Sobredispersão em Modelos de Contagem.

Modelagem Estatística - FGV EMAp.

Professor: Luiz Max Fagundes de Carvalho.

Aluno	Curso	Matrícula
Pedro Santos Tokar	Ciência de Dados e Inteligência Artificial, 5º período	231708008

Introdução

Este documento é referente ao trabalho auxiliar passado na matéria de Modelagem Estatística, e sua motivação é fornecer uma experiência prática com o ajuste de modelos estatísticos de contagem. Durante o desenvolvimento deste trabalho, uma base de dados reais será analisada e diferentes modelos estatísticos serão ajustados aos dados desta base. Os resultados de cada modelo serão discutidos e usados para exemplificar conceitos conhecidos da modelagem estatística.

1. Entendendo os dados

Iremos trabalhar com a base de dados *RecreationDemand*, disponível no pacote `Applied Econometrics with R`. Ela contém informações sobre o número de viagens recreativas de barco para o Lago Somerville que foram feitas por donos de barcos de lazer no leste do Texas, em 1980. O dataset conta com 8 variáveis, e iremos analisar cada uma para entender seu significado.

```
RangeIndex: 659 entries, 0 to 658
Data columns (total 8 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   trips    659 non-null    int64
 1   quality  659 non-null    int64
 2   ski      659 non-null    object
 3   income   659 non-null    int64
 4   userfee  659 non-null    object
 5   costC    659 non-null    float64
 6   costS    659 non-null    float64
 7   costH    659 non-null    float64
```

A primeira variável é a `trips`. Ela representa a quantidade de passeios de barco que os proprietários fizeram para o Lago Sommervile, e é a nossa **variável dependente**, pois é ela que temos interesse em regredir. A utilidade dessa regressão está em entender que fatores (principalmente econômicos) podem ou não influenciar a decisão de uma pessoa de visitar ou não o lago a passeio. Suas estatísticas e distribuição são:

```
'Média: 2.244309559939302 | Mediana: 0.0 | Variância:
39.59523732651941 | Desvio Padrão: 6.292474658393104'
```

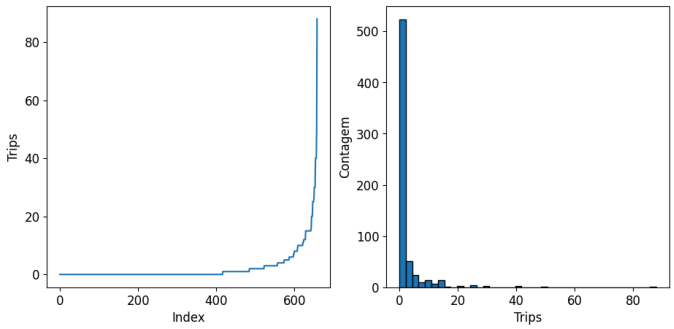

Figure 1: distribuição de `trips`

Percebemos aqui que o dataset está ordenado seguindo essa variável, e que ela contém apenas observações positivas, como o esperado. Sua média é 2,24 viagens e sua variância é 39. Notamos logo de cara que, de quase 700 registros, mais de 400 deles tem valores 0. Para observar melhor a distribuição da variável, é possível plotar ignorando valores 0:

```
'Contagem de 0s: 417'
```

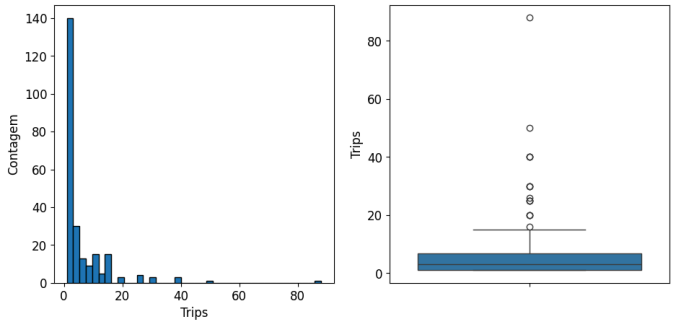

Figure 2: distribuição de `trips` sem zeros

Observamos que a grande maioria dos valores está abaixo de 20, com números menores de observações com valores mais altos, o valor máximo na casa do 80 (perto de alcançar a terceira casa decimal). Essa característica, aliada ao fato do domínio da variável ser os números naturais, torna adequado o uso de modelos como o de Poisson, que é capaz de modelar dados de contagens. Mais da metade das observações ser 0 pode ser um problema para a regressão, mas a principio isso será ignorado.

A segunda variável do dataset é a `quality`. Ao contrário da maioria das **variáveis independentes** do dataset, ela não é relacionada a um indicador econômico, e sim a avaliação que quem visitou o lago deu a ele. Quando observamos sua distribuição, vemos algo claro: temos novamente diversos valores 0, que são relacionados em sua maioria a quem não visitou o lago, mas também incluem pessoas que visitam o lago e não deram uma nota. Mais notavelmente ainda, existem alguns casos de pessoas que não visitaram mas deram uma nota (há menos zeros nessa coluna do que na de trips!).

```
'Contagem de 0s: 374 | Média contando 0s:
1.4188163884673748 | Média sem 0s: 3.280701754385965 |
Variância sem 0s: 1.477267111440573'
```

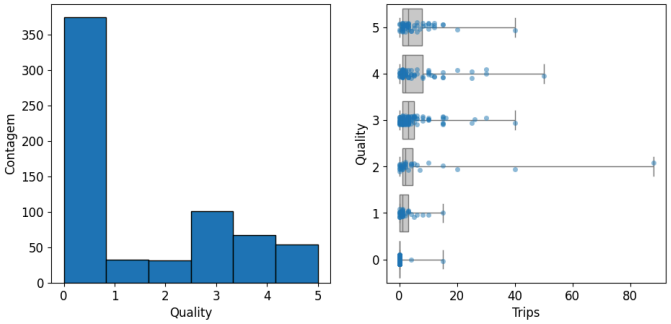

Figure 3: distribuição de `quality` e sua relação com `trips`

A média dessa variável é melhor interpretada quando calculamos sem contar valores de 0, já que indicam que o lago não foi avaliado. Nesse caso, temos média 3,28, indicando que há muitas avaliações intermediárias (confirmado pelo histograma).

A terceira variável, `ski`, também não é relacionada a fatores economicos/financeiros, e sim uma variável binária indicando se o entrevistado esquiava no lago enquanto passeava de barco. Naturalmente, para essa variável funcionar bem no modelo, precisamos convertê-la para valores numéricos (0 para não e 1 para sim). Analisando os histogramas, percebemos que pessoas que esquiam acabam fazendo mais viagens ao lago, sendo que a maioria dos valores pequenos de visitas são associados a pessoas que não visitaram tanto o lago.

```
Value counts de ski
0    417
1    242
```

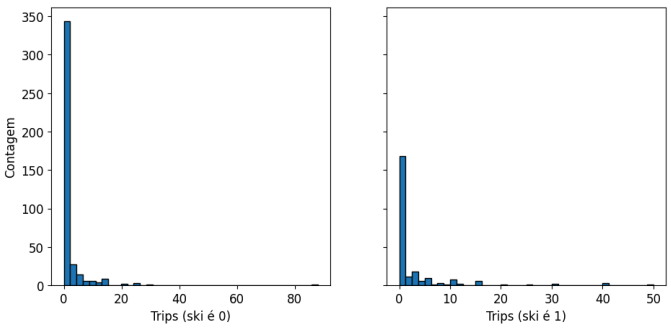

Figure 4: distribuição de `trips` dado o valor de `ski`

A quarta variável, `income`, corresponde à renda familiar do entrevistado. Os valores não são exatos e estão divididos em intervalos de 1000 dólares, o que na prática faz ela ser uma variável categórica. Ainda assim, poderemos tratar como variável contínua na interpretação dos resultados, já que as divisões são baseadas em valores contínuos, e de certa forma valores intermediários entre as categorias têm interpretação válida.

Essa variável não tem valores nulos ou faltantes, e suas estatísticas, em conjunto com seu histograma, indicam que não há outliers e que seu intervalo é bem definido (1 a 9).

```
'Média: 3.8528072837632776 | Mediana: 3.0 | Variância:
3.429669158852641 | Desvio Padrão: 1.8519365968770747'
```

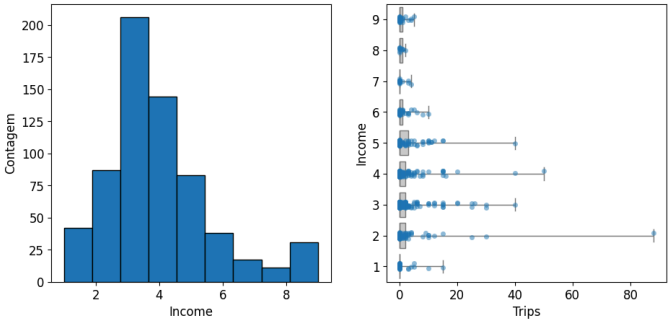

Figure 5: distribuição de `income` e sua relação com `trips`

Observamos, pelo scatterplot, que as pessoas com mais visitas ao lago não necessariamente tem mais renda familiar.

A variável `userfee` também é uma variável binária, indicando se o dono de barco pagou uma taxa anual de uso do lago. Novamente se faz necessário dar um tratamento adequado para a variável, convertendo seus valores para 0 e 1. O que observamos do plot dos histogramas e das contagens é que apenas 13 entrevistados pagaram essa taxa, e o número de passeios que eles fizeram tem uma range ampla, mas acima de 3.

```
Value counts de userfee
0    646
1     13
```

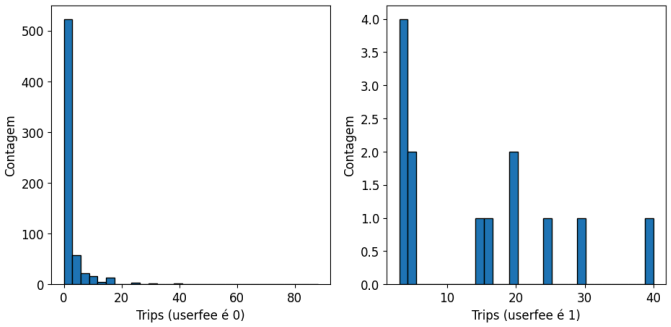

Figure 6: distribuição de `trips` dado o valor de `userfee`

As últimas variáveis do dataset tem funções semelhantes: `costC`, `costS` e `costH` indicam os **custos de oportunidade**, estimados em dólares, para cada entrevistado ir ou ao lago Conroe, ou ao lago Sommerville ou ao lago Houston, todos localizados no Texas com distância de carro entre eles abaixo de 2 horas.

Em outras palavras, são estimativas de quanto a pessoa gastaria para ir a algum desses três lagos, acrescidos de quanto ela "deixaria de ganhar" indo aos outros. Esse é um conceito de economia bem conhecido, e é dado como melhor do que estimar

apenas o gasto em dinheiro que ocorreria com a ida. A utilidade dessas variáveis em nossa regressão é entender se o custo de oportunidade de ir aos outros lagos influencia na quantidade de viagens ao lago de interesse. Porém, ao analisá-las, observamos que elas tem a distribuição muito próxima.

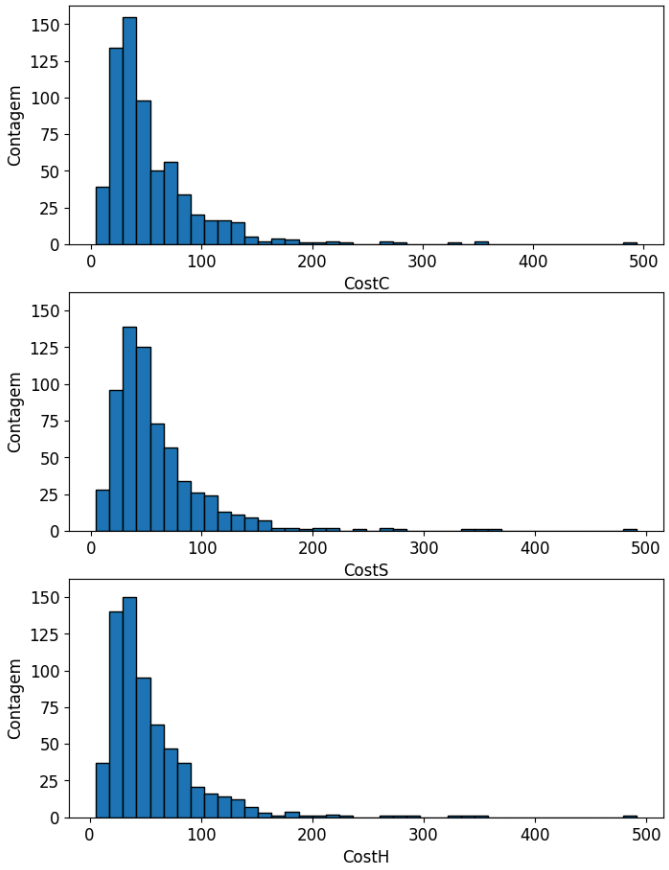

Figure 7: distribuição de `costC`, `costS` e `costH`

Podemos inspecionar melhor como essas variáveis se relacionam entre si e entre a variável de interesse plotando alguns scatterplots que mostrem as distribuições entre cada uma delas e delas com a variável de interesse.

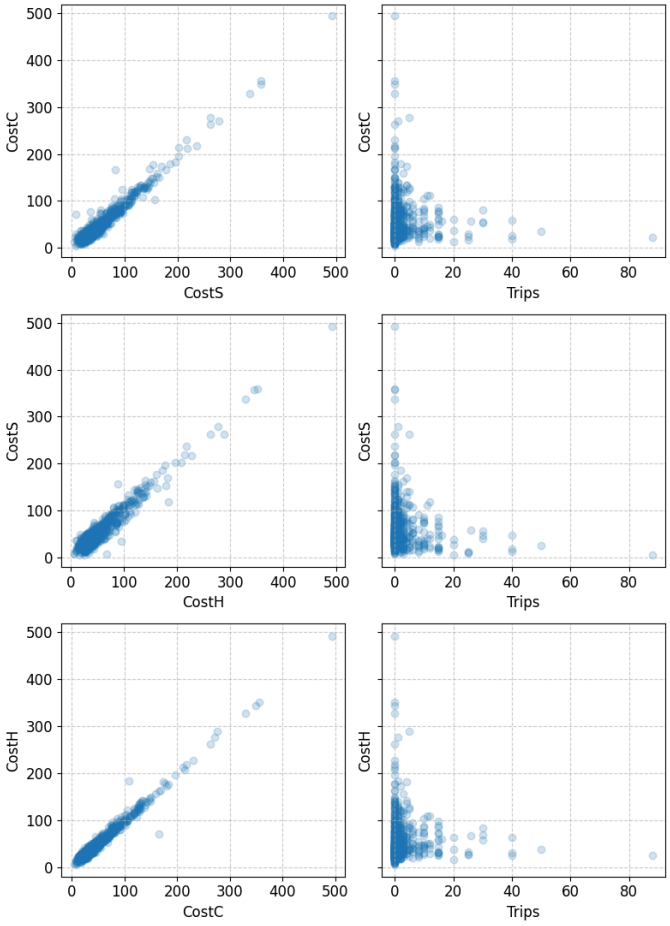

Figure 8: relacionamentos das variáveis `cost` com `trips` e com elas mesmas

Após inspecionar esses scatterplots, esperamos que:

- A correlação entre as variáveis de custo seja **muito** alta;
- A correlação delas com a variável de interesse seja baixa em módulo, já que a relação não parece ser linear.

Verificaremos isso fazendo uma matriz de correlação entre as variáveis do nosso dataset:

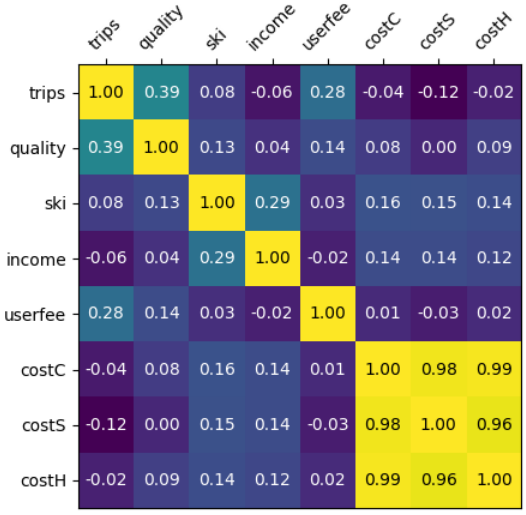

Figure 9: correlação entre as variáveis da base

Como a matriz mostra, as três variáveis de custo têm correlações maiores do que 95% entre elas. Além disso, a variável de custo ao lago de interesse (Sommerville) tem correlação com a variável de interesse maior em módulo do que a das outras duas, mas ainda assim com módulo muito baixo. Isso indica que, entre elas, essa pode ser a mais importante para o modelo. Além da matriz de correlação, uma métrica muito útil para definirmos se essas variáveis entrarão na regressão é o Fator de Inflação de Variância (em inglês, VIF).

	0	1	2	3	4	5	6
coluna	quality	ski	income	userfee	costC	costS	costH
VIF	1.717	1.792	3.190	1.065	131.397	61.328	93.752

Todas elas tem VIF muito maior do que os recomendados 5 pontos para a remoção de um modelo. Introduzir elas no modelo pode causar uma instabilidade numérica em sua avaliação. Mesmo se tratando de um GLM, que usa métodos iterativos para maximizar a verossimilhança e não calcula a matriz $(X^T X)^{-1}$, ainda assim ter variáveis muito próximas da colinearidade prejudica os cálculos da convergência do modelo.

Todas essas análises sobre as últimas variáveis levam à uma preocupação quanto a contribuição delas para o modelo que será ajustado. Existe a possibilidade delas introduzirem mais problemas do que ajudar na construção do modelo, levando a conclusões errôneas.

Uma possível forma de mitigar os efeitos negativos ao ajuste que a inclusão dessas variáveis traria é remover duas delas e deixar apenas uma. Nessa abordagem, o significado prático das variáveis leva à conclusão de que seria melhor manter a varíavel `costS` e remover as outras duas, já que o custo que mais influenciaria a ida ao lago seria o dele mesmo.

Porém, vale observar que essa abordagem tira do modelo uma capacidade que pode ser interessante: inferir se a diferença entre essa variáveis influencia na decisão (ou seja, se um lago ter um custo menor de ida que o outro muda a quantidade de viagens que uma pessoa fez a ele). Mantendo apenas uma variável, o potencial de comparação é perdido. Ao mesmo tempo, manter duas delas segue fazendo o VIF delas ser alto:

	0	1	2	3	4	5
coluna	quality	ski	income	userfee	costS	costH
VIF	1.712	1.785	3.115	1.062	42.9	40.775

Pensando em manter o potencial de interpretação que essas variáveis podem trazer e ao mesmo tempo evitar os problemas de colinearidade, é possível fazer uma transformação que mantenha informações úteis para inferência e remova as colunas sem muita perda de informação. Tendo em vista o significado das variáveis e visando não aumentar muito a complexidade da base, uma transformação possível é introduzir uma coluna `isCheapest`, com valores binários:

- 1 (verdadeiro) caso `costS` < `costC` e `costS` < `costH`.
- 0 (falso) caso contrário.

```
Value counts de isCheapest
0    505
1    154

Correlacao entre trips e isCheapest: 0.4399
```

	0	1	2	3	4	5
coluna	quality	ski	income	userfee	costS	isCheapest
VIF	1.899	1.785	3.042	1.081	2.352	1.554

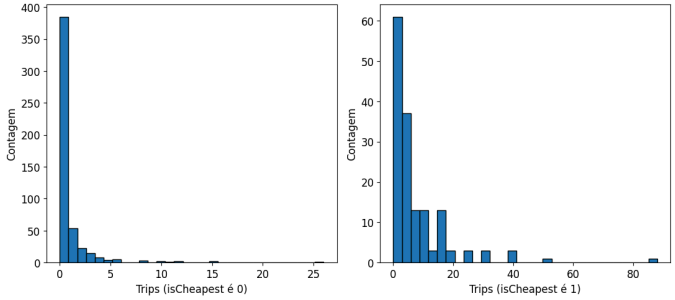

Figure 10: distribuição de `trips` dado `isCheapest`

A análise dessa nova variável parece mostrar que ela pode ser uma boa escolha para o ajuste do modelo, por mitigar os problemas causados pelas outras sem "jogar fora" totalmente o potêncial informativo que elas tinham. As distribuições do número de viagens para entradas em que ela é verdadeira mostra que ela pode ser bem impactante na decisão de uma pessoa de visitar ou não o lago. A correlação de 44% também corrobora com esse pensamento.

Para por a prova o que foi levantado nesta análise, é possível ajustar mais de um modelo, com diferentes escolhas de variáveis, e analisar qual se ajusta melhor aos dados.

2. Ajustando modelos Poisson

Aqui, serão ajustados 3 modelos lineares generalizados da família Poisson, devido à natureza da variável dependente. A função de ligação desse modelo é o logarítmo natural $log(.)$. O ajuste será feito de maneira frequentista, usando o Estimador de Máxima Verossimilhança. Devido à natureza desse modelo (e dos GLMs num geral), o ajuste é feito por métodos numéricos iterativos, que convergem para a solução de máxima verossimilhança.

Todos eles serão ajustados com as variáveis `quality`, `ski`, `income` e `userfee`, pois nenhuma delas apresentou sinais de que não tinha influência nenhuma na variável de interesse (a variável `quality` tinha algumas entradas suspeitas, mas a correlação dela com a variável dependente pode ser um sinal de que ela será uma boa preditora). Já as variáveis de custo terão um tratamento especial: o primeiro modelo será ajustado com as três variáveis, o segundo apenas com a `costS` e o terceiro com a `costS` e a nova variável `isCheapest`.

Generalized Linear Model Regression Results

Dep. Variable:	trips	No. Observations:	659
Model:	GLM	Df Residuals:	651
Model Family:	Poisson	Df Model:	7
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1529.4
Date:	Wed, 04 Jun 2025	Deviance:	2305.8
Time:	01:29:46	Pearson chi2:	4.10e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.9789
AIC:	3074.8626	BIC:	-1919.6755

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2650	0.094	2.827	0.005	0.081	0.449
quality	0.4717	0.017	27.602	0.000	0.438	0.505
ski	0.4182	0.057	7.313	0.000	0.306	0.530
income	-0.1113	0.020	-5.683	0.000	-0.150	-0.073

userfee	0.8982	0.079	11.371	0.000	0.743	1.053
costC	-0.0034	0.003	-1.100	0.271	-0.010	0.003
costS	-0.0425	0.002	-25.466	0.000	-0.046	-0.039
costH	0.0361	0.003	13.335	0.000	0.031	0.041

Observamos que o modelo com todas as variáveis apresenta um pseudo-R^2_{CS} de aproximadamente $0,98$. Essa métrica é o pseudo-R^2 de Cox e Snell, usado em modelos lineares generalizados e parametrizado como $R^2_{CS} = 1 - \exp\{\frac{2}{n}(\ln(L_0) - \ln(L_M))\}$. Ele é uma generalização do R^2 já conhecido para modelos que não são ajustados por mínimos quadrados, e sim por maxima verossimilhança.

Nos GLM, ele não é interpretado como a porcentagem da variância explicada, e sim como uma melhoria no ajuste do modelo (tanto que ele pode nem chegar a um para regressões logísticas), e ainda mantém a propriedade de quanto maior, melhor.

Podemos ver que todos as estimativas de regressores, com excessão do associado à variável `costC`, são estatísticamente significativos. Esse resultado pode parecer bom, mas é preciso tomar cuidado e se ter em mente que o VIF das variáveis de custo eram bem altos, e que isso pode interferir nas estatísticas e testes do modelo.

Generalized Linear Model Regression Results

Dep. Variable:	trips	No. Observations:	659
Model:	GLM	Df Residuals:	653
Model Family:	Poisson	Df Model:	5
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1720.3
Date:	Wed, 04 Jun 2025	Deviance:	2687.5
Time:	01:29:46	Pearson chi2:	5.82e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.9624
AIC:	3452.5604	BIC:	-1550.9591

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5861	0.092	6.377	0.000	0.406	0.766
quality	0.5408	0.016	33.924	0.000	0.510	0.572
ski	0.4542	0.056	8.044	0.000	0.344	0.565
income	-0.1578	0.020	-8.093	0.000	-0.196	-0.120
userfee	1.1015	0.080	13.786	0.000	0.945	1.258
costS	-0.0153	0.001	-15.098	0.000	-0.017	-0.013

Removendo as variáveis de custo que não são referentes ao lago Sommerville, observamos uma piora em algumas métricas: o AIC e Deviance são maiores (o que é indesejável) e o pseudo-R^2 é menor, ainda que por pouca diferença. Também observamos que a verossimilhança é menor, indicando uma menor compatibilidade entre os dados reais e as previsões.

Esse aumento pode indicar que a remoção das variáveis trouxe uma perda de poder preditivo, mesmo que estabilizasse a convergência do modelo e suas estatísticas.

Generalized Linear Model Regression Results

Dep. Variable:	trips	No. Observations:	659
Model:	GLM	Df Residuals:	652
Model Family:	Poisson	Df Model:	6
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1434.1
Date:	Wed, 04 Jun 2025	Deviance:	2115.1
Time:	01:29:46	Pearson chi2:	4.04e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.9842
AIC:	2882.1280	BIC:	-2116.9007

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4750	0.108	-4.418	0.000	-0.686	-0.264
quality	0.4259	0.018	24.026	0.000	0.391	0.461
ski	0.4920	0.055	8.882	0.000	0.383	0.601
income	-0.0768	0.019	-4.099	0.000	-0.113	-0.040
userfee	0.6083	0.081	7.508	0.000	0.450	0.767
costS	-0.0093	0.001	-9.947	0.000	-0.011	-0.007
isCheapest	1.5037	0.067	22.541	0.000	1.373	1.634

O terceiro modelo apresenta todas as métricas melhores: seu AIC e sua Deviance são menores que dos outros, indicando que mesmo com menos features, o modelo pode se ajustar melhor aos dados. O pseudo-R^2 é maior, também reforçando a ideia de que muito da variância foi explicada.

É possível ver que todas as estimativas de parâmetros são estatísticamente significantes, e para esse modelo esses resultados são mais confiáveis, principalmente quando se leva em conta que o VIF de todas as variáveis independentes usadas era menor do que 5. A log verossimilhança também é menor, o que é mais um bom sinal.

Para auxiliar na análise da bondade de ajuste desses modelos, podemos fazer plots de valor predito por resíduo de Pearson. Os resíduos de Pearson são definidos por $r_i^P = \frac{y_i - \hat{\mu}_i}{sqrt(\hat{\mu}_i)}$ e são padronizados, levando em conta qual seria a variância para um preditor da média. Por não estarmos lidando com um modelo linear, não é possível esperar que esses resíduos se distribuam uniformemente, mas é esperado, pela padronização, que eles fiquem distribuidos perto de 0 e a que a grande maioria fique no range (-2, 2).

O objetivo dessa padronização é se adaptar ao fato de que, na regressão Poisson, a variância dos dados aumenta junto com a média, e tornar esses resíduos mais uniformes mesmo com o aumento da média (que aqui é a previsão do modelo).

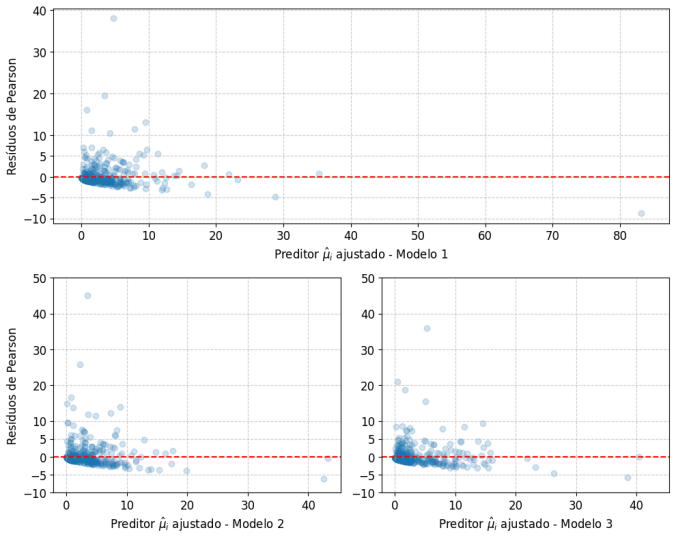

Figure 11: plot dos resíduos para os três modelos Poisson

Dos plots, três observações saltam aos olhos:

- O primeiro e o segundo modelos tem resíduos padronizados que parecem ser menos dispersos a medida em que o valor do preditor aumenta, o que não acontece tão claramente com o terceiro modelo, que tem a distribuição desses resíduos mais consistente mesmo com as mudanças do preditor;
- Todos os modelos tem vários dos resíduos longe do 0, indo contra o que seria esperado da distribuição desses resíduos, que idealmente teriam desvio padrão 1;
- Os resíduos menores que 0 parecem se concentrar mais, enquanto os maiores que zero são mais espalhados e amplos.

A primeira observação corrobora com a ideia de que o terceiro modelo conseguiu se ajustar melhor aos dados, mantendo uma distribuição dos resíduos de Pearson mais próxima do que se espera. Já a segunda observação é um forte indício de que todos os modelos sofreram de sobredispersão.

Os resíduos estarem muito espalhados, mais do que o esperado, indica que os valores reais estão muito distantes das previsões do modelo. Essa distância pode se dar pela sobredispersão: o modelo não foi pensado para lidar com dados com variância maior do que a média, e então não faz previsões de acordo com a variância real.

É possível conduzir um teste estatístico formal para atestar se há ou não sobredispersão no modelo. Para conduzir este teste, é necessário escolher um dos modelos. Devido às métricas apresentadas serem melhores e a distribuição dos resíduos de Pearson serem mais próximas do esperado, o teste será conduzido com o terceiro modelo, que faz uso da nova variável criada para o dataset.

3. Testando a sobredispersão

Para esse teste, queremos saber se rejeitamos ou não a hipotese nula: para uma parametrização da variância como $Var(Y) = \mu + \alpha\mu^2$, a hipotese nula é $H_0 : \alpha = 0$. Os resultados da regressão de Z_i em $\hat{\mu}_i$ obtidos foram:

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.010
Model:	OLS	Adj. R-squared (uncentered):	0.009
Method:	Least Squares	F-statistic:	6.888
Date:	Wed, 04 Jun 2025	Prob (F-statistic):	0.00888
Time:	01:29:47	Log-Likelihood:	-3572.5
No. Observations:	659	AIC:	7147.
Df Residuals:	658	BIC:	7151.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	1.2061	0.460	2.624	0.009	0.304	2.109

Estatísticamente, obtivemos uma estatística de teste t igual à $2,624$, cujo p-valor é $0,009$. Esse resultado indica significância estatística, já que o p-valor está abaixo de 0,05. Logo, rejeitamos a hipótese H_0, levando a conclusão de que existe sobredispersão. O resultado do teste formal condiz com o que observamos nos resíduos dos modelos na etapa anterior.

Tendo conhecimento da sobredispersão, não é vantajoso usar um modelo que não é pensado para lidar com ela para fazer inferências e interpretações sobre os dados. Por isso, é importante usar modelos que são pensados para lidar com esse tipo de fenômeno.

4. Modelo Binomial Negativo

O modelo em questão, que consegue levar a sobredispersão em conta em seu ajuste, é um modelo da família Binomial Negativa. Esse modelo, assim como o de Poisson, modela dados cujo domínio é o conjunto dos números naturais. A parametrização desse modelo lembra o teste que foi feito acima:

$$\begin{cases} \mathbb{E}[Y_i \mid X_i] = \mu_i = \exp\{x_i^T \beta\} \\ Var[Y_i \mid X_i] = \mu_i + \alpha\mu_i^2 \end{cases}$$

Essa variância é o que modela a sobredispersão: o valor α é estimado junto com o β, e então o ajuste leva em conta a sobredispersão, e o parâmetro α ganha uma estimativa $\hat{\alpha}$ com intervalo de confiança. Como estamos lidando com dados inteiros positivos, e como os dados de contagem geralmente abrangem várias magnitudes (como vimos no nosso dataset também), a função de link deste modelo é o logarítmo natural, função inversa da exponencial.

NegativeBinomial Regression Results

Dep. Variable:	trips	Df Residuals:	652
Model:	NegativeBinomial	Df Model:	6
Method:	MLE	Deviance:	439.04
Date:	Wed, 04 Jun 2025	Pseudo R-squ.:	0.2142
Time:	01:29:47	Log-Likelihood:	-836.62
converged:	True	LL-Null:	-1064.7
Covariance Type:	nonrobust	LLR p-value:	2.260e-95
No. Observations:	659	AIC:	1689.2337
BIC:	1725.1595		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6855	0.228	-7.393	0.000	-2.132	-1.239
quality	0.7132	0.046	15.582	0.000	0.624	0.803
ski	0.6324	0.153	4.143	0.000	0.333	0.932
income	-0.0595	0.045	-1.317	0.188	-0.148	0.029
userfee	0.7272	0.364	1.996	0.046	0.013	1.441
costS	-0.0057	0.002	-2.899	0.004	-0.010	-0.002
isCheapest	1.7376	0.156	11.154	0.000	1.432	2.043
alpha	1.4219	0.150	9.454	0.000	1.127	1.717

Observamos que após o ajuste, o resultado contém as estimativas para os regressores e também para o parâmetro da sobredispersão. A métrica do pseudo-R^2 se mostrou bem menor em relação ao que estava sendo visto na regressão Poisson, mas isso é esperado: o Modelo Poisson ignorava a sobredispersão e ela não era levada em conta para fazer o ajuste e nem calcular as métricas. Já o modelo Binomial Negativa, por levar em conta a sobredispersão, incorpora o ajuste à esse fenômeno em sua verossimilhança e por consequência no cálculo de métricas como o pseudo-R^2.

Isso se reflete em um valor mais baixo, mas mais real e útil para comparação de possíveis modelos com essa parametrização (ele não está jogado para perto do 1, o que pode facilitar comparações). Já o valor da Deviance é bem menor que o final dos outros modelos, indicando um melhor ajuste geral aos dados, com menos erros de previsão. Assim como feito nos modelos Poisson, é possível analisar o plot dos resíduos de Pearson:

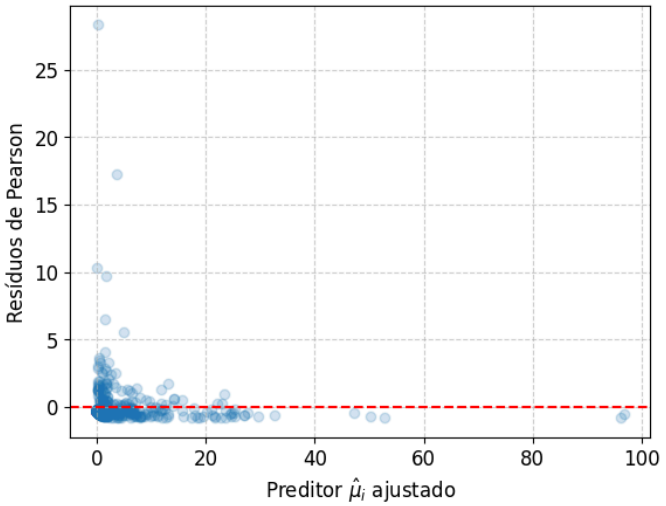

Figure 12: plot dos resíduos para o modelo Binomial Negativa

Vemos que agora os resíduos se concentram dentro do range (-2, 2), ao contrários dos resíduos dos modelos Poisson treinados, que eram muito mais dispersos. Isso indica que o modelo conseguiu lidar melhor com a sobredispersão dos dados, fazendo previsões mais "amplas". Esse resultado é positivo, já que se deseja modelar o comportamento real dos dados e todos os seus fenômenos intrinsecos.

É preciso ressaltar, porém, que os resíduos ainda apresentam comportamentos estranhos. Assim como ocorreu com os modelos Poisson, os resíduos negativos se concentram no intervalo esperado, enquanto alguns positivos são muito espalhados.

Esse resultado é de certa forma esperado, já que o domínio dos dados impede que o erro seja muito baixo (significaria uma previsão negativa). O que causa estranheza é esse comportamento ser especialmente predominante em predições que foram próximas de 0: o plot forma uma espécie de "L" espelhado, com muitos valores preditos errôneamente como zeros. Disso concluímos que:

- O modelo está errando mais para baixo do que para cima, ou seja, está predizendo muitos valores abaixo do esperado.
- Algumas dessas previsões tem erros que jogam elas para outras casas de magnitude abaixo do esperado.

Esses são sintomas de que o modelo está predizendo valores abaixo dos que deveria predizer muito frequentemente, fugindo do comportamento que se esperaria ver normalmente. Uma das possíveis causas para isso é a quantidade de entradas com zeros na base de dados.

5. Modelo inflado de zeros

Como reportado no ínicio, muitos registros da base de dados original tem a variável `trips` como 0. Os modelos treinados também fazem bastante previsões zero:

```
'Zeros na variável trip: 417'
```

```
'Zeros previstos pelo modelo Binomial Negativo: 458'
```

```
'Zeros previstos pelo terceiro modelo Poisson: 339'
```

Como a saída dos modelos são números reais, contei como zero qualquer valor abaixo de 0,5. Os zeros presentes na base de dados estão em muito maior quantidade do que se espera de uma distribuição Poisson com a média que os dados tem. Isso pode ser um sinal de **zeros estruturais**.

O conceito de zero estrutural é associado à ideia de dividir as observações 0 de uma base de dados em duas categorias: zeros gerados pela distribuição modelada pelo modelo de contagem e zeros gerados por uma outra distribuição binária. Esses segundos zeros não seriam gerados pela disstribuição do modelo inicial, e por isso fazer um ajuste puro dele poderia ignorar essa característica dos dados.

Uma das formas de lidar com esse tipo de observação é usando modelos inflados de zeros. Eles consistem em modelos de contagem já conhecidos combinados com modelos de classificação binária. Assim, o modelo de classificação binária modela a probabilidade de uma observação zero ser estrutural, e o treino do modelo de contagem leva isso em conta.

Formalmente, sendo π a probabilidade de uma observação ser um zeros estrutural e $1-\pi$ a probabilidade de uma observação ter sido gerada de um modelo poisson com parâmetro μ, temos:

$$p(y_n \mid \pi, \mu) = \begin{cases} \pi + (1-\pi)\ \text{Poisson}(0 \mid \mu) \text{ se } y_n = 0; \\ (1-\pi)\ \text{Poisson}(y_n \mid \mu) \quad \text{caso contrário .} \end{cases}$$

O valor π é também expresso como um modelo em função dos dados que fornece diferentes valores de π para diferentes registros dos dados, por meio também de um GLM (com função de ligação logit ou probit, já que o interesse é ter $\pi \in [0, 1]$). A estimativa de π pode ser feita utilizando um conjunto diferente de colunas do que o usado no modelo de contagem, e essa abordagem pode ser melhor caso seja possível separar intuitivamente quais variáveis influenciariam na observação ser um zero estrutural ou não.

No caso dos dados de passeios, para ajustar o modelo de inflação usarei apenas a variável `quality` e um intercepto. Essa escolha se deu pela descrição do dataset, que deixa explicito que valores 0 na variável `quality` indicam pessoas que não visitaram o lago.

ZeroInflatedPoisson Regression Results

Dep. Variable:	trips	No. Observations:	659
Model:	ZeroInflatedPoisson	Df Residuals:	652
Method:	MLE	Df Model:	6
Date:	Wed, 04 Jun 2025	Pseudo R-squ.:	0.3457
Time:	01:29:47	Log-Likelihood:	-1146.6
converged:	True	LL-Null:	-1752.5
Covariance Type:	nonrobust	LLR p-value:	1.333e-258
AIC:	2311.1265	BIC:	2351.5430

	coef	std err	z	P>|z|	[0.025	0.975]
inflate_const	5.2892	0.806	6.564	0.000	3.710	6.869
inflate_quality	-6.6033	0.976	-6.768	0.000	-8.516	-4.691
const	1.3046	0.127	10.241	0.000	1.055	1.554
quality	0.0606	0.024	2.576	0.010	0.014	0.107
ski	0.4822	0.056	8.619	0.000	0.373	0.592
income	-0.0868	0.019	-4.557	0.000	-0.124	-0.049
userfee	0.5426	0.080	6.782	0.000	0.386	0.699
costS	-0.0089	0.001	-9.265	0.000	-0.011	-0.007
isCheapest	1.1095	0.065	16.988	0.000	0.982	1.238

Observamos que o modelo apresenta todos os estimadores como estatisticamente significativos. Também observamos um valor de pseudo-R^2 maior do que o atingido com o modelo da família Binomial Negativa. Nos modelos inflados de zeros, não é possível

usar a Deviance para comparações, já que esses modelos não são GLMs por definição, e portanto essa métrica não é adequada.

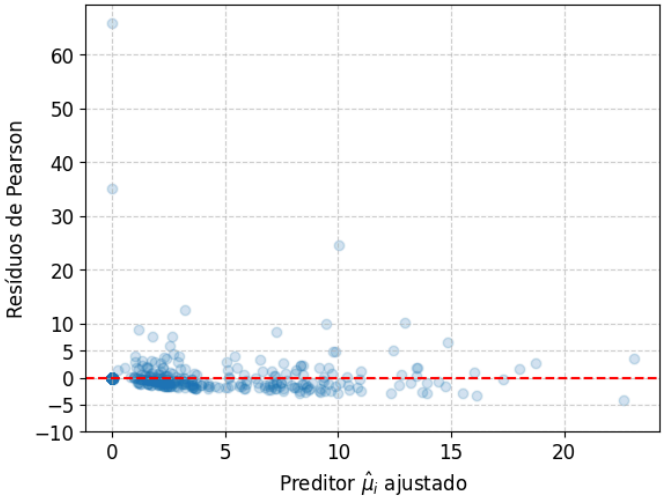

Figure 13: plot dos resíduos para o modelo Poisson inflado de zeros

Analisando os resíduos, vemos que agora temos várias observações próximas de zero e com resíduos também próximos de zero. Elas são associadas aos casos em que o modelo classificador retornou uma alta probabilidade do registro ser um zero estrutural. Nesse plot, porém, observamos algo que havia nos primeiros modelos Poisson: resíduos mais dispersos do que o esperado, indicando que a sobredispersão existe mesmo filtrando corretamente os zeros estruturais.

Para tentar reduzir esse fenômeno, podemos usar um modelo inflado de zeros baseado na família Binomial Negativa:

ZeroInflatedNegativeBinomialP Regression Results

Dep. Variable:	trips	No. Observations:	659
Model:	ZeroInflatedNegativeBinomialP	Df Residuals:	652
Method:	MLE	Df Model:	6
Date:	Wed, 04 Jun 2025	Pseudo R-squ.:	0.3160
Time:	01:29:48	Log-Likelihood:	-728.23
converged:	True	LL-Null:	-1064.7
Covariance Type:	nonrobust	LLR p-value:	4.175e-142
AIC:	1476.4690	BIC:	1521.3762

	coef	std err	z	P>|z|	[0.025	0.975]
inflate_const	3.9179	0.479	8.172	0.000	2.978	4.858
inflate_quality	-5.4033	0.780	-6.924	0.000	-6.933	-3.874
const	0.8416	0.269	3.130	0.002	0.315	1.369
quality	0.1165	0.053	2.199	0.028	0.013	0.220
ski	0.4823	0.136	3.552	0.000	0.216	0.748
income	-0.0922	0.043	-2.125	0.034	-0.177	-0.007
userfee	0.5994	0.281	2.136	0.033	0.049	1.149
costS	-0.0043	0.002	-2.398	0.016	-0.008	-0.001
isCheapest	1.1739	0.142	8.248	0.000	0.895	1.453
alpha	0.8399	0.093	9.026	0.000	0.658	1.022

Quando observamos os valores estimados, percebemos que para as variáveis do modelo logístico as estimativas são parecidas, o que pode indicar que nos dois métodos o modelo se ajustou para interpretar os zeros estruturais de maneira semelhante.

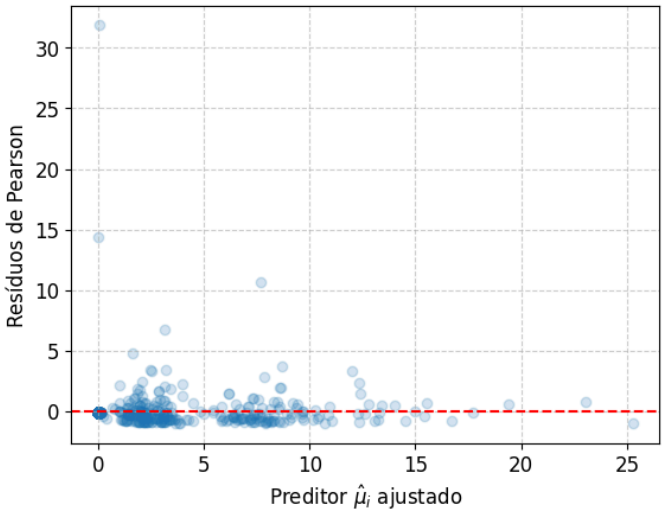

Figure 14: plot dos resíduos para o modelo Binomial Negativa inflado de zeros

O plot dos resíduos também apresenta uma grande concentração de pontos ao redor de $(0, 0)$, mas não apresenta resíduos muito dispersos ao longo dos valores dos preditores. A estimativa do valor α indica que existia sobredispersão, que foi capturada pelo modelo. Uma métrica que pode corroborar para o argumento de que este modelo se saiu melhor é o AIC, que é visivelmente mais baixo que o do modelo ZIP.

Tendo em vista os plots de resíduos e a métrica citada acima, o modelo que eu escolheria para fundamentar análises do assunto seria o **modelo da família Binomial Negativa inflado de zeros**. Ele se mostrou eficiente em ajustar a sobredispersão dos dados ao mesmo tempo em que foi eficiente em identificar os zeros estruturais dos dados.

6. Considerações finais e direções futuras

O número de modelos disponíveis na modelagem estatística é grande. Além dos modelos testados neste trabalho, seria possível testar modelos de hurdle, que usam uma aproximação levemente diferente para a ideia dos zeros estruturais, mas que acaba por mudar a forma como eles são ajustados. Também seria ideal testar mais transformações entre as variáveis, a fim de chegar à um conjunto de variáveis independentes robusto para o problema. Isso requer um trabalho de testes e exploração extenso, e se beneficia de conhecimentos extras sobre a origem dos dados.

Observação

Para melhor adequação ao formato impresso da entrega do trabalho, o código usado para fazer as visualizações e os modelos foi omitido, e alguns outputs foram reformatados e combinados com outros para tornar o layout mais agradável. O notebook `ipynb` usado para gerar o relatório e que contém os outputs originais está disponível no GitHub: https://github.com/pedrotokar/statsmodeling-A2.