zero-inflated

Rafael d'Angelo Reis e Maria do Carmo Teixeira

2023-01-15

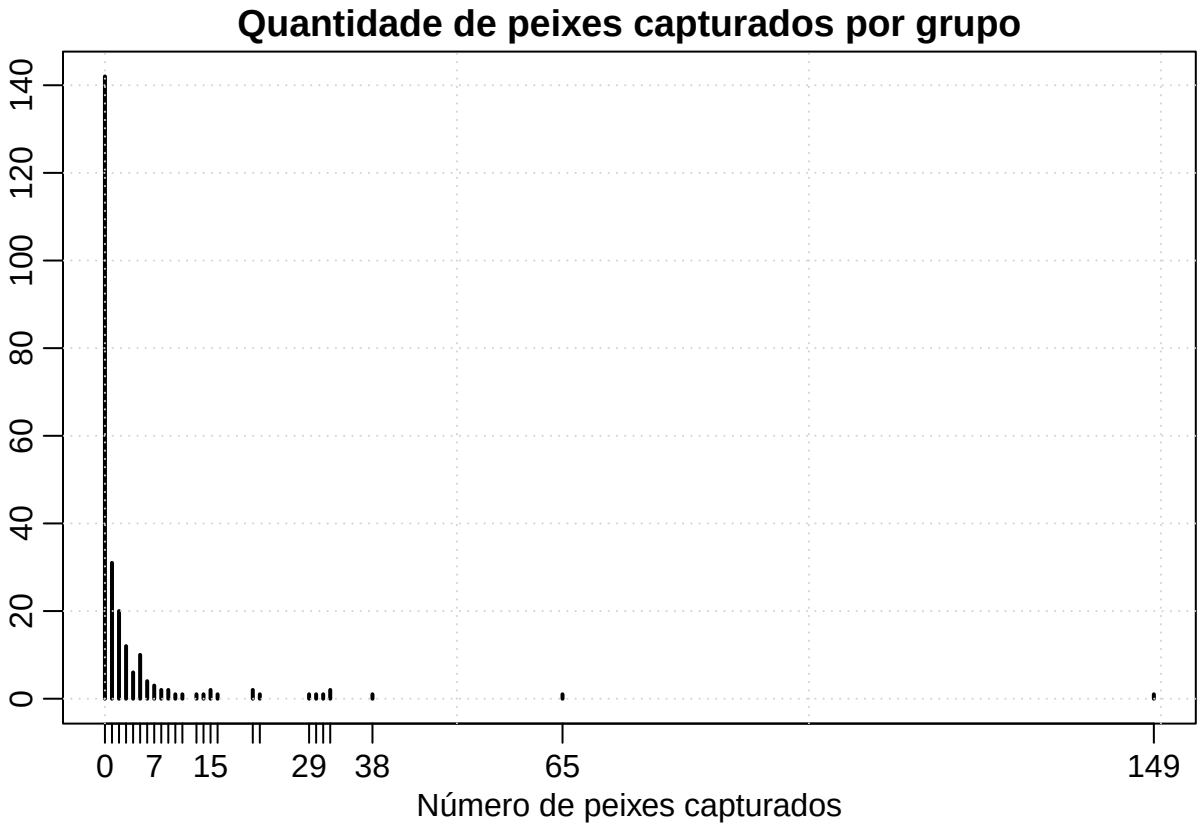
Descrição dos dados

Os biólogos de vida selvagem do estado da Califórnia, nos Estados Unidos, queriam modelar quantos peixes estavam sendo capturados por pescadores em um parque estadual. Os visitantes foram questionados quanto tempo ficaram, quantas pessoas estavam no grupo, se havia crianças no grupo e quantos peixes foram pescados. Alguns visitantes que visitaram o parque não pescaram, mas não há dados sobre se uma pessoa pescou ou não, e como alguns visitantes que pescaram não pegaram nenhum peixe, há excesso de zeros nos dados por causa das pessoas que não pescaram.

Foram 250 grupos que foram ao parque e cada grupo foi questionado sobre quantos peixes pescaram (variável resposta count), quantas crianças haviam no grupo (variável explicativa child), quantas pessoas haviam no grupo (variável explicativa persons) e se trouxeram ou não um trailer para o parque (variável explicativa camper).

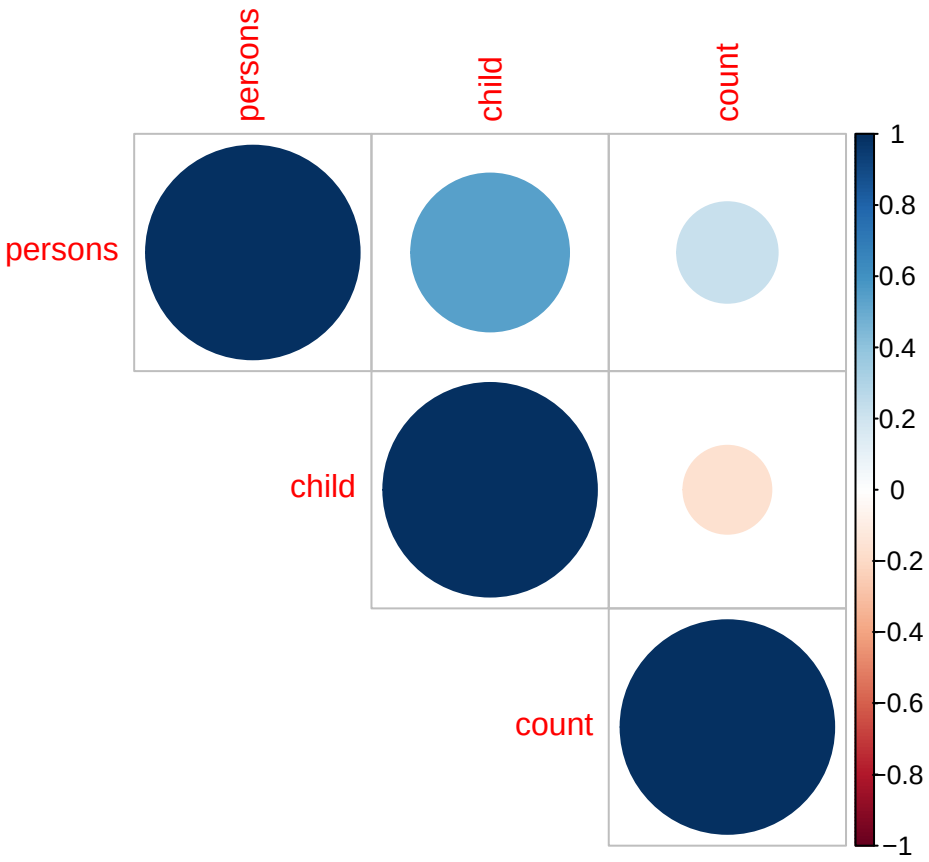
Análise Exploratória

Fez-se necessário um estudo inicial das variáveis do banco de dados. Realizou-se então uma análise exploratória, a fim de avaliar suas principais características e seu comportamento.

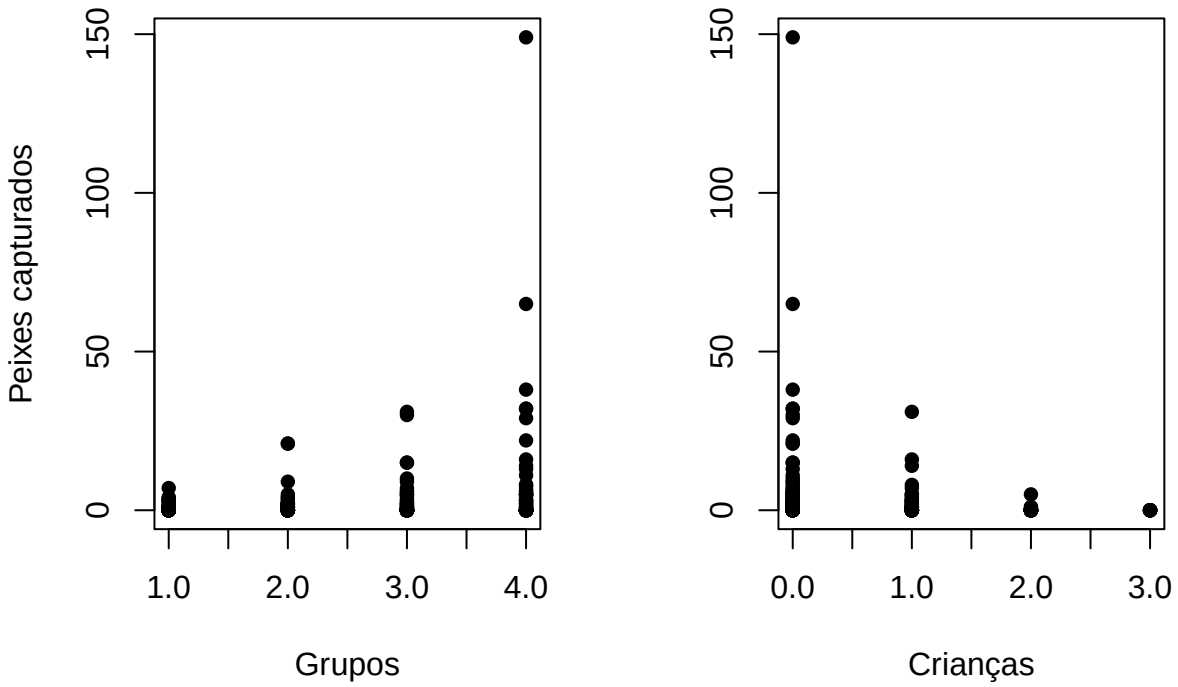


Podemos ver pelo gráfico 1 que apesar do grande número de grupos de pescadores, a quantidade de grupos que não pegaram nenhum peixe é consideravelmente maior que a daqueles que conseguiram alguns. Esses dados com quantidades inflacionadas de 0 afetam diretamente o estudo de contagens ao utilizarmos os modelos de poisson e binomial negativa que estamos acostumados, então algo deve ser feito para encontrarmos o modelo ideal.

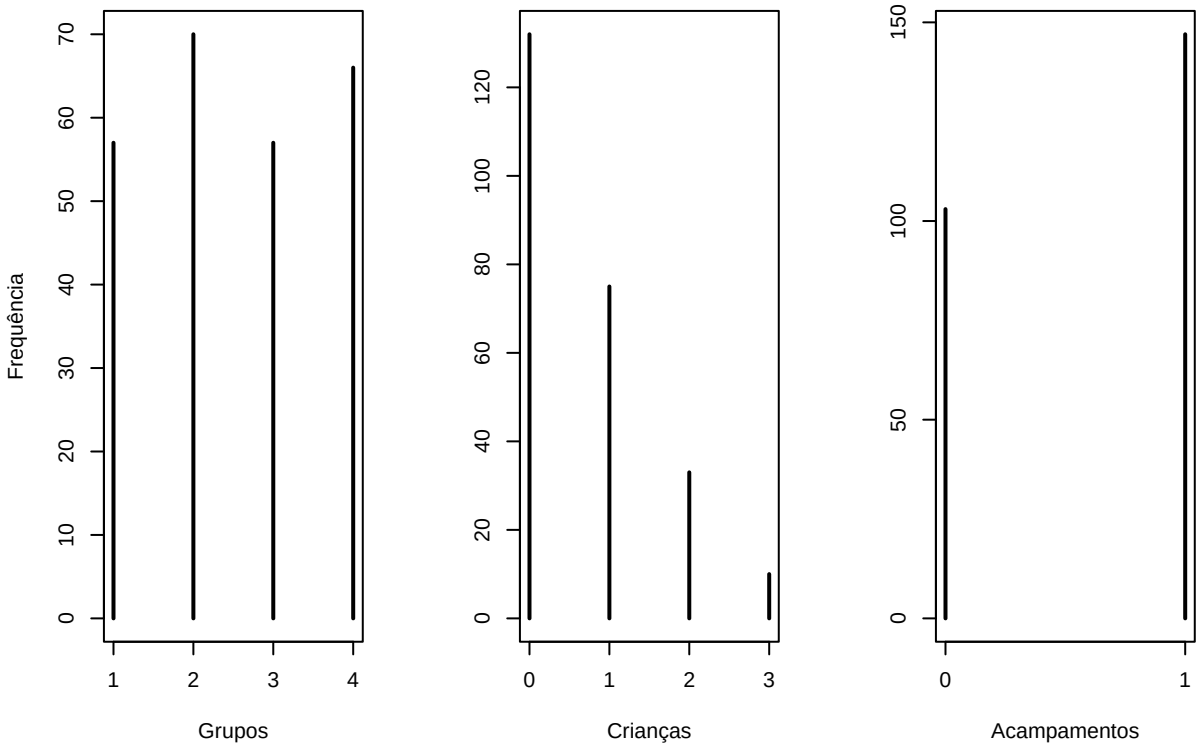
Também pelo gráfico podemos perceber alguns pontos fora da curva no que alguns grupos pegaram quantidades significamente altas de peixes comparados com os demais, então já imaginamos que eles serão observações a serem analisadas separadamente.



Vemos no gráfico 2 as correlações entre nossa variável resposta (count) e as duas variáveis numéricas do modelo. Já por ele podemos imaginar que ocorrerá uma relação positiva entre o número de pessoas nos grupos de pescadores e o número de peixes pescados, da mesma forma que existirá uma relação negativa entre o número de crianças levadas e a quantidade de peixes conseguidos.



Pelos gráficos 3 e 4 podemos confirmar as suspeitas que vimos no gráfico 2 com a análise de correlações, inclusive como é curioso como os grupos que levaram 3 crianças não conseguiram pescar nenhum peixe. É de se notar que os gráficos ainda são também afetados pela quantidade significante de zeros na variável resposta, o que torna difícil a análise de alguns fatores.



Por fim no gráficos 5, 6 e 7 foram feita uma análise das 3 variáveis que serão usadas no estudo com excesso de zeros para podermos entendê-las melhor. Analisamos a frequência com que os grupos consistiam de 1 a 4 pessoas, frequência da quantidade de crianças levadas por cada grupo, assim como a distribuição de quantos grupos levaram um trailer para acampar ou não. Os gráficos 5 e 7 parecem estar equilibrados, enquanto o 6 mostra como menos grupos escolheram levar uma quantidade maior de crianças para pescar.

Testando famílias padrões

Para melhor justificar o uso dos modelo ZIP e ZINB, primeiro tentamos utilizar os modelos de família Poisson e Binomial Negativa que estamos acostumados para analisarmos nossos dados.

Para ambos nossos modelos nós utilizamos todas as variáveis explicativas mencionadas (camper, persons e child) e todos os modelos, assim como as variáveis dentro deles, se demonstraram significativos a um nível de significância de 5%.

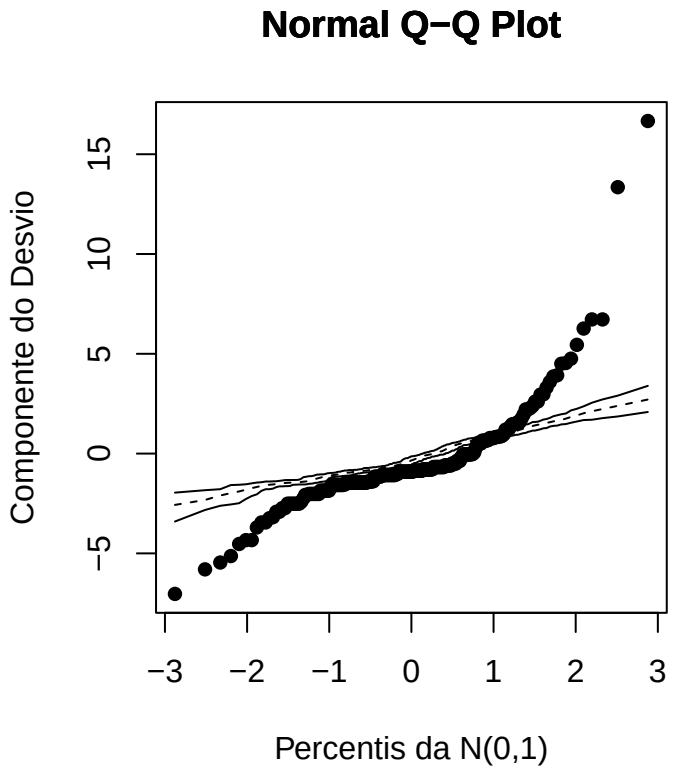
Para ambos, os links que deram o melhor resultado foram o link log, os quais nos deram as seguintes saídas e envelopes:

Poisson

	Coeficientes	Erro Padrão	P-valor	
Intercepto	-1.98183	0.15226	<2e-16	***
camper1	0.93094	0.08909	<2e-16	***
persons	1.09126	0.03926	<2e-16	***

	Coeficientes	Erro Padrão	P-valor	
child	-1.68996	0.08099	<2e-16	***

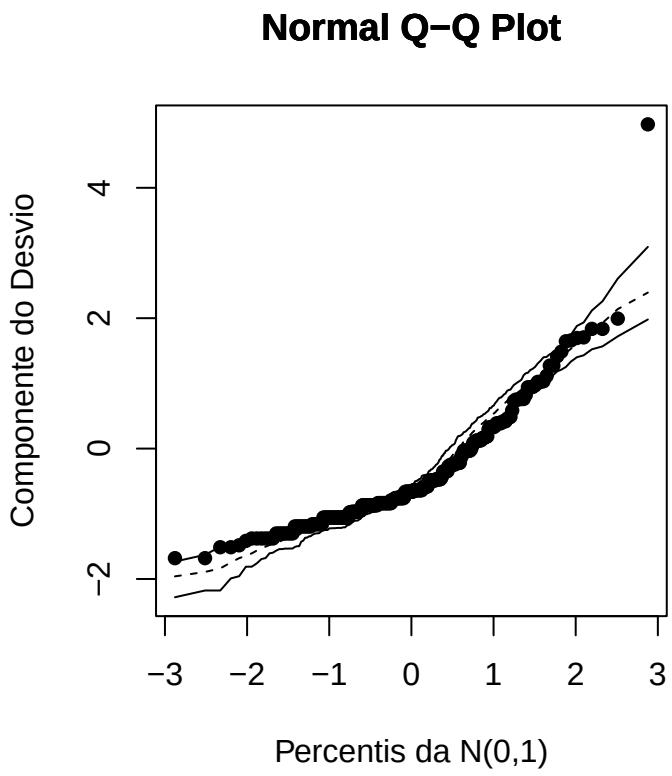
AIC: 1682.1



Binomial Negativa

	Coeficientes	Erro Padrão	P-valor	
Intercepto	-1.6250	0.3304	8.74e-07	***
camper1	0.6211	0.2348	0.00816	**
persons	1.0608	0.1144	<2e-16	***
child	-1.7805	0.1850	<2e-16	***

AIC: 820.44



Como podemos ver, nosso primeiro envelope pela família Poisson ficou péssimo, enquanto o da Binomial Negativa, apesar de melhor que o anterior, ainda ficou longe do desejado.

Modelos pra Zero Inflacinado

Partimos então para modelos feitos especificamente para cenários como este, em específico os modelos ZIP (regressão de Poisson Inclacionado de Zeros) e ZINB (regressão Binomial Negativa Inclacionada de Zeros)

Foi utilizado o pacote *pscl*, que nos permitiu criar os modelos necessários para análises que desejamos. Para ambos os modelos ZIP e ZINB foram utilizados as mesmas variáveis, as quais são específicadas nas tabelas 3 e 4. Todos nossos modelos, assim como suas variáveis, se apresentaram significatívos a um nível de significância de 5%.

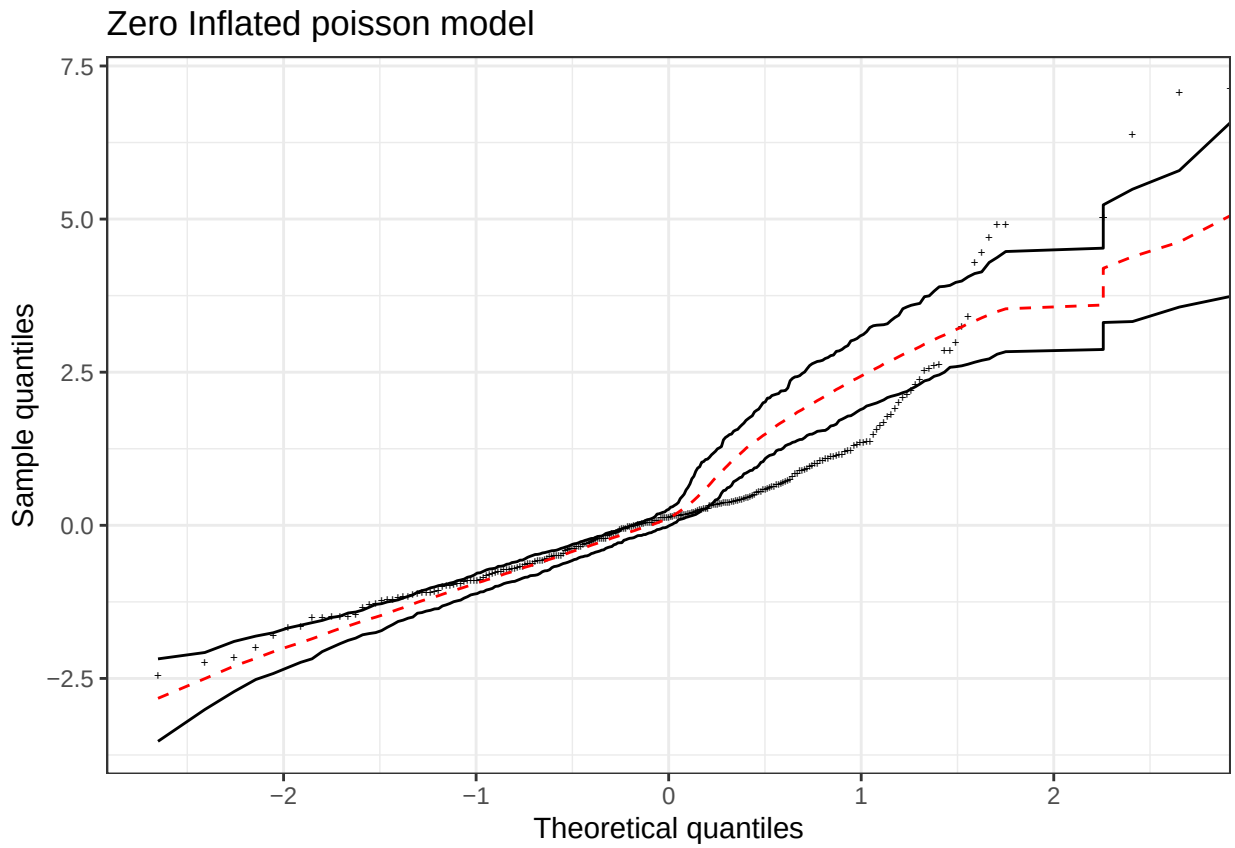
Coeficientes para o modelo de Contagem				
	Coeficientes	Erro Padrão	P-valor	
Intercepto	1.59789	0.08554	<2e-16	***
camper1	0.83402	0.09363	<2e-16	***
child	-1.04284	0.09999	<2e-16	***

Coeficientes para o modelo Inflacionado de Zeros				
	Coeficientes	Erro Padrão	P-valor	

Coeficientes para o modelo Inflacionado de Zeros				
Intercepto	1.2974	0.3739	0.000520	***
persons	-0.5643	0.1630	0.000534	***

A saída apresentada na tabela 3 se parece muito com a saída de duas regressões OLS. No primeiro bloco de saída, encontramos os coeficientes da regressão de Poisson para cada uma das variáveis, juntamente com erros padrão, e os p-valores para os coeficientes. Segue-se um segundo bloco que corresponde ao modelo de inflação. Isso inclui coeficientes logit para prever zeros em excesso junto com seus erros padrão, e p-valores.

Utilizamos do pacote *iccCounts*, que nos permitiu gerar os gráficos dos envelopes dos modelos Inflacionados de Zeros. Nosso modelo de Poisson Inflacionado de Zero pode ser visto no gráfico 8.



O envelope que conseguimos nos diz que nosso modelo não é o ideal para os dados em questão, como podemos ver uma quantidade considerável dos pontos fora do envelope.

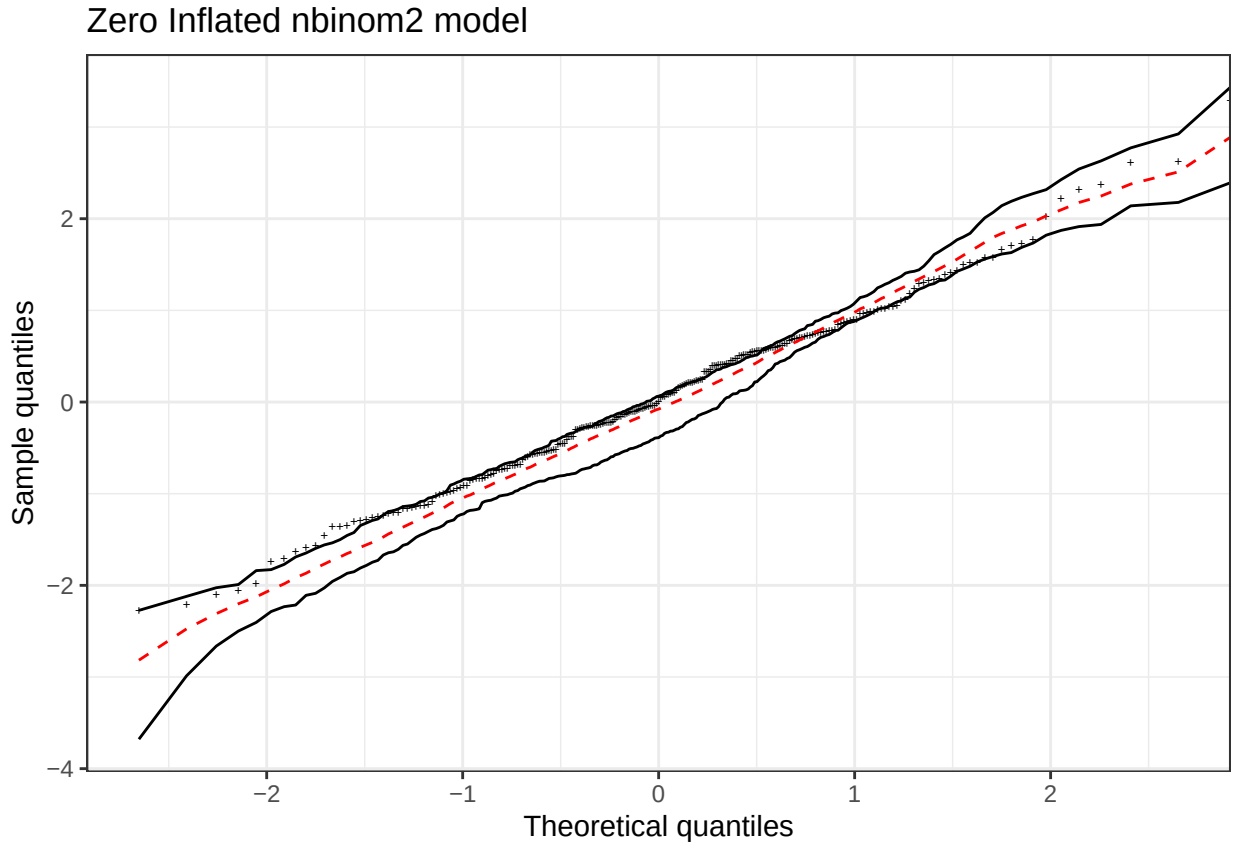
Partimos então para um modelo de Binomial Negativa Inflacionada de Zeros, apresentado na tabela 4.

Coeficientes para o modelo de Contagem				
_____-	Coeficientes	Erro Padrão	P-valor	
Intercepto	1.3710	0.2561	8.64e-08	***
camper1	0.8791	0.1956	9.41e-15	***
child	-1.5153	0.2693	0.0011	**
Log(theta)	-0.9854	0.1760	2.14e-08	***

Coeficientes para o modelo Inflacionado de Zeros				
_____-	Coeficientes	Erro Padrão	P-valor	
Intercepto	1.6031	0.8365	0.0553	.
persons	-1.6666	0.6793	0.0142	**

A saída apresentada na tabela 4 se parece muito com a saída de duas regressões OLS. No primeiro bloco de saída, encontramos os coeficientes da regressão Binomial Negativa para cada uma das variáveis, juntamente com erros padrão e os p-valores para os coeficientes. Segue-se um segundo bloco que corresponde ao modelo de inflação. Isso inclui coeficientes logit para prever zeros em excesso junto com seus erros padrão e p-valores.

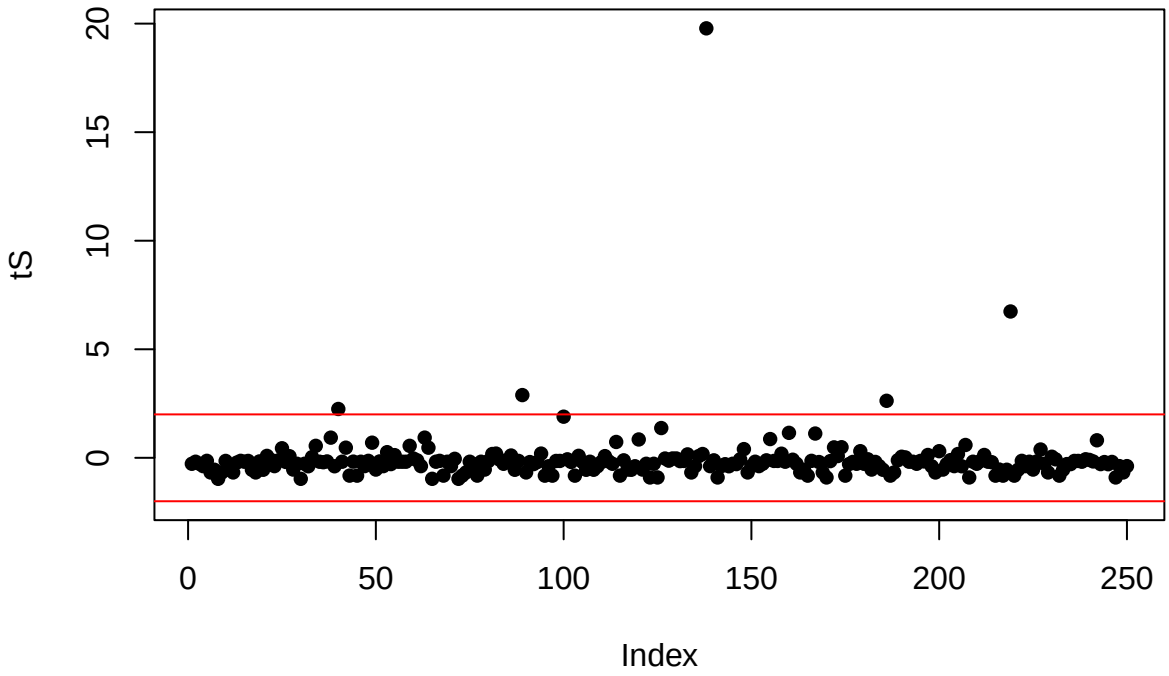
Novamente utilizamos do pacote *iccCounts* para desenvolver o gráfico 9.



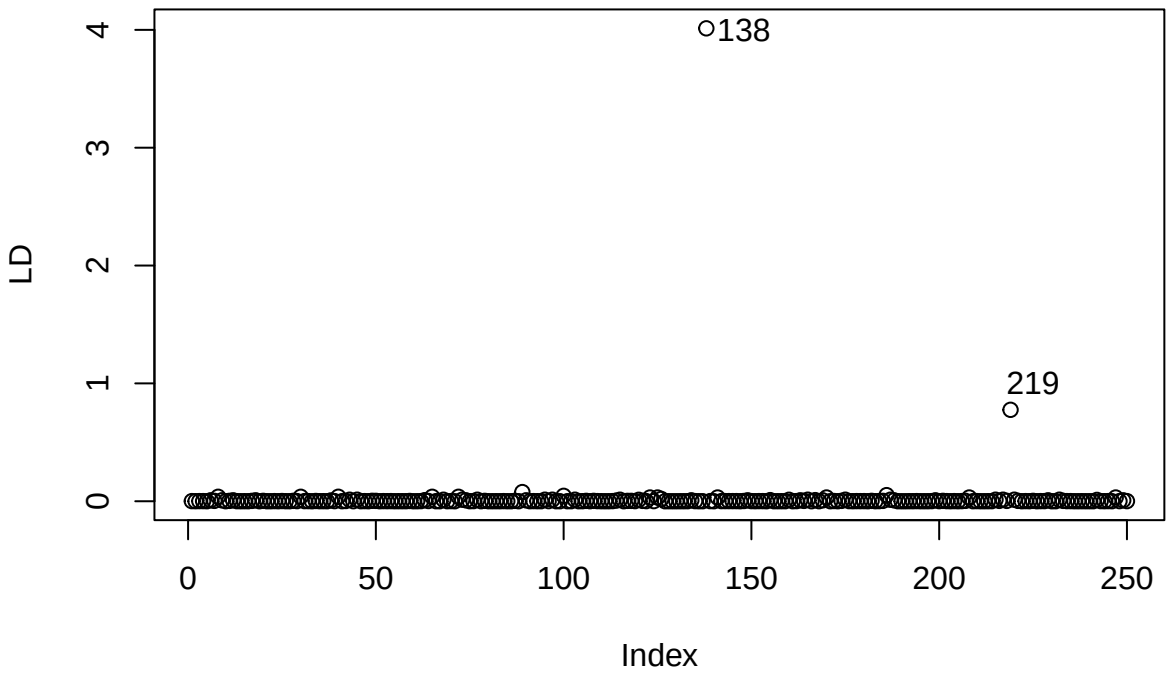
Podemos notar pelo gráfico 9 que o envelope do modelo binomial negativo inflacionado de zeros é consideravelmente melhor para os nossos dados, com pouquíssimos pontos fora do envelope relativo ao total de observações.

Análise de Resíduos e Diagnóstico

Com nosso modelo ideal encontrado, o próximo passo ideal seria analisarmos seus resíduos e realizarmos uma análise de diagnóstico para validarmos nossa escolha. O gráfico 10 apresentam os resíduos do nosso modelo Binomial Negativo Inflacionado de Zeros.



Apesar dos pontos muito distantes de 0, que serão discutidos em nossa análise de diagnóstico, a quantidade de pontos fora do intervalo de confiança de -2 a 2 não é significativo para a quantidade de observações que temos. Da mesma forma, nossos pontos não apresentaram nenhum padrão visível, o que nos faz acreditar que os resíduos estão em boas condições.



Ao realizarmos uma análise de diagnósticos apresentada no gráfico 11, notamos duas observações que nos chamam atenção entre todas, o que nos faz ter de analisá-las individualmente.

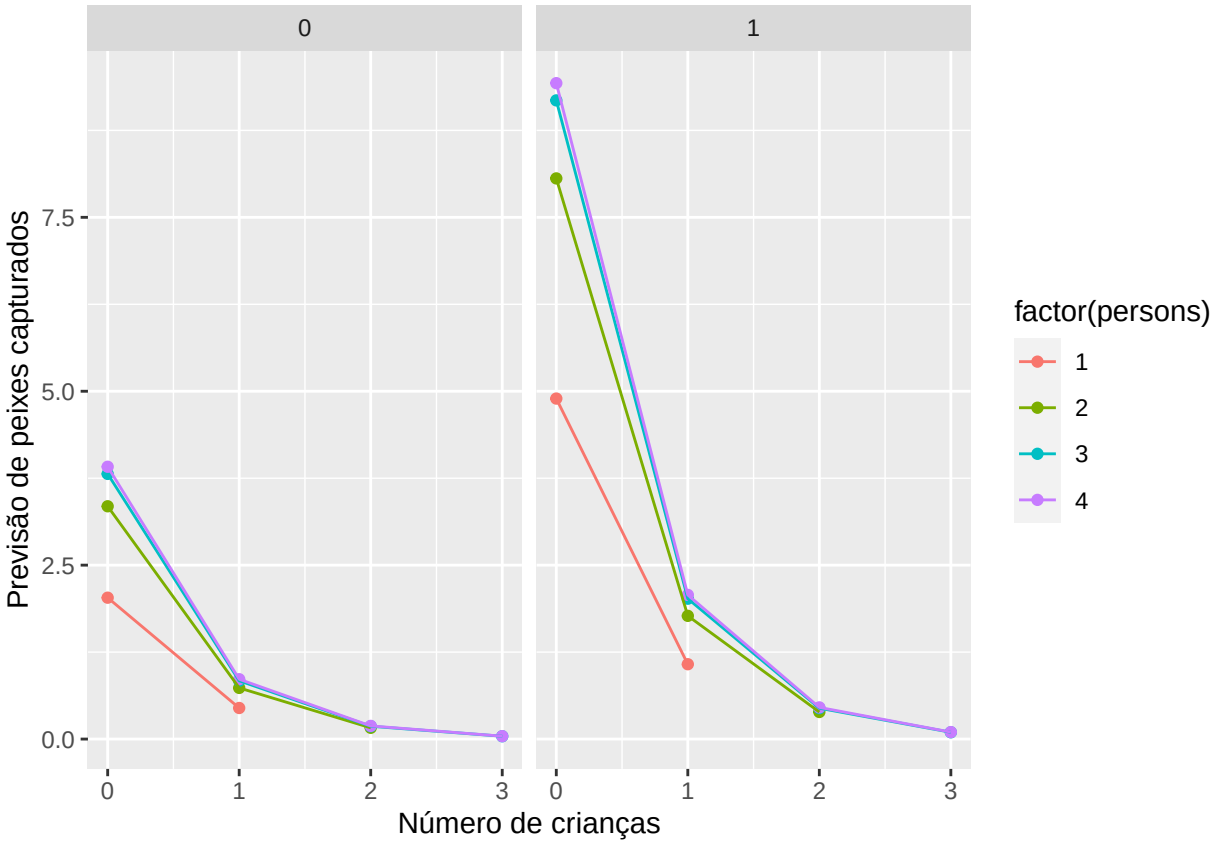
O grupo de pescadores 138 foi um dos grupos que pescou a maior quantidade de peixes, mas diferente dos demais ele era consistido de apenas 3 pessoas, enquanto todos outros que pescaram por volta dessa quantidade consistiam de 4. O grupo também possuía uma criança entre eles, o que também os difere de todos outros que pescaram em grandes quantidade. E por fim eles não levaram um trailer para o parque, outra característica vista apenas neles e em nenhum outro grupo que pescou tanto.

O grupo 219, por outro lado, pescou apenas 5 peixes, mas o que os difere tanto do resto é que eles levaram 2 crianças para o parque. Considerando que a média de peixes capturados por grupos que levaram 2 crianças é consideravelmente próxima de 0, isso fez com que eles se destacassem entre as observações.

Aplicação

Agora com os melhores modelos selecionados e estudados, podemos aplicá-los em nossas análises para melhor prever e estimar a quantidade de peixes que um grupo conseguiria pegar em uma ida para o parque, considerando as variáveis estudadas.

O gráfico 12 nos apresenta essa previsão. O gráfico é dividido no quesito se o grupo foi ou não para o parque em um trailer, o eixo X indica a quantidade de crianças que consistiam o grupo, a diferentes linhas nos mostram os tamanhos dos grupos em si e por fim nosso eixo Y indica a previsão de quantos peixes o grupo conseguiria pescar.



É justo dizer que a maior probabilidade de ser conseguir um número grande de peixes é indo com um grupo de 4 pessoas, utilizando-se de um trailer, e sem a presença de crianças entre os integrantes.