Universidade Federal do Paraná - Departamento de Estatística

Especialização em Data Sciece e Big Data

Prof. Cesar Augusto Taconeli

Avaliação - 01/09/2018

Vamos considerar a aplicação de um modelo linear generalizado com resposta binomial e função de ligação logito (regressão logística). Os dados são referentes a 699 nódulos de mama. O objetivo é ajustar um modelo preditivo, que permita classificá-los em benignos ou malignos com base num conjunto de covariáveis. Segue a descrição das variáveis usadas na análise:

- **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 numérica, com valores 0, 1, 2,..., 10. Além disso, a variável resposta foi codificada, para a análise, de forma que $y = 0$, se o tumor é benigno, e $y = 1$, caso o tumor seja maligno. Assim, vamos modelar a probabilidade (ou melhor, a chance) de um tumor ser maligno.

Na sequência são apresentadas as dez primeiras linhas da base, a título de ilustração.

```
breast <- read.csv2('breast.csv')[,-1]
head(breast, 10)
```

```
##    CT MA BC     Class
## 1   5  1  3    benign
## 2   5  5  3    benign
## 3   3  1  3    benign
## 4   6  1  3    benign
## 5   4  3  3    benign
## 6   8  8  9 malignant
## 7   1  1  3    benign
## 8   2  1  3    benign
## 9   2  1  1    benign
## 10  4  1  2    benign
```

Para avaliar a capacidade preditiva do modelo a ser ajustado, a base foi dividida, aleatoriamente, em duas novas bases: a primeira, com 500 observações, para o ajuste; a segunda, com 199 linhas, para validação. Na sequência é apresentado o resumo do modelo ajustado.

```
set.seed(232)
ordem <- sample(1:nrow(breast))
breast_aj <- breast[ordem[1:500],] ### Base de ajuste.
breast_pred <- breast[ordem[501:nrow(breast)],] ### Base de validação.
```

```r
ajuste <- glm(Class ~ CT + MA + BC, data = breast_aj, family = binomial)
summary(ajuste)
```

```
##
## Call:
## glm(formula = Class ~ CT + MA + BC, family = binomial, data = breast_aj)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5333  -0.1925  -0.0702   0.0367   2.4677
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.9033     1.0396  -9.526  < 2e-16 ***
## CT            0.9118     0.1268   7.193 6.34e-13 ***
## MA            0.4917     0.1078   4.562 5.07e-06 ***
## BC            0.8925     0.1460   6.113 9.75e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 636.98  on 499  degrees of freedom
## Residual deviance: 129.33  on 496  degrees of freedom
## AIC: 137.33
##
## Number of Fisher Scoring iterations: 7
```

O modelo foi usado na classificação dos dados de validação. Os seguintes resultados foram obtidos:

```r
preds <- predict(ajuste, newdata = breast_pred, type = 'response')
t1 <- table(preds < 0.5, breast_pred$Class)
dimnames(t1)[[1]] <- c('Teste positivo', 'Teste negativo')
dimnames(t1)[[2]] <- c('Tumor benigno', 'Tumor maligno')
t1
```

```
##
##                  Tumor benigno Tumor maligno
##   Teste positivo             3            64
##   Teste negativo           122            10
```

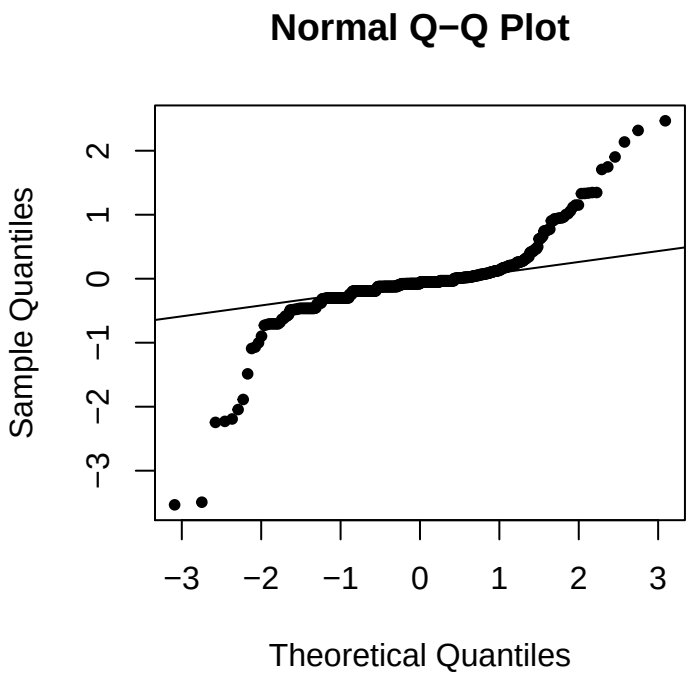

Figura 1: Gráfico quantil-quantil normal para os resíduos componentes da deviance

```
qqnorm(resid(ajuste), pch = 20)
qqline(resid(ajuste))
```

Com base nos resultados apresentados, responda as seguintes questões.

1. Para uma unidade a mais no escore de CT (fixados os valores de MA e BC), a chance de um tumor maligno:

a) É acrescida em 0.9118;
b) Fica multiplicada por 0.9118;
c) É acrescida em 2.4888;
d) Fica multiplicada por 2.4888;
e) Não se altera.

2. Para nódulos com CT=4, MA=4 e BC=5, a probabilidade estimada do tumor ser maligno é:

a) 0.1732;

b) 0.8268;

c) 0.5432

d) 0.4568;

e) 1.

3. O limite inferior de um intervalo de confiança 95% para o parâmetro correspondente à variável CT é:

a) 0.2138;

b) 0.6632;

c) 0.7850;

d) 0.9118;

e) 0.

4. A sensibilidade e a especificidade do modelo são estimadas, respectivamente, por:

a) 0.9552 e 0.9242;
b) 0.8648 e 0.9760;
c) 0.3216 e 0.6130;
d) 0.7692 e 0.9346;
e) 1 e 1.

5. Com base apenas no gráfico de resíduos apresentado, podemos afirmar que:

a) Os resíduos não têm distribuição normal, e portanto o modelo não está bem especificado;
b) Os resíduos têm distribuição normal, e portanto o modelo está bem especificado;
c) Os resíduos não têm distribuição binomial, e portanto o modelo não está bem especificado;
d) Os resíduos têm distribuição binomial, e portanto o modelo está bem especificado;
e) Nenhuma das afirmações é necessariamente verdadeira.