Fundação Getúlio Vargas
Escola de Matemática Aplicada

Ciência de Dados e Inteligência Artificial

Modelos de Contagem: Recreation Demand

Kaiky Eduardo Alves Braga

Rio de Janeiro
Junho / 2025

Sumário

1 Introdução **3**

2 Metodologia **4**
2.1 Modelos 4
2.2 Ajuste 4
2.3 Avaliação 5
2.3.1 Comparação de Modelos 5
2.3.2 Comparação Interna do Modelo 5
2.4 Seleção de Covariáveis 6

3 Contextualização **6**
3.1 Distribuição de Valores 6
3.2 Análise Exploratória 7
3.3 Resultados 8
3.4 Modelo de Poisson 8
3.5 Teste de Sopredispersão 9
3.6 Modelo Binomial Negativa 10
3.7 Zero-Inflated Poisson (ZIP) 11
3.8 Comparação entre modelos e particularidades 13

4 Discussão e Conclusão **17**

5 Referências **17**

1 Introdução

O conjunto de dados Recreation Demand oferece uma base valiosa para estudos sobre demanda por atividades recreativas, especialmente no contexto de modelagem de dados de contagem. Ele contém informações coletadas em 1980 por meio de uma pesquisa com 2.000 proprietários de embarcações de lazer registrados em 23 condados do leste do Texas, EUA. O foco principal é o número de viagens recreativas realizadas ao Lago Somerville, permitindo análises detalhadas sobre os fatores que influenciam o comportamento recreativo dos indivíduos.

Neste relatório, propomos a análise da variável `trips` como variável dependente, a partir da construção e avaliação de modelos de contagem ajustados com base em preditores selecionados. Serão exploradas diferentes especificações de modelos com o objetivo de identificar a abordagem estatística mais adequada para explicar a variabilidade observada nas decisões de visita ao Lago Somerville.

Composto por 659 observações e 8 variáveis (todos com campos populados), o conjunto de dados inclui informações como:

- **trips:** número de viagens recreativas ao Lago Somerville;
- **quality:** avaliação subjetiva da qualidade das instalações do lago (escala de 1 a 5);
- **ski:** indicador se o indivíduo praticou esqui aquático no lago;
- **income:** renda anual do domicílio do respondente (em milhares de dólares);
- **userfee:** indicador se o indivíduo pagou uma taxa anual de uso no Lago Somerville;
- **costC, costS, costH:** despesas associadas a visitas aos Lagos Conroe, Somerville e Houston, respectivamente.

Recreation Demand							
trips	quality	ski	income	userfee	costC	costS	costH
0.00	0.00	yes	4.00	no	67.59	68.62	76.80
0.00	0.00	no	9.00	no	68.86	70.94	84.78
0.00	0.00	yes	5.00	no	58.12	59.47	72.11
0.00	0.00	no	2.00	no	15.79	13.75	23.68
0.00	0.00	yes	3.00	no	24.02	34.03	34.55
0.00	0.00	yes	5.00	no	129.46	137.38	137.85

Tabela 1: Amostra de 6 observações da Tabela Recreation Demand.

Observarção: Optamos por converter previamente as colunas "ski"e "userfree"para dados binários na nossa análise ("yes"/"no"$\rightarrow$ 0/1).

2 Metodologia

2.1 Modelos

Antes de tudo, precisamos entender o conceito de Modelos Lineares Generalizados (GLM), que são modelos que relacionam a média condicional da variável resposta Y a um conjunto de preditores $X = (X_1, X_2, \ldots, X_p)$ por meio de uma função de ligação $g(\cdot)$. A estrutura básica do GLM é dada por:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

onde $\mu_i = E(Y_i|X_i)$ é a média condicional de Y_i, η_i é o preditor linear, e $g(\cdot)$ é a função de ligação que conecta a média à combinação linear dos preditores. A variável resposta Y_i é assumida como pertencente a uma família de distribuições, o que permite modelar diferentes tipos de dados.

Para nosso problema de contagem, utlizamos os seguintes modelos:

- Poisson: Modelo clássico para dados de contagem, assume que a média e a variância da variável resposta são iguais. É útil em contextos simples, mas pode ser inadequado quando há sobredispersão (variância maior que a média).
- Binomial Negativa: Modelo alternativo ao de Poisson quando há sobredispersão. Introduz um parâmetro extra para modelar a variabilidade adicional, tornando-se mais flexível para dados reais com maior dispersão.
- ZIP: O modelo Poisson Inflacionado de Zeros (Zero-Inflated Poisson) é usado quando há excesso de zeros nos dados, ou seja, mais zeros do que o esperado sob uma distribuição Poisson. Esse modelo combina uma parte que gera apenas zeros com outra parte que segue uma Poisson.

2.2 Ajuste

Para o ajuste, optamos por utilizar a máxima verossimilhança, uma abordagem amplamente utilizada para estimar os parâmetros dos modelos de regressão. A função de verossimilhança foi especificada de acordo com a distribuição assumida (Poisson, Binomial Negativa ou ZIP), todos resultados obtidos à partir de códigos em R. Por padrão, a linguagem usou os seguintes métodos numéricos para determinação dos hiperparâmetros de cada modelo:

Tabela 2: Resumo dos Métodos Numéricos Usados (Padrão)

Modelo	**Método Numérico Principal**	**Algoritmo Específico**
Poisson (glm)	Máxima Verossimilhança	Iteração Fisher scoring (Newton-Raphson)
Binomial Negativa (glm.nb)	Máxima Verossimilhança + otimização numérica para dispersão	Iteração Fisher scoring + otimização numérica (Newton)
Zero-Inflated Poisson (zeroinfl)	Máxima Verossimilhança via otimização direta	Otimização com `optim()` (BFGS)

2.3 Avaliação

Na nossa modelagem, possuímos dois tipos de avaliações distintas:

2.3.1 Comparação de Modelos

- **AIC:** O Critério de Informação de Akaike (AIC) foi utilizado para comparar modelos distintos. Ele penaliza modelos com muitos parâmetros e favorece modelos mais parcimoniosos que explicam bem os dados. Modelos com menor AIC são preferidos. Além disso, sua utlização é interessante pois permite a comparação de modelos não aninhados.

- **Previsões pontuais:** Analisamos a distribuição de desfechos para verificar o quão bem o modelo captura a distribuição observada dos dados. Essa análise permite avaliar a qualidade das previsões e identificar possíveis falhas estruturais nos modelos.

- **Resíduos:** Representam a diferença entre os valores observados e os valores previstos pelo modelo. A análise dos resíduos permite avaliar o quão bem o modelo se ajusta aos dados, identificando possíveis desvios sistemáticos, outliers ou falhas na suposição do modelo. Quanto mais concentrados em torno de zero, melhor o desempenho do modelo.

2.3.2 Comparação Interna do Modelo

- **Teste t:** Utilizado para verificar a significância estatística dos coeficientes estimados no modelo. Um valor absoluto elevado da estatística t sugere uma associação significativa entre a covariável e a variável resposta.

- **P-Valor:** O p-valor associado a cada coeficiente indica a evidência contra a hipótese nula de que o coeficiente é zero. Valores menores que 0.05 são geralmente interpretados como estatisticamente significantes.

- **Teste z:** Utilizado para avaliar a significância estatística dos coeficientes em modelos cuja distribuição assintótica dos estimadores é normal, como em modelos lineares generalizados. Um valor absoluto elevado da estatística z indica uma forte evidência de que o coeficiente associado é diferente de zero.

- **Deviance média residual:** A razão entre o desvio residual e os graus de liberdade é uma medida que indica o quão bem o modelo se ajusta aos dados. Em modelos de contagem, o ideal é que esse valor esteja próximo de 1. Isso significa que a variância observada nos dados está em linha com o esperado pelo modelo. Valores muito maiores que 1 indicam que os dados têm mais variabilidade do que o modelo considera (sobredispersão), enquanto valores muito menores sugerem que a variabilidade é menor do que o esperado (subdispersão). Portanto, esse índice ajuda a diagnosticar se o modelo está adequado ou se há necessidade de ajustes.

2.4 Seleção de Covariáveis

Para os modelos de contagem (Poisson e Binomial Negativa), utilizamos o método Stepwise Forward para selecionar as variáveis mais relevantes. O procedimento começa com um modelo nulo, que contém apenas o intercepto, e adiciona uma variável por vez, escolhendo sempre aquela que mais melhora o modelo segundo o AIC.

A cada etapa, o algoritmo compara os modelos possíveis com a adição de uma nova variável e mantém aquela que gera a maior redução do AIC. O processo continua até que nenhuma adição melhore significativamente o ajuste do modelo.

Este método busca equilibrar qualidade de ajuste e simplicidade, evitando incluir variáveis irrelevantes.

Após tal processamento, verificamos o nível de significância de cada coeficiente, e selecionamos aqueles com nível de significância de até 0.05, considerado um valor aceitável.

3 Contextualização

3.1 Distribuição de Valores

Observando o comportamento das contagens da coluna Trips, percebemos que a distribuição dos dados está fortemente concentrada no valor zero, com poucas ocorrências associadas a valores mais altos. Esse padrão indica uma assimetria acentuada e uma grande proporção de zeros, o que pode dificultar o desempenho de alguns GLMs em capturar adequadamente os extremos da distribuição. Para lidar com esse problema, adotaremos futuramente um modelo com inflação de zeros, que é mais apropriado para esse tipo de estrutura de dados.

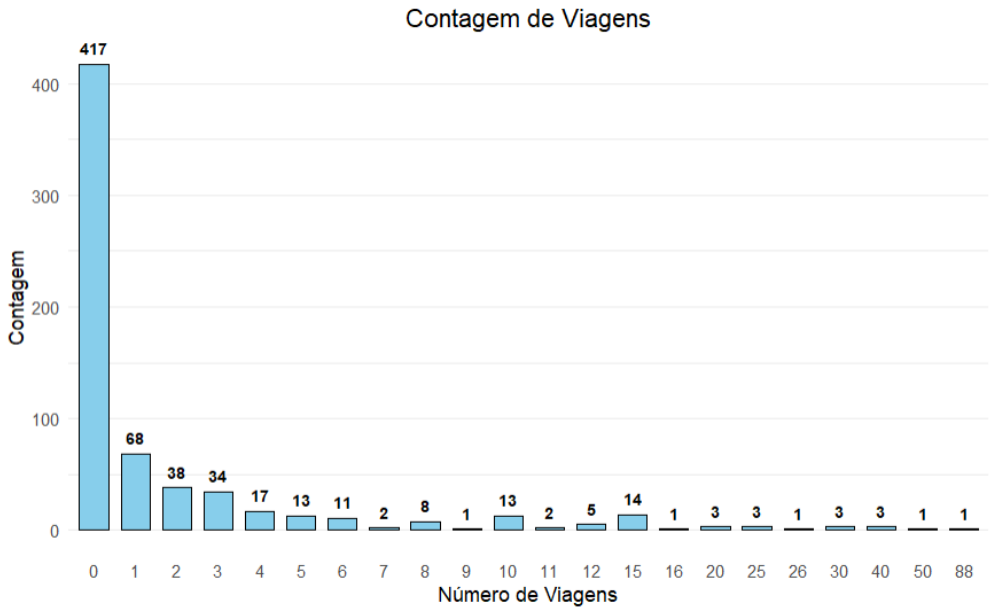

Figura 1: Distribuição de Trips

3.2 Análise Exploratória

Podemos observar os dados de cada coluna, presente em nosso conjunto de dados. Logo de cara, é perceptível que a nossa varíavel de interesse a ser usada como desfecho (trips), possui uma super variância se comparada com a sua média (um forte indiício do fenômeno conhecido como sobredispersão que exploraremos posteriormente). Além disso, as covariáveis de custo, possuem medidas de resumo muito parecidas. Além disso, possuímos covariáveis binárias em nossa base. Perceba que a média de userfree, por ser tão baixa e ser uma variável binária, indica que o evento de userfree costuma ser raro, aparecendo em cerca de 2 por cento de nossa base.

Resumo Estatístico das Variáveis

Estatística	trips	quality	ski	income	userfee	costC	costS	costH
mean	2.24	1.42	0.37	3.85	0.02	55.42	59.93	55.99
sd	6.29	1.81	0.48	1.85	0.14	46.68	46.38	46.13
min	0.00	0.00	0.00	1.00	0.00	4.34	4.77	5.70
max	88.00	5.00	1.00	9.00	1.00	493.77	491.55	491.05
var	39.60	3.28	0.23	3.43	0.02	2,179.27	2,150.80	2,128.27

Tabela 3: Resumo estatístico das colunas dos dados de Recreation Demand

Observando a correlação entre as variáveis, percebemos que as variáveis de custo

estão altamente correlacionadas. Além disso, temos um indicador forte de que a informação de trips é influenciada por quality e userfree, por terem correlações acima das demais.

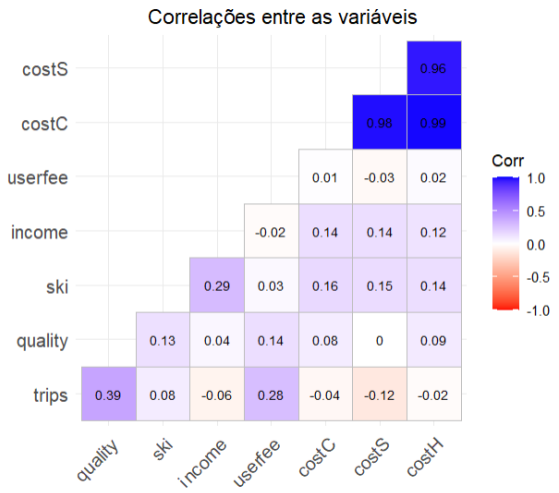

Figura 2: Correlação entre as variáveis do conjunto Recreation Demand

3.3 Resultados

3.4 Modelo de Poisson

Após o processamento e seleção de covariáveis, como mencionado nas metodologias, obtivemos os seguintes resultados:

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.268288   0.093783   2.861  0.00423 **
quality      0.469157   0.016917  27.732  < 2e-16 ***
ski          0.403249   0.055554   7.259 3.91e-13 ***
income      -0.111785   0.019577  -5.710 1.13e-08 ***
userfee      0.900718   0.078877  11.419  < 2e-16 ***
costS       -0.043091   0.001571 -27.427  < 2e-16 ***
costH        0.033632   0.001543  21.790  < 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: 2307.0  on 652  degrees of freedom
AIC: 3074.1

Number of Fisher Scoring iterations: 6
```

Figura 3: Outup do R, modelo de Poisson

Como mencionado, todos os coeficientes apresentam valores de z muito baixos,

indicando forte evidência contra a hipótese nula de que esses coeficientes seriam iguais a zero.

O AIC do modelo final Poisson foi 3074.1. Além disso, o valor da Deviance média residual foi de aproximidamente 3.53, o que aponta a presença de sobredispersão nos dados.

3.5 Teste de Sopredispersão

O modelo anterior testado indica um forte indício de sobredispersão nos dados. Assim, Cameron e Trivedi (1990) propuseram um teste estatístico para detectar a presença de *sobredispersão* em modelos de contagem, como o modelo de Poisson. A ideia é verificar se a variância dos dados é maior do que a esperada sob a suposição de que $\text{Var}(Y) = \mu$.

Para isso, assume-se a seguinte forma alternativa para a variância:

$$\text{Var}(Y) = \mu + \alpha \cdot \mu^2$$

As hipóteses do teste são:

- H_0: $\alpha = 0$ (sem sobredispersão, modelo de Poisson adequado)
- H_1: $\alpha > 0$ (há sobredispersão, o modelo de Poisson subestima a variância)

O teste é realizado em três etapas:

1. **Ajuste do modelo Poisson:** estima-se um modelo GLM com família Poisson e obtêm-se os valores ajustados $\hat{\mu}_i$.
2. **Cálculo da estatística de teste:** para cada observação, calcula-se:

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

3. **Regressão auxiliar:** regressa-se Z_i em $\hat{\mu}_i$ (sem intercepto) e realiza-se um teste t para o coeficiente da regressão. Um coeficiente significativamente diferente de zero indica a presença de sobredispersão.

Se o p-valor do teste t for pequeno (geralmente menor que 0.05), rejeita-se a hipótese nula e conclui-se que o modelo de Poisson não é adequado, sendo recomendável o uso de modelos que acomodem sobredispersão, como o modelo binomial negativo.

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

Residuals:
    Min      1Q  Median      3Q     Max
 -45.64   -2.44   -0.19   -0.11 1412.58

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
mu   1.3125     0.4387   2.992  0.00288 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.71 on 658 degrees of freedom
Multiple R-squared:  0.01342,   Adjusted R-squared:  0.01192
F-statistic:  8.95 on 1 and 658 DF,  p-value: 0.002879
```

Figura 4: Teste de Sobredispersão no modelo Poisson

Observando os resultados da imagem acima, vemos que o coeficiente estimado para μ é 1.3125, com um valor-p significativo (0.00288), indicando que há uma associação estatisticamente significativa entre a média prevista e a variabilidade dos dados.

No entanto, o valor do coeficiente maior que 1 (em vez de próximo de 1) sugere que a variabilidade dos dados é maior do que o esperado pela suposição inicial do modelom ou seja, há sobredispersão. Isso significa que a variância dos dados é maior do que a média esperada, o que viola a suposição padrão do modelo.

3.6 Modelo Binomial Negativa

A distribuição de Poisson assume que a média e a variância dos dados são iguais, uma propriedade chamada equidispersão. Porém, em muitos conjuntos de dados reais, especialmente em contagens, a variância é maior que a média, situação conhecida como sobredispersão. Quando aplicamos um modelo de Poisson nesses casos, ele tende a subestimar a variabilidade dos dados.

A Binomial Negativa, por outro lado, é uma generalização da Poisson que inclui um parâmetro extra para modelar a variância de forma mais flexível. Ela permite que a variância seja maior que a média, capturando assim a sobredispersão. Esse parâmetro adicional ajusta a dispersão dos dados, tornando o modelo mais robusto e adequado para situações onde a variância não é igual à média, melhorando a precisão das estimativas e a qualidade do ajuste.

Nesse caso, a função de ligação log foi escolhida para o modelo Binomial Negativa porque assegura que as previsões da média da variável resposta sejam sempre positivas, condição essencial para dados de contagem. O link log proporciona previsões coerentes e uma interpretação intuitiva dos coeficientes em termos de efeitos multiplicativos sobre a média esperada. Essa característica torna o link log a opção mais adequada para modelar dados de contagem e capturar sua variabilidade de forma consistente.

Matematicamente, o modelo assume a forma:

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

onde μ representa a média esperada da variável resposta e β_i são os coeficientes associados às variáveis explicativas x_i.

Obtivemos os seguintes resultados após o processo descrito nas metodologias:

```
Call:
glm.nb(formula = trips ~ quality + ski + costC + costS + costH,
    data = df, link = log, init.theta = 0.7125577474)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.194127   0.165300  -7.224 5.05e-13 ***
quality      0.728901   0.040340  18.069  < 2e-16 ***
ski          0.597068   0.148124   4.031 5.56e-05 ***
costC        0.052004   0.009223   5.638 1.72e-08 ***
costS       -0.096924   0.006560 -14.774  < 2e-16 ***
costH        0.038998   0.007762   5.024 5.06e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.7126) family taken to be 1)

    Null deviance: 1227.99  on 658  degrees of freedom
Residual deviance:  424.23  on 653  degrees of freedom
AIC: 1669.3

Number of Fisher Scoring iterations: 1


              Theta:  0.7126
          Std. Err.:  0.0725

 2 x log-likelihood:  -1655.3420
```

Figura 5: Output do R, modelo Binomial Negativa

Como mencionado, todos os coeficientes apresentam valores de z muito baixos, indicando forte evidência contra a hipótese nula de que esses coeficientes seriam iguais a zero.

O AIC do modelo final foi 1669.3, menor que o de Poisson. Além disso, o parâmetro de dispersão foi estimado em aproximadamente 0.713, valor inferior ao infinito esperado para o modelo de Poisson, o que confirma a presença de sobredispersão nos dados. Esse ajuste pela binomial negativa permite modelar melhor a variabilidade observada, característica que não foi capturada pelo modelo de Poisson.

Por fim, observamos que a deviance residual média (0.649) do modelo foi bastante reduzida em relação ao modelo de Poisson, indicando a melhora do impacto da sobredispersão.

3.7 Zero-Inflated Poisson (ZIP)

Assim como observado na Figura 1, há uma quantidade excessiva de zeros na variável resposta, superior ao que seria esperado em modelos tradicionais de contagem, como o Poisson ou a Binomial Negativa. Diante desse padrão, é apropriado

considerar o uso de um modelo com inflação de zeros (ZIP – Zero-Inflated Poisson), que permite distinguir entre diferentes tipos de zeros presentes nos dados. Esses zeros podem ser classificados da seguinte forma:

- **Zeros reais:** são os zeros que aparecem naturalmente em um conjunto de dados devido ao processo que está sendo medido ou observado. Eles são zeros "verdadeiros", ou seja, fazem parte da distribuição natural dos dados.
- **Zeros inflados (ou excessivos):** aparecem quando o número de zeros observados é maior do que o esperado pelo modelo estatístico que você usaria normalmente para modelar os dados, como uma distribuição de Poisson ou Binomial Negativa. Esses zeros extras podem surgir por algum processo diferente do que gera os demais valores, e não apenas como parte da distribuição comum do fenômeno.

O modelo ZIP (Zero-inflated model) visa conhecer essa estrutura de zeros construindo a seguinte distribuição:

$$f_{\text{zeroinfl}}(y) = \pi \cdot I_{\{0\}}(y) + (1 - \pi) \cdot f_{\text{count}}(y; \mu_i)$$

onde agora μ_i e π_i são modelados como funções das covariáveis disponíveis. Utilizando o link canônico para a parte de contagem, temos $\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, enquanto para a parte binária adotou-se $g(\pi_i) = \mathbf{z}_i^\top \boldsymbol{\gamma}$, com g sendo a função logit, que é o link canônico da família binomial (Bernoulli neste caso particular). A escolha do logit deve-se à sua interpretação direta em termos de odds e à sua maior sensibilidade a variações nos dados em comparação ao link probit, o que pode torná-lo mais informativo em contextos com covariáveis bem especificadas. Ainda assim, o link probit permanece uma alternativa comum na literatura.

Nesse modelo, temos que selecionar dois conjuntos de covariáveis, uma para a parte do modelo de contagem natural e outro conjunto que acreditamos influenciar na inflação de zeros do modelo. Vale ressaltar que esse conjunto de covariáveis pode ser diferente e não contido um no outro.

Neste estudo, diferentemente do algortimo stepwise, incluímos inicialmente todas as covariáveis utilizadas no modelo de Poisson anterior para ambos os conjuntos. Em seguida, excluímos aquelas consideradas estatisticamente insignificantes com base na estatística z. As covariáveis finais selecionadas foram:

- **Modelo Poisson (Contagem):** ski, income, userfee, costS, costH
- **Modelo Binário (Veridito do Tipo de zero):** quality, costS, costH

Obtemos os seguintes resultados:

```
Call:
zeroinfl(formula = trips ~ ski + income + userfee + costS + costH | quality + costS + costH, data = df,
    dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-6.1181 -0.3570 -0.1507 -0.1127 14.6027

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.224343   0.084593  26.295  < 2e-16 ***
ski          0.487047   0.056564   8.611  < 2e-16 ***
income      -0.090751   0.019997  -4.538 5.67e-06 ***
userfee      0.598737   0.079113   7.568 3.79e-14 ***
costS       -0.036248   0.001896 -19.119  < 2e-16 ***
costH        0.025649   0.001912  13.417  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.34394    0.39068   8.559  < 2e-16 ***
quality     -1.69827    0.15272 -11.120  < 2e-16 ***
costS        0.07223    0.01487   4.856 1.20e-06 ***
costH       -0.08137    0.01525  -5.335 9.58e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 19
Log-likelihood: -1167 on 10 Df
```

Figura 6: Output do R, modelo ZIP

3.8 Comparação entre modelos e particularidades

Para comparar os modelos, podemos levar em consideração algumas métricas.

Como de praxe, observamos o AIC obtido em cada resultado particular na seguinte tabela.

Perceba que o modelo de melhor AIC possui uma quantidade menor de parâmetros se comparado com ZIP. No entanto, vamos levar em consideração mais características para selecionar o veredito de escolha de modelo final.

AIC nos Modelos		
Modelo	Número de Parâmetros	AIC
Poisson	7	3,074.06
Binomial Negativa	7	1,669.34
ZIP	10	2,353.71

Tabela 4: Comparação de AIC dos modelos

A seguir, usamos outra métrica com a tabela abaixo, a linha "Observado"apresenta a frequência real com que cada valor da variável trips, de 0 a 9, aparece no conjunto de dados.

Assim, buscamos explicar o padrão de demanda por atividades recreativas, medido em número de viagens realizadas pelos indivíduos. O objetivo é compreender como essa variável se distribui na população analisada e identificar possíveis excessos

de zeros, variação excessiva ou outros desvios em relação ao que seria esperado por distribuições teóricas.

As demais linhas da tabela mostram as frequências esperadas para esses mesmos valores de trips, calculadas a partir dos diferentes modelos de contagem. Assim, com base em parâmetros ajustados aos dados, foi calculado quantas ocorrências de cada valor seriam previstas por suas respectivas distribuições teóricas. Desta maneira, a comparação entre essas linhas e a linha observada permite avaliar o quão bem cada modelo consegue reproduzir o padrão de contagens presente nos dados reais no geral.

Frequências Observadas e Esperadas										
Desfecho	0	1	2	3	4	5	6	7	8	9
Observado	417	68	38	34	17	13	11	2	8	1
Poisson	275	146	68	41	30	23	17	13	10	7
Binomial_Negativa	423	81	33	20	14	11	9	7	6	5
ZIP	419	22	28	29	28	25	22	18	15	11

Tabela 5: Comparação de frequências observadas e esperadas dos modelos

A seguir, o seguinte gráfico apresenta a relação entre os valores observados da variável trips e os valores previstos pelos modelos (na melhor das estimativas). Cada ponto representa uma amostra, mostrando como o modelo estimou o número de viagens em comparação com o valor real.

A linha tracejada $y = x$ indica a previsão perfeita, onde o valor previsto é igual ao observado. Pontos próximos a essa linha refletem boas previsões, enquanto pontos afastados indicam erros.

Os painéis separados mostram o desempenho dos modelos, facilitando a comparação entre eles. Assim, podemos avaliar se algum modelo tende a superestimar ou subestimar os dados e analisar a dispersão das previsões.

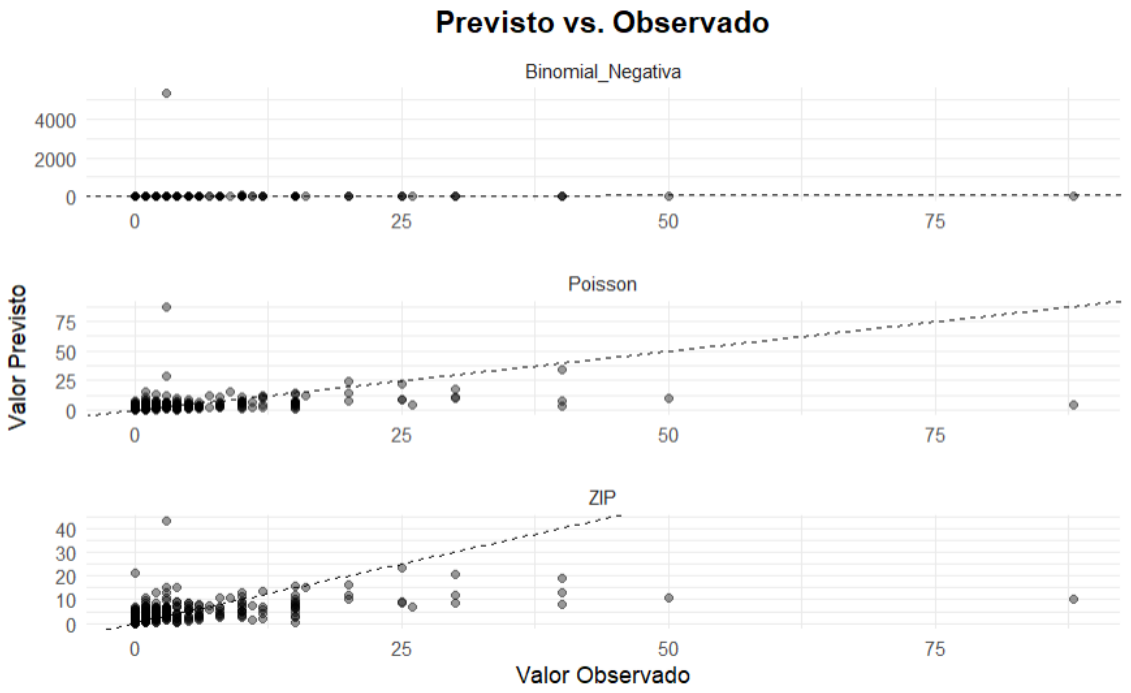

Figura 7: Comparação de Previsto vs. Observado (sem limitação de Outliers)

Em seguinta ampliamos o gráfico no mesmo range y limitando até o valor 80, para outliers não atraplharem na comparação.

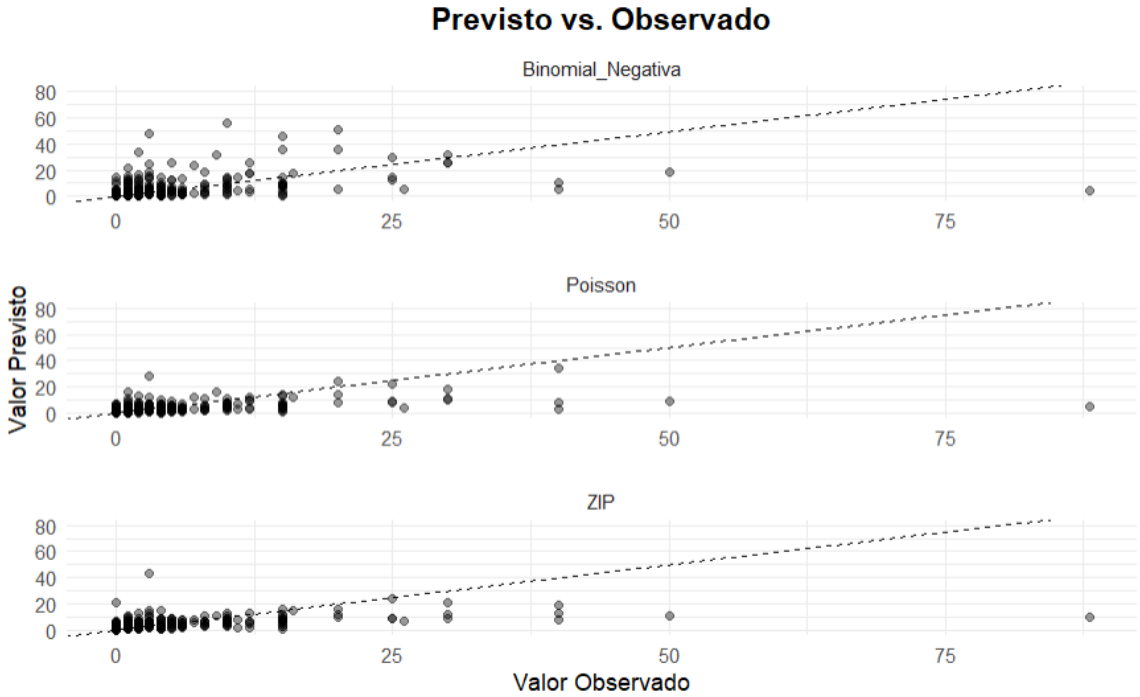

Figura 8: Comparação de Previsto vs. Observado (na mesma escala y)

Em seguida, foi plotada a distribuição dos resíduos de cada modelo. Quanto mais concentrados os resíduos estiverem em torno de zero, melhor o ajuste do modelo, pois isso indica que as previsões estão, em média, próximas dos valores observados.

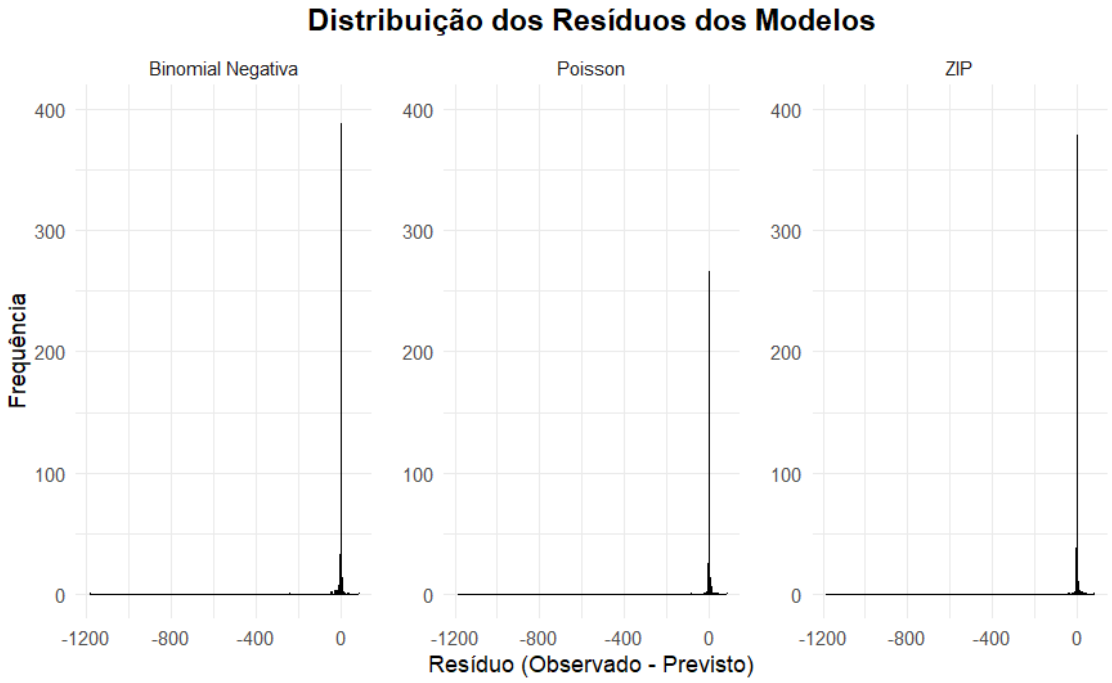

Figura 9: Distribuição dos Resíduos

Ampliamos com um zoom para verificar de melhor maneira sem os outliers.

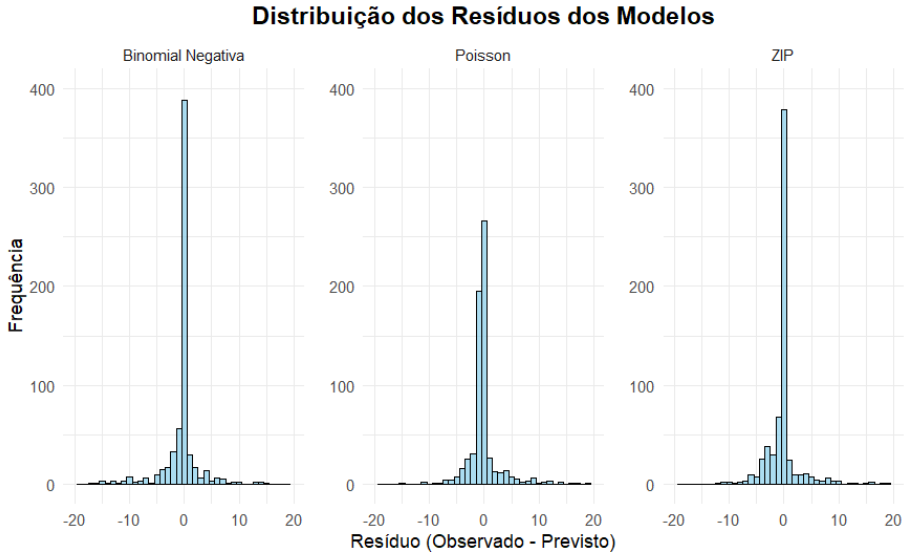

Figura 10: Distribuição dos Resíduos (Limitação do intervalo de resíduos)

4 Discussão e Conclusão

Como discutido na seção anterior sobre a comparação entre modelos, analisamos o AIC, a distribuição das previsões, os valores extremos e os resíduos gerados por cada modelo. De forma geral, o modelo ZIP apresentou um desempenho mais balanceado.

Embora a Tabela 5 sugira, à primeira vista, que o modelo de Binomial Negativa seja o mais adequado, ela se limita a apresentar a frequência dos desfechos previstos, sem refletir diretamente a acurácia do modelo. Medidas como os resíduos fornecem uma avaliação mais precisa da variabilidade explicada pelos modelos. Ainda assim, é possível observar que tanto o modelo de Binomial Negativa quanto o ZIP capturam melhor a característica de inflação de zeros em comparação ao modelo de Poisson.

Apesar da presença de alguns outliers nas previsões do modelo ZIP (Figura 7), essa ocorrência é justificável, já que todos os modelos apresentaram dificuldades semelhantes nesses mesmos pontos. Além disso, o modelo ZIP apresentou menor variabilidade nas previsões em comparação à Binomial Negativa (Figura 8). Para valores de "trips"acima de 25, todos os modelos mostraram maior dispersão, o que é compreensível dado o pequeno número de observações com valores tão elevados, conforme evidenciado também na Figura 8.

O modelo ZIP apresentou um AIC intermediário entre os modelos avaliados, mas conseguiu capturar de forma eficaz os valores médios, conforme mostrado na Tabela 4. É importante ressaltar que um AIC mais alto não implica necessariamente em pior desempenho em todos os aspectos, já que essa métrica avalia o equilíbrio entre ajuste e complexidade do modelo. Além disso, outras métricas ou critérios substantivos podem justificar a escolha de um modelo com AIC intermediário, especialmente se ele representar melhor os dados observados.

Ao analisarmos a distribuição dos resíduos, observamos uma forte concentração de valores próximos de zero, o que indica um bom ajuste geral (Figura 10). Essa característica também é observada no modelo de Binomial Negativa, embora com diferenças sutis.

Por fim, é importante destacar que a implementação do modelo ZIP, em comparação ao modelo de Poisson, trouxe melhorias significativas na modelagem dos zeros, conforme evidenciado nas comparações realizadas.

Como sugestão para aprofundamentos futuros, seria interessante explorar uma versão de Binomial Negativa com ajuste para excesso de zeros (zero-inflated), o que pode melhorar ainda mais a capacidade preditiva do modelo em contextos semelhantes.

5 Referências

[1] Gelman, A.; Hill, J.; Vehtari, A. *Regression and Other Stories*. Cambridge: Cambridge University Press, 2020. Acesso em: maio de 2025.

[2] Kleiber, C.; Zeileis, A. *RecreationDemand: Recreation Demand Data in AER*. Disponível em: `https://rdrr.io/cran/AER/man/RecreationDemand.html`. Acesso em: maio de 2025.

[3] Kleiber, C.; Zeileis, A. *Recreation Demand Data – R Documentation*. Disponível em: `https://search.r-project.org/CRAN/refmans/AER/html/RecreationDemand.html`. Acesso em: maio de 2025.

[4] Cameron, A. C.; Trivedi, P. K. Regression-Based Tests for Overdispersion in the Poisson Model. *Journal of Econometrics*, v. 46, n. 3, p. 347–364, 1990. Acesso em: maio de 2025.