Identificação e Alternativas para Sobredispersão em Modelos de Contagem

Gustavo Bianchi da Silva

Fundação Getúlio Vargas
Curso - Modelagem Estatística
5° período de Ciência de Dados e Inteligência Artificial

Rio de Janeiro
2025

Sumário

1 Introdução

Diante de problemas de regressão com dados de contagem, deve-se procurar alternativas ao modelo clássico de regressão linear com erros normalmente distribuídos. Problemas de contagem são aqueles que contam a frequência de eventos, de modo que o suporte da variável resposta deve estar restrito aos números inteiros não-negativos.

Uma abordagem intuitiva é a adoção de um modelo que segue a distribuição Poisson, a qual possui propriedades desejáveis para esse tipo de variável. A distribuição de Poisson é definida para uma variável aleatória $X \sim \text{Poisson}(\lambda)$, com $\lambda > 0$, onde a esperança e a variância são iguais:

$$E[X] = \lambda \quad \text{e} \quad \text{Var}(X) = \lambda \tag{1}$$

Com base na suposição de que os dados seguem essa distribuição, espera-se que, empiricamente, a média e a variância amostral sejam próximas. No entanto, em dados reais, raramente essa propriedade é observada. Quando a variância observada excede significativamente a média, temos o fenômeno conhecido como sobredispersão. Por outro lado, embora mais raro, também é possível encontrar subdispersão, quando a variância é menor que a média.

A presença de sobredispersão ou subdispersão indica que a distribuição de Poisson pode não ser adequada, o que compromete a qualidade da inferência estatística e o poder preditivo do modelo. Nesses casos, torna-se necessário recorrer a modelos alternativos, como o modelo binomial negativo ou modelos de Poisson com inflação de zeros.

No decorrer do trabalho, será ajustado um modelo Poisson em um conjunto de dados de contagem que possui uma grande quantidade de zeros estruturais - substituição por falta de dados. Neste contexto, a situação da sobredispersão e o ajuste do modelo são prejudicados. Diante da situação, busca-se, com a utilização de bibliotecas e funções do R, entender soluções para o problema da sobredispersão e entender possibilidades de incorporação das informações dos zeros estruturais no modelo, melhorando assim a predição.

2 Métodos

2.1 A Base de Dados

A base de dados analisada faz parte do conjunto `RecreationDemand` do pacote `AER` em R. Os dados referem-se ao número de viagens recreativas de barco para o Lago Somerville, Texas, em 1980, coletados através de pesquisa com 2.000 proprietários de barcos de lazer registrados em 23 condados do leste do Texas. A variável resposta de interesse é *trips*, que representa a quantidade de viagens recreativas.

Tabela 1: Descrição das Variáveis do Conjunto de Dados

Variável	Descrição
trips	Número de viagens recreativas de barco
quality	Classificação subjetiva da qualidade (0=não visitado, 1-5)
ski	Indicador de prática de esqui aquático (0=Não, 1=Sim)
income	Renda anual do domicílio (em milhares de USD)
userfee	Pagamento de taxa de usuário (0=Não, 1=Sim)
costC	Custo de visita ao Lago Conroe (em USD)
costS	Custo de visita ao Lago Somerville (em USD)
costH	Custo de visita ao Lago Houston (em USD)

O conjunto de dados apresenta zeros estruturais na variável *quality*, que assume valor 0 quando o lago não foi visitado (além da escala ordinal 1-5 para avaliação de qualidade). Intuitivamente, podem-se criar suspeitas de relações entre o estrutural com a variável resposta, já que $quality = 0$ pode ser um indício de poucas viagens. Esta hipótese será melhor visualizada nos resultados dos modelos.

Antes de proceder com a modelagem, é essencial entender os dados e avaliar a relevância estatística de cada variável explicativa para a análise, considerando:

- Medidas da variável objetivo
- Correlações entre preditores
- Relação das covariáveis com a variável resposta
- Balanceamento das variáveis categóricas

2.2 Análise Exploratória

Como já esperado pelo tema do estudo, a variável objetivo *trips* possui uma grande quantidade de zeros, o que implica numa discrepância entre a média e a variância amostrais. A distribuição pode ser melhor analisada na imagem 1.

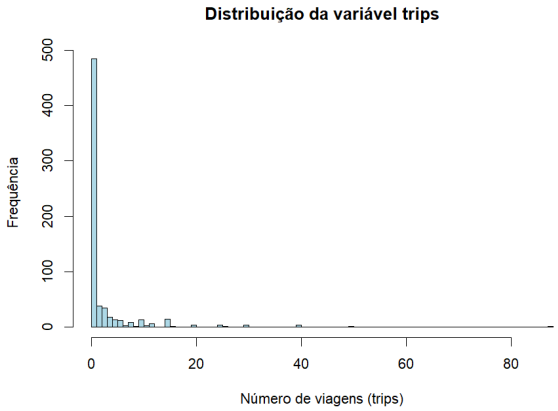


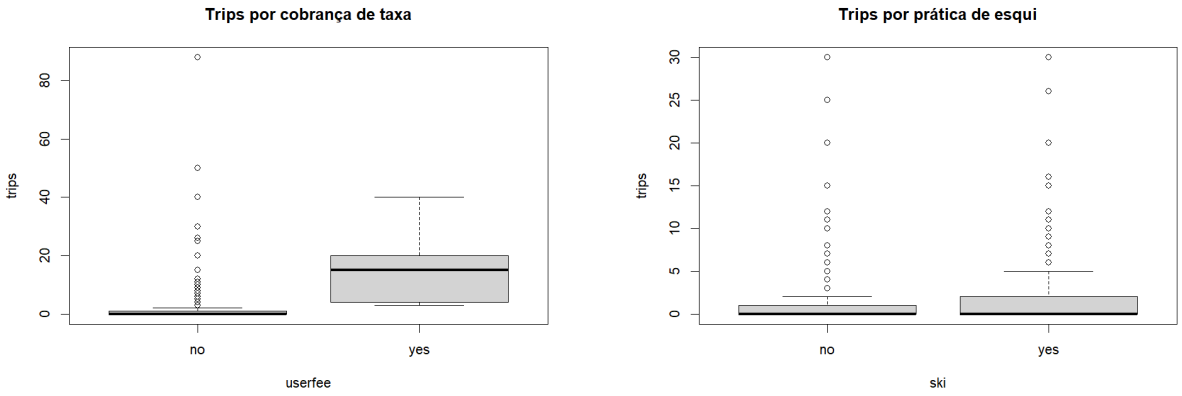
Figura 1: Enter Caption

- Média: 2.24
- Variância: 39.60
- Excessivo de zeros (417 casos = 0 vs 242 > 0)

Embora evidenciada a sobredispersão, seguiremos com a implementação do modelo Poisson para posteriormente lidarmos com o rigor e testes necessários. O segundo passo da análise é visualizar como se comporta a variável *trips* com relação às variáveis independentes.

2.2.1 Análise Gráfica

Começando pelas variáveis binárias *ski* e *userfee*, temos um destaque importante: *userfee* é quase sempre 0. Há apenas 13 registros de pagamento da taxa (2% apenas). Embora isso seja alarmante, podemos ver pelo boxplot e por intuição que o pagamento da taxa pode estar associado à ocorrência de pelo menos uma viagem, então manteremos a variável no modelo.

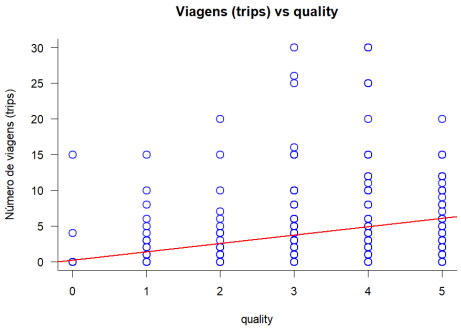


(a) Relação entre *trips* e *userfee*

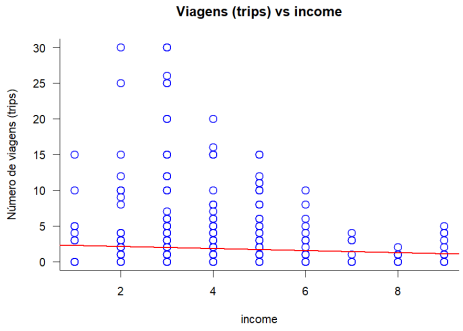
(b) Relação entre *trips* e *ski*

Figura 2: Distribuição de viagens por variáveis binárias

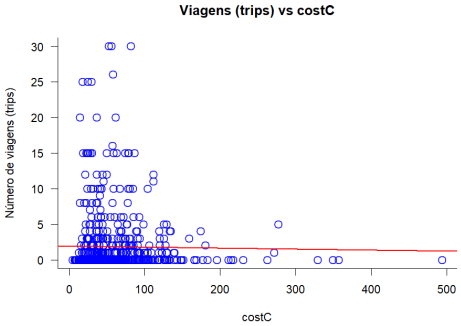
Para os gráficos das outras variáveis, por serem scatterplots, decidiu-se omitir valores altos de *trips* (os 5 maiores, com valores acima de 30) para melhor visualização das correlações.



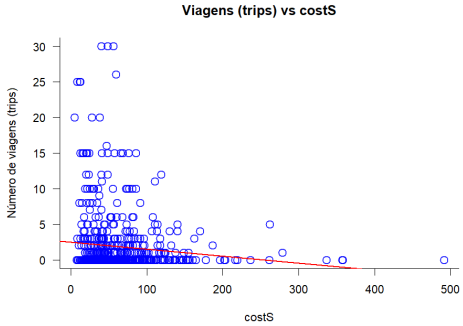
(a) Relação entre trips e Quality



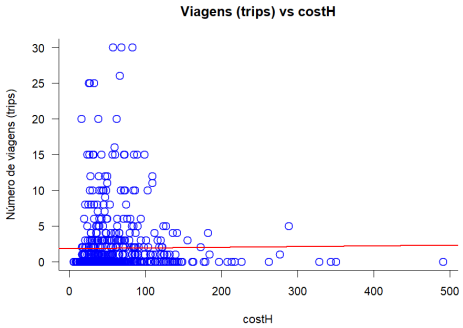
(b) Relação entre trips e Income



(c) Relação entre trips e costC



(d) Relação entre trips e costS



(e) Relação entre trips e costH

Figura 3: Análise de correlações entre trips e outras variáveis

2.2.2 Análise das Variáveis de Custo

As variáveis *costC*, *costH* e *costS* exibem padrões gráficos muito similares, como era esperado dado que representam custos de visita a lagos que são substitutos próximos geograficamente. Essa suspeita de redundância foi confirmada através da matriz de correlação de Pearson:

Tabela 2: Matriz de correlação entre variáveis de custo

	costH	**costC**	**costS**
costH	1.000	0.986	0.965
costC	0.986	1.000	0.977
costS	0.965	0.977	1.000

Os valores da matriz variam entre -1 e 1 e valores extremos indicam que há um padrão linear entre as variáveis. Como todos os valores estão próximos de 1, isso indica que a informação de apenas uma das variáveis contém poder descritivo suficiente para explicar as outras duas. Essa redundância informacional justifica a exclusão de *costC* e *costH* do modelo final, mantendo apenas *costS* (Lago Somerville) como representante do construto "custo de visita". Essa decisão preserva o poder preditivo, reduz a complexidade do modelo e elimina problemas de estimação causados pela multicolinearidade.

2.3 Ajuste do Modelo Poisson

Sob a hipótese de que os dados seguem uma distribuição Poisson ($Y \sim \text{Poisson}(\theta)$), que pertence à família exponencial, podemos utilizar Modelos Lineares Generalizados (GLM) para estimar os parâmetros. A função de densidade de probabilidade é dada por:

$$P(Y = y) = \frac{e^{-\theta}\theta^{y}}{y!}, \quad y = 0, 1, 2, ... \tag{2}$$

2.3.1 Forma Canônica da Família Exponencial

A distribuição Poisson pode ser escrita na forma canônica da família exponencial:

$$f(y|\theta) = \exp\{y \log\theta - \theta - \log(y!)\} \tag{3}$$

Onde identificamos:

- Parâmetro natural: $\eta = \log\theta$
- Função de ligação canônica: $g(\theta) = \log(\theta)$ (ligação logarítmica)

2.3.2 Estimação por Máxima Verossimilhança e Fisher Scoring

O modelo de regressão Poisson com ligação logarítmica é especificado como:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{quality}_i + \beta_2 \text{ski}_i + \beta_3 \text{userfee}_i + \beta_4 \text{income}_i + \beta_5 \text{costS}_i, \tag{4}$$

onde $\mu_i = E(Y_i \mid x_i)$ representa a média condicional da variável resposta dado o vetor de covariáveis x_i. A estimação dos parâmetros β é realizada por máxima verossimilhança, buscando maximizar a função log-verossimilhança dada por

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i (x_i^\top \beta) - \exp(x_i^\top \beta) - \log(y_i!) \right]. \tag{5}$$

Diferentemente do modelo linear normal, cujo estimador de máxima verossimilhança coincide com o estimador de mínimos quadrados e possui solução fechada, o modelo Poisson exige métodos iterativos para encontrar o vetor $\hat{\beta}$ que maximiza $\ell(\beta)$. Entre as técnicas numéricas disponíveis, destacam-se o método de Newton-Raphson e o método de Fisher Scoring.

O método de Fisher Scoring é uma variante do Newton-Raphson que utiliza a informação esperada da matriz Hessiana (informação de Fisher) ao invés da matriz Hessiana observada. Essa abordagem tende a ser mais estável e garantir melhor convergência, especialmente em modelos de regressão generalizada como o Poisson, onde a informação de Fisher é mais simples de calcular e positiva definida. Além disso, o Fisher Scoring reduz a sensibilidade a irregularidades na superfície de verossimilhança, tornando-o mais eficiente em termos computacionais e numéricos.

A atualização dos parâmetros na iteração $t+1$ do algoritmo de Fisher Scoring é dada por:

$$\beta^{(t+1)} = \beta^{(t)} + \left[\mathcal{I}(\beta^{(t)}) \right]^{-1} U(\beta^{(t)}), \tag{6}$$

com:

$$U = \left. \frac{\partial \ell(\beta)}{\partial \beta} \right|_{\beta=\beta^{(t)}} \quad \text{(vetor score)}$$

$$\mathcal{I} = -E \left[\left. \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} \right|_{\beta=\beta^{(t)}} \right] \quad \text{(matriz de informação)}$$

Esse procedimento itera até que $\|\beta^{(t+1)} - \beta^{(t)}\| < \epsilon$, para ϵ pré-definido.

2.3.3 Resultados do Ajuste no R

O modelo foi ajustado utilizando a função `glm()` do R, que, neste caso, convergiu após 7 iterações do método numérico Fisher Scoring. A partir da função `summary()`, é possível visualizar as estimativas dos coeficientes, seus erros padrões e as estatísticas de teste baseadas na estatística Z — que avalia a significância individual dos coeficientes (similar ao teste t na regressão linear). Através do output do R, devemos procurar saber se todos os coeficientes possuem altas probabilidades de não serem nulos, sendo assim significativos.

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

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

Figura 4: Resultados do ajuste do modelo Poisson com variáveis quality, ski, income e costS

Os valores-p apresentados indicam a probabilidade de observarmos um coeficiente tão extremo sob a hipótese nula de que o coeficiente seja zero, ou seja, que a variável não tenha efeito significativo. Valores-p muito baixos, como os encontrados, indicam que é improvável que o coeficiente seja nulo, confirmando a relevância das variáveis no modelo.

Além da significância dos coeficientes, o resumo do modelo traz informações sobre seu ajuste geral. A comparação entre a null deviance (4849,7) e a residual deviance (2687,5) revela que o modelo com as variáveis explicativas melhora consideravelmente o ajuste em relação ao modelo que contém apenas o intercepto. Essa redução substancial na deviance indica que o modelo capta boa parte da variabilidade dos dados, embora não seja uma medida perfeita como o R^2 na regressão linear.

2.3.4 Análise dos Resíduos

A análise gráfica dos resíduos deviance revelou um padrão que se assemelha a uma curva, evidenciando o que já se suspeitava por conta da sobredispersão - a inadequação do modelo Poisson. Observa-se que, embora os dados possuam muitos zeros na variável resposta *trips*, o modelo não conseguiu prever corretamente essas ocorrências, o que se manifesta por resíduos negativos acentuados para observações com poucos *trips* (os dados estão ordenados em ordem crescente).

Nota-se que os resíduos apresentam um padrão, como se formassem uma curva e variação dos resíduos também aumenta conforme trips aumenta, indicando heterocedasti-

cidade. Esse padrão nos resíduos sugere que o modelo não ajusta bem algumas partes da distribuição dos dados. Outra grnade falha do modelo foi a não ingestão da informação dos zeros estruturais para prever os zeros de *trips*: se usarmos os coeficientes encontrados e utilizarmos o modelo nos dados, vamos prever um total de 261 zeros, o que diverge muito dos 417 zeros do dataset.

Podemos inferir que o modelo sofreu falhas de ajuste devido a problemas como sobredispersão — situação em que a variância observada excede a média, violando a suposição básica do modelo Poisson.

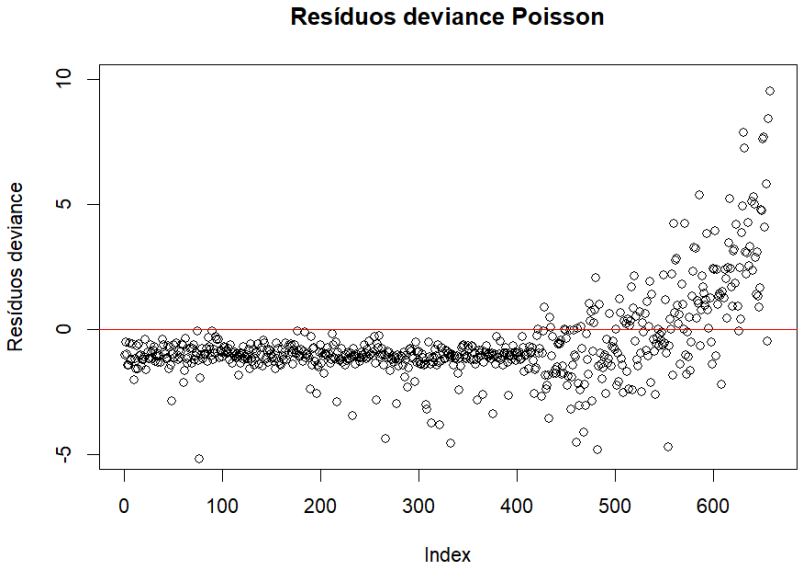


Figura 5: Gráfico dos resíduos deviance do modelo Poisson, ordenados pela variável resposta *trips*.

2.4 Teste de Sobredispersão

Para verificar formalmente a presença de sobredispersão nos dados, aplicamos o teste proposto por Cameron I& Trivedi (1990), que parametriza a variância como:

$$\text{Var}(Y) = \mu + \alpha \cdot \mu^2 \tag{7}$$

As hipóteses do teste são:

$$H_0 : \alpha = 0 \quad \text{(ausência de sobredispersão)}$$
$$H_1 : \alpha \neq 0 \quad \text{(presença de sobredispersão)}$$

2.4.1 Metodologia do Teste

O procedimento consiste em três etapas:

1. Ajustar um modelo Poisson e obter os valores preditos $\hat{\mu}_i$

2. Calcular a estatística de teste:

$$Z_i = \frac{(Y_i - \hat{\mu}_i)^2 - Y_i}{\hat{\mu}_i} \tag{8}$$

3. Realizar uma regressão linear de Z_i em $\hat{\mu}_i$ (sem intercepto) e testar a significância do coeficiente

2.4.2 Resultados do Teste

A aplicação do teste de sobredispersão de Cameron & Trivedi no R gerou os resultados ilustrados na Figura 6.

```
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
```

Figura 6: Resultados do teste de sobredispersão de Cameron & Trivedi

2.4.3 Interpretação

Os resultados fornecem evidências robustas de sobredispersão nos dados, conforme indicado por:

- O coeficiente estimado $\hat{\alpha} = 1.6406$ com erro padrão 0.7664 é significativamente diferente de zero ($p = 0.0327$), rejeitando a hipótese nula H_0 de ausência de sobredispersão ao nível de 5% de significância.

- A positividade do coeficiente sugere que a variância aumenta mais rápido que a média, confirmando a presença de variância extra que o modelo Poisson não captura.

Esta conclusão corrobora nossa análise exploratória inicial, onde observamos variância amostral (39.60) muito superior à média amostral (2.24). Com evidência estatística de sobredisperção nos dados, deve-se substituir o modelo Poisson, que apresenta problemas no caso, e utilizar alternativas mais robustas para os dados.

2.5 Modelos Alternativos

2.5.1 Binomial Negativa

A Binomial Negativa é a principal alternativa ao modelo Poisson para dados de contagem que apresentam sobredispersão. Diferentemente da Poisson, cuja variância é igual à média, a Binomial Negativa permite que a variância seja maior que a média, acomodando a sobredispersão observada nos dados. Para facilitar a modelagem em regressão, a distribuição é reparametrizada em termos do valor esperado μ e do parâmetro de dispersão θ, que é inversamente relacionado ao parâmetro tradicional r da Binomial Negativa clássica.

A parametrização clássica da Binomial Negativa define a variância como

$$\mathrm{Var}(Y) = \frac{pr}{(1-p)^2} \tag{9}$$

onde $r > 0$ e $p \in (0, 1)$ são os parâmetros originais da distribuição. Para a aplicação em modelos lineares generalizados, usa-se a reparametrização via média μ e dispersão θ, onde $\mu = r\frac{1-p}{p}$ e $\theta = r$. Assim, a variância pode ser expressa como:

$$\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\theta} = \mu + \alpha\mu^2, \tag{10}$$

onde $\alpha = \frac{1}{\theta} > 0$ é o parâmetro de dispersão. Quando $\alpha \to 0$ (ou $\theta \to \infty$), o modelo Binomial Negativa se reduz ao modelo Poisson clássico, em que a variância iguala-se à média.

Especificação do Modelo

Mantém-se a função de ligação logarítmica, igual à do modelo Poisson:

$$\log(\mu_i) = \beta_0 + \beta_1 \mathrm{quality}_i + \cdots + \beta_p \mathrm{costS}_i \tag{11}$$

A função de verossimilhança do modelo incorpora o parâmetro extra θ (ou $\alpha = 1/\theta$) por meio da seguinte distribuição de probabilidade:

$$P(Y = y) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\,\Gamma(y+1)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{y} \tag{12}$$

onde $y = 0, 1, 2, \ldots$, $\mu > 0$ é a média condicionada às covariáveis, e $\theta > 0$ é o parâmetro de dispersão que controla a sobredispersão do modelo.

Implementação e Resultados

O ajuste do modelo foi realizado no R utilizando a função `glm.nb()` do pacote `MASS`. Na Figura 7 apresenta-se o resumo dos resultados obtidos com esse ajuste.

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

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

Figura 7: Resultados do ajuste do modelo Binomial Negativo

A primeira observação importante é que os coeficientes estimados pelo modelo binomial negativo diferem bastante daqueles obtidos no modelo Poisson. Diferentemente do modelo Poisson, aqui a variável `income` não foi considerada estatisticamente significativa, indicando uma provável nulidade do parâmetro associado. Essa diferença não é motivo de preocupação, pois a principal vantagem do modelo binomial negativo é justamente lidar melhor com a sobredispersão presente nos dados, proporcionando estimativas mais realistas e coerentes com o comportamento observado.

Além disso, as métricas de ajuste confirmam a superioridade do modelo binomial negativo. O valor do AIC foi reduzido quase pela metade, o que indica um ajuste muito melhor. Também houve uma queda expressiva nas deviance — a deviance nula caiu de 956,34 para 453,89 uma grande redução em comparação com o modelo Poisson.

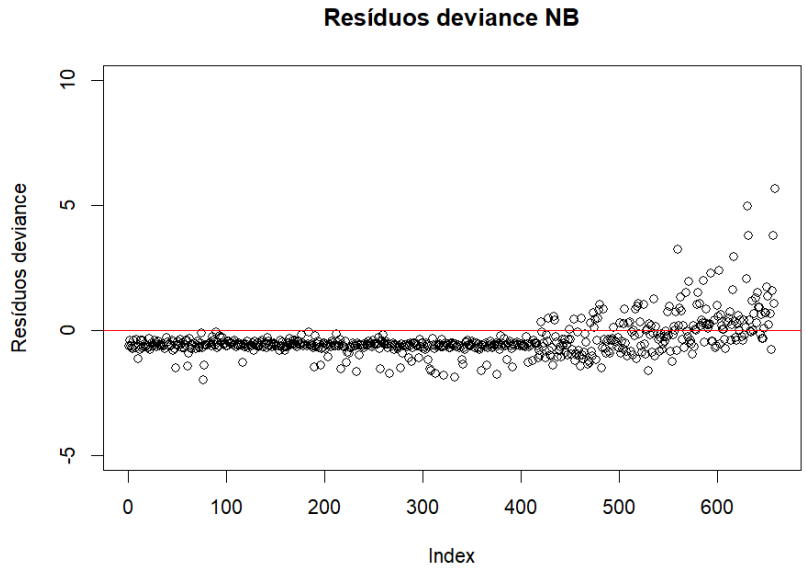


Figura 8: Distribuição dos resíduos do modelo binomial negativo

Ao analisarmos os resíduos (Figura 8), percebe-se que a variância está significativamente mais controlada em comparação ao modelo Poisson e em geral os resíduos estão mais próximos de zero, porém nosso modelo ainda não conseguiu identificar que a grande maioria dos nossos dados são de fato nulos.

2.5.2 Excesso de Zeros – Hurdle Model

Os dados apresentam uma quantidade excessiva de zeros, o que, em contextos de variáveis de contagem, resulta em sobredispersão e dificulta o bom ajuste dos modelos tradicionais como Poisson. Para lidar melhor com os zeros, outra alternativa é utiliazar um modelo inflado de zeros.

Uma maneira de lidar com essa característica é dividir o problema em duas etapas. Primeiramente, modela-se a chance de uma observação ser zero por meio de uma regressão logística (modelo binário). Para as observações cuja contagem prevista é maior que zero, utiliza-se um modelo de contagem truncado em zero — ou seja, essa segunda parte do modelo não permite gerar novos zeros. A segunda parte pode ser poisson ou binomial negativa.

A regressão logística (primeira parte do modelo) é definida como:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = x^\top \beta$$

onde $\pi(x)$ representa a probabilidade de uma contagem ser positiva, x é o vetor de covariáveis e γ é o vetor de coeficientes da regressão logística.

Para as observações com contagem estritamente positiva ($y > 0$), utilizamos um modelo Poisson truncado em zero, embora também seja possível utilizar uma Binomial Negativa truncada. A função de probabilidade do modelo truncado é ajustada da seguinte forma:

$$f^+(y;\lambda) = \frac{f(y;\lambda)}{1 - f(0;\lambda)} \quad \text{para } y > 0$$

onde $f(y;\lambda)$ é a função de probabilidade da distribuição original (Poisson ou NB), e $f^+(y;\lambda)$ representa a distribuição truncada em zero. A truncação assegura que apenas valores positivos sejam gerados nesta segunda etapa do modelo.

Assim, o modelo Hurdle trata os zeros de forma separada dos valores positivos, o que proporciona maior flexibilidade e melhora a qualidade do ajuste em contextos com excesso de zeros.

Implementação e Resultados

O modelo Hurdle foi ajustado no software R utilizando a função `hurdle()`, do pacote `pscl`. O resumo dos coeficientes estimados pode ser visualizado na Figura 9.

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

Pearson residuals:
    Min      1Q  Median      3Q     Max
-4.0317 -0.2508 -0.2076 -0.1797 14.9000

Count model coefficients (truncated poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.552889   0.105651  24.164  < 2e-16 ***
quality      0.055034   0.023320   2.360   0.0183 *
skiyes       0.489607   0.058651   8.348  < 2e-16 ***
userfeeyes   0.775855   0.078844   9.840  < 2e-16 ***
income      -0.127710   0.020665  -6.180 6.41e-10 ***
costS       -0.015438   0.001092 -14.140  < 2e-16 ***
Zero hurdle model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.736e+00  3.785e-01  -7.228  4.9e-13 ***
quality      1.484e+00  1.010e-01  14.695  < 2e-16 ***
skiyes       2.366e-01  3.183e-01   0.743    0.457
userfeeyes   1.686e+01  1.447e+03   0.012    0.991
income      -3.556e-02  8.297e-02  -0.429    0.668
costS       -3.302e-03  2.947e-03  -1.120    0.263
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 15
Log-likelihood: -1258 on 12 Df
```

Figura 9: Resumo do modelo Hurdle (Poisson truncado em zero)

Os resultados obtidos são relevantes em diversos aspectos. Na parte logística do modelo (que estima a probabilidade de uma contagem ser maior que zero), conseguimos comprovar a suspeita de que os zeros da variável resposta podem ser explicados quase que inteiramente apenas pela variável `quality` (foi a única estatísticamente significativa). O modelo foi capaz de prever com alta precisão a ocorrência de zeros. Do total de 417 observações com valor zero na variável `trips`, o modelo previu corretamente 385 casos, o que corresponde a uma taxa de acerto de aproximadamente 92,33%. Esse desempenho é consideravelmente superior ao modelo de Poisson simples, que tende a subestimar a ocorrência de zeros em contextos de excesso de zeros.

No entanto, é importante destacar um aspecto técnico importante: ao utilizar uma distribuição Poisson truncada em zero para a parte de contagem, a variabilidade entre os valores positivos pode se tornar ainda mais evidente. Como resultado, o modelo pode apresentar dificuldades para capturar a variância dos dados positivos. Isso reflete um *trade-off* característico do modelo Hurdle — ele melhora a previsão de zeros, mas pode intensificar a sobredispersão na parte contínua (não-zero).Esse comportamento pode ser observado na Figura 10.

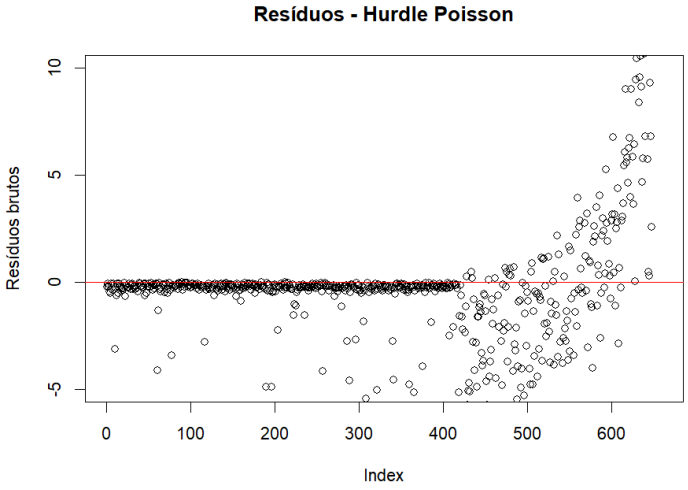


Figura 10: Resíduos brutos do modelo Hurdle (Poisson truncado em zero)

3 Resultados

Como vimos, ao ajustar modelos para dados de contagem com excesso de zeros, há um trade-off importante entre capacidade preditiva e modelagem estatística dos zeros estruturais. Enquanto o modelo de Binomial Negativa lida melhor com a variabilidade dos dados, ele apresenta dificuldades em prever corretamente os zeros. Por outro lado, o modelo *hurdle* consegue, quase perfeitamente, captar a informação dos zeros estruturais, mas o problema da sobredispersão torna-se ainda mais evidente nos dados truncados, resultando em previsões mais imprecisas.

Para avaliar os modelos, além da análise gráfica dos resíduos, utilizamos o AIC, o RMSE e a log-verossimilhança. O AIC (Critério de Informação de Akaike) mede a qualidade relativa de um modelo estatístico, penalizando a complexidade. Ele é calculado com base na log-verossimilhança, sendo dado por $AIC = 2k - 2\log(\hat{L})$, onde k é o número de parâmetros e $\log(\hat{L})$ é o valor da log-verossimilhança do modelo ajustado. RMSE (erro quadrático médio da previsão) avalia quão próximos os valores previstos estão dos valores observados, sendo uma métrica de desempenho preditivo.

A Tabela 3 apresenta a comparação dos três modelos ajustados:

Tabela 3: Comparação dos modelos ajustados com base em RMSE, AIC e log-verossimilhança

Modelo	**RMSE**	**AIC**	**Log-Verossimilhança**
Poisson	5,678	3452,560	-1720,280
Binomial Negativa	10,052	1802,703	-894,351
Hurdle	5,476	2625,465	-1302,732

Ajuste do modelo versus desempenho preditivo

Os resultados podem parecer contraditórios à primeira vista: o modelo com melhor ajuste estatístico (menor AIC e maior log-verossimilhança) apresenta o maior RMSE entre os três modelos. Essa aparente contradição se explica pela diferença entre ajuste estatístico e desempenho preditivo. O objetivo da modelagem aqui proposta é compreender o processo gerador dos dados, e não necessariamente prever com máxima precisão cada observação.

De fato, os modelos foram ajustados por máxima verossimilhança, que visa encontrar os parâmetros que tornam os dados observados mais prováveis, e não por minimização do RMSE. A Binomial Negativa demonstrou ser o modelo que melhor explica a estrutura dos dados de contagem com sobredispersão. Considerando a verossimilhança e a penalização por complexidade, espera-se que ela também ofereça melhor generalização para novos dados, mesmo que sua performance preditiva (RMSE) em dados observados seja inferior.

Portanto, mesmo com RMSE maior, a Binomial Negativa se mostra, neste caso, o modelo mais adequado.

4 Discussão

A sobredispersão dos dados mostrou-se ser um grande desafio para a modelagem. Mesmo tratando a excessiva ocorrência de zeros, com modelos truncados, o problema da variabilidade não foi resolvido. O modelo que melhor se ajustou à distribuição dos dados foi o modelo binomial negativo, que demonstrou capacidade de lidar com a variância além da capacidade do modelo Poisson. Seria também possível utilizar uma alternartiva ao modelo Hurdle, que se trata em utilizar a Binomial Negativa na parte de contagem para juntar o melhor dos dois mundos (boa modelagem de zeros e controle da variância). A tabela compara o novo modelo com os descritos anteriormente.

Tabela 4: Comparação estendida dos modelos ajustados

Modelo	**RMSE**	**AIC**	**Log-Verossimilhança**
Poisson	5,678	3452,560	-1720,280
Binomial Negativa	10,052	1802,703	-894,351
Hurdle (Poisson)	5,476	2625,465	-1302,732
Hurdle (Binomial Negativa)	5,566	1600,530	-789,265

O uso do modelo Hurdle com Binomial Negativa indica que há espaço para refinar a modelagem por meio de estruturas mistas e mais flexíveis.

A inclusão de modelos mais robustos e a consideração de abordagens bayesianas fortalecem a compreensão do processo gerador dos dados e permitem escolhas mais adequadas para fins preditivos ou explicativos.