Regressão logística

Francisco Caramelo

Wednesday, December 9, 2015

A regressão logística é um método de ajustamento de dados a uma curva, $y = \frac{1}{1-e^{g(z)}}$, quando a variável independente é qualitativa dicotómica. O método é bastante usado em problemas de classificação pois a equação obtida permite rapidamente cálcular da probabilidade de um dado rótulo tendo em consideração uma ou mais variáveis independentes.

Para exemplificar a forma como se pode ajustar um modelo logístico em R, vamos usar os dados constantes na base de dados *survey*.

```
library(MASS)
head(survey,6)
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72
##        M.I    Age
## 1   Metric 18.250
## 2 Imperial 17.583
## 3     <NA> 16.917
## 4   Metric 20.333
## 5   Metric 23.667
## 6 Imperial 21.000
```

O modelo vai ser ajustado tendo como variável dependente o género e como variáveis independentes as medidas realizadas sobre a mão dominante e a idade.

```
CompleteSurvey = survey[complete.cases(survey),]
SexLR = glm(Sex ~ Wr.Hnd + Age,  data=CompleteSurvey, family=binomial)
```

O comando cria um modelo linear generalizado na família binomial, cujo sumário é:

```
summary(SexLR)
```

```
##
## Call:
## glm(formula = Sex ~ Wr.Hnd + Age, family = binomial, data = CompleteSurvey)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90132  -0.70471   0.02541   0.62736   2.51526
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -21.790148   3.548387  -6.141 8.21e-10 ***
## Wr.Hnd        1.157422   0.185397   6.243 4.29e-10 ***
## Age           0.008816   0.039772   0.222    0.825
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 232.90  on 167  degrees of freedom
## Residual deviance: 149.74  on 165  degrees of freedom
## AIC: 155.74
##
## Number of Fisher Scoring iterations: 5
```

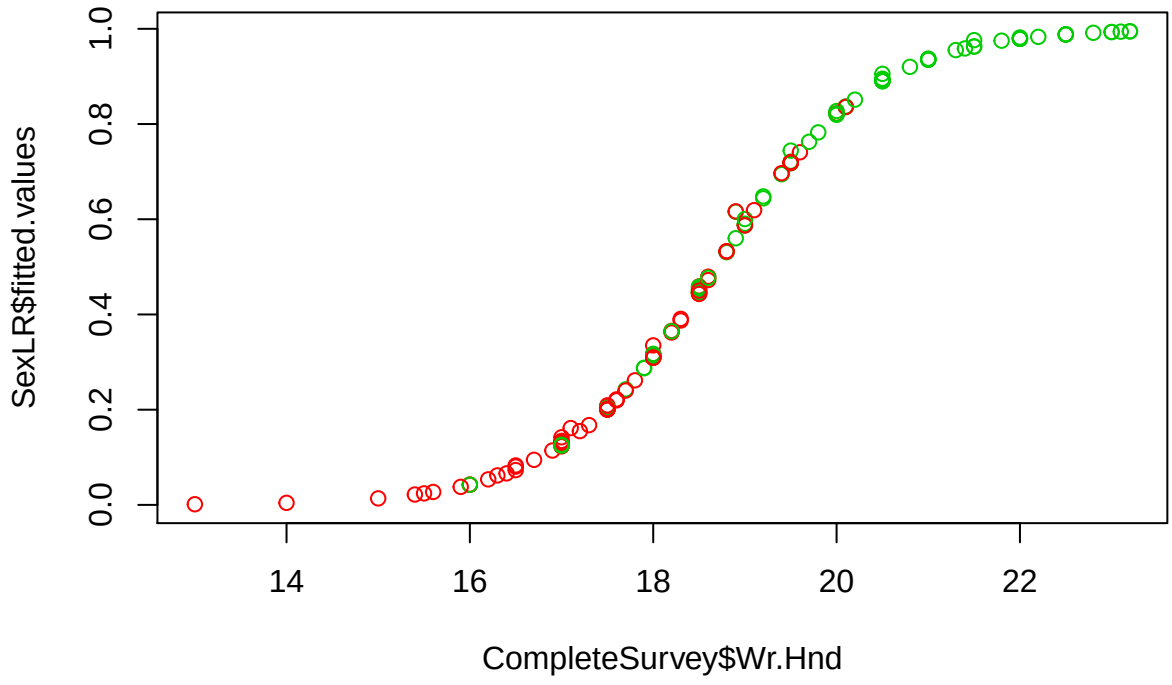
O modelo obtido pode então ser traduzido pela equação:

$$P = \frac{1}{1 + e^{-(-21.79 + 1.157 \times Wr.Hnd + 0.009 \times Age)}}$$

em que P é a probabilidade de ser do sexo masculino.

Todavia, como se pode observar a variável independente 'idade' não é estatisticamente significativa, pelo que pode ser retirada do modelo. O gráfico seguinte mostra a distribuição de valores observados no modelo ajustado.

```
Color = CompleteSurvey$Sex == 'Male'
plot(CompleteSurvey$Wr.Hnd,SexLR$fitted.values, col = Color+2)
```



Exercício

1. Usando a base de dados *mtcars* da livraria *datasets* faça uma regressão logística entre o tipo de transmissão (am) e a potência (hp) e o peso (wt).
2. Diga, justificando o que pode concluir do modelo logístico.