

Regressão Logística II

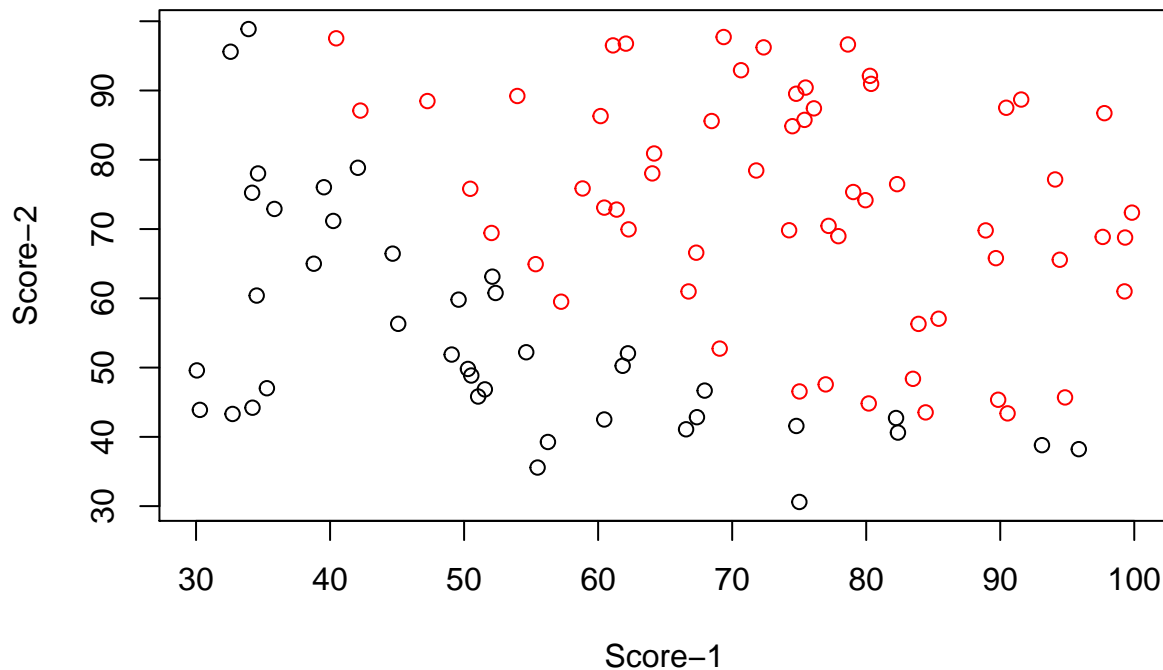
Francisco Caramelo

Wednesday, December 9, 2015

O exemplo que vamos usar para mostrar a avaliação de uma regressão logística faz uso de dados de dois testes e o resultado da admissão num dado curso.

O gráfico seguinte mostra os resultados na base de dados.

```
plot(MyData$score.1, MyData$score.2, col=as.factor(MyData$label), xlab="Score-1", ylab="Score-2")
```



Vamos então aplicar a regressão logística.

```
AdmiLR = glm(label ~ score.1 + score.2, data=MyData, family=binomial)
summary(AdmiLR)
```

```
##
## Call:
## glm(formula = label ~ score.1 + score.2, family = binomial, data = MyData)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19287  -0.18009   0.01577   0.19578   1.78527
```

```
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.16133     5.79836  -4.339 1.43e-05 ***
## score.1      0.20623     0.04800   4.297 1.73e-05 ***
## score.2      0.20147     0.04862   4.144 3.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 134.6  on 99  degrees of freedom
## Residual deviance:  40.7  on 97  degrees of freedom
## AIC: 46.7
##
## Number of Fisher Scoring iterations: 7
```

Os intervalos de confiança para os coeficientes obtidos são determinado por:

```
confint(AdmiLR)
```

```
## Waiting for profiling to be done...
```

```
##           2.5 %      97.5 %
## (Intercept) -38.9918822 -15.7757315
## score.1      0.1279764   0.3204597
## score.2      0.1221850   0.3168368
```

O ajuste do modelo pode ser avaliado por recurso a teste de Hosmer-Lemeshow.

```
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 3.1.3
```

```
## ResourceSelection 0.2-5    2015-11-06
```

```
HL <- hoslem.test(AdmiLR$y,fitted(AdmiLR), g = 10)
HL
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  AdmiLR$y, fitted(AdmiLR)
## X-squared = 4.0324, df = 8, p-value = 0.8542
```

Mas também se pode determinar os valores de R^2 (Cox-Snell e Naguelkerke). As fórmulas respectivas são

$$R^2 = 1 - \left(\frac{L_{null}}{L_k} \right)^{2/n}$$

$$R^2 = \frac{1 - \left(\frac{L_{null}}{L_k} \right)^{2/n}}{1 - (L_{null})^{2/n}}$$

```
AdmiLR_null <- glm(label ~ 1, data=MyData, family=binomial)
LL_null <- logLik(AdmiLR_null)
LL_k <- logLik(AdmiLR)
R_Cox <- 1 - (exp(LL_null[1])/exp(LL_k[1]))^(2/length(MyData$label))
R_Nag <- R_Cox/(1-(exp(LL_null[1]))^(2/length(MyData$label)))

print(sprintf('R2 Cox = %s',R_Cox))
```

```
## [1] "R2 Cox = 0.608992262746728"
```

```
print(sprintf('R2 Nagelkerke = %s',R_Nag))
```

```
## [1] "R2 Nagelkerke = 0.823266486222187"
```

Outra forma usual e avaliar a regressão logística é apresentando a matrix confusão.

```
library(knitr)
prob = predict(AdmiLR,type = c('response'),MyData)
confusion<-table(prob>0.5,MyData$label)
kable(confusion)
```

	0	1
FALSE	34	5
TRUE	6	55

Exercício

1. Usando a base de dados *survey* da livreria *Mass* faça uma regressão logística entre o género (*Sex*) e o número de batimentos por minuto (*pulse*).
2. Avalie o modelo, indicando os testes que achar apropriados.