Trabalho de Dados Categorizados 1/2024

Profa. Maria Tereza Leão Costa

Bruno Gondim Toledo (15/0167636)
Rafael de Acypreste (20/0060023)

14/07/2024

Sumário

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**[1]
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%)

[1]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.

Envolvimento nodal	**Não**, N = 60[1]	**Sim**, N = 42[1]
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)

[1]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.

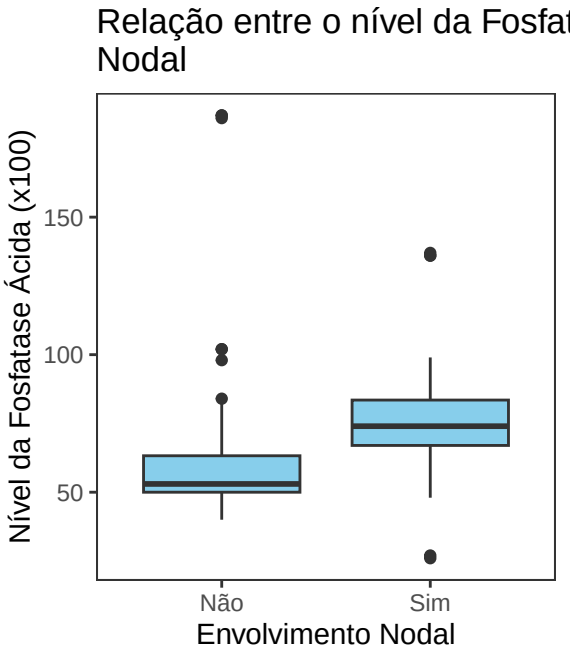

Figura 1: Relação entre o nível da Fosfatase Ácida e o Envolvimento Nodal.

2 Metodologia

2.1 Análise de Regressão Logística

A regressão logística é uma das técnicas disponíveis na família dos modelos lineares generalizados (MLG). Segundo Casella (2022), um MLG descreve uma relação entre a média de uma variável resposta Y e uma variável independente x. No caso do modelo de regressão logística, teremos três componentes, sendo eles: o aleatório, o sistemático e a função de ligação, em que:

- As variáveis resposta $Y_1, ..., Y_n$ são o componente aleatório. Elas são consideradas variáveis aleatórias independentes, cada uma com uma distribuição de uma família exponencial específica. Estes vetores não precisam ser identicamente distribuídos, mas cada um deles deve ter uma distribuição da mesma família.
- O componente sistemático é o modelo. Esta é a função da variável preditora x_i, linear **nos parâmetros**, que é relacionada à **média** de Y_i.
- Por fim, a função de ligação $g(\mu_i)$ estabelece ligação entre os dois componentes, afirmando que $g(\mu_i) = \beta_0 + \beta_i x_i$, em que $\mu_i = \mathbb{E}(Y_i)$

De forma geral, a função de ligação canônica para a regressão logística será a função *logito*, descrita como

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_{i(px1)} X_{i(px1)}, \tag{1}$$

em que $\beta_{i(p\times 1)} = [\beta_{i0}, \beta_{i1}, ..., \beta_{ip-1}]^T$, e $\mathbf{X_{i(px1)}} = [1, X_{i1}, X_{i2}, ..., X_{i\ p-1}]^T$, visto que para este modelo o vetor de respostas $Y_1, ..., Y_n$ são independentes e $Y_i \sim Bernoulli(\pi_i), \forall i$.

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, tal que a resposta esperada será igual a probabilidade $p = \mathbb{E}(Y)$ de um paciente apresentar envolvimento nodal. Segundo Morettin (2023) assim como no caso de modelos de regressão linear, o objetivo da análise é modelar a resposta esperada, que neste caso é uma probabilidade, como função da variável explicativa.

Numa primeira etapa, para se avaliar apenas a capacidade do nível de fosfatase ácida em predizer 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 (2).

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

em que π_i é a probabilidade de um paciente i apresentar envolvimento nodal, X_{3i} é o nível de fosfatase ácida do paciente i e β_0 e β_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^{β_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 mim, o modelo completo 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 (3).

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

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

Na regressão logistica, como não há mais uma conexão direta (por conta da função de ligação) entre Y_i e $\beta_{i(px1)}$ como existe no caso da regressão linear, a opção de ajustar o modelo pelo método de mínimos quadrados não é mais possível. Portanto, o ajuste do modelo dar-se-á pelo método de máxima verossimilhança, que pode ser descrito da seguinte forma:

Se considerarmos o modelo geral $Y_i \sim Bernoulli(\pi_i)$, onde $\pi(x) = F(\alpha + \beta x)$, para seguir a notação de Casella (2022), em que α representa o intercepto do modelo, que na matéria convencionamos chamar de β_0, tal que $\alpha = \beta_0$. Portanto, se considerarmos que $F_i = F(\alpha + \beta x_i)$, então a função de verossimilhança é:

$$L(\alpha, \beta|y) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} = \prod_{i=1}^{n} F_i^{y_i}(1-F_i)^{1-y_i}, \tag{4}$$

Com Log verossomilhança $\ell(\alpha, \beta|y)$ dada por:

$$\ell(\alpha, \beta|y) = \sum_{i=1}^{n}\{log(1-F_i) + y_i log(\frac{F_i}{1-F_i})\}. \tag{5}$$

Visto que o logarítmo é uma função monótona, a maximização desta verossimilhança pode ser obtida maximizando (5). Para isso, os estimadores de máxima verossimilhança de α e β correspondem à solução das equações de estimação $\sum_{i=1}^{n}\{y_i - \frac{exp(\hat{\alpha}+\hat{\beta}x_i)}{1-exp(\hat{\alpha}+\hat{\beta}x_i)}\} = 0$ e $\sum_{i=1}^{n} x_1\{y_i - \frac{exp(\hat{\alpha}+\hat{\beta}x_i)}{1-exp(\hat{\alpha}+\hat{\beta}x_i)}\} = 0$. Conforme Morettin (2023), esse sistema de equações não tem solução explícita, logo deve-se recorrer a métodos iterativos como o método de Newton-Raphson para uma solução numérica, que é o que os algoritmos implementados em ferramentas como SAS e R aplicam para estimar os modelos.

2.1.1 Métricas para seleção do modelo logístico

Para testar o ajuste aos dados, podemos aplicar técnicas como o teste da razão de verossimilhanças, o teste de Wald ou o teste score, que são descritos a seguir

- Teste da razão de verossimilhança

Seja $\ell(\hat{\beta}^{(0)})$ o valor do logarítmo da função de verossimilhança do modelo sob H_0), e $\ell(\hat{\beta}^{(1)})$ o valor do logarítmo da função de verossimilhança do modelo sob H_1), então a estatística de teste é dada por $G = -2(\ell(\hat{\beta}^{(0)}) - \ell(\hat{\beta}^{(1)}))$, que segue uma distribuição χ^2 com p graus de liberdade, em que p é o número de parâmetros adicionados ao modelo da hipótese alternativa.

- Teste de Wald

Seja o vetor de parâmetros do modelo ajustado $\hat{\beta}$ assintoticamente normal, tal que $\hat{\beta} \sim N_p(\beta, K^{-1})$, onde $K = \sigma^{-2}X^TX$ é a matriz de informação (Cordeiro, 2024, p. 79). Sob a hipótese nula $H_0)\beta_k = 0$, a estatística do teste é dada por $z^* = \frac{\beta_k}{s\{\beta_k\}} \sim N(0,1)$ no caso de um parâmetro, que é a situação onde este teste será aplicado.

- Teste de escore

O teste de escore de Rao é definida por Cordeiro (2024) a partir da função escore como $S_R = U_1^T(\tilde{\beta})\widetilde{Cov}(\hat{\beta}_1)U_1(\tilde{\beta})$, sendo $\widetilde{Cov}(\hat{\beta}_1)$ a matriz $Cov(\hat{\beta}_1)$ avaliada nos vetores de parâmetros dos modelos sob $H_0) $ e H_1), e $U_1(\tilde{\beta})$ o vetor de derivadas parciais da função de verossimilhança em relação aos parâmetros do modelo sob H_1). Essa estatística tem distribuição qui-quadrado, com número de graus de liberdade igual a quantidade de parâmetros adicionais da hipótese alternativa.

Uma das métricas possíveis para avaliação e seleção de modelos é a análise da *deviance* do modelo. Segundo Cordeiro et al. (2024), a análise de desvio (se referenindo à *deviance*) é uma generalização da análise de variância (ANOVA) para os modelos lineares generalizados, visando obter, a partir de uma sequência de modelos encaixados, cada modelo incluindo mais termos do que os anteriores, os efeitos de variáveis explanatórias, os fatores e suas interações. Usa-se o desvio como uma medida de discrepância do modelo e forma-se uma tabela de diferença de desvios, montando assim o que pode ser chamado de tabela ANODEV ("Analisys of Deviance), em analogia à tabela ANOVA. Buscaremos o modelo com menor valor de *deviance* entre os modelos que se ajustam aos dados.

A comparação entre modelos também pode ser feita utilizando o critério de informação de Akaike (AIC). Giolo (2017) define este critério como sendo: AIC = -2(log verossimilhança - número de parâmetros do modelo), em que buscaremos o modelo que minimiza este valor. Podemos utilizar ainda o Critério Bayesiano de Schwarz $SBC_p = -2\ell(b) + pln(n)$, ou o critério Log da verossimilhança: $-2\ell(b)$.

Com estas ferramentas, estamos preparados para analisar e selecionar um modelo útil que se ajuste bem aos dados.

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

[H]

Deviance	AIC	BIC	Log Likelihood
138.2	140.2	142.8	-69.1

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 (6).

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 \tag{6}$$

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

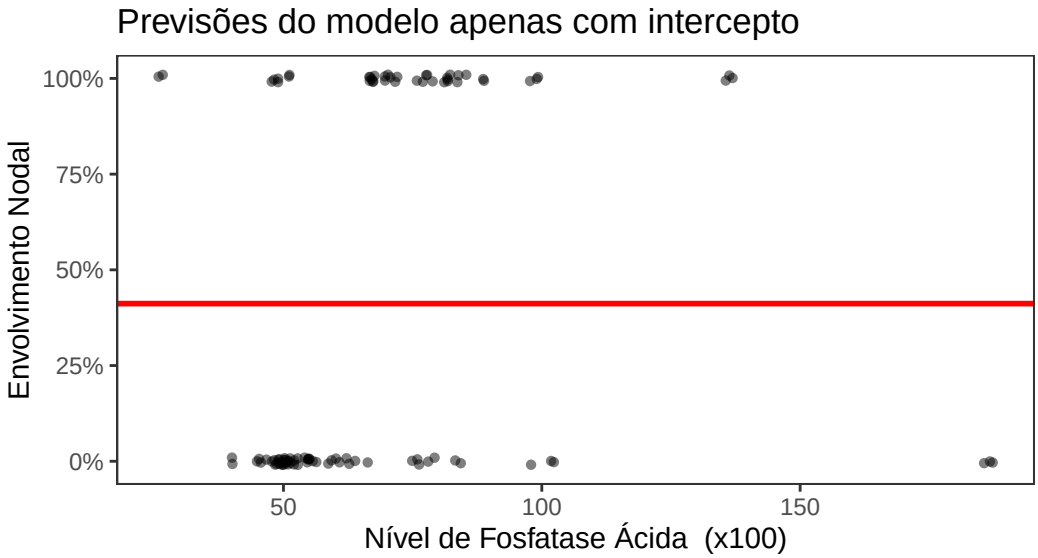

Figura 2: Modelo de regressão logística com intercepto. Os pontos foram ligeiramente agitados para facilitar a visualização.

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

Como o modelo inclui apenas o intercepto, a estimativa pontual é dada por 0.412, que é a proporção de pacientes com envolvimento nodal. O intervalo de confiança de 95% para essa proporção é [0.316; 0.507].

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

[H]

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

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 (2). As probabilidades previstas estão apresentadas na Figura 3.

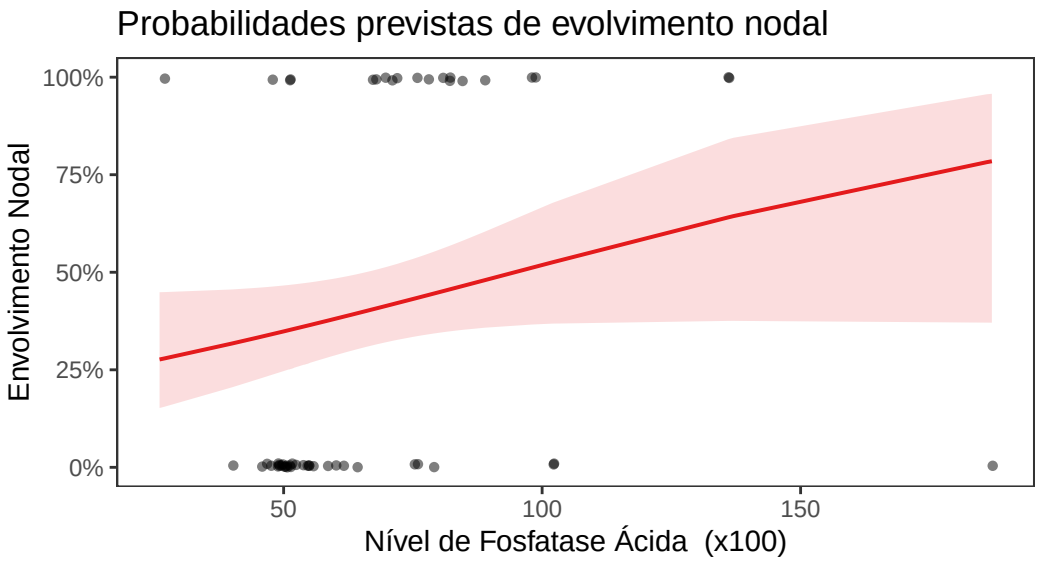

Figura 3: Modelo de regressão logística com o nível de Fosfatase Ácida como variável preditora. Os pontos foram ligeiramente agitados para facilitar a visualização.

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

Como o modelo apresenta uma variável não categórica como variável explicativa, a deviance residual não apresenta distribição aproximadamente qui-quadrado. Nesse caso, as maneiras tradicionais de análise dos resíduos (como o gráfico de resíduos de Pearson ou dos resíduos padronizados) não são apropriadas (Angresti, 2019, p. 131). Portanto, uma aproximação para avaliar a qualidade do ajuste é o teste de Hosmer-Lemeshow .

O teste de Hosmer-Lemeshow pode ser utilizado para avaliar a qualidade do ajuste. Para o modelo da apresentado na Tabela 4, a estatística χ^2 do teste de Hosmer-Lemeshow é de 26.48

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

[H]

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

com um p-valor de 8.7×10^{-4}. O teste sugere que o modelo ajustado não é adequado para os dados.

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.

[H]

Deviance	AIC	BIC	Log Likelihood
134.45	138.45	143.7	-67.225

Por fim, pela Tabela 4, o valor e_1^{β} fornece o incremento (ou redução) na razão de chances ao se aumentar uma unidade no Nível de Fosfatase Ácida é zero. O intervalo de confiança de 95% para essa proporção é [0.9998; 1.031]. Como o intervalo contem o 1 quando avaliado em 4 casas decimais, não há evidências de que o nível de Fosfatase Ácida influencia o envolvimento nodal ao nível de confiança desejado.

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.

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 Completo

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

Tabela 7: Estatísticas do modelo de regressão logística completo.

[H]

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

Tabela 7.

Novamente, como há uma variável quantitativa como variável explicativa, o teste de Hosmer-Lemeshow é sugerido pela literatura. Para o modelo completo a estatística do teste é de 22.05 com um p-valor de 0.00482. O teste sugere que o modelo completo ajustado não é adequado para os dados.

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

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

[H]

Deviance	AIC	BIC	Log Likelihood
86.1937	94.1937	104.6936	-43.0969

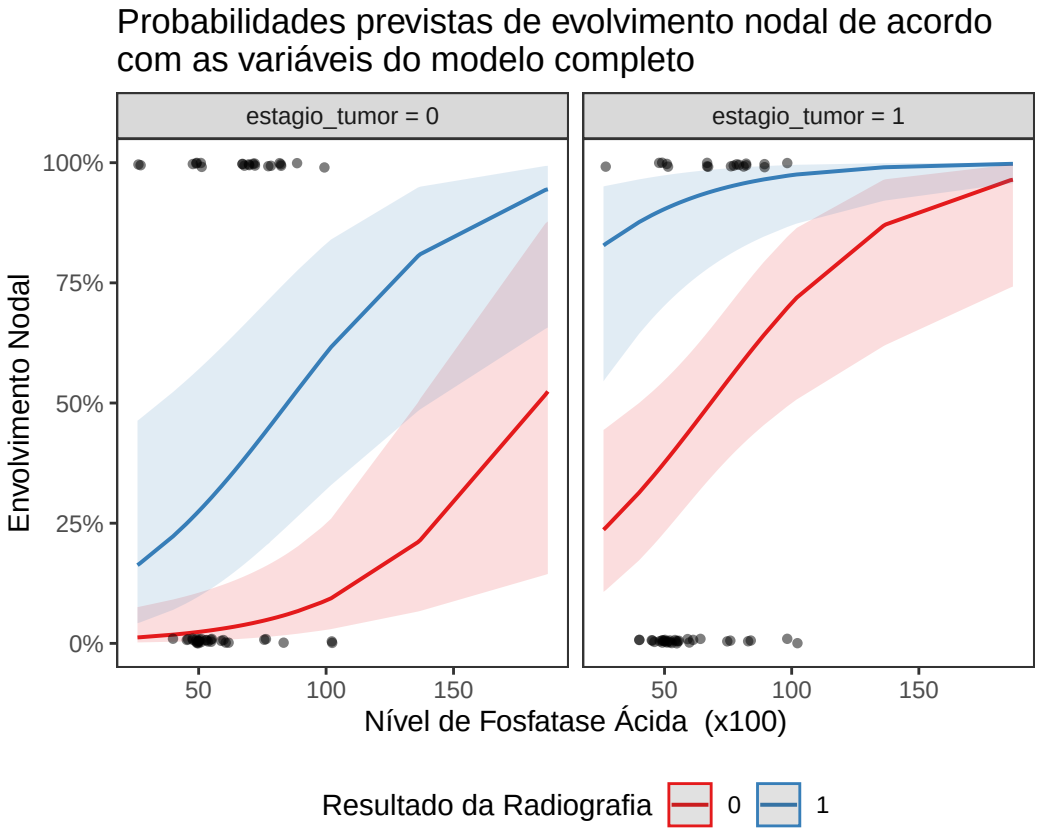

Figura 4: Modelo de regressão logística completo. Os pontos foram ligeiramente agitados para facilitar a visualização.

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

Portanto, pode-se visualizar as estimativas pontuais e seus respectivos intervalos de confiança para a razão de chances conforme a Tabela 9.

De acordo com os dados apresentados, a variável de Nível de Fosfatase Ácida se torna ligeiramente significativa para o modelo de regressão logística. A partir do modelo, o aumento em uma unidade no nível de fosfatase ácida aumenta a razão de chances de envolvimento nodal entre (1.011; 1.048) vezes.

Tabela 9: Estimativas pontuais e intervalos de confiança para a razão de chances do modelo completo.

[H]

Variável	Estimativa Pontual	2.5 %	97.5 %	P-valor
Resultado da Radiografia	15.5278	4.2459	82.7852	0.0002
Estágio do Tumor	24.7386	6.6540	137.7293	0.0000
Nível de Fosfatase Ácida	1.0282	1.0106	1.0480	0.0021

Tabela 10: Estatísticas de comparação entre os modelos de regressão logística completo e com intercepto.

[H]

Teste	Estatistica	GL	P-valor
Razao de Verossimilhanca	52.01516	3	0
Score	52.01516	3	0

Ademais, o resultado positivo na radiografia aumenta a razão de chances de envolvimento nodal entre (4.246; 82.785) vezes. Por fim, o estágio do tumor também aumenta a razão de chances de envolvimento nodal entre (6.654; 137.729) vezes.

3.3.1 Comparação dos modelos

A Tabela 10 apresenta as estatísticas de comparação entre o modelo completo e o modelo com intercepto. e com a variável "Nível de Fosfatase Ácida" (SE DER)

3.4 Demais modelos possíveis

Também é possível avaliar as demais combinações de modelos, cujas estatísticas de qualidade de ajuste estão na Tabela 11.

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.

Tabela 11: Estatísticas de qualidade dos modelos de regressão logística com as variáveis preditoras possíveis.

[H]

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

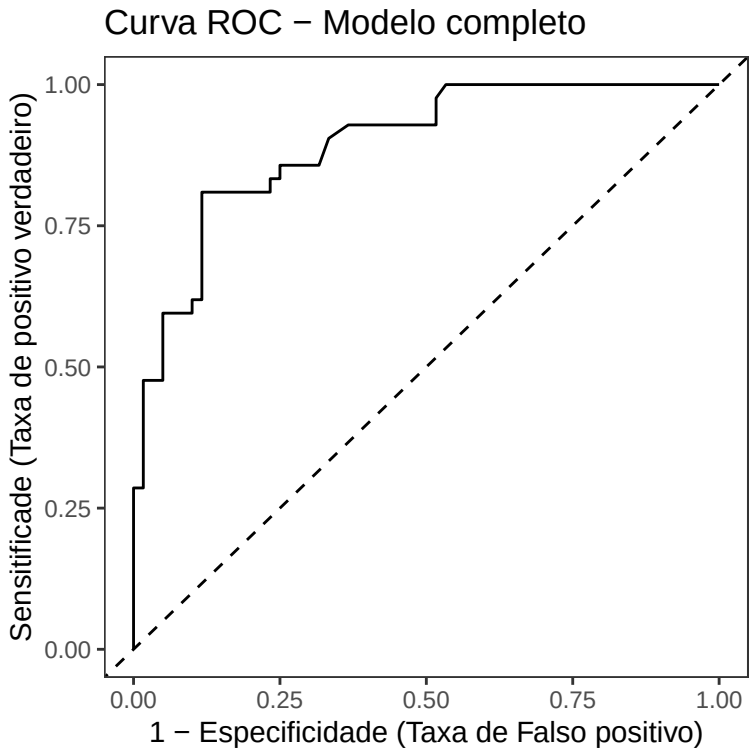

Figura 5: Curva ROC do modelo de regressão logística escolhido.

Tabela 12: Matriz de confusão do modelo de regressão logística aplicado ao conjunto de teste.

[H]

Observado\Predito	0	1
0	54	10
1	10	28

3.6 Aplicação do modelo escolhido ao conjunto de teste

O modelo escolhido é aplicado ao conjunto de teste. A Tabela 12 apresenta a matriz de confusão do modelo aplicado ao conjunto de teste. O valor de corte definido foi de 0.5.

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

4 Conclusão

Como os dados levantados pela pesquisa conseguem fornecer informações importantes para o diagnóstico do espalhamento do câncer de próstata para os lifonodos, pode-se sugerir a coleta dessas informações para novos pacientes de modo a auxiliar no diagnóstico de maneira mais direta, rápida e menos invasiva. O modelo escolhido apresentou uma taxa de acertos de 80.39%, o que confere um caráter interessante para previsão dos resultados e bons indicativos diagnósticos iniciais.

O modelo de regressão logística escolhido foi o modelo completo, que apresentou uma boa qualidade de ajuste. A curva ROC do modelo indica que o modelo é capaz de distinguir entre os grupos de interesse.

Além disso, de maneira individual, as três variáveis preditoras apresentaram um bom poder de auxílio na previsão das razões de chance. A variável "Nível de Fosfatase Ácida" foi a que apresentou o menor poder, sendo ligeiramente significativa. Ainda assim, como sua escala de variação é considerável, à medida que o nível se eleva, pode-se considerar um aumento na chance de apresentar envolvimento nodal. Já o resultado das variáveis de resultado positivo da radiografia e de estágio grave do tumor apresentam indicações elevadas sobre o aumento da razão de chances de um paciente apresentar envolvimento nodal.

5 Referências

- Agresti, Alan. "An introduction to categorical data analysis.". Third Edition John Wiley & Sons, Inc., 2019.

- Cordeiro, Gauss M.; Demétrio, Clarice G. B.; Moral, Rafael A. Modelos lineares generalizados e aplicações. Blucher. São paulo, 2024.
- George Casella, Roger L Berger. Inferência estatística. [tradução Solange Aparecida Visconde]. - São Paulo: Cengage Learning, 2022.
- Giolo, Suely Ruiz. Introdução à análise de dados categóricos com aplicações. Blucher. São paulo, 2017.
- Morettin, Pedro Alberto. Singer, Julio da Motta. Estatística e ciência de dados. 1ª ed. LTC. Rio de Janeiro, 2023.