

# Trabalho de Dados Categorizados 1/2024

Profa. Maria Tereza Leão Costa

Bruno Gondim Toledo (15/0167636)

Rafael de Acypreste (20/0060023)

13/07/2024

## Sumário

<b>1</b>	<b>Introdução</b>	<b>2</b>
1.1	Análise exploratória . . . . .	2
<b>2</b>	<b>Metodologia</b>	<b>4</b>
2.1	Análise de Regressão Logística . . . . .	4
<b>3</b>	<b>Resultados</b>	<b>4</b>
3.1	Modelo apenas com intercepto . . . . .	4
3.2	Modelo acrescido do Nível de Fosfatase Ácida como variável preditora . . . . .	5
3.2.1	Comparação dos modelos . . . . .	7
3.3	Modelo Saturado . . . . .	7
3.3.1	Comparação dos modelos . . . . .	8
3.4	Demais modelos possíveis . . . . .	9
3.5	Análise a partir do modelo escolhido . . . . .	9
3.6	Aplicação do modelo escolhido ao conjunto de teste . . . . .	10
<b>4</b>	<b>Conclusão</b>	<b>10</b>

# 1 Introdução

Ao se constatar que um paciente desenvolveu câncer, é fundamental, para se decidir qual tratamento utilizar, saber se o câncer já se espalhou para os linfonodos próximos. Um estudo foi realizado com o objetivo de medir a capacidade de predição para o envolvimento nodal em câncer de próstata de várias variáveis pré-operatórias cuja coleta é menos invasiva que uma cirurgia.

- Numa primeira etapa, desejava-se avaliar especificamente o efeito do nível de fosfatase ácida na predição para envolvimento nodal; e
- Na segunda etapa do estudo se considerou além desta variável as outras variáveis pré-operatórias.

Com isso, o presente estudo objetiva analisar os dados coletados e verificar a capacidade de predição para o envolvimento nodal em câncer de próstata.

## 1.1 Análise exploratória

Os dados avaliados consistem no acompanhamento clínico de 146 pacientes com câncer de próstata. Um resumo das variáveis pode ser visto na Tabela 1.

Tabela 1: Resumo dos dados.

Variável	N = 102 <sup>1</sup>
Resultado da Radiografia	
Negativo	72 (71%)
Positivo	30 (29%)
Estágio do Tumor	
- grave	47 (46%)
+ grave	55 (54%)
Nível da Fosfatase Ácida	62 (50, 79)
Envolvimento Nodal (x100)	
Não	60 (59%)
Sim	42 (41%)

<sup>1</sup>n (%); Median (IQR)

A variável resposta de interesse é de Envolvimento Nodal. Nesse sentido, a Tabela 2 apresenta a distribuição das variáveis coletadas por pessoas que apresentaram ou não o envolvimento.

Tabela 2: Distribuição das variáveis por envolvimento nodal.

<b>Envolvimento nodal</b>	<b>Não, N = 60<sup>1</sup></b>	<b>Sim, N = 42<sup>1</sup></b>
Resultado da Radiografia		
Negativo	52 (87%)	20 (48%)
Positivo	8 (13%)	22 (52%)
Estágio do Tumor		
- grave	39 (65%)	8 (19%)
+ grave	21 (35%)	34 (81%)
Nível da Fosfatase Ácida	53 (50, 63)	74 (67, 84)

<sup>1</sup>n (%); Median (IQR)

O que se pode perceber é que as variáveis relativas ao resultado da tomografia, do estágio do tumor e do nível da Fosfatase ácida são mais elevados entre os pacientes que apresentaram envolvimento nodal. Entretanto, é preciso usar métodos estatísticos adequados para verificar se essas diferenças são significativas.

A relação entre o Envolvimento Nodal e o nível da Fosfatase Ácida pode ser vista na Figura 1. Por ela, também é reforçada a ideia de que o nível da Fosfatase Ácida é mais elevado entre os pacientes que apresentaram envolvimento nodal.

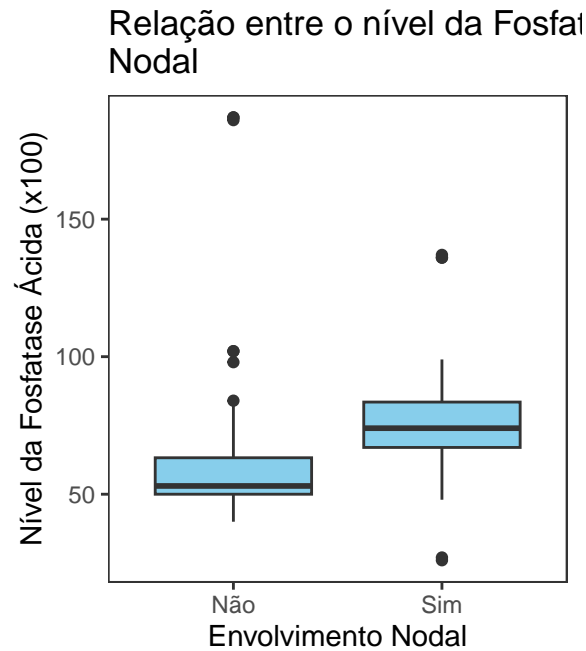


Figura 1: Relação entre o nível da Fosfatase Ácida e o Envolvimento Nodal.

## 2 Metodologia

### 2.1 Análise de Regressão Logística

Para a resposta às duas perguntas do problema, pode-se utilizar o modelo de regressão logística. Como a variável resposta, Envolvimento Nodal, é uma variável binária, pode-se codificá-la para que assuma o valor 1 se o paciente apresentou envolvimento nodal e 0 caso contrário.

Numa primeira etapa, para se avaliar apenas a capacidade do nível de fosfatase ácida em prever o envolvimento nodal, pode-se ajustar o modelo de regressão logística com a variável nível de fosfatase ácida como preditora. O modelo utilizado é dado pela equação (1).

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_{3i} \quad (1)$$

em que  $\pi_i$  é a probabilidade de um paciente  $i$  apresentar envolvimento nodal,  $X_{3i}$  é o nível de fosfatase ácida do paciente  $i$  e  $\beta_0$  e  $\beta_1$  são os coeficientes do modelo. A função de ligação logito é dada por  $\log \left( \frac{\pi_i}{1 - \pi_i} \right)$ , que tem como resposta o modelo linear visto na equação.

De maneira específica, o coeficiente  $e^{\beta_1}$  indica o quanto a razão de chances de um paciente apresentar envolvimento nodal aumenta para cada unidade de aumento no nível de fosfatase ácida.

Por fim, o modelo saturado inclui todas as variáveis disponíveis de Estágio do Tumor, Resultado da Radiografia e Nível da Fosfatase Ácida. Nesse caso, o modelo é dado pela equação (2).

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \quad (2)$$

em que  $X_{1i}$ ,  $X_{2i}$  e  $X_{3i}$  são as variáveis Nível da Fosfatase Ácida, Resultado da Radiografia e Estágio do Tumor, respectivamente.

## 3 Resultados

### 3.1 Modelo apenas com intercepto

Como estratégia inicial, pode-se avaliar o modelo com uma média única pra todos os níveis de Fosfatase Ácida. Nesse caso, o modelo é dado pela equação (3).

$$\log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 \quad (3)$$

O modelo com esse ajuste pode ser avaliado na Figura 2.

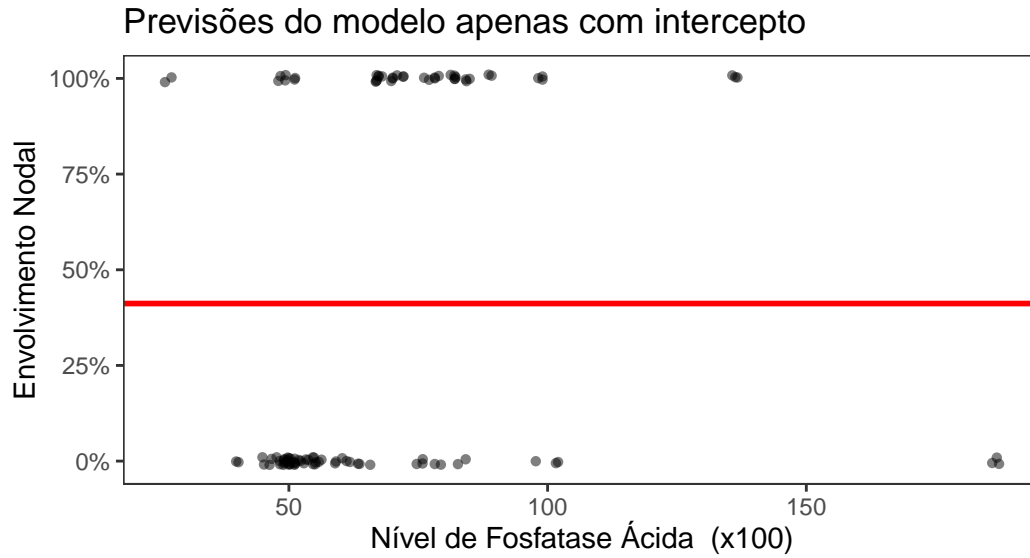


Figura 2: Modelo de regressão logística com intercepto. Os pontos foram ligeiramente agitados para facilitar a visualização.

As estatísticas de avaliação do modelo podem ser vistas na Tabela 3.

Tabela 3: Estatísticas do modelo de regressão logística com intercepto.

Deviance	AIC	BIC	Log Likelihood
138.2	140.2	142.8	-69.1

### 3.2 Modelo acrescido do Nível de Fosfatase Ácida como variável preditora

O modelo para o nível de Fosfatase Ácida como variável preditora é dado pela equação (1). As probabilidades previstas estão apresentadas na Figura 3.

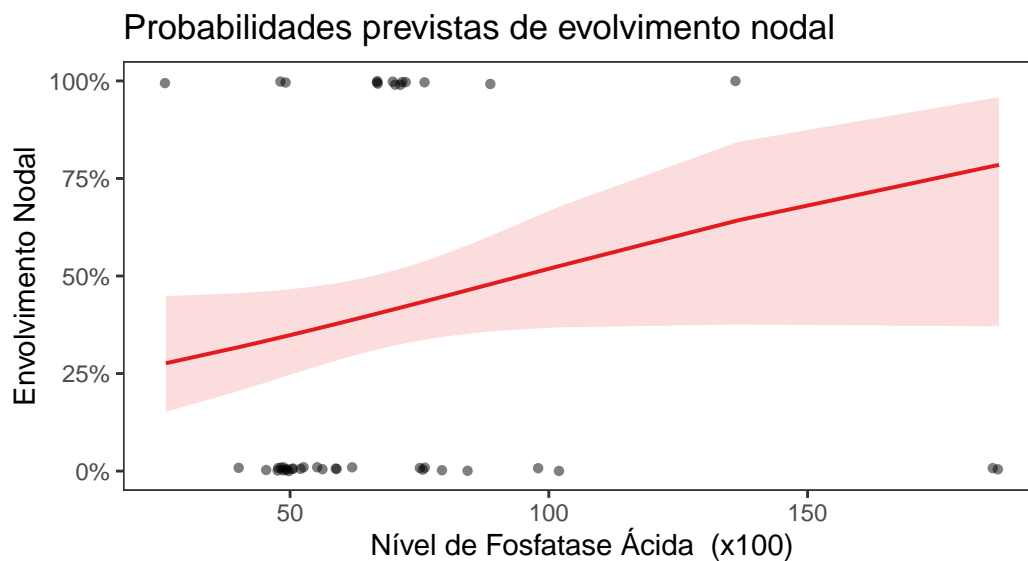


Figura 3: Modelo de regressão logística com o nível de Fosfatase Ácida como variável preditora. Os pontos foram ligeiramente agitados para facilitar a visualização.

As estatísticas do modelo quando é acrescido o nível de Fosfatase Ácida como variável preditora estão apresentadas na Tabela 4.

Tabela 4: Estatísticas do modelo de regressão logística com o nível de Fosfatase Ácida como variável preditora.

Variável	Estimativa	Erro Padrao	Valor Z	Pr(> z )
Intercepto	-1.326	0.565	-2.347	0.019
Nível de fosfatase ácida	0.014	0.008	1.828	0.068

O teste de Hosmer-Lemeshow pode ser utilizado para avaliar a qualidade do ajuste. Para o modelo da apresentado na Tabela 4, a estatística  $\chi^2$  do teste de Hosmer-Lemeshow é de 26.48 com um p-valor de  $8.7 \times 10^{-4}$ . O teste sugere que o modelo ajustado não é adequado para os dados.

Por fim, as estatísticas do modelo que inclui a variável “Nível de Fosfatase Ácida” como preditora estão apresentadas na Tabela 5.

Tabela 5: Estatísticas de qualidade do modelo de regressão logística com o nível de Fosfatase Ácida como variável preditora.

Deviance	AIC	BIC	Log Likelihood
134.45	138.45	143.7	-67.225

### 3.2.1 Comparação dos modelos

Uma forma de avaliar o modelo com a variável explicativa “Nível de Fosfatase Ácida” é comparar com o modelo apenas com intercepto. A Tabela 6 apresenta as estatísticas de comparação entre os modelos.

Tabela 6: Estatísticas de comparação entre os modelos de regressão logística com a variável “Nível de Fosfatase Ácida” e apenas com intercepto.

Teste	Estatística	GL	P-valor
Razao de Verossimilhanca	3.759	1	0.053
Wald	3.340	1	0.068
Score	3.759	1	0.053

Portanto, percebe-se que que nenhum dos 3 testes rejeitaram a hipótese nula a 5%. Isto é, o modelo não é significativamente melhor que o modelo que contém apenas o intercepto.

### 3.3 Modelo Saturado

As demais informações disponíveis podem auxiliar um melhor ajuste do modelo. Nesse caso, o modelo saturado é ajustado. As estatísticas do modelo saturado estão apresentadas na Tabela 7.

Tabela 7: Estatísticas do modelo de regressão logística saturado.

Variável	Estimativa	Erro Padrao	Valor Z	Pr(> z )
Intercepto	-5.1039	1.1454	-4.4559	0.0000
Resultado da Radiografia	2.7426	0.7360	3.7264	0.0002
Estágio do Tumor	3.2084	0.7543	4.2536	0.0000
Nível de Fosfatase Ácida	0.0278	0.0090	3.0762	0.0021

O teste de Hosmer-Lemeshow para o modelo saturado é de 22.05 com um p-valor de 0.00482. O teste sugere que o modelo saturado ajustado não é adequado para os dados.

Pode-se inspecionar as previsões do modelo saturado na Figura 4.

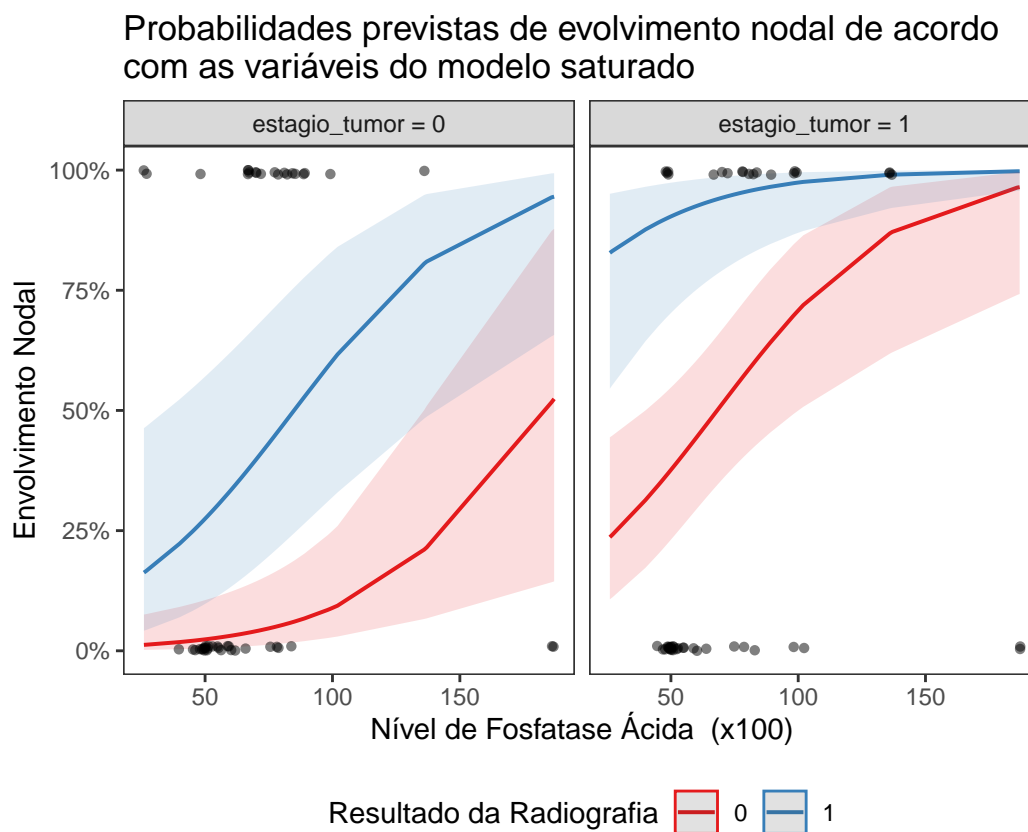


Figura 4: Modelo de regressão logística saturado. Os pontos foram ligeiramente agitados para facilitar a visualização.

As estatísticas de qualidade do modelo saturado estão apresentadas na Tabela 8.

Tabela 8: Estatísticas de qualidade do modelo de regressão logística saturado.

Deviance	AIC	BIC	Log Likelihood
86.194	94.194	104.69	-43.097

### 3.3.1 Comparação dos modelos

A Tabela 9 apresenta as estatísticas de comparação entre o modelo saturado e o modelo com intercepto, e com a variável "Nível de Fosfatase Ácida" (SE DER)



Tabela 9: Estatísticas de comparação entre os modelos de regressão logística saturado e com intercepto.

Teste	Estatística	GL	P-valor
Razao de Verossimilhanca	52.015	3	0
Score	52.015	3	0

### 3.4 Demais modelos possíveis

Também é possível avaliar as demais combinações de modelos, cujas estatísticas de qualidade de ajuste estão na Tabela 10.

Tabela 10: Estatísticas de qualidade dos modelos de regressão logística com as variáveis preditoras possíveis.

	Modelo	Deviance	AIC	BIC	Log Likelihood
1	$X4 \sim 1$	138.21	140.21	142.83	-69.10
3	$X4 \sim X1 + X2 + X3$	86.19	94.19	104.69	-43.10
4	$X4 \sim X1 + X2$	96.08	102.08	109.96	-48.04
5	$X4 \sim X1 + X3$	116.54	122.54	130.41	-58.27
6	$X4 \sim X2 + X3$	106.64	112.64	120.52	-53.32
8	$X4 \sim X1$	119.88	123.88	129.13	-59.94
7	$X4 \sim X2$	116.03	120.03	125.28	-58.01
2	$X4 \sim X3$	134.45	138.45	143.70	-67.22

O método *stepwise* pode ser utilizado para selecionar o melhor modelo. Nesse caso, considerando tal método e os dados apresentados de qualidade de ajustes dos modelos, a escolha indicada foi exatamente a do modelo completo.

Pode-se comparar o modelo completo com o modelo de segunda melhor \*deviance\*

### 3.5 Análise a partir do modelo escolhido

Para avaliar a qualidade do modelo, pode-se utilizar a curva ROC. A Figura 5 apresenta a curva ROC do modelo escolhido.

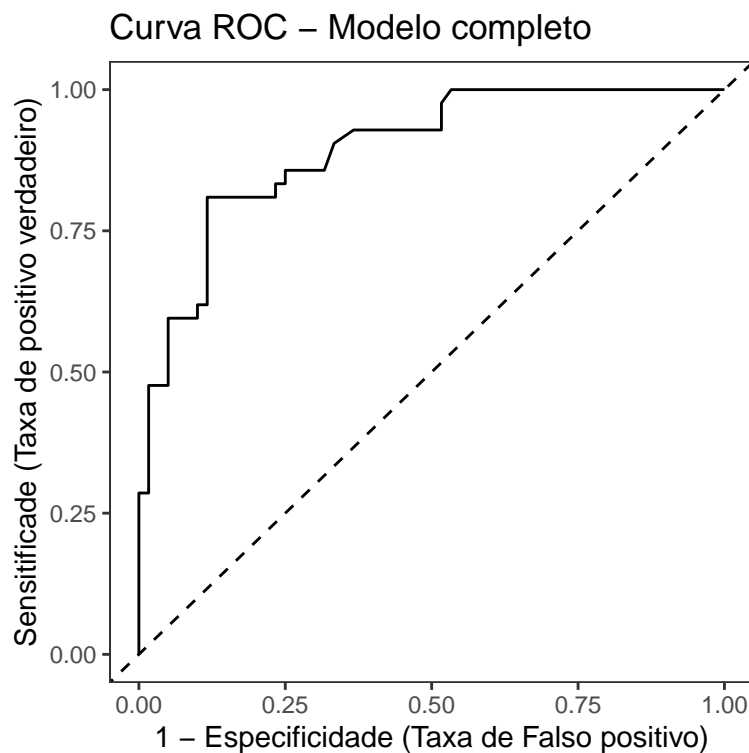


Figura 5: Curva ROC do modelo de regressão logística escolhido.

### 3.6 Aplicação do modelo escolhido ao conjunto de teste

O modelo escolhido é aplicado ao conjunto de teste. A Tabela 11 apresenta a matriz de confusão do modelo aplicado ao conjunto de teste. O valor de corte definido foi de 0.5.

Tabela 11: Matriz de confusão do modelo de regressão logística aplicado ao conjunto de teste.

Observado\Predito	0	1
0	54	10
1	10	28

Para esse caso específico, o modelo teve 80.39% de acertos, o que confere um caráter interessante para previsão dos resultados.

## 4 Conclusão