Universidade Federal do Paraná - Departamento de Estatística

Especialização em Data Science e Big Data

Disciplina: Modelos Estatísticos

Prof. José Luiz Padilha da Silva

Avaliação

Vamos considerar a aplicação de um modelo linear generalizado com resposta binomial e função de ligação logito (regressão logística). Os dados são referentes a uma amostra de 699 nódulos de mama, e estão disponíveis na página da disciplina no arquivo `breast.csv`. O objetivo é ajustar um modelo preditivo, que permita classificá-los em benignos ou malignos com base num conjunto de covariáveis. As variáveis disponíveis na base são as seguintes:

- **CL**: *Clump Thickness*;
- **MA**: *Marginal Adhesion*;
- **BC**: *Bare Nucleus*;
- **Class**: *benign*, para benigno; *malignant*, para maligno (variável resposta).

As três primeiras variáveis (variáveis explicativas) são expressas numa escala discreta, com valores 0, 1, 2,. . . , 10.

a) Faça uma análise descritiva dos dados, explorando, em particular, as associações entre a resposta e cada covariável;

Para fins preditivos, vamos separar, aleatoriamente, a base em duas (a primeira, com aproximadamente 75% dos dados, para o ajuste, e a outra, com os demais dados, para validação). Ajuste o modelo de regressão logística com os dados da primeira amostra.

b) Obtenha a acrácia, a sensibilidade e a especificidade para a regra de classificação em que o tumor é classificado como maligno se a probabilidade estimada for maior que $p_0 > 0.5$. Faça isso tanto para a base de ajuste quanto para a base de validação.

c) Repita o item anterior para os seguintes valores de p_0: 0.1, 0.3, 0.7 e 0.9;

d) Produza a curva ROC. Se admitirmos iguais custos de má-classificação, indique uma regra de classificação (valor de p_0) adequada;

e) Obtenha a melhor regra de classificação para os seguintes cenários:

i. Prevalência = 0.5; razão de custos =2;
ii. Prevalência = 0.2; razão de custos =2;
iii. Prevalência = 0.5; razão de custos =10;
iv. Prevalência = 0.2; razão de custos =2,

onde a prevalência é a proporção de tumores malignos na população, e a razão de custos refere-se a quantas vezes um falso negativo é mais "caro" que um falso positivo.

Para os itens seguintes, use a base completa e ajuste novamente o modelo de regressão logística.

f) Apresente a expressão do modelo ajustado nas escalas do preditor, da *odds* e da probabilidade de tumor maligno;

g) Interprete as estimativas obtidas para os parâmetros do modelo, com base em suas magnitudes, sinais e significância estatística;

h) Obtenha intervalos de confiança (95%) para os parâmetros do modelo;

i) Obtenha a chance e probabilidade estimadas de nódulo maligno para uma unidade em que $CT = 5$, $MA = 3$ e $BC = 4$;

j) Proceda o diagnóstico do ajuste com base na análise dos resíduos. O modelo parece se ajustar bem aos dados?