UNIVERSIDADE FEDERAL DO PARANÁ

Adriane Machado (GRR20149152),
Cinthia Zamin Cavassola(GRR20149075) e
Luiza Hoffelder da Costa(GRR20149107)

AJUSTE DE MODELO DE REGRESSÃO LOGÍSTICA REFERENTE À PRESENÇA DE CRESCIMENTO DE *ALICYCLOBACILLUS ACIDOTERRESTRIS* NO SUCO DE MAÇÃ

Curitiba
2017

Adriane Machado
Cinthia Zamin Cavassola
Luiza Hoffelder da Costa

AJUSTE DE MODELO DE REGRESSÃO LOGÍSTICA REFERENTES À PRESENÇA DE CRESCIMENTO DE *ALICYCLOBACILLUS ACIDOTERRESTRIS* NO SUCO DE MAÇÃ

Relatório técnico apresentado como atividade avaliativa na disciplina de Modelos Lineares Generalizados da Graduação em Estatística na Universidade Federal do Paraná.

Professor Dr. Cesar Augusto Taconeli

Curitiba
2017

RESUMO

Os dados se referem à presença ou ausência de crescimento da bactéria *Alicyclobacillus Acidoterrestris* CRA7152 no suco de maçã, levando em consideração as covariáveis pH, concentração de nisina, temperatura e Brix. Como a variável resposta foi apresentada de forma binária (presença/ausência), foi ajustado um modelo de regressão logística. O modelo de regressão logística foi utilizado para descrever o efeito das covariáveis referidas acima na probabilidade de crescimento da bactéria no suco de maçã, visando o controle de qualidade, desenvolvimento de processos industriais seguros, redução de contaminação da produção de suco de maçã e boas práticas na indústria de sucos em geral.

PALAVRAS-CHAVE: *ALICYCLOBACILLUS ACIDOTERRESTRIS.* SUCO DE MAÇÃ. PRESENÇA. REGRESSÃO LOGÍSTICA

SUMÁRIO

1. INTRODUÇÃO 05

2. MATERIAL E MÉTODOS..... 05

3. MODELAGEM ESTATÍSTICA..... 08

4. RESULTADOS E DISCUSSÃO 09

5. CONCLUSÃO 12

1. INTRODUÇÃO

O trabalho a seguir tem por objetivo analisar os fatores de influência na presença ou ausência de crescimento da bactéria *Alicyclobacillus Acidoterrestris* CRA7152 no suco de maçã, levando em consideração as covariáveis pH, concentração de nisina, temperatura e Brix, via ajuste de modelo de regressão logística.

Algumas bactérias produzem as bacteriocinas, que retardam ou inibem o crescimento de outras bactérias. A bacteriocina nisina, que é naturalmente produzida em vários alimentos fermentados, vem sendo consumida por humanos há séculos. A nisina tem seu uso aprovado em alimentos em mais de 50 países. Estudos têm mostrado a baixa toxicidade da nisina e sua alta eficiência como conservante em alimentos.

Brix é uma escala numérica de índice de refração de uma solução, comumente utilizada para determinar a quantidade de compostos solúveis numa solução de sacarose, utilizada geralmente para suco de fruta. A escala Brix é utilizada na indústria de alimentos para medir a quantidade aproximada de açúcares em sucos de fruta, vinhos e na indústria de açúcar, bem como outras soluções.

A variável resposta foi coletada no formato presença/ausência de crescimento de *Alicyclobacillus Acidoterrestris*. A critério dos pesquisadores, foram fixadas como covariáveis (variáveis que explicam o comportamento da resposta) pH, concentração de nisina, temperatura e Brix.

O objetivo final é quantificar as variáveis que influenciam o crescimento da bactéria por meio de um modelo de regressão logística, visando ao controle de qualidade de produção de suco de maçã.

2. MATERIAL E MÉTODOS

Utilizou-se uma base de dados obtida da URL:

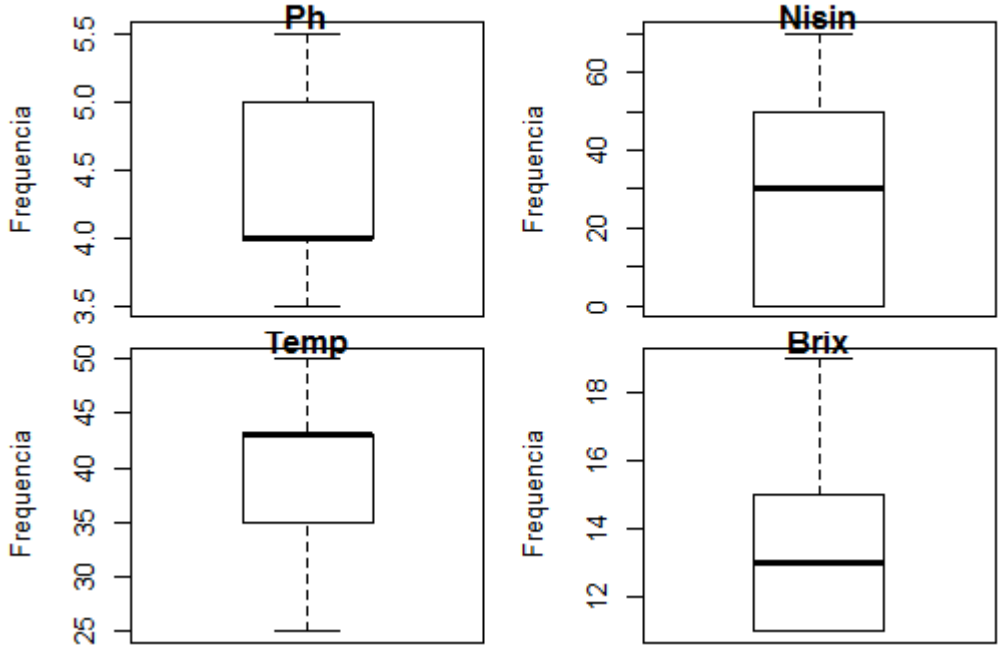
<http://www.stat.ufl.edu/~winner/datasets.html >

Esta base de dados apresenta 74 observações do fenômeno de presença ou ausência de crescimento de *Alicyclobacillus Acidoterrestris* CRA7152 como função de pH (3.5 a 5.5), Brix (11 a 19), temperatura (25 a 50°C) e concentração de nisina (0 a 70).

Foi feito uma análise exploratória visando avaliar preliminarmente se há relação entre essas variáveis. Verifica-se na análise gráfica abaixo que todas as covariáveis têm assimetria considerável e que, em se tratando de correlação

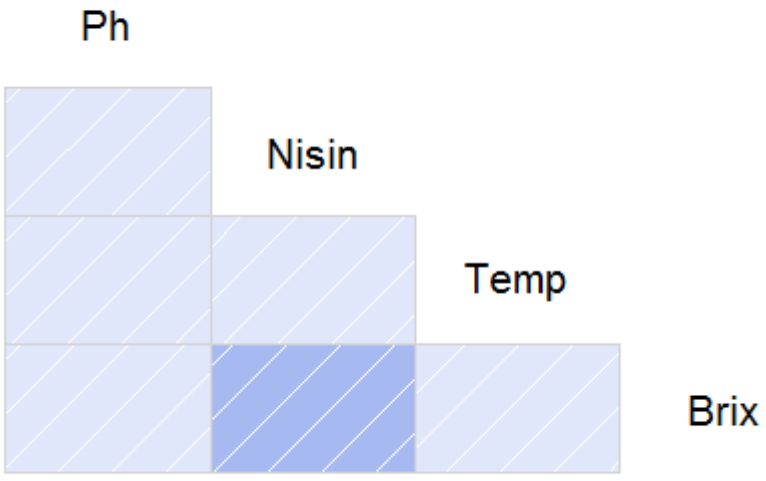
entre covariáveis, existe correlação entre todas elas, especialmente entre nisina e Brix:

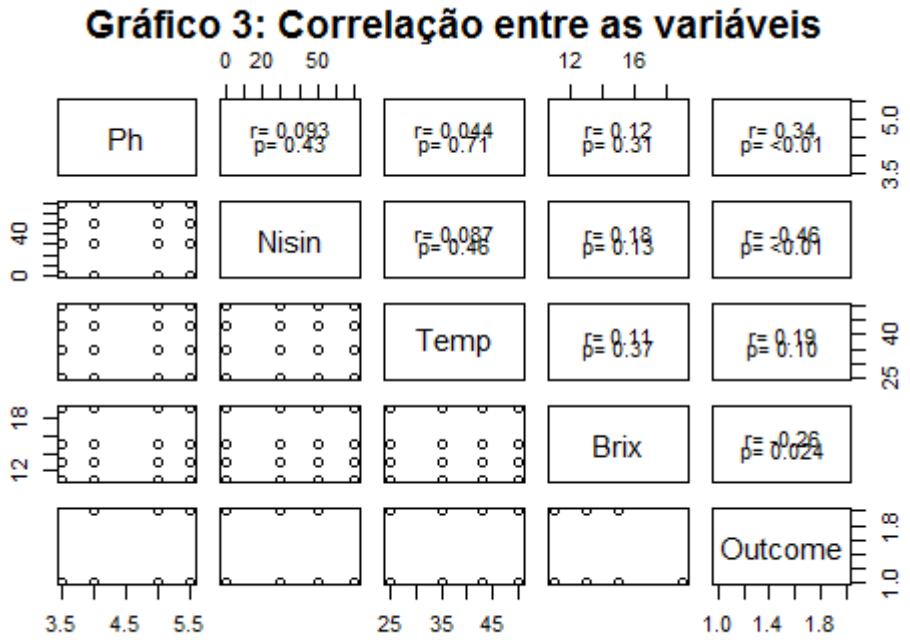
Gráfico 1: Box-plot referente as variáveis explicativas



Através do gráfico 1, verifica-se a assimetria de todas as covariáveis.

Gráfico 2: Correlação entre as variáveis





O gráfico 2 e o gráfico 3, apresenta a correlação entre as covariáveis. Nota-se que todas possuem correlação, e que a correlação entre Nisin e Brix é mais forte.

A tabela 1 abaixo apresenta as dez primeiras observações do conjunto de dados.

Tabela 1: Conjunto de Dados

	Brix	pH	Nisin	Temperatura	Outcome
	5.5	70	50	11	0
	5.5	70	43	19	0
	5.5	50	43	13	1
	5.5	50	35	15	1
	5.5	30	35	13	1
	5.5	30	25	11	0
	5.5	00	50	19	0
	5.5	00	25	15	1
	3.5	70	43	11	0
	3.5	70	35	13	0
Total:	**1.054**	**332**	**2.600**	**2.840**	

Dada a natureza binária da variável resposta (0 ou 1), utilizou-se os conceitos de modelos lineares generalizados. Mais especificamente, foi utilizado o modelo de regressão logística, que torna possível analisar e entender as relações entre a variável resposta e as variáveis explicativas. Para este tipo de dados a distribuição Binomial é a principal alternativa como componente aleatório do modelo, o componente sistemático é dado pela combinação linear das variáveis explicativas e para obtenção do modelo com o melhor ajuste possível, o primeiro passo foi verificar qual seria a melhor função

de ligação a se utilizar na regressão logística: para tal, foram ajustados modelos com todas as covariáveis, sem interação, testando-se as funções de ligação mais usuais para a regressão logística, quais sejam, logito (canônica), probito, e complemento log-log. Verificando os valores de AIC (Critério de Informação de Akaike), o qual indica um melhor ajuste de modelo quanto menor o seu valor, e de logverossimilhança, o qual indica um melhor ajuste de modelo quanto maior o seu valor, obtidos estes de cada modelo, optou-se pelo modelo com a função de ligação canônica, qual seja, a logito. A sua expressão geral é:

$$Yi\,|xi \sim Binomial(mi\,,\,\pi i)$$

$$g(\pi i) = \ln\left(\frac{\pi i}{1-\pi i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

3. MODELAGEM ESTATÍSTICA

Na sequência, estabelecida a função de ligação, foi efetuada a seleção de covariáveis, processo este que afere o impacto da inclusão ou exclusão de alguma das covariáveis no modelo, a fim de proporcionar o melhor ajuste possível. O algoritmo escolhido para este trabalho foi o “stepwise”, que faz uma inclusão e exclusão de variáveis no modelo utilizando como medida de seleção o AIC a cada iteração, ou seja, a cada modificação a maior ou a menor do número de covariáveis no modelo. Novamente, optou-se pelo modelo de menor AIC, permanecendo-se com o modelo acima.

Posteriormente, foi verificada a potencial significância de alguma interação, na medida em que os gráficos de correlação entre covariáveis apontavam a correlação entre nisina e Brix como especialmente alta, mas tal possibilidade foi descartada, tendo em vista que nenhuma interação era significativa o suficiente, considerando um alfa de 0,05, especialmente ao se considerar o aumento de um parâmetro, a perda de um grau de liberdade e o incremento de dificuldades de interpretação, o que vai contra o princípio da parcimônia. Ao final deste procedimento, a equação ajustada foi:

$$g(\pi i) = \ln\left(\frac{\pi i}{1-\pi i}\right) = -4.23845 + 1.29850pH - 0.05555Nisina + 0.10734Temp$$

Tabela 2: Resumo das Estimativas para o Modelo Ajustado

Variável	Parâmetro	Estimativa	Erro Padrão	Estatística Z	Pr (> z)
Intercepto	β_0	-4.23845	3.30824	-1.281	0.02048
pH	β_1	1.29850	0.56034	2.317	0.00443
Nisin	β_2	-0.05557	0.01953	-2.846	0.00443

Temperatura	β_3	0.10734	0.05401	1.987	0.04687
Brix	β_4	-0.35384	0.16968	-2.085	0.03704
Cod. Significância: AIC: 48.958	**‘*’ 0.01**	**‘**’ 0.001**			

Todas as covariáveis no modelo constaram como significativas a um α de 0,05, permanecendo no modelo:

O intercepto teve uma estimativa de -4,23845 com erro padrão de 3,30824; a covariável pH teve um coeficiente estimado em 1,2985 com um erro padrão de 0,56034 e p-valor de 0,02048, demonstrando significância para modelo frente a um α de 0,05; a covariável nisina teve um coeficiente estimado em -0,05557 com um erro padrão de 0,01953 e p-valor de 0,00443, demonstrando significância para modelo frente a um α de 0,05; a covariável temperatura teve um coeficiente estimado em 0.10734 com um erro padrão de 0.05401 e p-valor de 0.04687, demonstrando significância para modelo frente a um α de 0,05; por fim, a covariável Brix teve um coeficiente estimado em -0.35384 com um erro padrão de 0.16968 e p-valor de 0.03704, demonstrando significância para modelo frente a um α de 0,05.

O AIC do modelo foi calculado em 48.958. Ainda, a deviance do modelo caiu de 63.449 do modelo nulo (sem covariáveis e contando somente com o intercepto) para 38.958; como é desejável que a deviance seja cada vez mais reduzida em relação ao valor do modelo nulo quanto melhor seja o ajuste, temos bons indicativos de um ajuste satisfatório com covariáveis significativas.

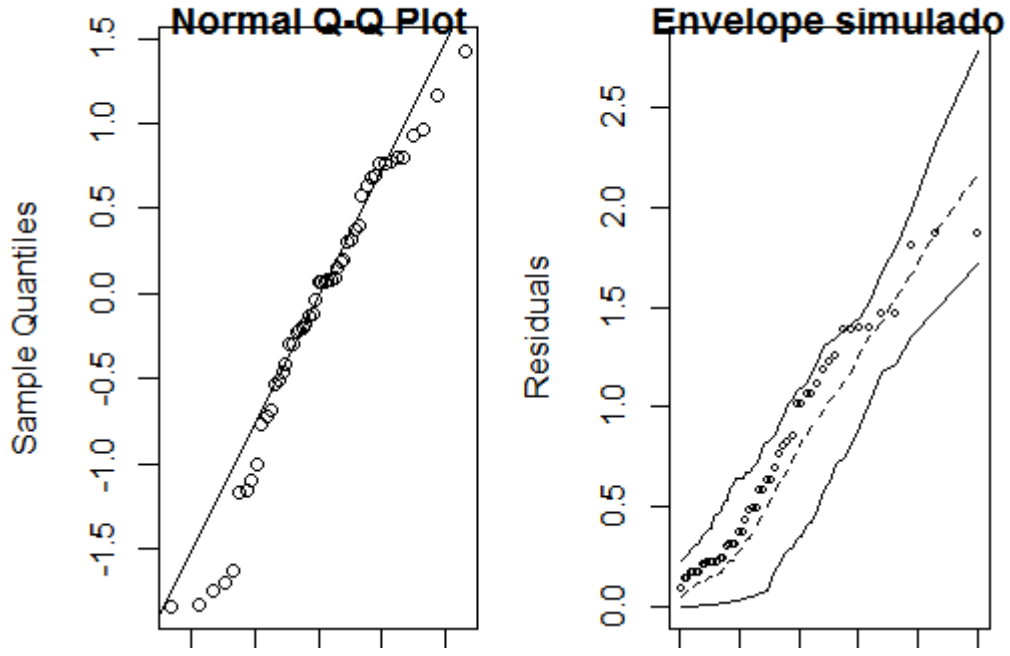
4. RESULTADOS E DISCUSSÃO

Definido o modelo, passa-se ao seu diagnóstico para verificação de qualidade de ajuste.

O gráfico de resíduos quantílicos é uma ferramenta valiosa de diagnóstico de ajuste, sendo desejável que os pontos alinhem-se consideravelmente à reta dos quantis teóricos, sem grandes distanciamentos; verifica-se que existe uma boa adesão à reta, o que indica um ajuste razoável.

Outro gráfico de grande valia para o diagnóstico do ajuste é o gráfico de envelope meio-normal: espera-se que os pontos representativos da resposta estejam dentro de um envelope que representa a banda de resíduos, o que é verificado no ajuste abaixo; há um único ponto quase que exatamente sobre a linha do envelope, o que não descaracteriza o indicativo de bom ajuste.

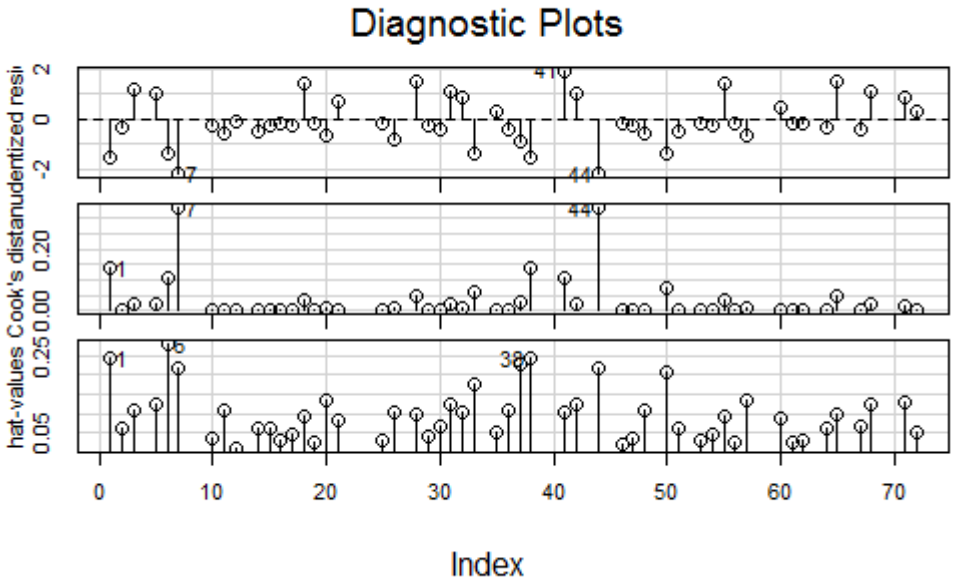
Gráfico 4: Análise de Diagnóstico do Modelo Proposto



Através do gráfico 4, verificamos um bom ajuste do modelo. Os resíduos encontraram-se próximos a reta e dentro do envelope simulado.

Na sequência foi promovida uma análise de influência para verificação de pontos atípicos que alterassem substancialmente os coeficientes da regressão, via gráfico de resíduos studentizados, distância de Cook e leverage; os gráficos apontaram algumas respostas sob suspeita (pontos 7, 38 e 44, principalmente, já que se repetiram em mais de um gráfico), porém a retirada de cada uma destas respostas do modelo não abalou os coeficientes tampouco sua significância, reforçando o modelo proposto.

Gráfico 5: Observações Influentes

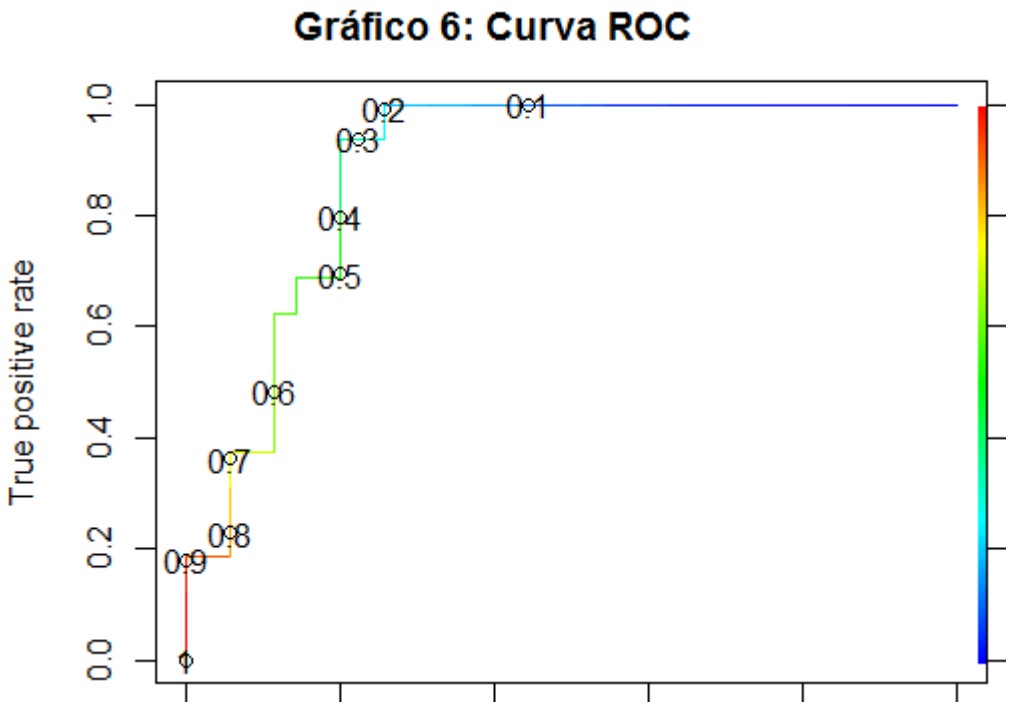


O gráfico 5, mostra os pontos sob suspeita de ser pontos atípicos. Foram devidamente testados e não altera os coeficientes.

Foi verificada ainda a curva ROC do modelo, a qual tem por objetivo verificar a proporção dos dados explicada pelo modelo proposto. Para se elaborar a curva ROC, é necessário definir um ponto de corte, ou um limiar de decisão, para se classificar e contabilizar o número de predições positivas e negativas.

Para cada ponto de corte são calculados valores de sensibilidade e especificidade, que podem então serem dispostos em um gráfico denominado curva ROC, que apresenta no eixo das ordenadas os valores de sensibilidade e nas abcissas, o complemento da especificidade.

A área abaixo da curva ROC está entre os valores 0 e 1, e quanto melhor o ajuste, melhor o índice de explicação das respostas pelo modelo, e consequentemente maior o índice da curva ROC; foi obtido um valor de 0,7, o que pode ser entendido como mais um fator de convencimento em prol do ajuste proposto. Conforme gráfico 6 abaixo:



5. CONCLUSÃO

Conclui-se que as covariáveis propostas para explicar a presença ou ausência de crescimento da bactéria *Alicyclobacillus Acidoterrestris* CRA7152 no suco de maçã explicam de forma abrangente o comportamento da resposta, que o modelo ajustado com base na regressão logística foi corretamente aplicado, com ajuste satisfatório e seleção parcimoniosa sem comprometer a profundidade da análise do fenômeno.