Regressão binária aplicada a dados pluviométricos

Breno Gabriel da Silva [1], **Naiara Caroline Aparecido dos Santos** [2], **Yana Miranda Borges** [3], **Rafaela Galo** [4], **Brian Alvarez Ribeiro de Melo**[5]

1 Introdução

Quando o intuito é analisar a possível relação entre uma ou mais variáveis, os modelos de regressão são uma das ferramentas que podem ser utilizadas. Essa técnica de modelagem pode ser classificada em: Regressão Linear Simples ou Regressão Linear Múltipla. Regressão Linear Simples é quando existe a relação de apenas uma variável regressora com a variável resposta e a Regressão Linear Múltipla é quando há a influência de mais de uma variável regressora na variável resposta. Os autores Nelder e Wedderburn (1972), sugeriram uma extensão dos modelos habituais lineares, denominados Modelos Lineares Generalizados (MLGs), nos quais a distribuição da variável resposta faz parte da família exponencial. Essa generalização permite a modelagem de dados de diferentes características: quando a variável resposta é binária, consideramos o modelo Binomial, caso y seja um dado de contagem, podemos utilizar a distribuição de Poisson e para variáveis contínuas positivas assimétricas, os modelos Gama e Normal Inversa.

De acordo com Bagolin e Bender (2014), quando a variável resposta é do tipo qualitativa dicotômica, é preciso utilizar técnicas de regressão binária para a análise do conjunto de dados, em que umas destas é utilizar o modelo binomial. Tem-se que um dos principais objetivos da regressão através do modelo binomial é estimar a probabilidade de ocorrência de determinado evento, ou seja, os resultados da variável resposta permitem a interpretação dos parâmetros em termos probabilísticos.

Neste trabalho foram utilizados dados de uma das estações meteorológica da região Sudeste do país (83743), compilados a partir das séries históricas de precipitação diárias obtidas no portal do Instituto Nacional de Meteorologia (INMET, 2019), órgão responsável por promover informações meteorológicas oficiais. Neste trabalho, considerou-se como variável resposta uma variável binária, que assume o valor 1 nos dias em que houve precipitação e o valor zero, caso contrário. Tem-se ainda que a variável resposta esta compreendida nos meses de janeiro dos anos de 2014 a 2017. As covariáveis consideradas foram: temperatura máxima, temperatura mínima, umidade relativa média e evaporação piche. O interesse é comparar as funções de ligação logit, probit e cauchit e verificar a influência das covariáveis listadas em relação à precipitação.

[1]Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá - UEM. e-mail: *omatematico.breno@gmail.com*

[2]Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá - UEM. e-mail: *naicaroline2@gmail.com*

[3]Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá - UEM. e-mail: *borges.yana@gmail.com*

[4]Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá - UEM. e-mail: *galo.rafaela@gmail.com*

[5]Professor do Departamento de Estatística, Universidade Estadual de Maringá - UEM. e-mail: *brian.rmelo@gmail.com*

2 Metodologia

2.1 Material

Para este trabalho foram utilizados dados de uma das estações meteorológica da região Sudeste do país (83743), compilados a partir das séries históricas de precipitação diárias obtidas no portal do Instituto Nacional de Meteorologia, INMET (2019) órgão responsável por promover informações meteorológicas oficiais. O Sudeste do Brasil é composto por quatro estados: Espírito Santo, Minas Gerais, Rio de Janeiro e São Paulo. A estação meteorológica considerada no estudo está localizada no estado do Rio de Janeiro e para este trabalho considerou-se como variável resposta a informação se houve ou não precipitação diariamente medida em mm, sendo esta categorizada, referente aos meses de janeiro dos anos 2014, 2015, 2016 e 2017. As covariáveis consideradas foram: temperatura máxima, temperatura mínima, umidade relativa do ar e evaporação piche.

2.2 Métodos

Advoga Demétrio (2001) que selecionar um modelo que melhor ajusta-se aos dados é sem dúvida uma das partes essenciais em qualquer análise estatística que se refere ao contexto de modelagem. Na literatura existem vários testes e critérios para ajuste e seleção de modelos. Neste trabalho, será utilizado as técnicas de Modelos Lineares Generalizados para explicar o comportamento da variável precipitação em função das covariáveis listadas. Em relação a escolha do modelo, será avaliadas a qualidade do ajuste dos modelos candidatos utilizando o Critério de Informação de Akaike (AIC). Em relação a significância dos parâmetros será utilizado o teste de Wald. As suposições do modelo serão avaliadas através do gráfico de envelope simulado. As análises foram realizadas no software R versão 3.5.0 (R CORE TEAM, 2018).

Resultados preliminares

Tabela 1: Análise descritiva dos dados

Variáveis	Precipitação	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Temp. máx	Sim	23,,80	28,30	31,00	30,60	32,60	36,80
$(^\circ C)$	Não	28,30	32,65	34,50	34,34	36,05	38,50
Evaporação	Sim	1,00	2,60	3,70	3,77	4,20	9,80
(mm)	Não	3,20	5,80	7,00	6,78	7,80	11,00
Um. relativa	Sim	49,25	72,44	77,25	76,38	83,06	92,75
do ar (g/Kg)	Não	45,75	58,88	63,75	64,26	69,62	84,25
Temp. mín	Sim	20,70	22,50	24,30	23,94	25,00	26,60
$(^\circ C)$	Não	22,10	25,10	26,10	25,94	27,00	28,50

Fonte: Os autores

Inicialmente, foi realizada uma análise descritiva da variável resposta (precipitação) em conjunto com as covariáveis listadas. Em relação aos resultados apresentados na Tabela 1, observamos que as covariáveis temperatura máxima, evaporação piche e temperatura mínima apresentam um índice mediano inferior aos dias que houve precipitação, podendo indicar que há uma relação inversamente proporcional a resposta, isto é, pode acarretar em uma menor probabilidade de chuva. Já a covariável umidade relativa do ar apresenta um índice mediano superior aos dias que houve precipitação, podendo indicar que há uma

relação diretamente proporcional a resposta, isto é, pode acarretar em uma probabilidade maior de chuva.

Observa-se na Tabela 2 que 73,39% dos dias em estudo não houve precipitação, ou seja, aproximadamente três vezes a quantidade de dias que houve precipitação.

Tabela 2: Análise descritiva da variável resposta

Precipitação	Frequência Relativa
Sim	26,61
Não	73,39

Fonte: Os autores

Considerou-se os modelos de regressão binária com as ligação *logit, probit* e *cauchit*, ajustando sempre a presença de precipitação em função de todas as variáveis descritas. Utilizando o método de seleção automática *stepwise*, segundo o Critério de Informação de Akaike (AIC), observamos que todos os modelos excluem a variável temperatura máxima e o que forneceu o melhor ajuste dos dados foi obtido utilizando-se a função de ligação *cauchit* conforme apresentado na Tabela 3.

Tabela 3: Valores dos AIC's para modelos de regressão binária com ligação *logit, probit* e *cauchit*

Função de ligação	AIC
logit	68,5
probit	70,64
cauchit	59,11

Fonte: Os autores

Em relação as suposições dos modelos, observa-se nas Figuras 1, 2 e 3 o gráfico de envelope simulado, em que este corrobora com o resultado da estatística AIC descrita anteriormente, ou seja, verifica-se uma adequação satisfatória em relação ao modelo binomial com ligação *cauchit* quando comparado com o modelo binomial com as ligações *logit* e *probit* e mais, os gráficos de Resíduos vs Ordem não apresentam nenhuma tendência, indicando assim que os resíduos são independentes, em que se pode concluir que os resíduos apresentam um comportamento adequado.

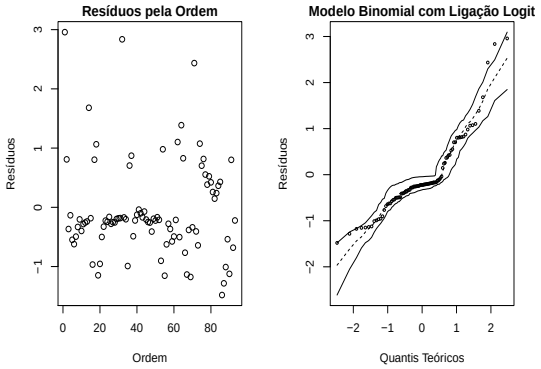

Figura 1: Resíduos pela Ordem e Gráfico de Envelope Simulado - Modelo Binomial com ligação *logit*
Fonte: Os autores

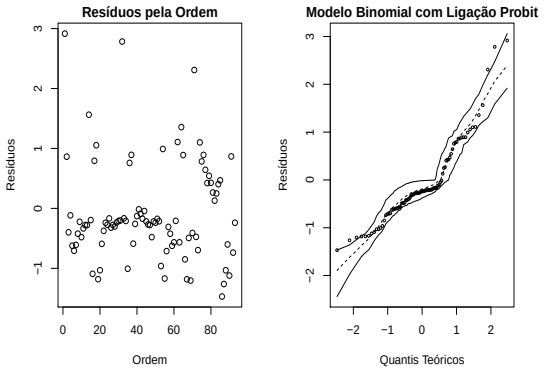

Figura 2: Resíduos pela Ordem e Gráfico de Envelope Simulado - Modelo Binomial com ligação *probit*
Fonte: Os autores

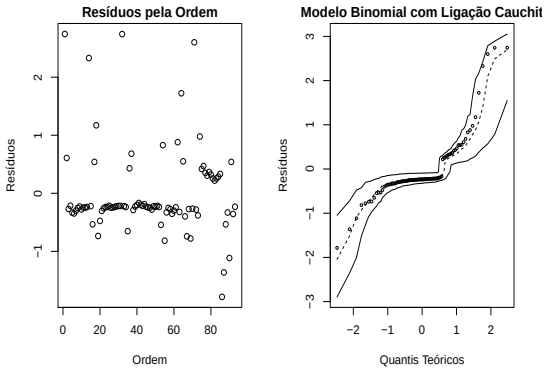

Figura 3: Resíduos pela Ordem e Gráfico de Envelope Simulado - Modelo Binomial com ligação *cauchit*
Fonte: Os autores

Assim, o modelo de regressão binária final é o que tem função de ligação *cauchit* que é dado na Equação 1.

$$cauchit(p_i) = \beta_0 + \text{TMin}\,\beta_1 + \text{Evaporaçõa}\,\beta_2 + \text{Um. relatica do ar}\,\beta_3 \quad (1)$$

Na Tabela 4 estão apresentadas as estimativas dos coeficientes do modelo, através da qual observa-se que as variáveis evaporação e umidade relativa do ar foram significativas a um nível de 10%.

Tabela 4: Estimativas dos coeficientes de regressão binária com função de ligação *cauchit*

Efeito	Estimativas	Erro Padrão	Estatística z	p-valor
β_0	13,815	20,025	0,690	0,490
β_1	-1,207	0,935	-1,287	0,198
β_2	-1,493	0,782	-1,909	0,056
β_3	0,292	0,152	1,925	0,054

Fonte: Os autores

E mais, nota-se que a evaporação apresenta estimativa negativa do parâmetro, indicando que têm-se mais probabilidade para ocorrer a precipitação quando à quantidade de evaporação é menor. Por exemplo, supondo que a medida da temperatura minima e da umidade relativa do ar esteja fixa em suas médias, a chance de ocorrer a precipitação num dia em que a evaporação é 5 *mm* quando comparado a um dia com 10 *mm* é de aproximadamente 3 vezes mais.

Já a variável umidade relativa do ar têm a estimativa do parâmetro positiva, o que indica uma maior probabilidade para ocorrer a precipitação quando essa medida é maior. Por exemplo, a chance de ocorrer a precipitação nos dias em que a umidade relativa do ar é de $80g/Kg$ tem aproximadamente 6 vezes mais chance para ocorrência da precipitação em relação aos dias que têm-se 40 g/Kg, contando que quantidade de evaporação e a medida de temperatura minima no dia estejam na média.

Agradecimentos

Agradecemos a CAPES pelo suporte financeiro.

Referências

BAGOLIN, I. P.; BENDER, R. Determinantes da permanência na condição de pobreza crônica na cidade de porto alegre: Aplicação do modelo logit multinomial. *Ensaios FEE (Impresso)*, 2014.

DEMÉTRIO, C. G. B. *Modelos lineares generalizados em experimentação agronômica*. [S.l.]: USP/ESALQ, 2001.

INMET. *Instituto Nacional de Meteorologia*. 2019. Disponível em: ⟨http://www.inmet.gov.br⟩. Acesso em: 14 mar. 2019.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: ⟨https://www.R-project.org/⟩.