

Universidade Estadual de Maringá
Modelos Lineares Generalizados

Modelos de regressão para dados de contagem e proporção: aplicação aos dados Somerville e IDHM

Wesley Oliveira Furriel RA:61493

Professora Dra. Rosangela Gentirana Santa

Maringá
2016

Conteúdo

1	Introdução	2
2	Modelos Lineares Generalizados	2
3	Modelo Beta	3
3.1	Análise exploratória	5
3.2	Seleção da função de ligação	10
3.3	Análise dos resíduos	12
3.4	Modelo de regressão beta ajustado	16
4	Modelo de regressão para dados de contagem	19
4.1	Análise exploratória	19
4.2	Ajuste dos modelos Poisson e Binomial Negativa	22
4.3	Ajuste dos modelos ZIP e ZINB	25
4.4	Modelo de contagem final	30
5	Considerações finais	33
6	ANEXOS R	34

1 Introdução

No presente trabalho nos ocupamos em realizar dois modelos de regressão linear generalizada, um modelo para dados discretos oriundos de contagem, pertencente a família exponencial canônica e um modelo para dados contínuos, de proporção, pertencente a família exponencial biparamétricas. No que diz respeito ao modelo de contagem, realizamos a modelagem dos dados banco de dados Somerville, de 1980, contendo 659 observações, em que buscamos estudar as variáveis que explicam o número de visitas ao Lago. Quanto ao modelo contínuo, analisamos o IDHM(Índice de Desenvolvimento Humano Municipal) a partir de indicadores socioeconômicos e de desigualdade social, os dados neste caso são de 2010 e o banco contém 4063 observações. Em ambos os casos foram aplicadas técnicas de estatística descritiva, para visualização do comportamento dos dados que compõe o banco, bem como, alguns critérios de seleção para a escolha do modelo que melhor representaram os dados.

Para atingir os objetivos desejados contamos com auxílio dos pacotes estatísticos Rstudio 1.0.136, com os pacotes betareg e tidyverse e SAS 9.4 com as proc genmod e proc sgplot, os resultados são apresentados nas seções seguintes.

2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados(MLG) foram propostos para aplicações onde a variável resposta y_i pode ser representada por alguma distribuição da família de exponencial, univariada como por exemplo as distribuições Normal, Binomial, Binomial Negativa, Gama, Poisson e Normal Inversa. Sendo uma extensão dos Modelos Lineares simples os MLG envolvem uma variável resposta canônica, variáveis explicativas e uma amostra aleatória de n observações. Para o ajuste de um MLG precisamos levar em consideração 3 componentes:

- i) **Componente Aleatório:** é a variável resposta do modelo que deve ter uma distribuição pertencente à família exponencial na forma canônica.

Sendo representado por um conjunto de variáveis aleatórias independentes Y_1, \dots, Y_k provenientes de uma mesma distribuição que faz parte da família exponencial canônica com médias μ_1, \dots, μ_k , ou seja

$$E(Y_i) = \mu_i, i = 1, 2, \dots, n \quad (1)$$

um parâmetro constante de escala, conhecido $\phi > 0$ e que depende de um único parâmetro θ_i , chamado parâmetro canônico ou **parâmetro de locação**. A f.d.p. de Y_i é dada por

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{1}{a_i(\phi)} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi) \right\} \quad (2)$$

sendo $b(\cdot)$ e $c(\cdot)$ funções conhecidas. Em geral, $a_i(\phi) = \frac{\phi}{w_i}$ sendo pesos a priori

- ii) **Componente Sistemático:** são as variáveis explicativas, que entram na forma de uma soma linear de seus efeitos.

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j = \mathbf{x}_i^T \boldsymbol{\beta} \quad \text{ou} \quad \eta = \mathbf{X} \boldsymbol{\beta} \quad (3)$$

sendo $\mathbf{X} = (x_1, \dots, x_n)^T$ a matriz do modelo. $\beta = (\beta_1, \dots, \beta_p)^T$ o vetor de parâmetros e $\eta = (\eta_1, \dots, \eta_n)^T$ o preditor linear. De modo geral podemos definir o preditor linear como

$$\eta = \beta_0 + \beta_1 x_1, \dots, \beta_k x_k \quad (4)$$

- iii) **Função de ligação:** a ligação entre os componentes aleatório e sistemático é feita através de uma função (por exemplo, logarítmica para os modelos log-lineares). Desse modo, uma função que liga o componente aleatório ao componente sistemático, ou seja, relaciona a média ao preditor linear, isto é,

$$\eta_i = g(\mu_i) = g(\beta_0 + \beta_1 x_{i1}, \dots, \beta_k x_{ik}) \quad (5)$$

sendo $g(\cdot)$ uma função monótona (preserva a relação de ordem), derivável.

Precisamos que conjunto o de parâmetros β_1, \dots, β_p sejam uma combinação linear igual a alguma função do valor esperado de Y_i conjunto menor de parâmetros tais que uma combinação linear dos seja igual a alguma função do valor esperado de Y

- Distribuição da variável resposta;
- matriz modelo;
- função de ligação

Se a função de ligação é escolhida de tal forma que $g(\mu) = \theta$ o preditor linear modela diretamente o parâmetro canônico e tal função de ligação é chamada ligação canônica. Assim, as estatísticas dessa ligação são garantidamente suficientes. As funções de ligação canônicas são apresentadas na tabela abaixo

3 Modelo Beta

O modelo de regressão Beta é indicado para situações em que a variável resposta Y é medida continuamente num intervalo $0 < T < 1$, ou em intervalos em que seja possível identificar valores limitados por um máximo e um mínimo. Ferrari e Cribari-Neto (2004) propõe uma reparametrização da distribuição Beta para sua utilização no modelo de regressão. Já que, na análise de regressão, geralmente é mais útil modelar resposta média, então, a fim de obter uma estrutura de regressão para tal, juntamente com um parâmetro de dispersão, é interessante trabalhar com uma parametrização diferente da densidade beta original. O modelo segue as propriedades da Beta, sendo então, adequado para casos em que a variável resposta Y é medida continuamente no intervalo $0 < Y < 1$. Seguindo a nova parametrização, $\mu = \alpha/(\alpha + \beta)$ e $\phi = \alpha + \beta$, ou seja, $\alpha = \mu\phi$ e $\beta = (1 - \mu)\phi$. Isto posto, a média e a variância de Y são dadas pelas seguintes expressões,

$$\mathbb{E}(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \frac{V(\mu)}{(1 + \phi)} \quad (6)$$

em que $V(\mu) = \mu(1 - \mu)$, de modo que μ é a média da variável resposta e ϕ pode ser interpretado como um parâmetro de dispersão ou precisão. Tendo em vista essa nova parametrização em termos de μ e ϕ , Y tem-se uma função de densidade Beta dada por:

$$f(y | \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1 - \mu)\phi)} y^{\mu\phi-1} (1 - y)^{(1 - \mu)\phi-1}, \quad 0 < y < 1 \quad (7)$$

com $0 < \mu < 1$ e $\phi > 0$.

Assim como na parametrização original é possível obter diferentes formas no comportamento da distribuição de acordo com os valores de seus parâmetros μ e ϕ . Ferrari e Cribari-Neto (2004) apontam que a distribuição pode ser simétrica quando $\mu = 1/2$ e assimétrica quando $\mu \neq 1/2$. Além disso, a dispersão da distribuição, para um μ fixado diminui quando os valores de ϕ aumentam.

Sendo Y_1, \dots, Y_n uma a.a. da resposta, que segue uma densidade Beta com média μ e parâmetro de dispersão desconhecido ϕ , o modelo de regressão beta tem a seguinte forma:

$$\eta_t = g(\mu_t) = \sum_{i=1}^k x_{ti}\beta_i = x'_t\beta \quad (8)$$

em que $\beta = (\beta_1, \dots, \beta_k)'$ é um vetor de parâmetros desconhecido e x'_t é um vetor de observações fixas e conhecidas de k covariáveis. No que tange a função de ligação ela é contínua e duas vezes diferenciável, pertencendo ao intervalo $(0, 1)$ e η_t preditor linear. Além disso, o parâmetro de precisão ϕ_t

$$h(\phi_t) = \sum_{j=1}^q z_{tj}\gamma_j = \nu_t \quad (9)$$

em que ν_t é o preditor linear.

No modelo de regressão beta, pode-se considerar o método de máxima verossimilhança, em que os parâmetros são estimados através da maximização da função de verossimilhança.

$$L(\beta) = \prod_{t=1}^n f(y_t|\beta) \quad (10)$$

em que β é um vetor de parâmetros a ser estimado, y_t , $t = 1, 2, \dots, n$ são valores observados da variável resposta Y e $f(y_t|\theta)$ é a função densidade de probabilidade para a variável aleatória Y .

3.1 Análise exploratória

Nesta seção iremos realizar uma análise exploratória dos dados do banco IDHM, dessa forma, as variáveis investigadas podem ser observadas na tabela abaixo.

Tabela 1: Dicionário de dados do banco IDHM

Código	Label
IDHM	Índice de Desenvolvimento Humano Municipal
PINDCRI	Crianças extremamente pobres
AGUA_ESGOTO	Abastecimento de água e esgotamento sanitário inadequados
T_FLSUPER	Taxa de frequência líquida ao ensino superior
MORT1	Mortalidade até um ano de idade
GINI	Coeficiente de Gini
regiao	Grandes regiões do Brasil

Como é possível constatar para investigar o IDHM iremos utilizar seis covariáveis que descrevem características socioeconômicas de desigualdade, saúde e saneamento básico dos municípios brasileiros, com exceção da variável região que é categórica, todas as demais são contínuas e limitadas entre 0 e 1.

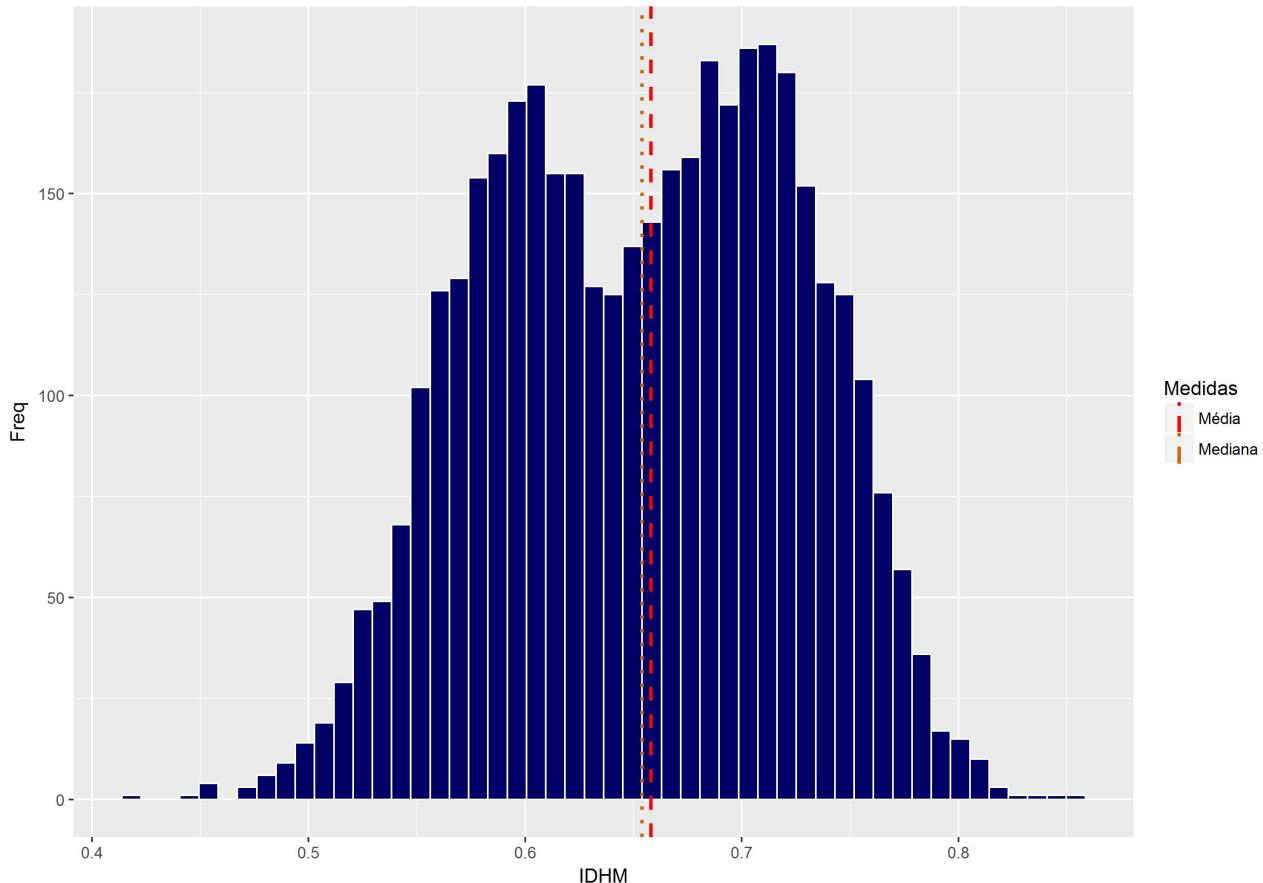


Figura 1: Histograma de frequência para o IDHM

No que tange o comportamento da variável resposta IDHM sem considerar as covariáveis, verifica-se que sua distribuição é bimodal apresentando dois pontos de máximo, sendo a média para o Brasil de 0.65 e a mediana 0.658, valores bastante próximos.

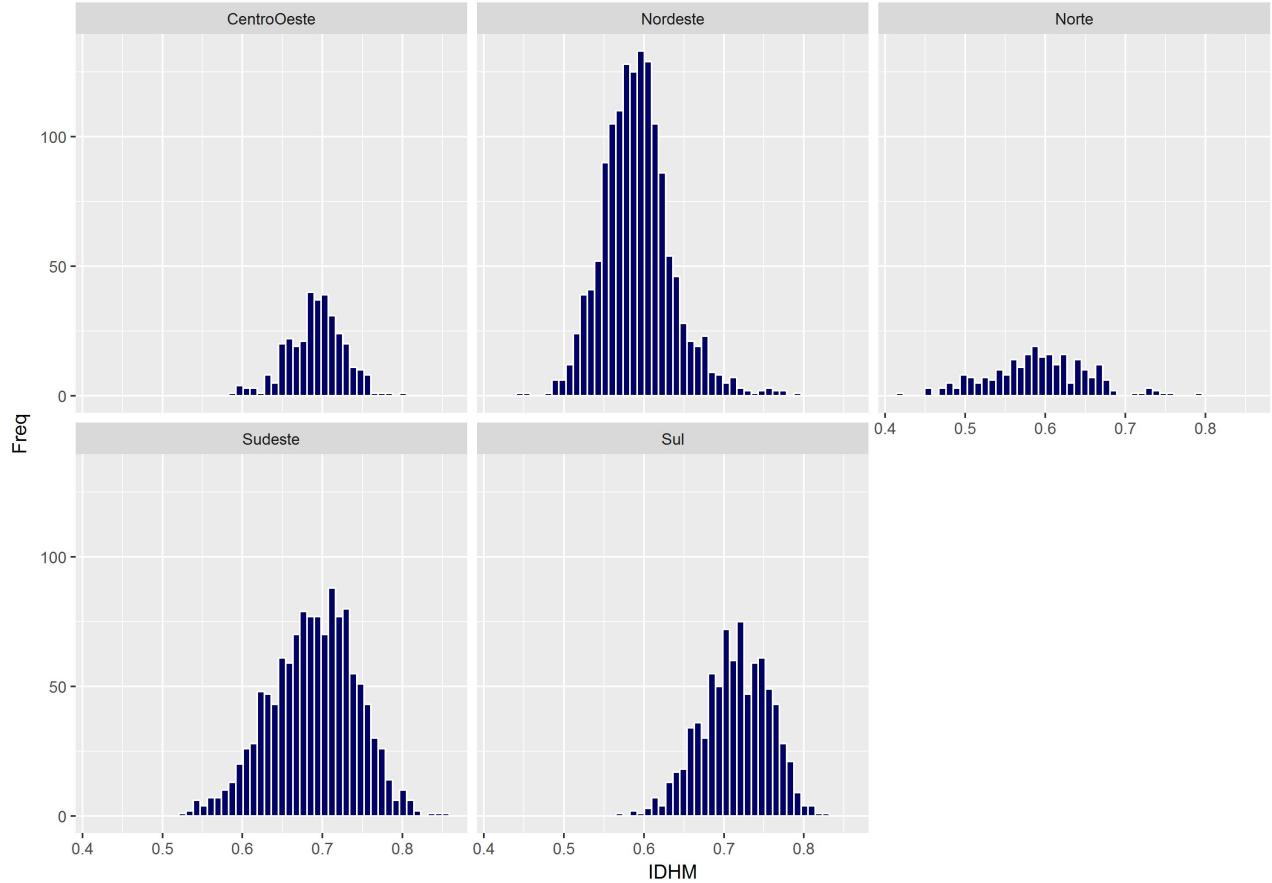


Figura 2: Histograma de frequência para o IDHM por região

Nos histogramas expostos na figura 2 adicionamos o fator de divisão do banco, ou seja, as grandes regiões do país. Pelos comportamentos observados é possível visualizar possíveis ajustes a partir da distribuição Beta. Tendo isso em vista utilizamos a função fitdistr dos pacotes MASS do R, foi possível obter os parâmetros de média e precisão da Beta reparametrizada para cada uma das curvas que serão apresentada abaixo.

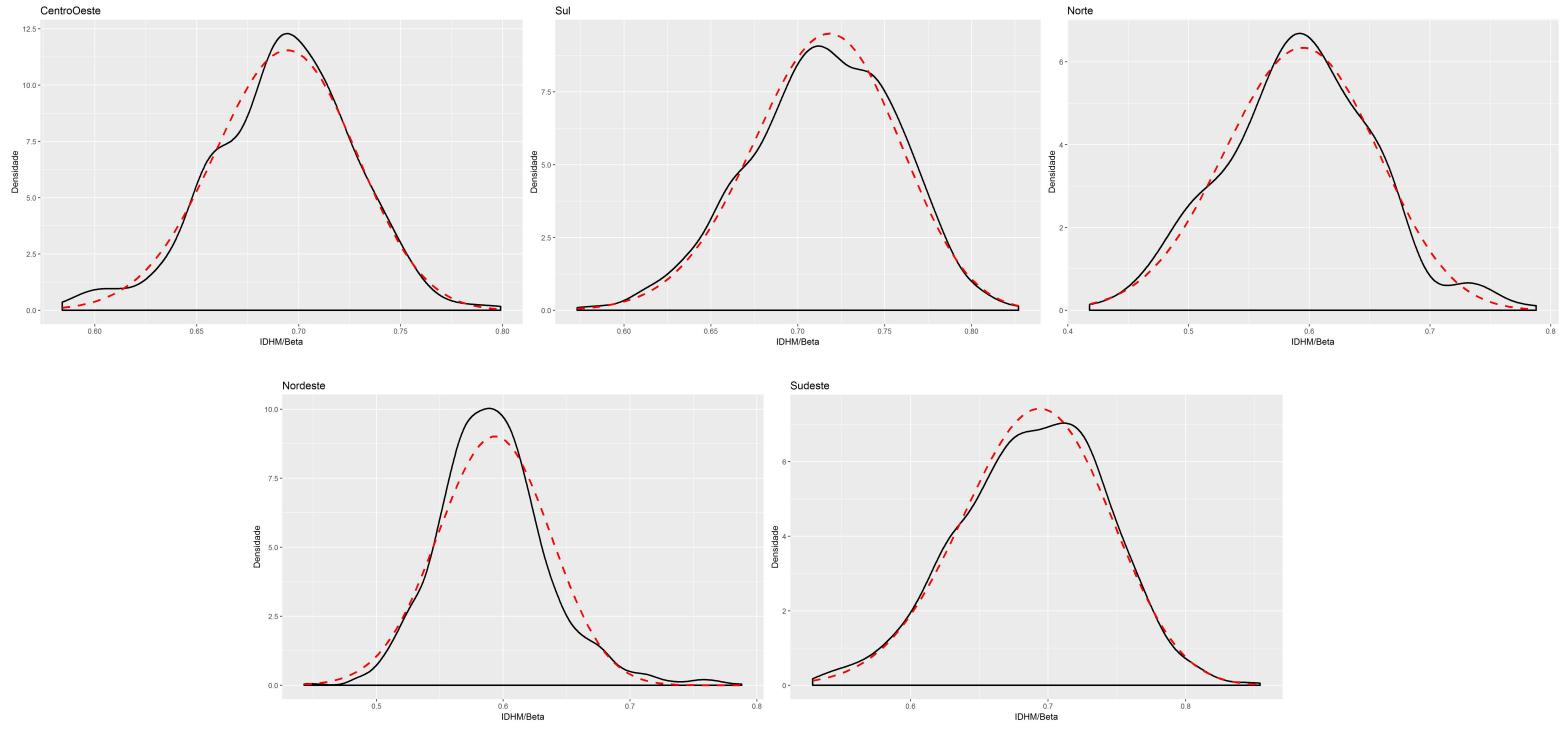


Figura 3: Ajuste teórico para beta segundo as regiões

Como é possível verificar por 3 a curva teórica se ajusta relativamente bem a observada para maior parte das regiões, com exceção do Nordeste em que ela não conseguir captar de forma precisa a forma assumida. Mesmo assim, de modo geral as curvas teóricas e observadas apresentam tendências bastante semelhantes, isto é, a distribuição Beta é uma boa escolha para o ajuste destes dados.

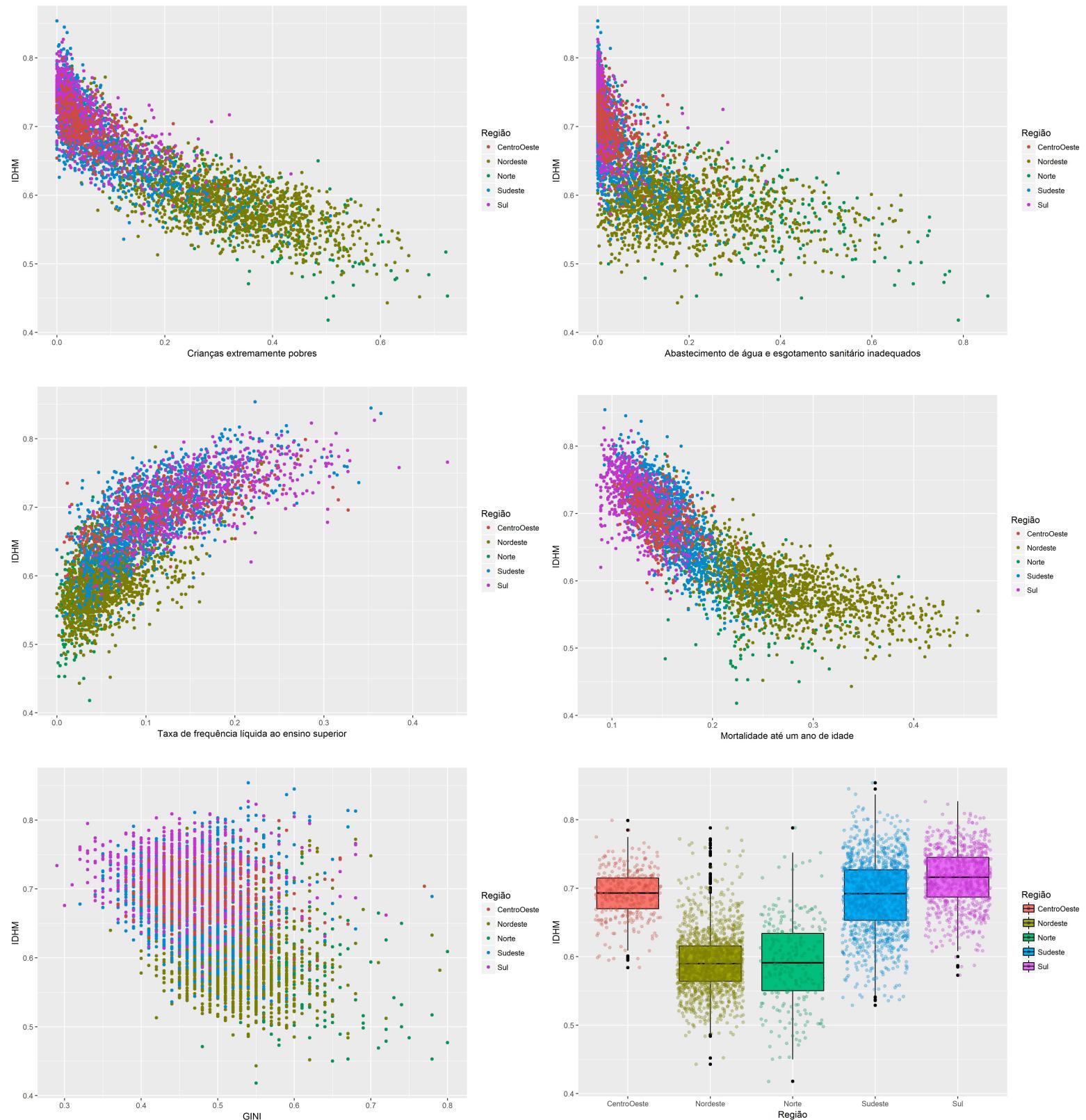


Figura 4: Resposta vs covariáveis

A figura 4 apresenta os gráficos de dispersão para a resposta vs as covariáveis. Pelos resul-

tados pode-se pontuar que nos gráficos para as Crianças extremamente pobres e Mortalidade até um ano de idade vs o IDHM, há uma relação negativa entre os dados, ou seja, conforme diminui o IDHM estas duas variáveis aumentam seus índices. Além disso, observamos que as regiões Sul, Centro-Oeste e Sudeste apresentam maiores valores do IDHM e menores valores as covariáveis de vulnerabilidade infantil. O abastecimento de água e esgoto inadequado segue um comportamento similar. Já no que tange a frequência ao ensino superior observa-se que quanto maior o IDHM, maior o acesso ao educação superior, além disso, Sul, Centro-Oeste e Sudeste mostram-se novamente com melhores indicadores quando comparados ao norte e nordeste. Aliás, as condições destas duas regiões diante do IDHM podem ser claramente vistas pelo gráfico de BoxPlots, em aparecem com a média bastante inferior as demais regiões.

3.2 Seleção da função de ligação

Antes de iniciarmos a análise do modelo de regressão Beta, foi realizado um estudo com o intuito de identificar a função de ligação mais adequada para os nossos dados, para tal foi realizada uma análise gráfica dos resíduos dos modelos segundo cada uma das ligações e por fim, o teste RESET(regression specification error test, Ramsey 1969).

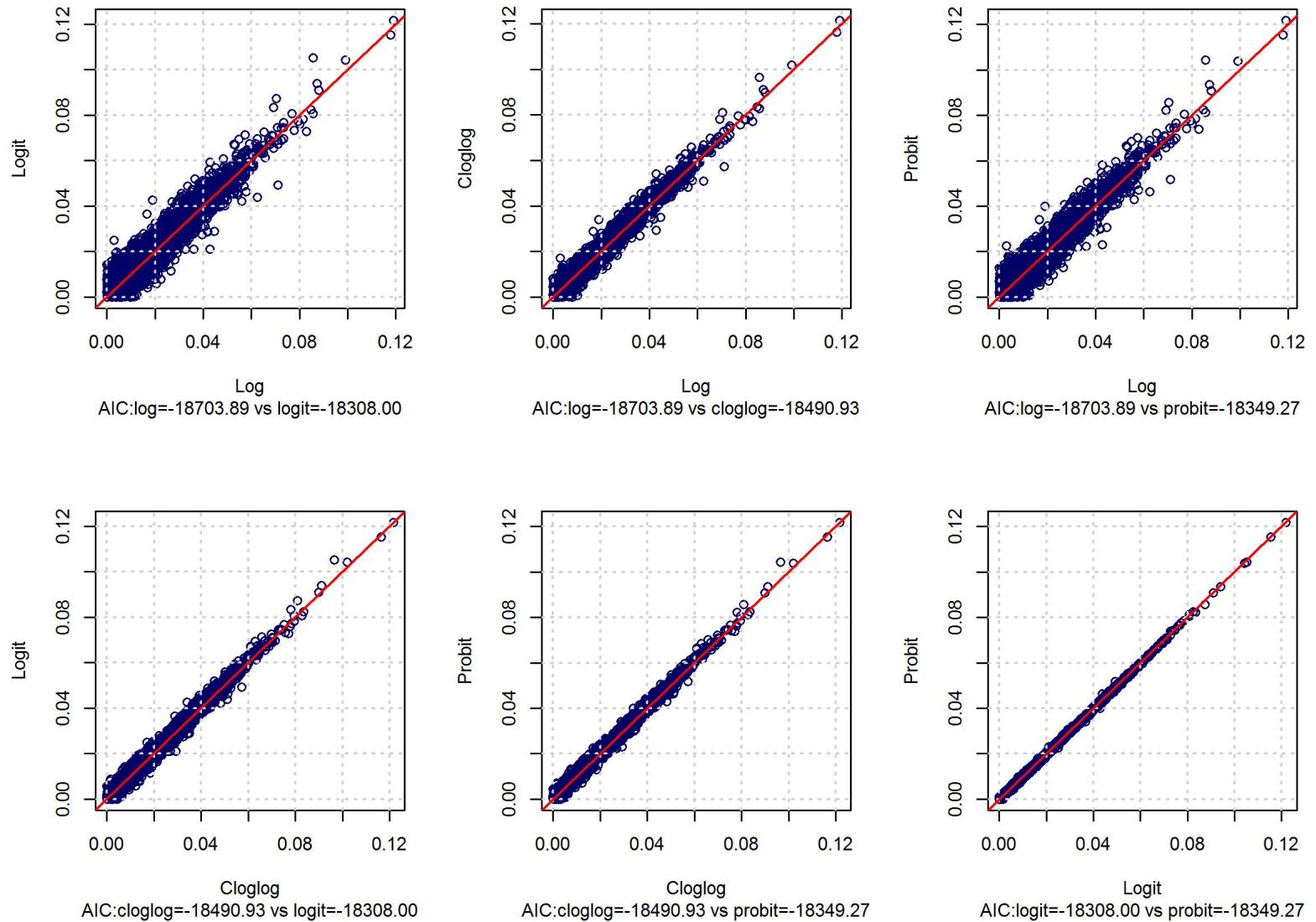


Figura 5: Comparação dos resíduos segundo as funções de ligação

Nos gráficos expostos na figura 5 foram analisados os resíduos dos modelos com cada uma das funções de ligação disponível, gerando-se assim, comparações dois a dois. A ideia aqui é verificar qual das ligações produziu erros maiores e tal efeito é facilmente verificado quando realizamos uma linha de corte passando pela origem do plano cartesiano. A partir disso, pode-se verificar o eixo que apresentou uma maior concentração de pontos, este será o modelo com a ligação que retornou maiores erros. Desse modo, constatamos que as ligações

"cloglog" e "log" mostraram os menores erros quando comparadas as demais.

Tabela 2: Teste RESET

Ligaçāo	ℓ	X^2	Valor p
Log	9367.9	191.66	<0.0001
Cloglog	9261.5	1342.7	<0.0001
Logit	9190.6	531.36	<0.0001
Probit	9190.6	1342.7	<0.0001

Pelo teste RESET de Ramsey constatamos que todas as ligações selecionadas foram consideradas como mal específicas pelo teste. Provavelmente, tal resultados se dá devido ao número bastante grande de observações, gerando graus de liberdade elevados para a comparação com a X^2 . Desse modo, para verificar qual destas ligações fornece o melhor ajuste, realizamos gráficos com envelopes de confiança para a comparação.

Comparando os gráficos expostos na figura 6 é possível notar que o ajuste do modelo Cloglog é um pouco superior as demais, pois apresenta uma menor quantidade de pontos fora da banda de confiança. Dessa forma, seguimos com as análises considerando o modelo de regressão Beta com função de ligação complemento log-log em que $g(\mu) = \log(-\log(1-\mu)) = X\beta$, sendo assim $\mu = 1 - \exp(-\exp(X\beta))$.

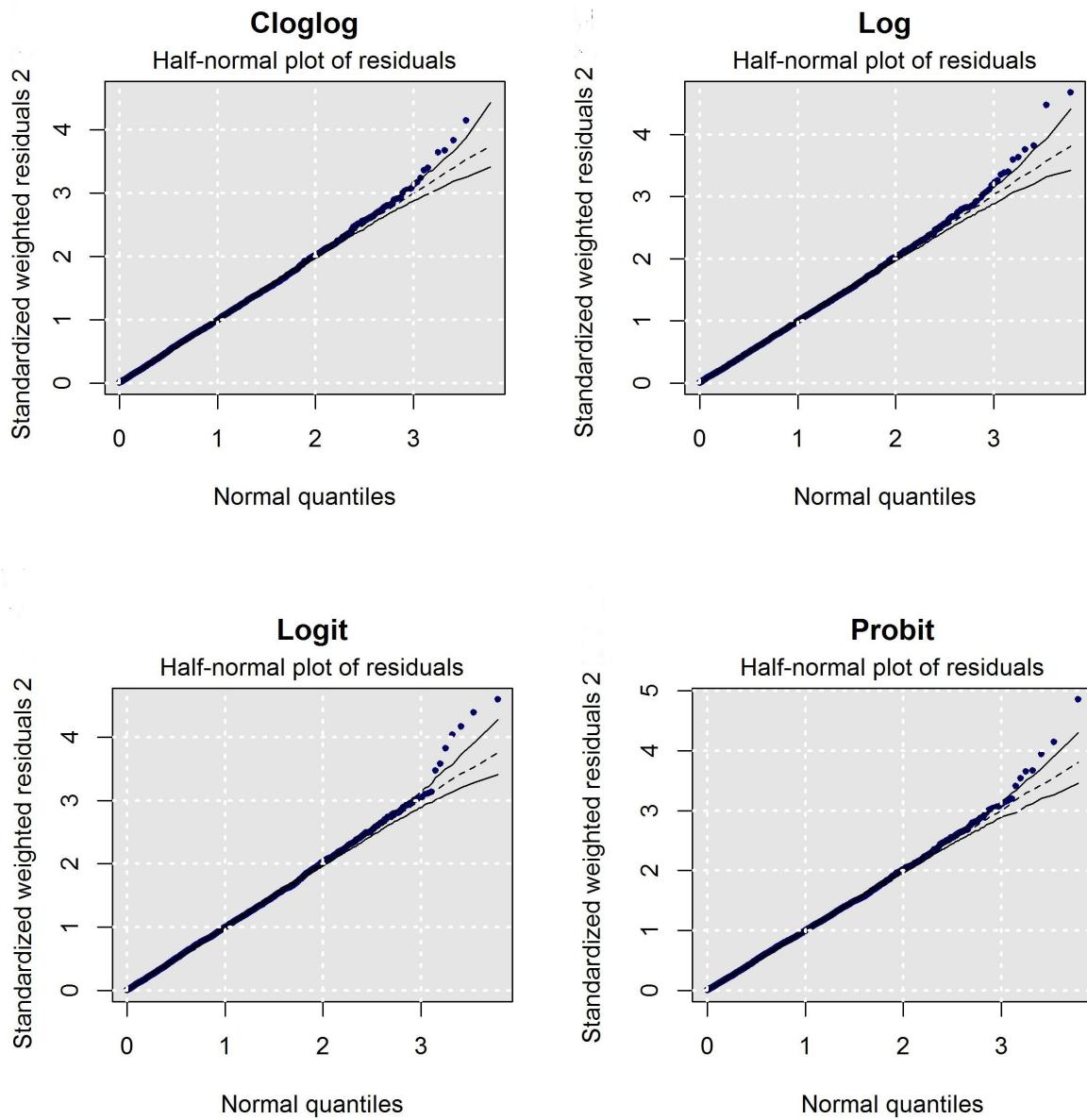


Figura 6: Comparação do ajuste dos modelo via envelope simulado

3.3 Análise dos resíduos

Nesta subseção nos ocupamos em realizar uma discussão sobre a adequabilidade do modelo com base nos resíduos. Sendo uma das fases mais importantes do processo de modelagem, uma vez que, as suposições que serão feitas sobre o modelo ajustado precisam ser validadas para que as inferências sejam confiáveis. Para a investigação gráfica dos resíduos utilizamos os resíduos propostos por Espinheira, Ferrari, and Cribari-Neto (2008b) chamados de standardized weighted residual 2, dados por:

$$r_i^{SW2} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\nu}_i(1 - h_{ii})}} \quad (11)$$

em que $y_i = \log(y_i/(1 - y_i))$ e $\mu_i = F(\mu_i, \phi) - F((1 - \mu_i), \phi)$, sendo $F(\cdot)$ uma função digamma.

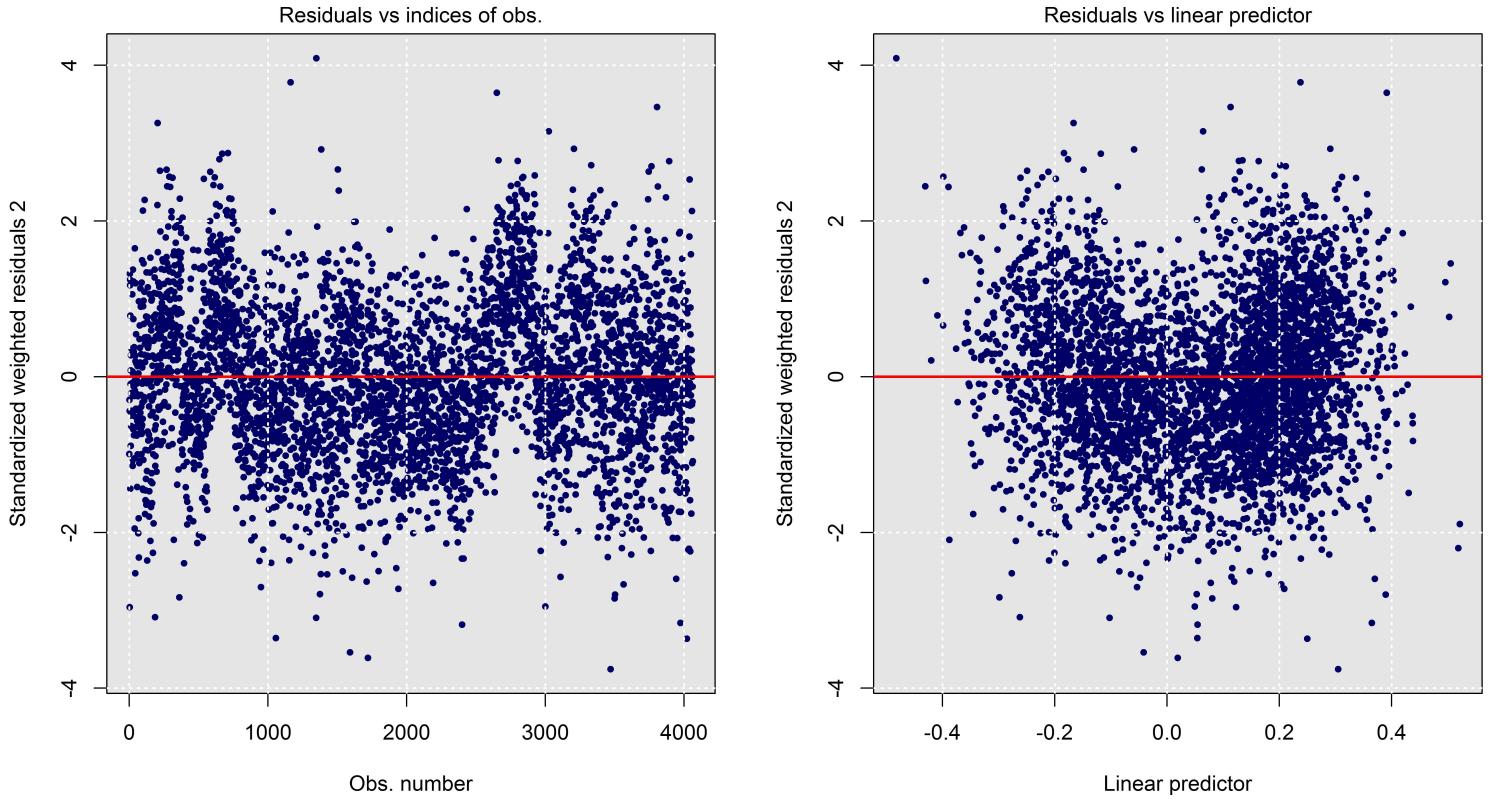


Figura 7: Resíduos de sw2

Partindo disso, pelos gráficos expostos na figura 7 pode-se notar que para gráfico o preditor linear vs os resíduos ponderados padronizados verificamos uma distribuição aparentemente aleatória em torno de zero, ou seja, não há uma tendência clara nos resíduos. Quanto ao primeiro gráfico, mesmo que neste caso, não faça muito sentido verificar se os resíduos são independentes, foi elencada está técnica gráficas para fins de aprendizagem. A seguir, temos o diagnóstico de independência por essas duas formas. Desse modo, podemos concluir que eles se organizam de forma aleatória, apontando para a não existência de homoscedasticidade.

Pelo gráfico de Leverage observamos que existem alguns pontos influentes de alavancagem que precisam ser investigados e pela distância de Cook observamos 4 pontos que podemos considerar como outliers. Para verificar se estas observações influenciam nas estimativas e conclusões obtidas, realizamos sua exclusão do banco e as comparamos com os resultados do modelo completo, contendo todas as observações. Os resultados obtidos foram bastante similares, desse modo, não constatamos a necessidade de sua retirada na modelagem.

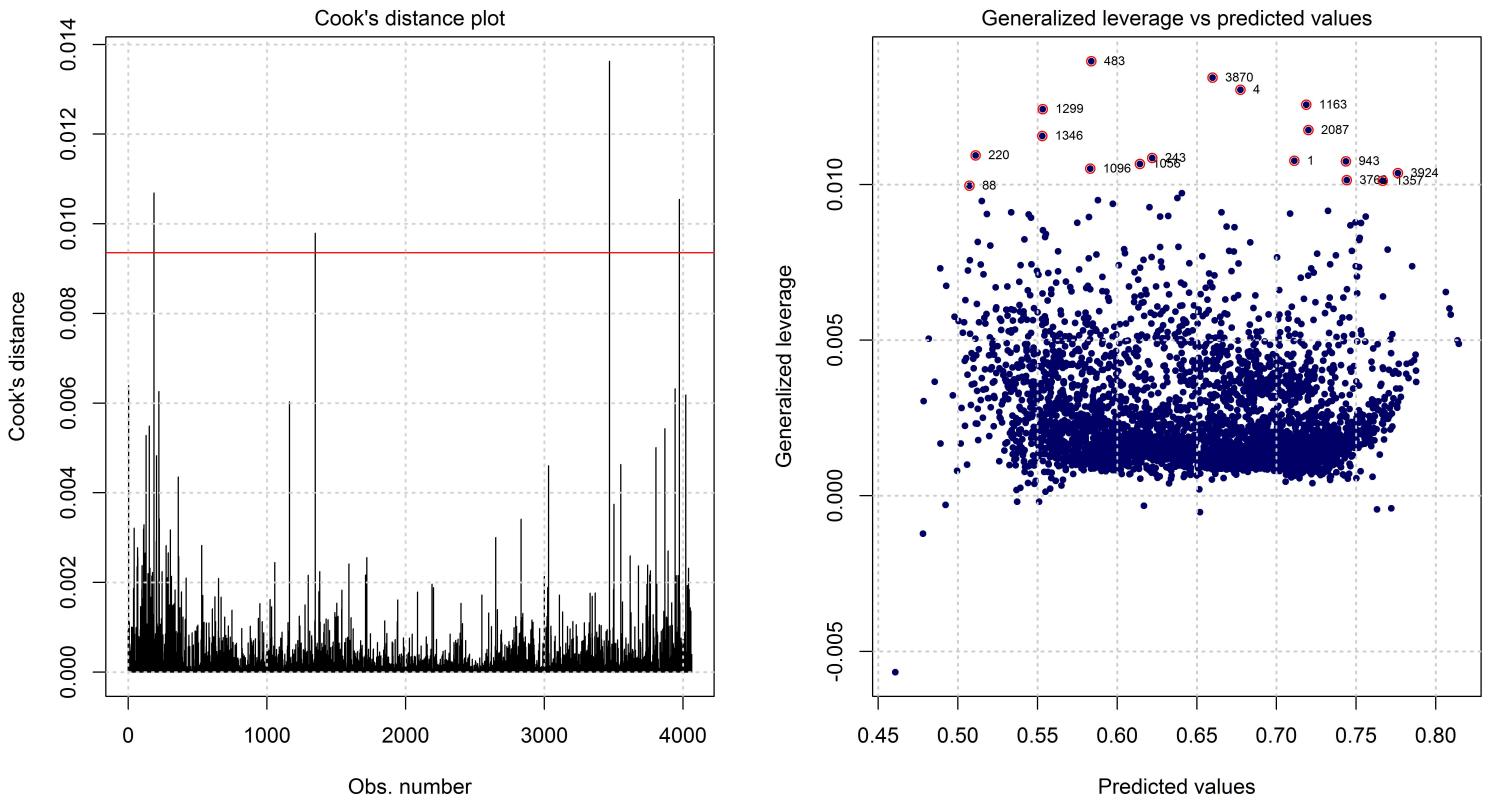


Figura 8: Distância de Cook e Leverage

Aqui, temos o gráfico dos valores observados vs os ajustados e o envelope simulado de confiança, expostos na figura 9. Pelos resultados, constata-se que os valores ajustados seguem próximos aos observados, em uma tendência aparentemente linear.

E pelo gráfico de envelope simulado, constatamos que há alguns pontos que se distanciam da banda de confiança ao fim da reta apenas, em decorrência dos valores de alavancagem listados acima.

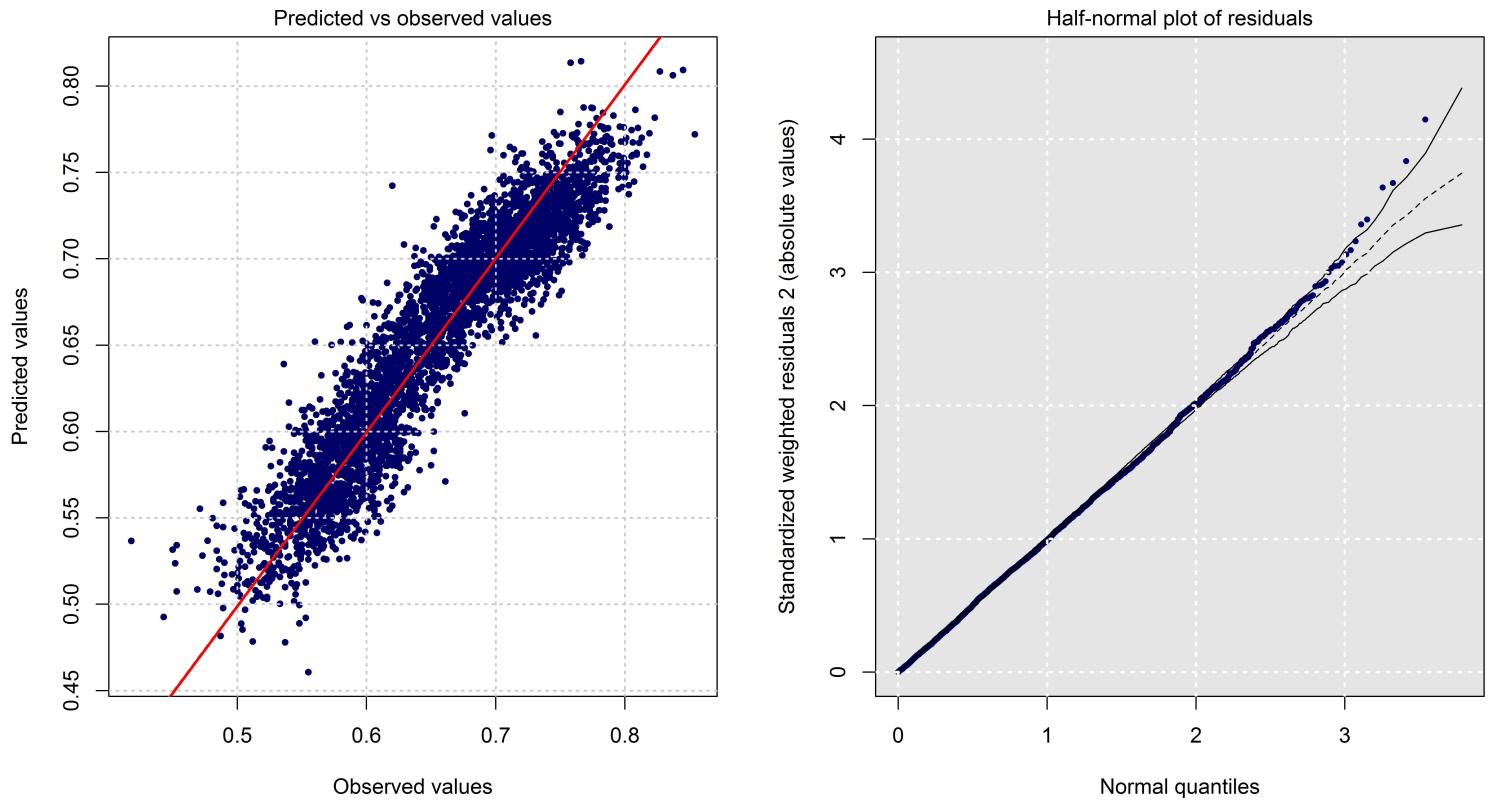


Figura 9: Envelope simulado e observados vs ajustados

Por fim, gráfico visto em 10 em que verificamos o envelope simulado sem os pontos de alavancagem, nota-se que o ajuste é um pouco melhor quando comparado ao anterior. Mesmo diante disso, seguiremos com o modelos em que consideramos o banco completo, desse modo, temos indícios de que a distribuição beta possa está adequada para explicar nossos dados.

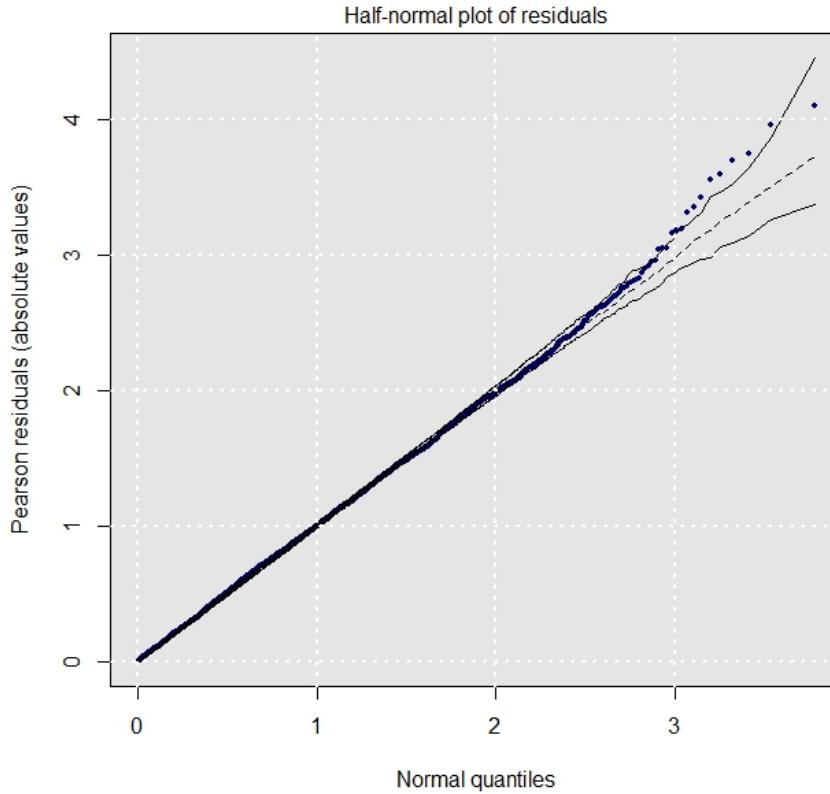


Figura 10: Envelope simulado sem pontos de alavancagem

3.4 Modelo de regressão beta ajustado

Como expõe Almeida e Souza(2015) o modelo de regressão beta, pode admitir estrutura de regressão para a parâmetro de precisão. Dessa forma, é preciso identificar se a precisão é fixa, ou seja, se existe estrutura de regressão para o parâmetro de precisão ϕ . Tendo isto em vista, foi realizado TRV sobre um modelo com precisão variável versus em que a precisão é fixa, $H_0 : \phi_1 = \dots = \phi_q$ o teste é dado por $TRV = -2\ln(\ell_1) + 2\ln(\ell_2)$, em que ℓ_1 representa o modelo com precisão variável adicionada e ℓ_2 o modelo reduzido ou com precisão fixa.

Tabela 3: TRV para o parâmetro de precisão

Precisão	#gl	LogLik	gl_{x^2}	x^2	valor p
ϕ fixo	11	9241.34			
ϕ var	20	9301.99	9	121.29	<0.0001

Pelos resultados, foi possível rejeitar a hipótese nula ao nível de significância nominal de de 0,05. Dessa forma, além de modelar a média, foi preciso modelar o parâmetro de precisão. Assim, nosso modelo assume a seguinte forma:

$$cloglog(\mu_t) = \beta_0 + \beta_1(AGUAESGOTO) + \beta_2(MORT1) + \beta_3(TFLSUPER) + \beta_4(PINDCRI) + \beta_5(Norte) + \beta_6(Nordeste) + \beta_7(Sudeste) + \beta_8(Sul)$$

$$\log(\phi_t) = \gamma_0 + \gamma_1(AGUAESGOTO) + \gamma_2(MORT1) + \gamma_3(TFLSUPER) + \gamma_4(PINDCRI) + \gamma_5(Norte) + \gamma_6(Nordeste) + \gamma_7(Sudeste) + \gamma_8(Sul)$$

É válido ressaltar que utilizamos a região Centro-Oeste como referência .Partindo disso, iremos verificar as estimativas obtidas, bem como seu efeito e contribuição para a variável resposta IDHM.

Tabela 4: Modelo para a média com ligação Cloglog

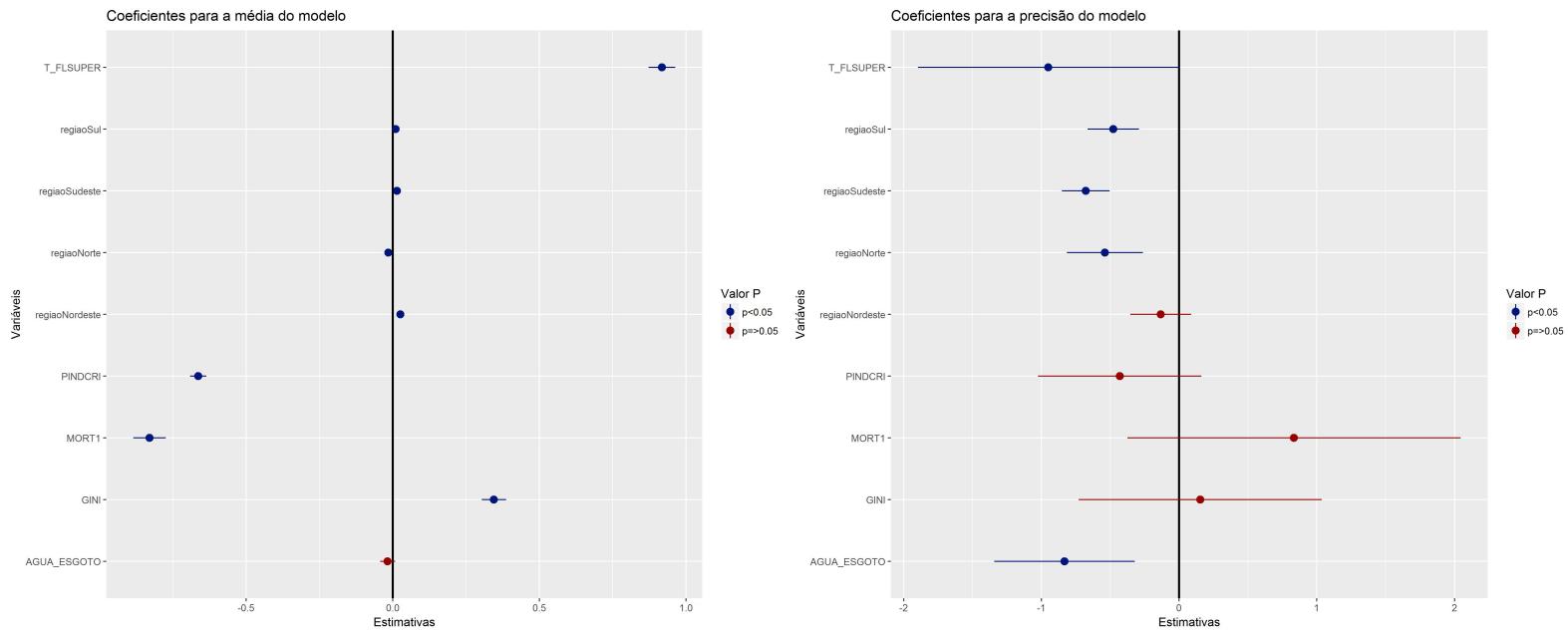
Parâmetros	Estimativa	Erro padrão	Z	Valor P
(Intercept)	0.0623	0.0118	5.2708	<0.0001
PINDCRI	-0.6629	0.0143	-46.4782	<0.0001
AGUA_ESGOTO	-0.0173	0.0130	-1.3252	0.1851
T_FLSUPER	0.9180	0.0232	39.6484	<0.0001
MORT1	-0.8290	0.0282	-29.4103	<0.0001
GINI	0.3452	0.0212	16.2861	<0.0001
regiaoNordeste	0.0262	0.0048	5.4787	<0.0001
regiaoNorte	-0.0152	0.0069	-2.2096	0.0271
regiaoSudeste	0.0150	0.0037	4.0568	<0.0001
regiaoSul	0.0103	0.0040	2.5862	0.0097
Pseudo R^2	0.875	Log-verossimilhança	9302	

Das covariáveis selecionadas para o modelo apenas AGUAESGOTO não foi significativa, no caso das demais foi possível rejeitar $H_0 : \beta_k = 0$ ao nível de significância nominal de 5%. No que diz respeito aos efeitos das estimativas continuas verificamos que PINDCRI e MORT1 apresentam efeitos negativos, ou seja, quanto menor o valor desses indicadores maior o IDHM, já T_FLSUPER e GINI retornam efeitos positivos. Quanto as regiões, constatamos que ser do Norte implica em diminuição no IDHM quando comparado a CentroOeste, nas demais regiões verificamos efeitos positivos do IDHM quando comparado a região de referência. Além disso, é possível verificar que o pseudo R^2 apresenta um valor bastante próximo de 1, apesar de não ter a mesma interpretação do que na regressão linear, essa medida também nos indica a capacidade do ajuste do modelo.

Apresentamos também as estimativas do parâmetros de precisão exposta na tabela 5, analisando os resultados observa-se para os casos contínuos AGUA_ESGOTO e T_FLSUPER retornaram valores negativos, isto é, os municípios que apresentam valores menores para estes indicadores tendem a ter respostas mais precisas. Quanto a região verifica-se que o fato de ser do Sul, Sudeste e Norte quando comparado ao CentroOeste diminui a precisão.

Tabela 5: Modelo para a precisão com ligação Log

Parâmetros	Estimativa	Erro padrão	z value	Valor P
(Intercept)	6.2899	0.2555	24.6185	0.0000
PINDCRI	-0.4311	0.3026	-1.4248	0.1542
AGUA_ESGOTO	-0.8322	0.2602	-3.1979	0.0014
T_FLSUPER	-0.9502	0.4828	-1.9680	0.0491
MORT1	0.8340	0.6172	1.3513	0.1766
GINI	0.1529	0.4499	0.3398	0.7340
regiaoNordeste	-0.1340	0.1130	-1.1857	0.2357
regiaoNorte	-0.5393	0.1408	-3.8314	0.0001
regiaoSudeste	-0.6785	0.0881	-7.7048	0.0000
regiaoSul	-0.4781	0.0949	-5.0366	0.0000



Para facilitar a visualização dos efeitos, bem como os intervalos de confiança e a rejeição da hipótese nula, foram realizados gráficos dos coeficientes estimados expostos na figura ???. O ponto mais interessante deste gráfico é verificar a não rejeição da hipótese nula que aparece para as estimativas em vermelho, uma vez que, fica claro observar que o zero está presente no intervalo, pois ele toca a linha vertical fixada em zero.

4 Modelo de regressão para dados de contagem

Modelos probabilísticos para v.a. discretas, pertencente ao conjunto de números inteiros não-negativos, são indicados para a análise de dados de contagens. Algumas alternativas para tal modelagem são as distribuições Binomial, Poisson, binomial negativa entre outras, assim foram consideradas estas componentes aleatórias para a modelagem do número de visitas ao Lago Somerville.

4.1 Análise exploratória

Tabela 6: Medidas de posição e dispersão

id	Variáveis	Min	\bar{x}	M_d	Max	s^2	CV %
Visitas	Número anual de visitas ao lago Somerville	0	2,2	0	88,0	39,6	280,4
Rank	Ranking de qualidade do lago Somerville	0	1,4	0	5,0	3,3	127,7
Renda	Renda anual	1,0	3,9	3,0	9,0	3,4	48,1
costCon	Gastos quando visita o lago Conroe	4,3	55,4	41,2	493,8	2179,3	84,2
costSom	Gastos quando visita o lago Somerville	4,8	59,9	47,0	491,5	2150,8	77,4
costHoust	Gastos quando visita o lago Houston	5,7	56,0	42,4	491,0	2128,3	82,4
Esqui	Pratica esqui aquático no lago Somerville	Sim	63,28	Não	36,72		
Socio	Sócio do parque Somerville	Sim	98,03	Não	1,97		

Na tabela 6 temos as medidas descritivas das variáveis do banco Somerville, como é possível observar, as variáveis Visitas e Rank apresentam mínimo e mediana 0, ou seja, se mostram com um número excessivo de zeros, além disso, apresentam coeficientes de variação muito grandes. Indicando problema de sobredispersão, uma vez que, $s^2 > \bar{x}$.

A variável Renda que representa a renda não apresentou uma variabilidade tão alta como as demais expostas na tabela. Quanto às variáveis categóricas, nota-se que no caso de Socio a grande maioria dos usuários do parque é sócio totalizando 98,03%.

Quanto ao comportamento da variável resposta pode-se verificar pelo gráfico da figura 11. Nele, notamos o problema mencionado acima, do inflacionamento de zero na distribuição dos dados, a barra que representa a frequência de zero, sendo referente a quando o indivíduo não visitou o parque naquele ano é maior que a soma das demais frequências, representando 63,28 do total.

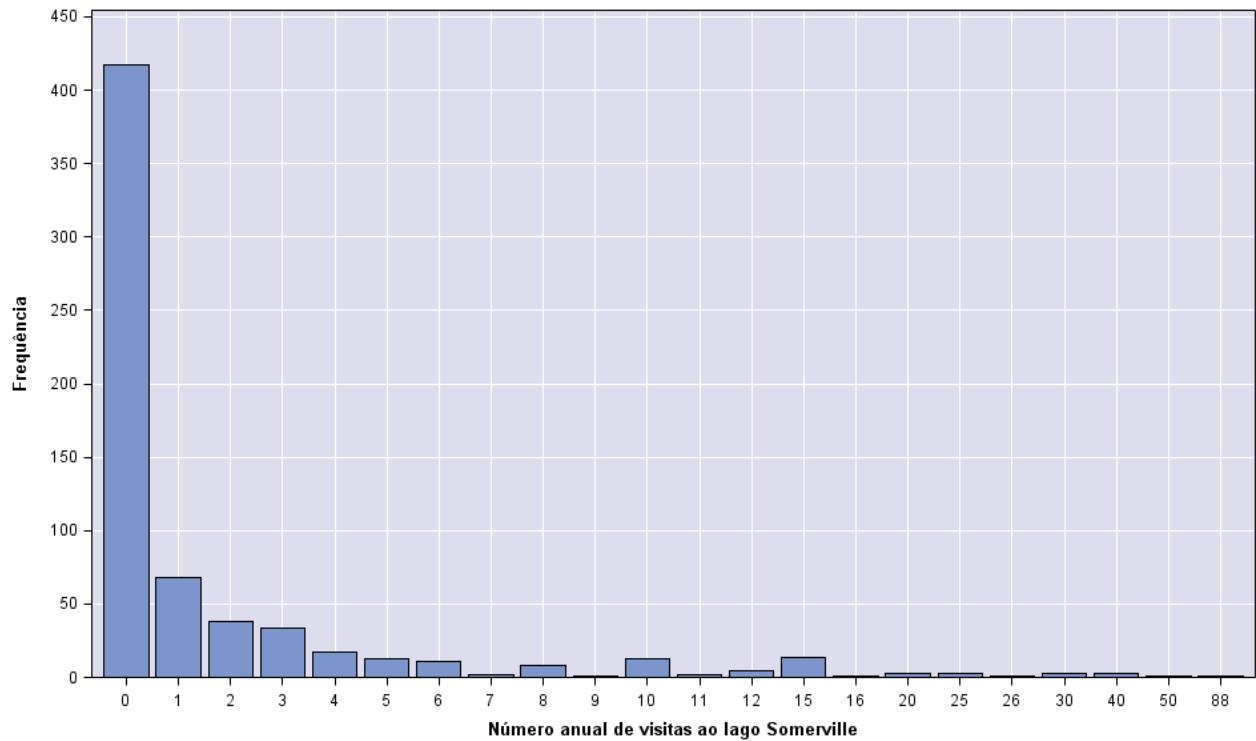


Figura 11: Frequênci a do número de visitas ao lago Somerville

Na figura 12 temos o comportamento das covariáveis segundo a resposta, observando os resultados constatamos que as variáveis Sócio do parque Somerville e Gastos quando visitam Somerville, Houston e Conroe apresentam grandes concentrações quando o número de visitas ao Parque Somerville é igual a 0. Desse modo, quando iniciarmos a modelagem dos zeros para os modelos ZIP e ZINB será preciso levar em consideração tais comportamentos.

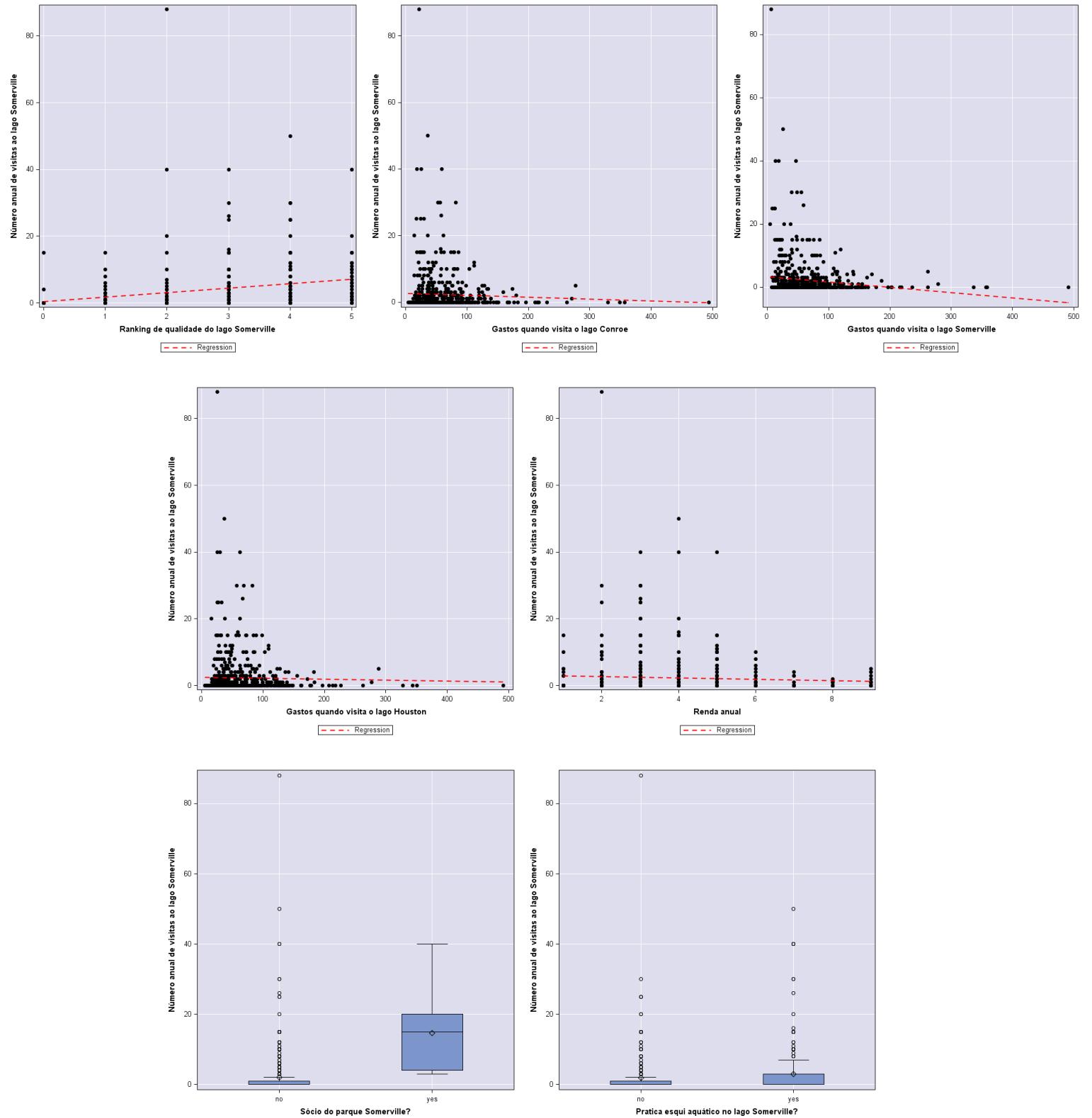


Figura 12: Resposta vs covariáveis

4.2 Ajuste dos modelos Poisson e Binomial Negativa

Dados de contagem são classificados como valores inteiros e positivos normalmente apresentando poucas frequências em cada valor único. Sendo a variável de interesse o número de indivíduos que visitam anualmente o lago Somerville, ou seja, uma contagem e nosso interesse é verificar os fatores que a afetam. Para tal, iremos empregar modelos de regressão para dados de contagem utilizando as componentes aleatória Poisson e Binomial Negativa.

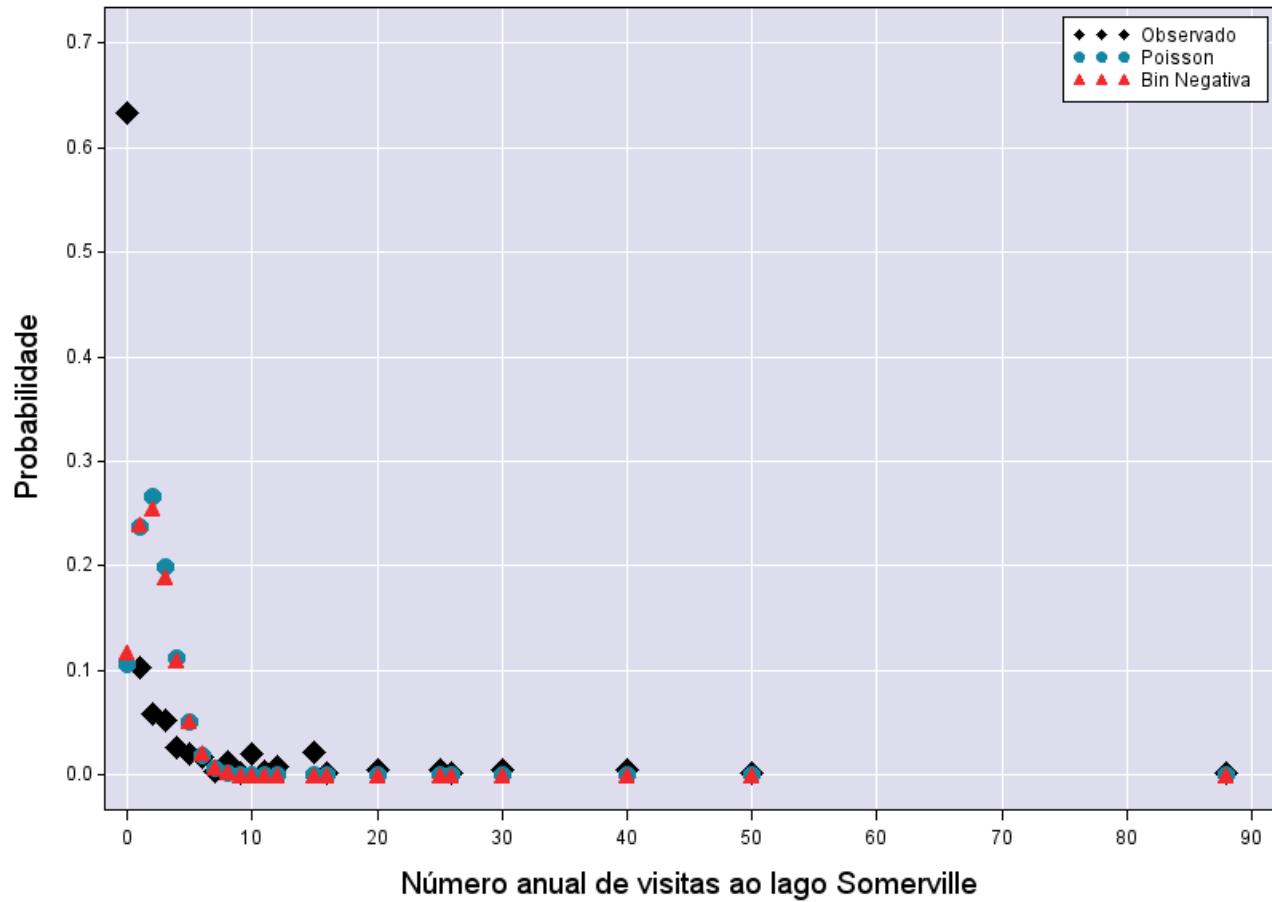


Figura 13: Observado vs Esperado Poisson e Bin. Negativa

Antes de ajustar o modelo é interessante verificar se as distribuições escolhidas para descrever nossos dados se ajustam bem a eles, isto é, verificar de forma gráfica se as frequências observadas do número de visitas ao lago são próximas as frequências esperadas da Poisson e da Binomial negativa. Pelo gráfico visto em 13 nota-se que as distribuições teóricas não conseguem realizar um bom ajuste, pois não conseguem captar a alta concentração de zeros observados. Esse comportamento se dá devido ao problema de sobredispersão indicado acima.

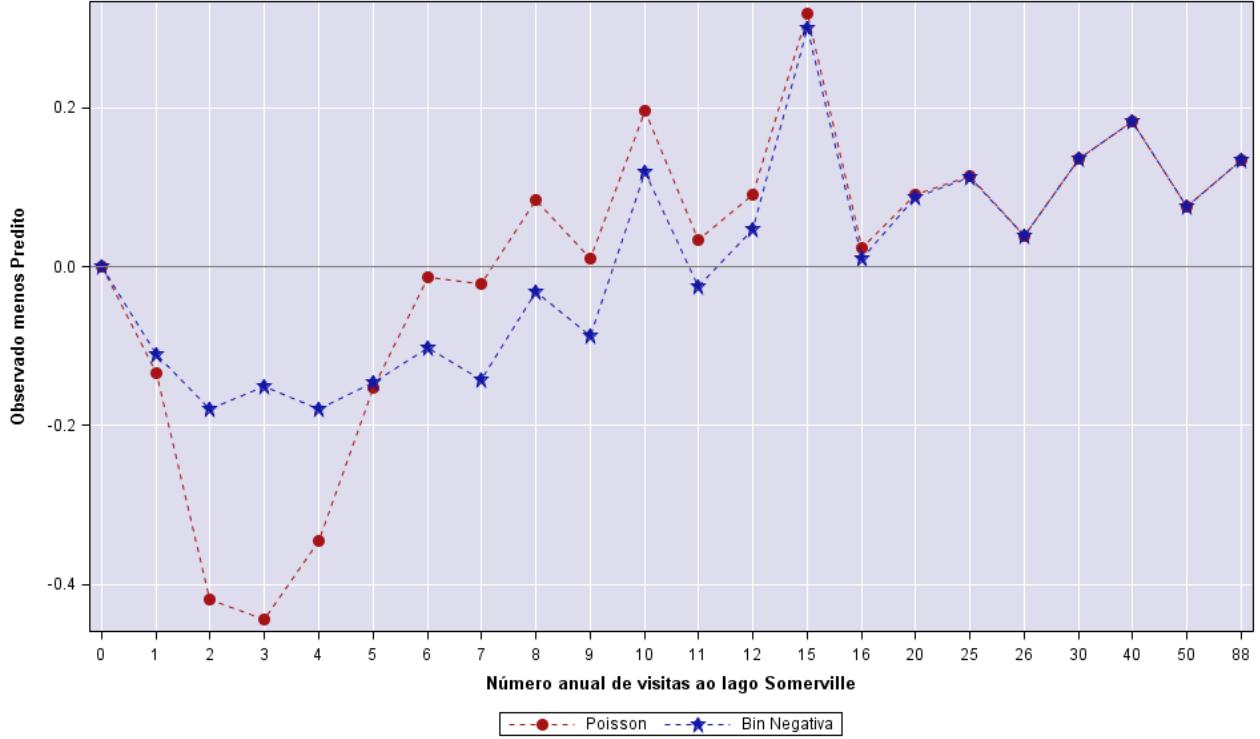


Figura 14: Frequênci a observada menos a esperada segundo a resposta

Para reforçar os apontamentos expostos acima, foi realizado um gráfico em que colocamos no eixo y a frequências dos observados menos os valores teóricos e no x a frequência dos observados. Observando o comportamento das curvas das distribuição Poisson e Binomial Negativa notamos que ambas demonstram considerável variação em torno de zero, ou seja, as distribuições não se ajustaram bem aos dados de interesse.

$$\log(\mu_i) = \gamma_0 + \gamma_1(Rank) + \gamma_2(Renda) + \gamma_3(costCon) + \gamma_4(costSom) + \gamma_5(costHoust) + \gamma_6(Esqui_1) + \gamma_7(Esqui_2) + \gamma_8(Socio_1) + \gamma_9(Socio_2)$$

Aqui apresentamos a componente sistemático para os modelos Poisson e Binomial negativa que ajustamos, na tabela abaixo são expostos os resultados das estimativas para ambas as distribuições.

Tabela 7: Parâmetros estimados por Máxima verossimilhança

Distribuições	Poisson		Binomial Negativa		
Parâmetros	Estimativa	Erro padrão	Estimativa	Erro padrão	
Intercept	1.5814*	0.1272	0.1594	0.4403	
Ranking de qualidade do lago Somerville	0.4717*	0.0171	0.7220*	0.0453	
Sócio do parque Somerville	Não	-0.8982*	0.0790	-0.6692	0.3614
Sócio do parque Somerville	Sim	0.0000	0.0000	0.0000	0.0000
Pratica esqui aquático no lago Somerville	Não	-0.4182*	0.0572	-0.6121*	0.1504
Pratica esqui aquático no lago Somerville	Sim	0.0000	0.0000	0.0000	0.0000
Renda anual		-0.1113*	0.0196	-0.0261	0.0452
Gastos quando visita o lago Conroe		-0.0034	0.0031	0.0480*	0.0160
Gastos quando visita o lago Somerville		-0.0425*	0.0017	-0.0927*	0.0083
Gastos quando visita o lago Houston		0.0361*	0.0027	0.0388*	0.0117
Scale		1.0000	0.0000	1.3713	0.1454

Na tabela 7 estão expostas as estimativas e seu respectivos erros para o modelo em que consideramos as distribuições Poisson e Binomial Negativa. Apesar de apresentarem efeitos similares notamos que as estimativas dos dois modelos apresentaram resultados distintos, além disso, também há distinções nas que se mostraram significativas e no tamanho dos erros das estimativas.

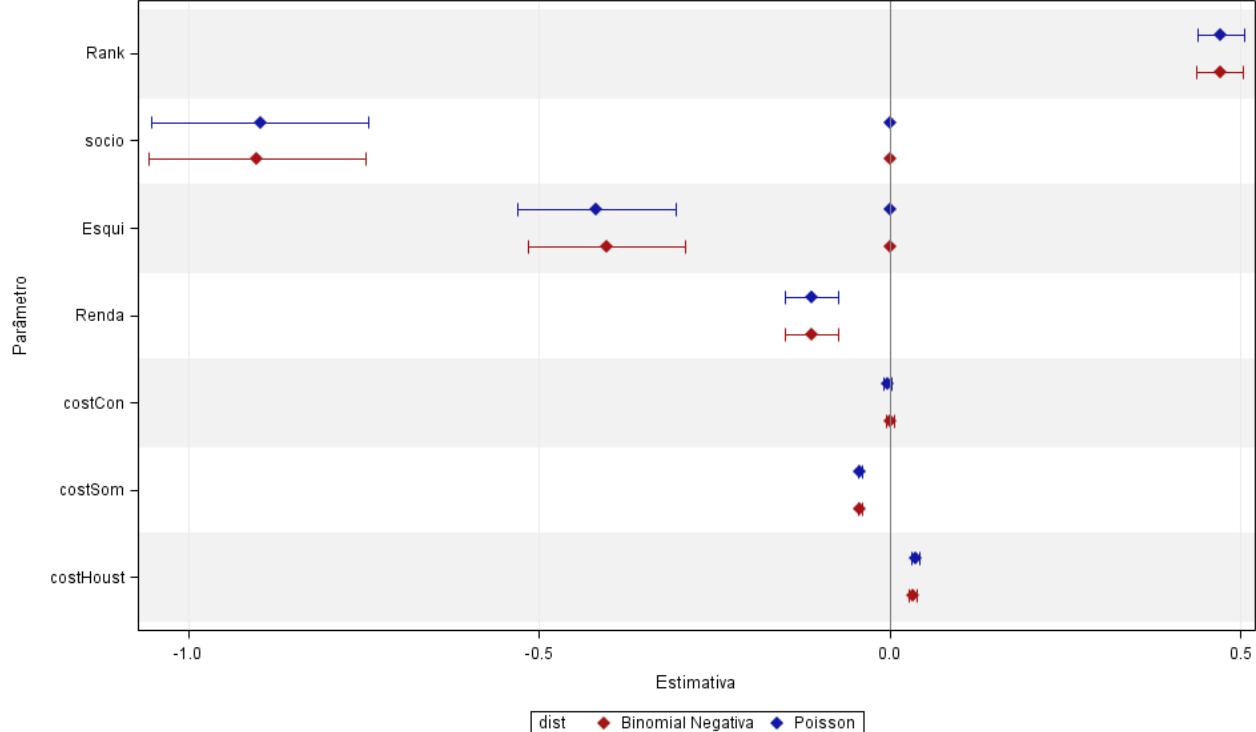


Figura 15: Comparaçāo das estimativas e efeitos dos modelos

Na figura 15 temos a representação gráfica dos parâmetros estimados e seus efeitos. Ele

nos fornece uma forma mais rápida e intuitiva de visualizar os resultados de um modelo de regressão. Bem, como a capacidade de comparar os resultados dos dois modelos ajustados.

Tabela 8: Qualidade do ajuste

Distribuição	gl	Deviance	Deviance/gl	
Poisson	651	2305.7856	3.5419	
Binomial negativa	651	425.4153	0.6535	
Distribuição	gl	$PearsonX^2$	$PearsonX^2/gl$	Valor P
Poisson	651	4100.0934	6.2981	<.0001
Binomial negativa	651	1003.8931	1.5421	<.0001

Na tabela 8 temos os valores da Deviance e do Pearson X^2 para os dois modelos, como é possível verificar a Deviance para o modelo Poisson é 3 vezes e meia maior que um, ou seja, indica problemas de sobredispersão, já para o modelo considerando a binomial negativa temos um valor abaixo de 1. No que tange a estatística de Pearson X^2 foi realizado um teste sobre a hipótese nula comparando o valor calculado com uma X_{n-p}^2 . Como podemos garantir a propriedade assintótica, já que nosso $n=659$ pode ser considerado grande, assim:

$$H_0 : \text{Não há sobredispersão};$$

$$H_1 : \text{Existe sobredispersão}$$

Pelos resultados constatamos que para ambos os modelos, tanto pela parte gráfica, quanto para o teste exposto na tabela 8 há fortes evidências de sobredispersão devido a alta frequência de zeros apresentada pela variáveis resposta. Desse modo, iremos empregar nas análises que seguiram modelos que nos permitiram modelar os zeros, ou seja, que irão dividir nossas modelagem em duas partes.

4.3 Ajuste dos modelos ZIP e ZINB

Dados de contagem que tem alta incidência de zeros podem ser modelados utilizando distribuições para dados inflacionados de zeros, para nosso problema iremos utilizar os modelos ZIP(zero-inflated Poisson) e ZINB(zero-inflated negative binomial). Nesses modelos, a população é considerada como sendo constituída por dois tipos de indivíduos. O primeiro é modelado por uma Poisson ou Binomial Negativa desconsiderando os zeros. E o segundo considera apenas a contagem dos zeros. Assim suponha λ é a média da distribuição e ω a proporção de indivíduos com zero. O parâmetro ω é chamado de probabilidade do inflacionamento de zeros, sendo a probabilidade de contar zeros em excessos no modelo.

O modelo Poisson inflacionada de zeros foi proposto por Lambert(1992) e o qual apresenta um parâmetro de média λ_i , com probabilidade ω_i para os zeros estruturais, assim:

$$P(Y = y) = \begin{cases} \omega + (1 - \omega)e^{-\lambda_i} & \text{para } y = 0 \\ (1 - \omega)\frac{\lambda^y e^{-\lambda}}{y!} & \text{para } y = 1, 2, \dots \end{cases} \quad (12)$$

o valor esperado e a variância de y_i são dados por $E(y_i|x_j) = (1 - \omega_i)\lambda_i$ e $V(y_i|x_j) = \lambda_i(1 - \omega_i)(1 + \lambda_i\omega_i)$.

O modelo Binomial Negativo inflacionado de zeros (Cheung,2002) de modo análogo detém o parâmetro de média λ_i , probabilidade ω_i para os zeros estruturais $1 - \omega_i$ para os zeros amostrais, sendo o parâmetro de dispersão dados por k .

$$P(Y = y) = \begin{cases} \omega_i + (1 - \omega_i)(1 + k\lambda_i)^{-1/k} & \text{para } y = 0 \\ (1 - \omega_i) \frac{\Gamma(y + 1/k)}{\Gamma(y + 1)\Gamma(1/k)} \frac{(k\lambda_i)^y}{(1 + k\lambda_i)^{y+1/k}} & \text{para } y = 1, 2, \dots \end{cases} \quad (13)$$

desse modo, o valor esperado e a variância de y_i são dados por $E(y_i|x_j) = (1 - \omega_i)\lambda_i$ e $V(y_i|x_j) = \lambda_i(1 - \omega_i)(\lambda_i\omega_i + 1 + \lambda_i/k)$.

Partindo disso, vamos verificar o gráfico da frequência observada vs a teórica. Pelo gráfico



Figura 16: Observado vs Esperado ZIP e ZINB

exposto em 16 constatamos que a distribuição ZIP apesar de conseguir captar a concentração excessiva de zeros, não consegue se ajustar ao restante dos dados, descendo de forma abrupta criando uma elevação seguida de um decréscimo. Já a ZINB apresenta um ajuste bastante bom, conseguindo captar praticamente todos os valores observados.

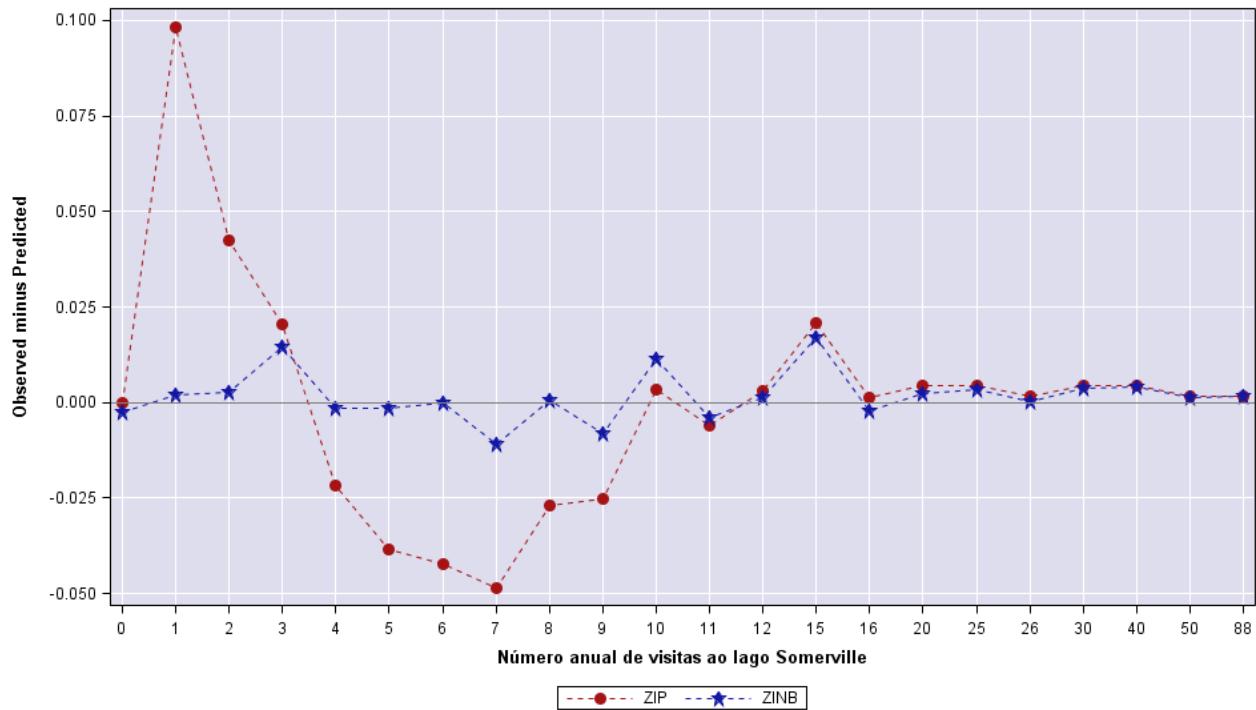


Figura 17: frequência observada menos a esperada segundo a resposta

No que tange ao gráfico dos valores observados pelos teóricos, reforçamos as observações feitas acima, ou seja, a distribuição ZINB se apresenta melhor capacidade para descrever o número de visitas ao lago Somerville, pois pode-se verificar que seus pontos estão bastante próximo de zero em todo decorrer da reta, assumindo valores inferiores a -0.025 e 0.025. Partindo, desta discussão iremos ajustar os modelos para ZIP e ZINB.

Tabela 9: Parâmetros estimados por Máxima verossimilhança

Distribuições	ZIP		ZINB		
Parâmetros	Estimativa	Erro padrão	Estimativa	Erro padrão	
Intercept	2.7799*	0.1673	2.9471*	0.1499	
Ranking de qualidade do lago Somerville	0.1718*	0.0348	0.1578*	0.0283	
Sócio do parque Somerville	Não	-0.6552*	0.0803	-0.6469*	0.0797
Sócio do parque Somerville	Sim	0.0000	0.0000	0.0000	0.0000
Pratica esqui aquático no lago Somerville	Não	-0.4715*	0.0592	-0.4872*	0.0582
Pratica esqui aquático no lago Somerville	Sim	0.0000	0.0000	0.0000	0.0000
Renda anual		-0.1021*	0.0211	-0.1040*	0.0205
Gastos quando visita o lago Conroe		-0.0018	0.0037	-0.0055	0.0035
Gastos quando visita o lago Somerville		-0.0379*	0.0020	-0.0365*	0.0020
Gastos quando visita o lago Houston		0.0285*	0.0033	0.0288*	0.0031
Scale/Dispersion	1.0000	0.0000	2.7183	0	

A tabela 9 apresenta os resultados das estimativas e seus respectivos erros para os modelos ZIP ZINB. Assim como no modelo em que não consideramos a modelagem dos zeros os resultados dos modelos foram bastante próximos

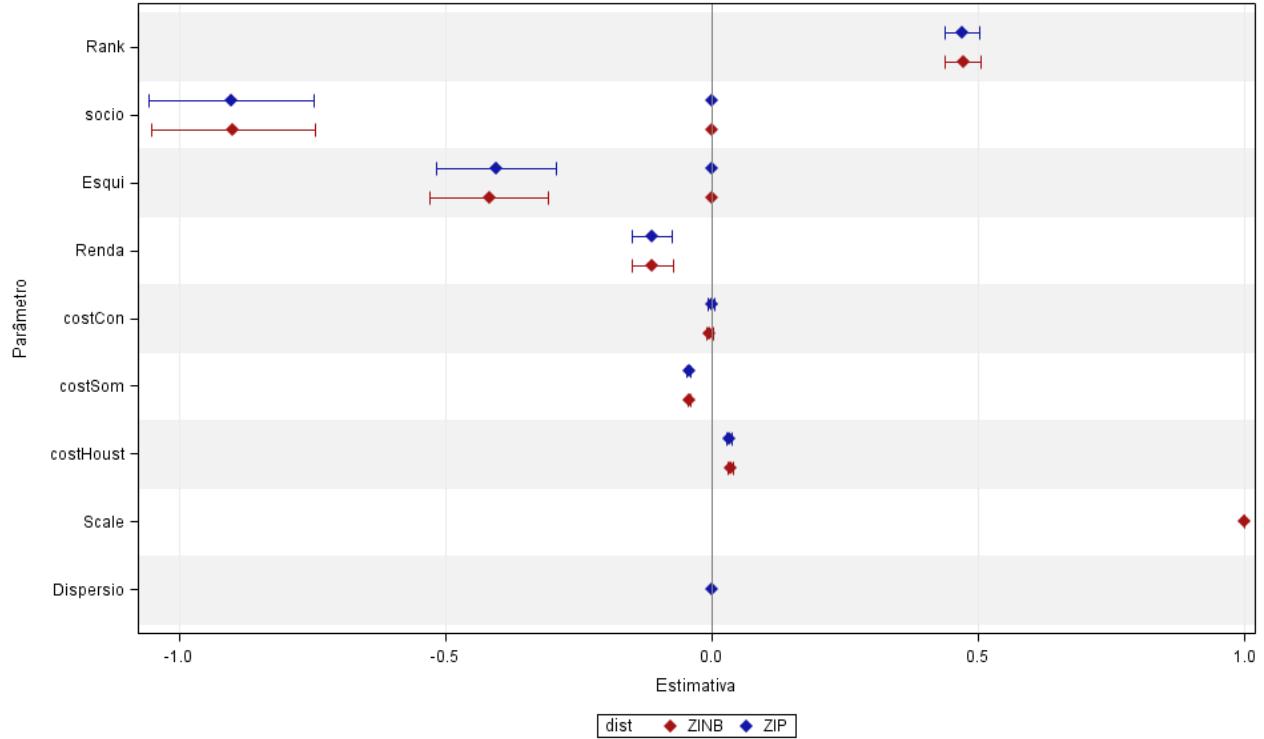


Figura 18: Comparaçāo das estimativas e efeitos dos modelos

A figura 18 nos mostra os resultados das estimativas, seus efeitos, os intervalos de confiança e os casos em que foi possível verificar que $\gamma_k \neq 0$, como mostrados nos modelos anteriores.

Tabela 10: Estimativas para o modelo inflacionado de zeros

Distribuições	ZIP		ZINB	
Parâmetros	Estimativa	Erro padrão	Estimativa	Erro padrão
Intercept	-20.7249*	0.2683	-20.7249*	0.2683
Sócio do parque Somerville	no	21.6861	30427.88	21.6861
Sócio do parque Somerville	yes	0.0000	0.0000	0.0000
Gastos quando visita o lago Conroe		-0.1456*	0.0266	-0.1456*
Gastos quando visita o lago Somerville		0.1405*	0.0165	0.1405*
Gastos quando visita o lago Houston		-0.0218	0.0176	-0.0218

Por fim, é possível verificar a estimativas para o modelos dos zeros. Pelos retornos constata-se que as variáveis relacionadas aos gastos em Somerville e Houston se demonstraram significativas para a explicação dos zeros no número de visitas.

Tabela 11: Qualidade do ajuste modelos inflacionado de zero

Distribuição	gl	Deviance	Deviance/gl	Valor P
ZIP	.	2683.6297	.	.
ZINB	.	1645.0183	.	.
Distribuição	gl	Pearson X^2	Pearson X^2/gl	Valor P
ZIP	646	1618.8480	2.5060	<.0001
ZINB	646	718.0369	1.1115	0.0254

Foram realizados novamente teste para verificar a existência de sobredispersão no modelo, como é possível verificar, mesmo com o emprego das distribuições ZIP e ZINB e teste rejeita hipótese nula de não existência de sobredispersão. Tendo em vista as proximidades observadas nos resultados destes dois modelos, vimos a necessidade de compará-los para constatarmos se de fato eles são iguais. Para tal, foi realizado o teste Score ou Multiplicador de Lagrange. O teste é calculado da seguinte forma

$$X^2 = \frac{s^2}{V} \quad (14)$$

em que s^2 é componente do vetor Score no máximo restrito e $V = I_{11} - I_{12}I_{22}^{-1}I_{21}$ a matriz informação, 1 se refere ao parâmetro restrito e 2 aos restante dos parâmetros. Sobre certas condições de regularidade está estatística é assintótica X^2 com 1 gl.

Tabela 12: Multiplicador de Lagrange ou Score

Parâmetro	X^2	Valor P
Dispersão	167.8521	<.0001

Como queremos comparar a Binomial Negativa com a Poisson iremos verificar se $k=0$ na Binomial negativa, neste caso s é a estatística score de Cameron e Trivedi (1998) para testar

a sobredispersão em um modelo Poisson contra a forma alternativa $V(\mu) = \mu + k\mu^2$. Pelos resultados da tabela 12 rejeitamos H_0 , ou seja, o modelo Binomial Negativo não é equivalente ao Poisson.

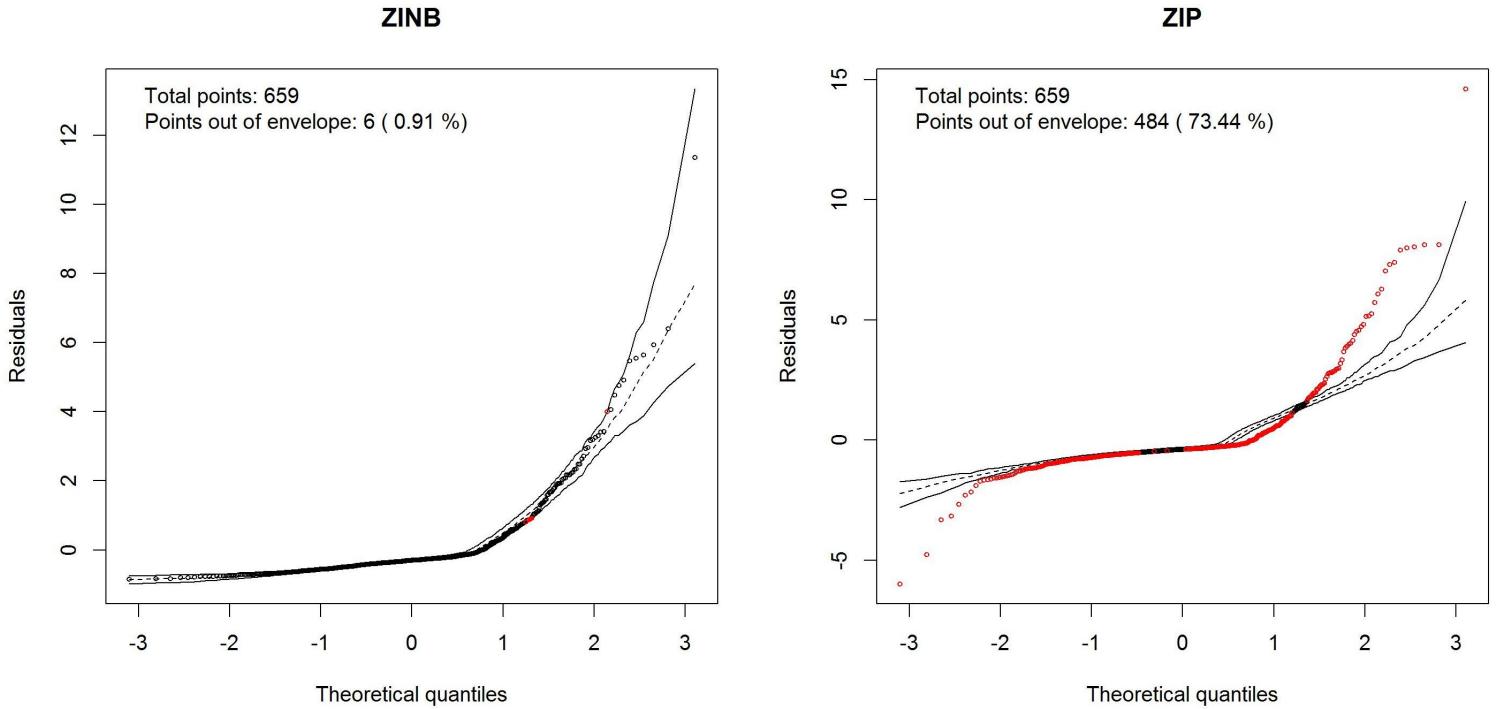


Figura 19: Envelope simulado para ZIP e ZINB

Para avaliar a qualidade do ajuste dos modelos realizamos os gráficos dos envelopes simulados para ambos. Verificando os resultados nota-se que o resíduos da ZINB estão muito melhores que os da ZIP, contendo apenas 6 pontos dos 659, fora da banda de confiança. Entretanto é preciso ressaltar que nenhum dos modelos seguem uma reta, ou seja, os resíduos não seguem uma $N(0,1)$. Desse modo, mesmo diante de tais restrições, elencamos a distribuição ZINB para modelar nosso dados, a partir disso, iremos realizar a seleção de variáveis e a análise de resíduo pelo modelo ZINB.

4.4 Modelo de contagem final

Nessa seção apresentamos os resultados, as discussões e análise dos resíduos do modelo final, ou seja, o modelo que escolhemos para explicar o número de visitas ao lago Somerville. Antes de verificarmos as estatísticas serão apresentados alguns critérios de ajuste, a fim de comparar os modelos encaixados inicial e o final. É válido ressaltar que o TRV para modelo encaixados foi aplicado em várias combinações de variáveis, além disso, foram avaliados os resíduos após a retirada das variáveis e constatamos que a ausência delas pouco altera o modelo, desse modo, sua retirada foi por conveniência do autor.

Tabela 13: Critérios para verificar qualidade do ajuste

Critério	Modelo inicial	Modelo Final
Deviance	1650.6140	1524.0824
Pearson X^2 /gl	1.5493	1.2791
Log Likelihood	-843.7750	-762.0412
AIC	1678.6140	1548.0824
AICC	1679.2662	1548.5653
BIC	1741.4841	1601.9710

Pela tabela 13 constatamos que o modelo final apresenta medidas de ajuste melhores do que o modelo que nominamos de inicial. Verificamos que as medidas de AIC, AICC e BIC, ou seja, os critérios de informação, são menores para o segundo modelo, bem como a Deviance. Dessa forma, é possível constatar que o modelo que selecionamos como o final possui ajuste melhor que o modelo dito como inicial, compostos por todas as variáveis disponíveis.

Tabela 14: Estimativas para o modelo ZINB

Parâmetro	Exp(Estim.)	Estimativa	Erro pad.	IC 95%	X^2	Valor p
Ranking de qualidade	2.0704	0.7278	0.0453	0.6389 0.8166	257.74	<.0001
Pratica esqui aquático	0.5432	-0.6102	0.1482	-0.9007 -0.3198	16.96	<.0001
Gastos no lago Conroe	1.0581	0.0564	0.0161	0.0250 0.0879	12.37	0.0004
Gastos no lago Somerville	0.9050	-0.0999	0.0081	-0.1158 -0.0839	150.21	<.0001
Gastos no lago Houston	1.0382	0.0374	0.0117	0.0145 0.0604	10.24	0.0014

Quanto as resultados de modelo verificamos pela tabela 14 que as variáveis que contribuíram para a explicação do número de visitas ao lago Somerville foram, Ranking de qualidade atribuído ao lago, prática de esqui e gastos nos lagos Somerville, Houston e Conroe. No que diz respeito ao efeitos verifica-se que a prática de esqui e aumento no gasto em Somerville diminuem as chances de visita ao lago, enquanto que, aumento nos gastos nos lagos Conroe e Houston e avaliações mais positivas a Somerville, elevam as chances de visitá-lo. No caso das avaliações por exemplo, verificamos que cada elevação no Rank aumenta em mais de duas vezes a chance do número de visitas a Somerville aumentar.

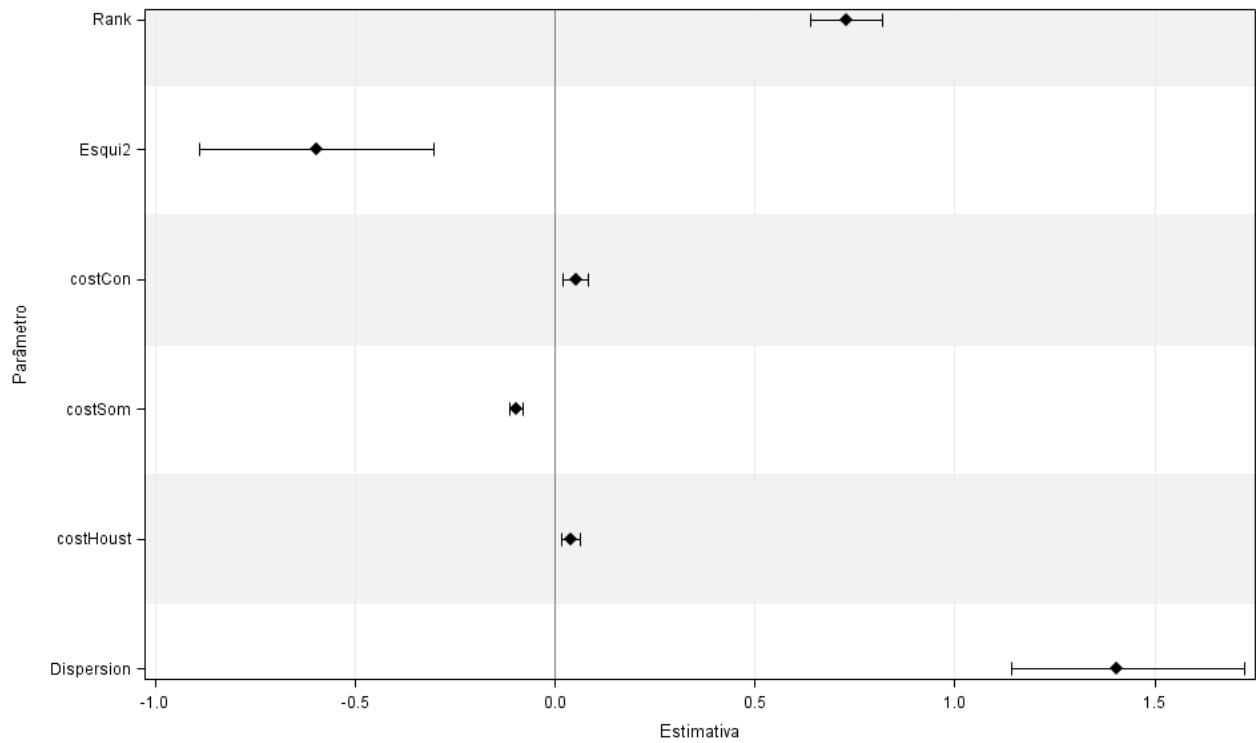


Figura 20: Estimativas e efeitos do modelo ZINB final

Pode-se verificar os efeitos de modo mais direto, bem como seus respectivos intervalos de confiança e tamanho da contribuição para o número de visitas ao Somerville, pelo gráfico de efeitos exposito em 20.

Modelo ZINB Final

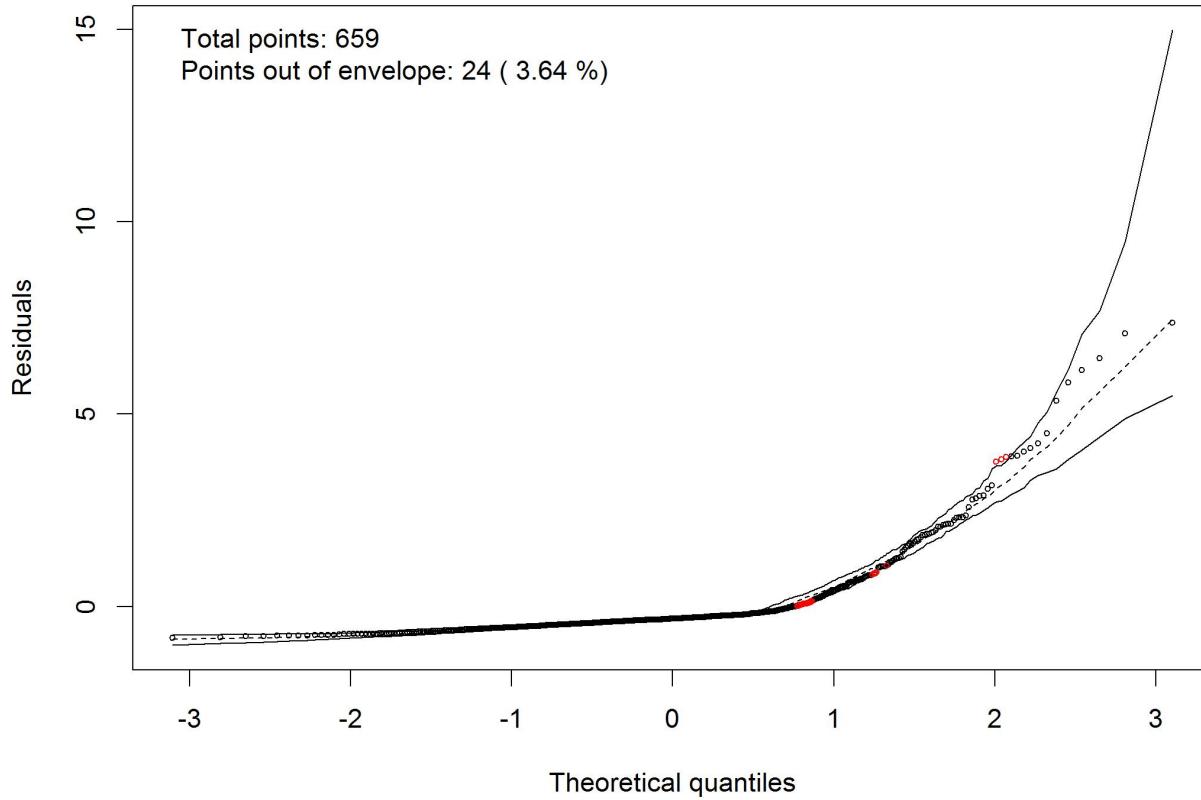


Figura 21: Envelope para o modelo final

Por fim, podemos realizar uma avaliação final pelo gráfico de envelope simulado, nele observa-se que apenas 24 pontos aparecem fora das bandas de confiança, além disso, verificamos que os pontos, encontram-se dentro da banda de confiança mesmo no final da reta teórica. Mesmo diante de tais resultados é preciso ressaltar que Entretanto é preciso ressaltar que os resíduos uma reta, ou seja, os resíduos não seguem a reta dos quantis teóricos $N(0,1)$.

5 Considerações finais

Como foi exposto anteriormente nosso objetivo neste trabalho foi analisar os dados de contagem referente ao número de visitas ao Lago Somerville, e os dados de proporção referentes ao IDHM dos municípios brasileiros. No que diz respeito aos dados de contagem verificamos que o modelo Binomial Negativo Inflacionado de Zeros foi o que melhor se ajustou aos nossos dados, deste modo, pontuamos que o modelo final contou com as variáveis prática de esqui, ranking e qualidade atribuído ao parque e gastos nos parques Somerville, Conroe e Hounston. Sendo que o fato do individuo praticar esqui e elevações nos gastos em Somerville diminuem o número de visitas ao lago e avaliações positivas, seguidas de mais gastos nos lagos Conroe e Hounston elevam o número de visitas a Somerville.

Quanto ao modelo de regressão Beta referente ao IDHM foi realizada a seleção da função

de ligação, na qual realizamos a modelagem à partir da ligação complementar log-log, além disso, verificamos a necessidade de admitir uma estrutura de regressão para o parâmetro de precisão, considerando todas as variáveis empregadas no modelo para a média. Quanto as covariáveis verificou-se que para o modelo das médias, apenas o abastecimento de água e esgoto inadequados não foi significativo para a explicação do IDHM. No que tange a modelagem da precisão verificamos que o índice de pobreza infantil, Gini e a região Nordeste, não nos permitiram rejeitar a hipótese nula.

Referências

- [1] Dobson, Annette J., and Adrian Barnett. An introduction to generalized linear models. CRC press, 2008.
- [2] Sileshi, G. Selecting the right statistical model for analysis of insect count data by using information theoretic measures. Bulletin of entomological research 96.5 (2006): 479-488.
- [3] Casella, George, and Roger L. Berger. Statistical inference. Vol. 2. Pacific Grove, CA: Duxbury, 2002.
- [4] Paula, A. Gilberto. Modelos de regressão com apoio computacional, USP, 2013.
- [5] Agresti, Alan, and Barbara Finlay. Métodos estatísticos para as ciências sociais. 2012.
- [6] Cameron, A. Colin, and Pravin K. Trivedi. Regression analysis of count data. Vol. 53. Cambridge university press, 2013.
- [7] Ferrari, Silvia, and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. Journal of Applied Statistics 31.7 (2004): 799-815.
- [8] Demétrio, Clarice Garcia Borges. Modelos lineares generalizados em experimentação agronômica. USP/ESALQ, 2001.
- [9] de Almeida Junior, Pedro M., and Tatiene C. Souza. "Estimativas de votos da presidente Dilma Rousseff nas eleições presidenciais de 2010 sob o âmbito do Bolsa Família." Ciência e Natura 37.1 (2015).

6 ANEXOS R