UNIVERSIDADE ESTADUAL DE MARINGÁ

Modelos Lineares Generalizados

II Avaliação

ARTHUR CESAR ROCHA

RA:94361

ARTHUR CESAR ROCHA
RA:94361

Conteúdo

ARTHUR CESAR ROCHA
RA:94361

1 Introdução

Solicitado pela professora Dra. Clédina Regina Lonardan Acorsi, esta análise consiste na avaliação final da disciplina de Modelos Lineares Generalizados. O intuito é desenvolver uma análise completa baseada nos conhecimentos passados em sala de aula.

Para este caso, o banco de dados disponibilizado contava com 19 observações de 4 variáveis:

- calor: tempo de aquecimento das placas.
- imersão: tempo de imersão das placas.
- inadequado: número de placas inadequadas no estudo.
- teste: número de placas usadas em cada teste.

O objetivo era modelar a proporção de placas inadequadas.

Primordialmente, o modelo binomial com função de ligação canônica foi utilizado por conta da natureza do problema. Sucessivamente investigou-se a adequação do modelo, notando que a interação entre as covariáveis não era significativa para a regressão. Além disso, foi preciso saber se a função de ligação era adequada. Também foi necessária a procura de pontos influentes e discrepantes nos dados.

Por fim, considerando a inflação de 0 no conjunto, ajustou-se um modelo de quasi-binomial para fins de comparação.

2 Metodologia

Os modelos Lineares Generalizados são uma extensão dos modelos de regressão ordinários, em que a variável resposta pode ter alguma outra distribuição de probabilidade diferente da distribuição normal, desde que ela pertença a uma classe de distribuições denominada **Familia Exponencial**. Em sua construção relaciona a distribuição aleatória da variável resposta com uma parte sistemática, não aleatória, a partir de uma função chamada função de ligação.

Para a avaliação do ajuste de um modelo, é essencial fazermos testes sobre seus parâmetros. Estes testes utilizam as distribuições assintóticas dos parâmetros β's para sua construção.

Outra forma de se comparar modelos é a partir do o Critério de Informação de Akaike (AIC), calculado por:

$$AIC = -2log(L) + 2[(p+1)+1] \tag{1}$$

Sendo que L é a função de máxima verossimilhança do modelo e p é o número de variáveis explicativas consideradas.

É importante compará-lo com o teste de razão de verossimilhança para uma maior certeza dos resultados.

Esse teste considera as seguintes hipóteses:

$$\begin{cases} H_0 : D_a = D_b \\ H_1 : D_a \neq D_b \end{cases} \tag{2}$$

Sendo que D representa a função desvio dos modelos a serem testados (aninhados).

3 Resultados

3.1 Análise Exploratória

Antes de qualquer análise de cunho inferencial é interessante observar os dados de forma descritiva a fim de entender seu comportamento geral.

Entendendo melhor o comportamento da resposta do problema (π = proporção de placas inadequadas), nota-se pela Figura 1 que a maior concentração de proporções encontradas estão entre 0.13 e 0.28 . Além disso, nota-se uma quantidade razoável de 0's no banco de dados, o que pode complicar a modelagem. É possível perceber ainda que não parecem existir pontos discrepantes para essa variável.

ARTHUR CESAR ROCHA
RA:94361

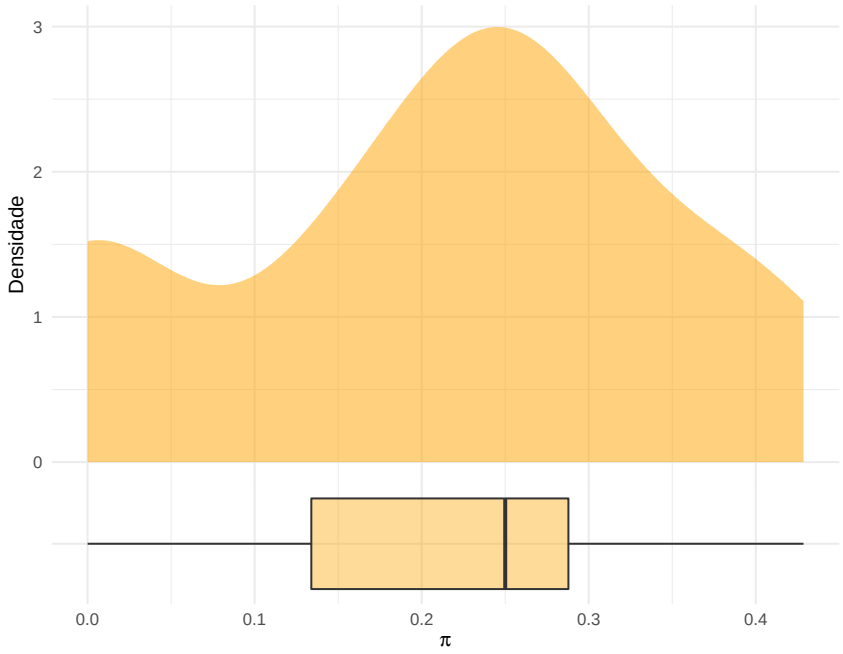

Figura 1: Distribuição da proporção de placas inadequadas.

Observando melhor as variáveis do conjunto de dados, observa-se pela Figura 2 e Figura 3 que as proporções parecem ser diferentes conforme mudam os valores das duas covariáveis.

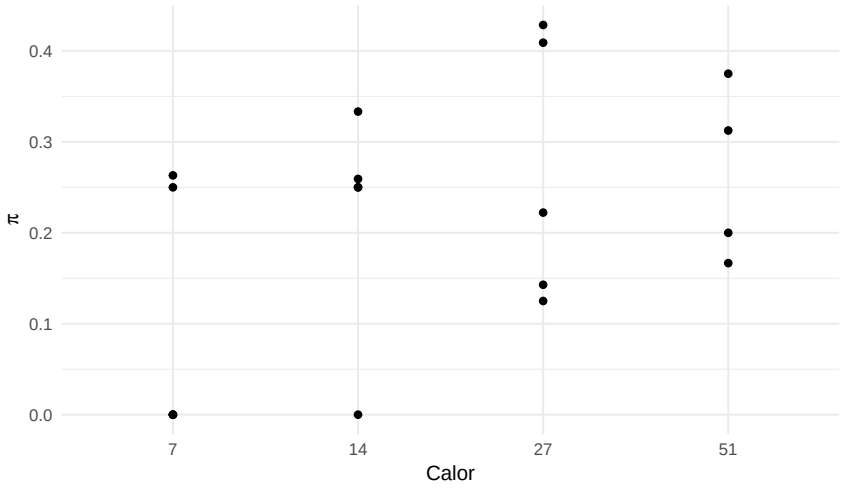

Figura 2: Distribuição da proporção de placas inadequadas conforme níveis de calor.

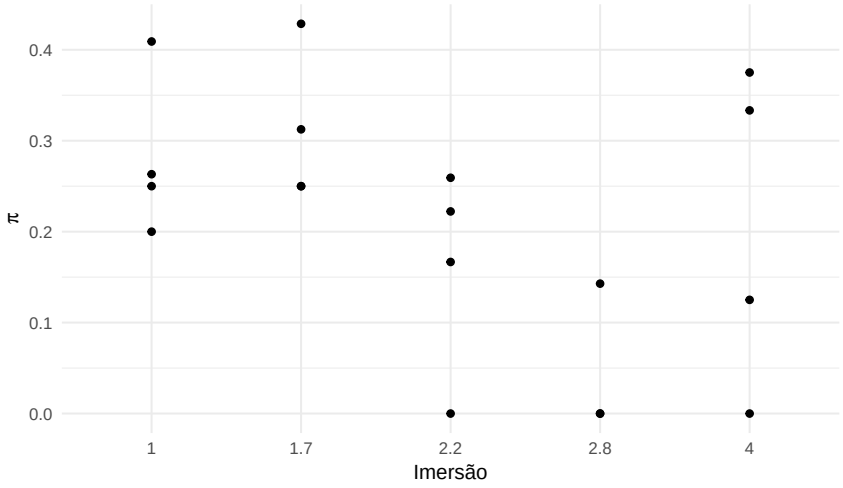

Figura 3: Distribuição da proporção de placas inadequadas conforme níveis de imersão.

Para tentar identificar se existe algum tipo de interação entre as covariáveis, criou-se a Figura 4, nela é possível perceber que talvez possa existir uma interação entre as variáveis independentes, pois o comportamento em relação à resposta não é constante. Mas isso será investigado mais tarde.

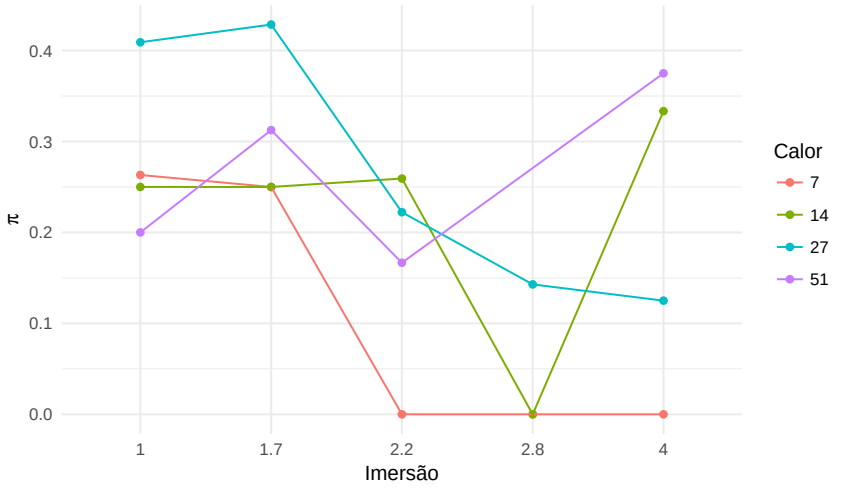

Figura 4: Distribuição da proporção de placas inadequadas conforme calor e imersão.

3.2 Modelagem

A princípio, por se tratar de um problema em que o interesse é modelar uma proporção, uma escolha natural é utilizar o modelo Binomial com função

de ligação canônica. Logo, o primeiro modelo testado foi:

$$log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = \beta_0 + \beta_1 Calor + \beta_2 Imersao \quad (3)$$

As estimativas e os valores-p dos testes sobre os parâmetros são dadas pela tabela a seguir:

Tabela 1: Estimativas do primeiro modelo (aditivo).

Efeito	Estimativa	Erro P.	Valor-p
β_0	-0.8254	0.4346	0.0575
β_1	0.0171	0.0105	0.1035
β_2	-0.3380	0.1709	0.0480
Deviance	19.833		
AIC	65.748		

Como visto na sessão anterior (análise exploratória), há indícios de existência de interação entre as variáveis, logo ajustou-se um modelo considerando essa interação e fez-se o teste de razão de verossimilhança entre as funções desvios do primeiro modelo e desse segundo, a fim de saber se existe ganho significativo em considerar a interação.

As estimativas do segundo modelo são dadas pela Tabela 2.

Tabela 2: Estimativas do segundo modelo (interação).

Efeito	Estimativa	Erro P.	Valor-p
β_0	-0.184039	0.678604	0.7862
β_1	-0.009265	0.023869	0.6979
β_2	-0.665259	0.325804	0.0412
β_3	0.013041	0.010568	0.2172
Deviance	18.328		
AIC	66.244		

Nota-se que houve uma pequena mudança na função Deviance, porém o coeficiente da interação não foi significativo, se considerarmos um nível de significância de 11%. Além disso, ao se fazer o teste da razão de verossimilhança dos dois modelos, chegou-se ao Valor-p de 0.2199, indicando

que não existe diferença significativa nas deviances. Além disso, o AIC do modelo aditivo foi menor. Portanto o primeiro modelo foi o escolhido.

3.2.1 Envelope Simulado

Para continuar a investigação da adequação do modelo, fez-se o gráfico de envelope simulado, representado aqui pela Figura 5. Percebe-se que todos os pontos estão contidos nas bandas do envelope, indicando um bom ajuste do modelo.

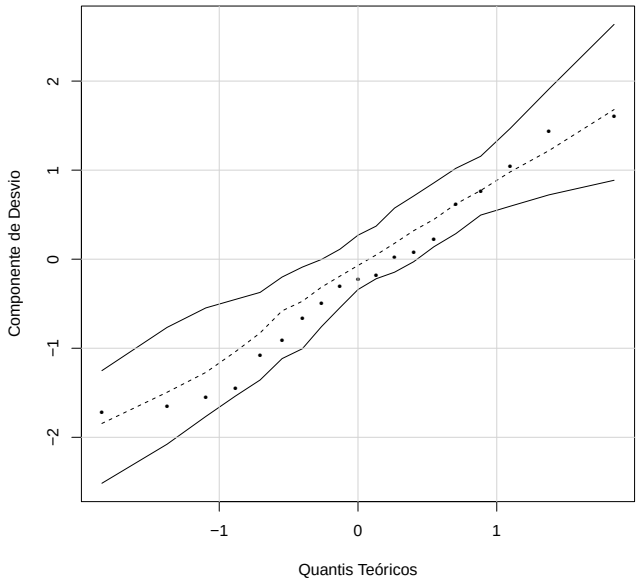

Figura 5: Envelope simulado para os componentes de desvio do modelo 1.

3.2.2 Teste de função de ligação

Antes de prosseguir com a análise de resíduo é interessante verificar se a função de ligação é adequada. Para isso, adicionou-se no modelo o preditor linear elevado ao quadrado e assim, foi verificado se houve diminuição na Deviance, além de mudanças nas estimativas. Um resumo disso é dado na Tabela 3.

Tabela 3: Estimativas do modelo com preditor linear ao quadrado.

Efeito	Estimativa	Erro P.	Valor-p
β_0	-0.679752	0.448430	0.130
β_1	-0.060529	0.292272	0.836
β_2	0.004998	0.015064	0.740
β_3	-0.296722	0.268435	0.269
Deviance	18.531		
AIC	66.447		

É perceptível que, apesar de uma pequena diminuição na Deviance, nenhum dos coeficientes foi significativo, indicando que não houve ajuste. Desta forma entende-se que a função de ligação escolhida está adequada.

3.2.3 Análise de Resíduos

Seguindo em frente com a análise, observando a Figura 6, nota-se que não parecem haver problemas graves em relação a pontos aberrantes. Por outro lado, parecem existir alguns pontos influentes, porém não causaram problemas inferenciais.

Ao observar atentamente o quarto gráfico da Figura 6, nota-se que parece haver um comportamento no Resíduo, podendo indicar que seja necessário modelar a variância dos dados.

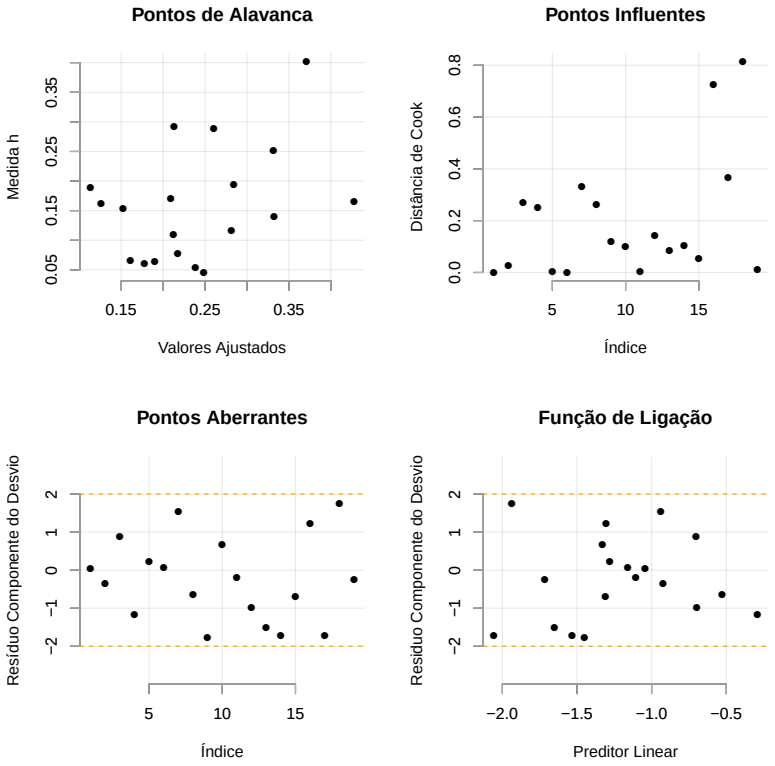

Figura 6: Diagnósticos do modelo aditivo.

3.2.4 Modelo de Quasi-Binomial

A partir da hipótese levantada de um comportamento na variância e pela quantidade razoável de 0 nos dados, ajustou-se um modelo de Quasi-Verossimilhança com função de ligação $logit$ e variância $\mu(1-\mu)$, caracterizando um modelo de Quasi-Binomial.

As estimativas são dadas pela Tabela 4.

Tabela 4: Estimativas do modelo de Quasi-Verossimilhança.

Efeito	Estimativa	Erro P.	Valor-p
β_0	-0.82535	0.43837	0.0780 .
β_1	-0.33797	0.0676	0.740
β_2	0.01710	0.01059	0.1261
Deviance	18.531		
AIC	-		

ARTHUR CESAR ROCHA
RA:94361

Os gráficos de adequação e diagnóstico são dados a seguir:

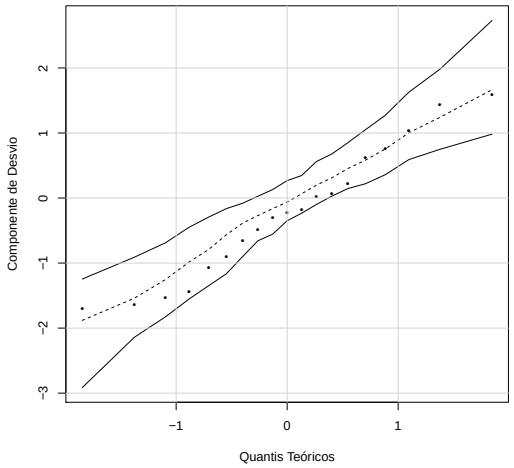

Figura 7: Envelope simulado para os componentes de desvio do modelo Quasi-Binomial.

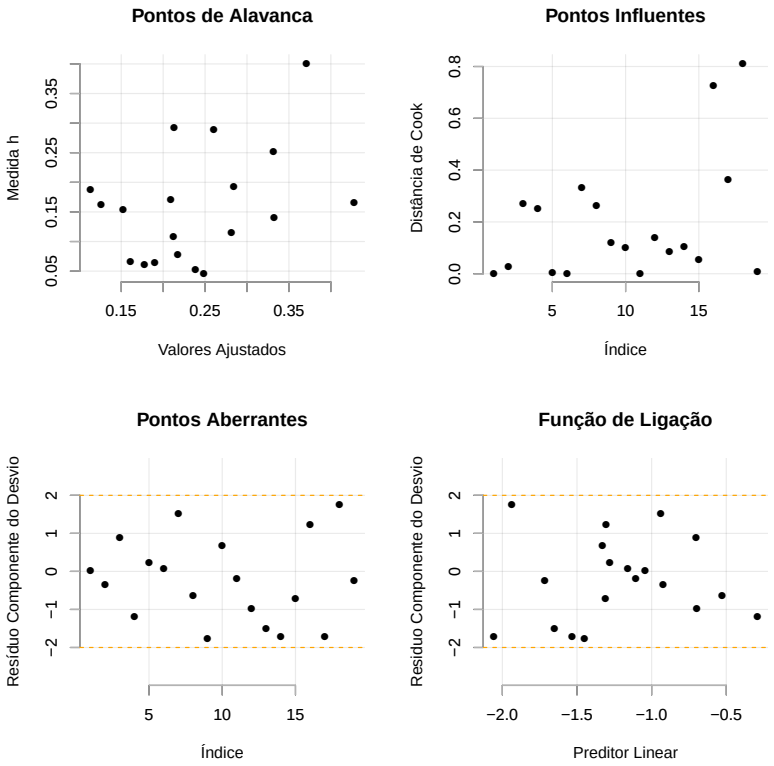

Figura 8: Diagnósticos do modelo Quasi-Binomial.

Nota-se que o modelo estudado teve desempenho similar ao modelo mais simples. Portanto, é preferível usar o primeiro modelo, pois é mais simples e robusto.

3.2.5 Modelo final

O modelo final ajustado foi:

$$log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = -0.8254 + 0.0171Calor - 0.3380Imersao \tag{4}$$

e portando, a função para estimar a proporção média é encontrada se $\hat{\pi}_i$ for "isolado". Desta forma, desenvolvendo:

$$exp\{log(\hat{\pi}_i) - log(1-\hat{\pi}_i)\} = exp\{-0.8254 + 0.0171Calor - 0.3380Imersao\} \tag{5}$$

$$\hat{\pi}_i - (1-\hat{\pi}_i) = exp\{-0.8254 + 0.0171Calor - 0.3380Imersao\} \tag{6}$$

$$2\hat{\pi}_i = 1 + exp\{-0.8254 + 0.0171Calor - 0.3380Imersao\} \tag{7}$$

$$\hat{\pi}_i = \frac{1 + exp\{-0.8254 + 0.0171Calor - 0.3380Imersao\}}{2} \tag{8}$$

4 Conclusão

A partir do conjunto de dados, modelou-se a proporção de placas inadequadas e chegou-se ao ajuste por um modelo binomial com função de ligação canônica (logit), considerando as duas covariáveis estudadas, sendo que a variável Imersão pareceu ter mais significância na explicação dessa proporção.

ARTHUR CESAR ROCHA
RA:94361

Referências

ACTION, P. **Portal action**. [S.l.]: Fonte: PORTAL ACTION: http://www. portalaction. com. br.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **Sao Paulo**, 2008.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004.

TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados-da teoria à prática. In: **VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa**. [S.l.: s.n.], 2000.