quality
Model 2: trips ~ income + costS + userfee + ski + quality
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   6 -1720.28
2   7  -894.35  1 1651.9  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como o p-valor do teste foi pequeno, o modelo binomial negativa é significativamente melhor que o de Poisson.

Conclusão

Como o modelo binomial negativo teve menor AIC e o teste de razão de verossimilhança indicar melhoria significativa, então ele deve ser preferido ao modelo de Poisson para os dados RecreationDemand.

Avaliação do Modelo

- **Significância dos parâmetros:** Análise dos p-valores no `summary()`.
- **Critério de Informação de Akaike (AIC):** Modelos com menor AIC são preferíveis.
- **Parâmetro de dispersão θ:** Um valor pequeno de θ indica alta variabilidade extra, reforçando a inadequação do modelo de Poisson.

Excesso de zeros

Na análise exploratória inicial, observamos que uma grande proporção dos barcos no conjunto de dados reportou **zero viagens**. Isso pode indicar **excesso de zeros**, um fenômeno comum em dados de contagem, onde a **frequência de zeros é maior**

do que a esperada pelo modelo de Poisson ou até mesmo pelo modelo de Binomial Negativo.

Comparando a quantidade de zeros observados e esperados

```{r}
# Quantidade de zeros observada
zeros_observados <- sum(RecreationDemand$trips == 0)

# Previsão do modelo de Poisson
mu_poisson <- predict(poisson_model, type = "response")
zeros_estimados_poisson <- sum(dpois(0, lambda = mu_poisson))

c(zeros_observados = zeros_observados,
  zeros_estimados_poisson = zeros_estimados_poisson)

```

```
        zeros_observados         zeros_estimados_poisson
                417.0000                        275.2735
```

Como a quantidade de zeros estimada foi substancialmente menor do que a observada, temos uma forte indicação de que o modelo de Poisson não está capturando adequadamente o excesso de zeros.

Modelo de Poisson Inflacionado de Zeros (ZIP)

Na presença de **excesso de zeros**, o modelo de Poisson pode ser inadequado. Para lidar com esse fenômeno, podemos utilizar o **modelo de Poisson Inflacionado de Zeros (ZIP)**.

Esse modelo combina dois processos:

- Um modelo **logístico** que estima a probabilidade de ocorrência de zeros estruturais (ou seja, observações que sempre terão valor zero);

- Um modelo **Poisson** para os demais valores (inclusive para zeros que não são estruturais).

Formulação Matemática

No modelo ZIP, a distribuição da variável resposta Y_i é definida como uma mistura de duas componentes:

$$\mathbb{P}(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i}, & \text{se } y_i = 0 \\ (1-\pi_i)\dfrac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, & \text{se } y_i > 0 \end{cases}$$

onde:

- $\pi_i = \mathbb{P}(Y_i = 0 \text{ estrutural})$ é a probabilidade de zero estrutural, modelada por regressão logística:
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i^\top \boldsymbol{\gamma}$$
- λ_i é o parâmetro da distribuição de Poisson para os não-zeros estruturais, com:
$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

Esse modelo é ajustado por máxima verossimilhança, tratando separadamente o processo de geração de zeros e o processo de contagem, o que permite maior flexibilidade quando há excesso de zeros nos dados.

Ajuste do Modelo ZIP

Utilizaremos a função `zeroinfl()` do pacote `pscl`, especificando o modelo de contagem e o modelo logístico separadamente.

```
Call:
zeroinfl(formula = trips ~ income +
costS + userfee + ski + quality | income + ski,
data = RecreationDemand)

Pearson residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.87822 -0.64327 -0.59521 -0.03182 17.70152

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.375415   0.114898  20.674  < 2e-16 ***
income      -0.130309   0.021321  -6.112 9.86e-10 ***
costS       -0.015136   0.001033 -14.660  < 2e-16 ***
userfeeyes   0.792024   0.078935  10.034  < 2e-16 ***
skiyes       0.473575   0.058322   8.120 4.66e-16 ***
quality      0.105718   0.028036   3.771 0.000163 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.66123    0.20305   3.256  0.00113 **
income      -0.01510    0.05179  -0.292  0.77061
skiyes      -0.48828    0.18324  -2.665  0.00770 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 18
Log-likelihood: -1515 on 9 Df
```

Note que a fórmula y x1 + x2 | z1 + z2 indica:

À esquerda do |: covariáveis do modelo de contagem Poisson;

À direita do |: covariáveis do modelo logístico para o excesso de zeros.

O modelo ZIP assume dois processos:

Um processo binário que determina se a observação está na parte "zero-inflada"(ou seja, sempre zero).

Um processo de Poisson que gera contagens (inclusive zeros que não são estruturais).

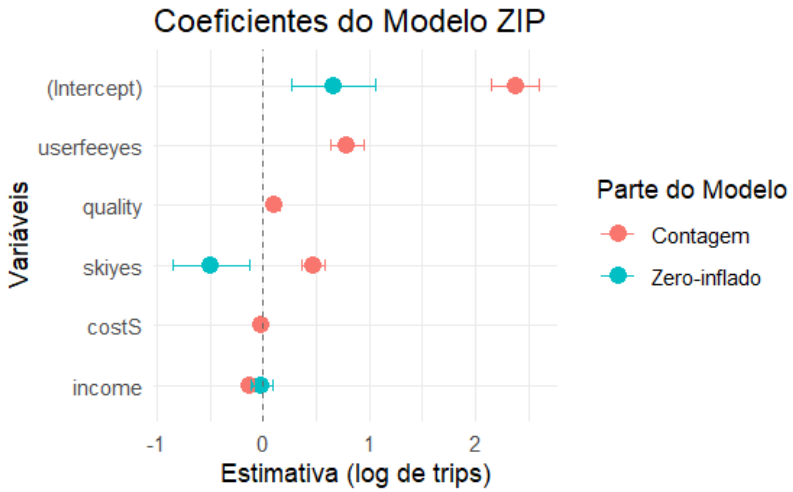


Figura 7: Coeficiente previstos e Intervalo de confiança

Modelo	df	AIC
poisson_model	7	3452.560
nb_model	8	1802.703
zip_model	9	3048.226

Tabela 2: Comparação do número de parâmetros e AIC entre os modelos ajustados

```{r}
# Log-verossimilhanças
logLik(poisson_model)
logLik(nb_model)
logLik(zip_model)

```

```
'log Lik.' -1720.28 (df=6)
'log Lik.' -894.3514 (df=7)
'log Lik.' -1515.113 (df=9)
```

```{r}
# Comparando número de zeros observados e estimados
```

```
observed_zeros <- sum(RecreationDemand$trips == 0)
predicted_zeros_poisson <- sum(dpois(0, lambda = predict(poisson_model,
type = "response")))
predicted_zeros_zip <- sum(predict(zip_model, type = "prob")[,1])

c(
  Observados = observed_zeros,
  Poisson = round(predicted_zeros_poisson),
  ZIP = round(predicted_zeros_zip)
)
```

```
Observados    Poisson        ZIP
    417      275.2735        416
```

4 Conclusão e trabalhos futuros

Este trabalho explorou a modelagem estatística de dados de contagem, com foco no problema da sobredispersão e no excesso de zeros no conjunto de dados RecreationDemand. Inicialmente, ajustamos um modelo de regressão de Poisson, que se mostrou inadequado devido à evidente sobredispersão nos dados, confirmada pelo teste de Cameron e Trivedi. A análise revelou que a variância da variável resposta trips era significativamente maior que sua média, violando a suposição fundamental do modelo de Poisson. Como alternativa, utilizamos a regressão Binomial Negativa, que incorpora um parâmetro de dispersão adicional, proporcionando um ajuste mais adequado aos dados e reduzindo significativamente o AIC em comparação ao modelo Poisson. Além disso, investigamos a presença de excesso de zeros, que levou à aplicação do modelo ZIP (Poisson Inflacionado de Zeros). Embora o modelo ZIP tenha capturado melhor a estrutura de zeros observada, o modelo Binomial Negativo apresentou melhor desempenho geral em termos de AIC e log-verossimilhança, sugerindo que a sobredispersão, e não apenas o excesso de zeros, era o principal desafio neste conjunto de dados.

Para trabalhos futuros, seria interessante explorar outras extensões dos modelos de contagem, como o modelo ZINB (Binomial Negativo Inflacionado de Zeros), que combina a flexibilidade da Binomial Negativa com a capacidade de modelar excesso de zeros. Além disso, a inclusão de efeitos aleatórios poderia ser investigada para capturar possíveis heterogeneidades não observadas nos dados.

Referências

[1] P. Lima, "Sobredispersao em modelos contagem" GitHub. Disponível em: https://github.com/PAULA-123/StatMod_Sobredispersao_em_modelos_contagem.

[2] Cameron, A. C., & Trivedi, P. K. (1990). *Regression-based tests for overdispersion in the Poisson model*. Journal of Econometrics, 46(3), 347-364.

[3] Zeileis, A., Kleiber, C., & Jackman, S. (2008). *Regression models for count data in R*. Journal of Statistical Software, 27(8), 1-25.

[4] Science and statistics. Journal of the American Statistical Association, 71(356), 791-799; Box (1976)..