Aprendizaje multivariable

Antonio Falcó

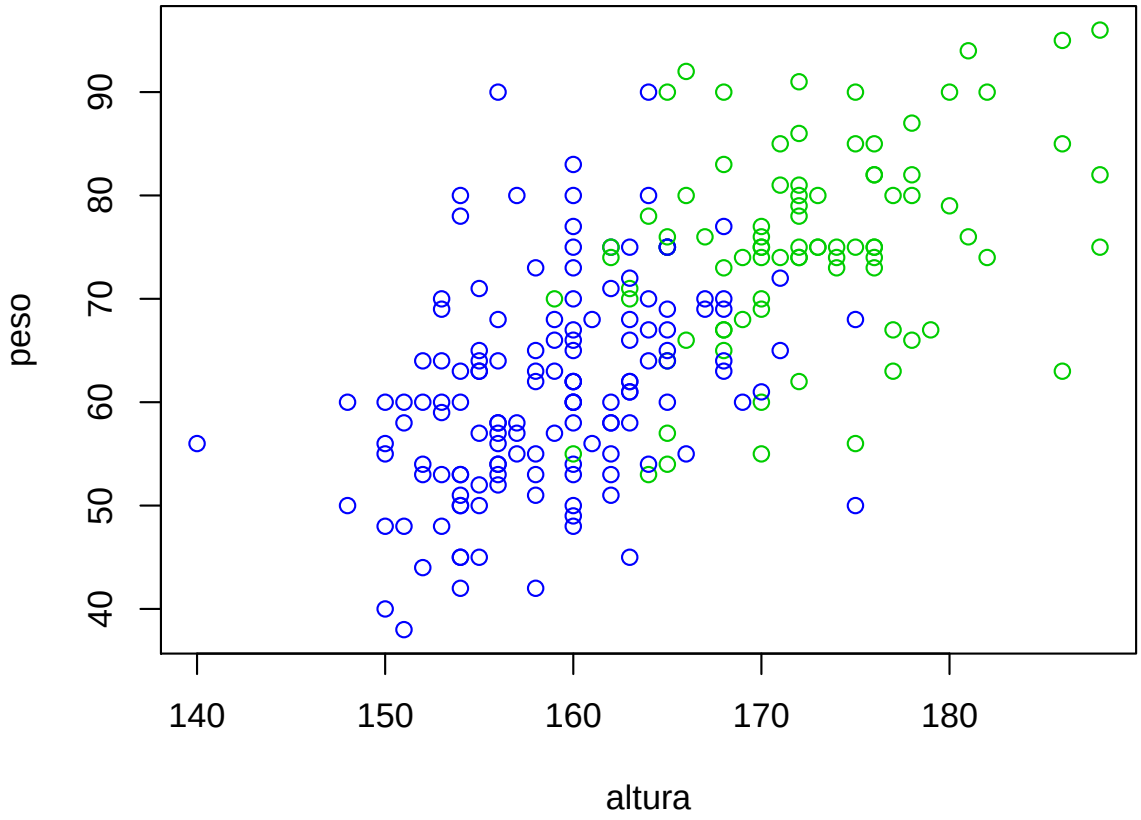
Seminario 8

sexo	altura	peso
1	151	58
1	162	60
1	162	75
1	154	45
1	154	50
1	159	66
1	160	66
1	163	66
1	154	60
1	160	77
1	175	68
1	165	75
1	158	53
1	155	63
1	154	80

sexo	altura	peso
0	166	80
1	159	57
1	157	55
0	165	57
1	156	90
0	175	90
1	161	68
0	168	83
0	168	90
1	156	56
0	170	70
1	162	58
0	172	80
1	165	75
0	171	81

sexo	altura	peso
1	160	54
0	176	82
0	164	53
1	156	58
1	153	70
0	159	70
0	170	60
0	178	87
0	172	86
0	186	95
1	168	77
0	181	94
0	172	62
0	160	55
1	155	50

La muestra es de 226 individuos.



0 = verde y 1 = azul

Cuestión científica

¿Podemos predecir el sexo de la persona en función de su altura y peso?

Modelo

Construir una función

$$\mathbb{E}(\texttt{sexo}|(\texttt{peso} = x, \texttt{altura} = y)) = \begin{cases} 0 & \text{si} \quad \texttt{sexo} = \text{H}, \\ 1 & \text{si} \quad \texttt{sexo} = \text{M}. \end{cases}$$

que es una variable aleatoria binomial, entonces se tiene que cumplir

$$\begin{aligned} \mathbb{E}(\mathbb{E}(\texttt{sexo}|(\texttt{peso}, \texttt{altura}))) &= \mathbb{P}(\mathbb{E}(\texttt{sexo}|(\texttt{peso}, \texttt{altura})) = 1) \\ &= \mathbb{P}(\text{M}). \end{aligned}$$

Idea intuitiva

Construir una función (llamada función discriminante):

$$f(\texttt{peso}, \texttt{altura}) = \widehat{\beta}_1 \frac{\texttt{peso} - \texttt{mean(peso)}}{\texttt{sd(peso)}} + \widehat{\beta}_2 \frac{\texttt{altura} - \texttt{mean(altura)}}{\texttt{sd(altura)}}$$

empleamos las variables tipificadas para evitar problemas con la suma de unidades de diferente naturaleza, de forma que

$$\begin{cases} \widehat{\beta}_1 \frac{\texttt{peso}-\texttt{mean(peso)}}{\texttt{sd(peso)}} + \widehat{\beta}_2 \frac{\texttt{altura}-\texttt{mean(altura)}}{\texttt{sd(altura)}} > 0 & \text{si } \mathbb{P}(\text{H}) > \mathbb{P}(\text{M}) \\ \widehat{\beta}_1 \frac{\texttt{peso}-\texttt{mean(peso)}}{\texttt{sd(peso)}} + \widehat{\beta}_2 \frac{\texttt{altura}-\texttt{mean(altura)}}{\texttt{sd(altura)}} < 0 & \text{si } \mathbb{P}(\text{H}) < \mathbb{P}(\text{M}) \end{cases}$$

Para ello calcularemos los valores β_1 y β_2 que minimizan el error cuadrático medio:

$$\mathbb{E}\left(\left(\texttt{sexo} - \beta_1 \frac{\texttt{peso} - \texttt{mean(peso)}}{\texttt{sd(peso)}} - \widehat{\beta}_2 \frac{\texttt{altura} - \texttt{mean(altura)}}{\texttt{sd(altura)}}\right)^2\right)$$

Ecuación discriminante

$$\begin{cases} \widehat{\beta}_1 \frac{\texttt{peso}-\texttt{mean(peso)}}{\texttt{sd(peso)}} + \widehat{\beta}_2 \frac{\texttt{altura}-\texttt{mean(altura)}}{\texttt{sd(altura)}} > 0 & \text{entonces} \quad \text{H} \\ \widehat{\beta}_1 \frac{\texttt{peso}-\texttt{mean(peso)}}{\texttt{sd(peso)}} + \widehat{\beta}_2 \frac{\texttt{altura}-\texttt{mean(altura)}}{\texttt{sd(altura)}} < 0 & \text{entonces} \quad \text{M} \end{cases}$$

Consideramos la igualdad

$$\widehat{\beta}_1 \frac{\texttt{peso} - \texttt{mean(peso)}}{\texttt{sd(peso)}} + \widehat{\beta}_2 \frac{\texttt{altura} - \texttt{mean(altura)}}{\texttt{sd(altura)}} = 0$$

con la ecuación (**ecuación discriminante**) equivalente

$$\texttt{peso} = -\frac{\widehat{\beta}_1}{\widehat{\beta}_2} \frac{\texttt{sd(peso)}}{\texttt{sd(altura)}} \texttt{altura} + \frac{\widehat{\beta}_1}{\widehat{\beta}_2} \frac{\texttt{sd(peso)}}{\texttt{sd(altura)}} \texttt{mean(altura)} + \texttt{mean(peso)}.$$

Ecuación discriminante

Entonces dados $(\texttt{altura}, \texttt{peso})$ de una persona si se cumple

$$\texttt{peso} > -\frac{\widehat{\beta}_1}{\widehat{\beta}_2}\frac{\texttt{sd(peso)}}{\texttt{sd(altura)}}\texttt{altura} + \frac{\widehat{\beta}_1}{\widehat{\beta}_2}\frac{\texttt{sd(peso)}}{\texttt{sd(altura)}}\texttt{mean(altura)} + \texttt{mean(peso)}$$

entonces el sexo de la persona será H, en caso contrario

$$\texttt{peso} < -\frac{\widehat{\beta}_1}{\widehat{\beta}_2}\frac{\texttt{sd(peso)}}{\texttt{sd(altura)}}\texttt{altura} + \frac{\widehat{\beta}_1}{\widehat{\beta}_2}\frac{\texttt{sd(peso)}}{\texttt{sd(altura)}}\texttt{mean(altura)} + \texttt{mean(peso)}$$

el sexo de la persona será M. Si se da la igualdad no podemos decidir.

```r
X <- (altura-mean(altura))/sd(altura)
Y <- (peso-mean(peso))/sd(peso)
library('MASS')
r <- lda(sexo~(X+Y),prior=c(1/2,1/2))
r$scaling
```

```
##          LD1
## X -1.2569970
## Y -0.3685788
```

```r
d0 <- r$scaling[1]
d1 <- r$scaling[2]
```

Resultado

Obtenemos los valores

$$\widehat{\beta}_1 = -1.256997 \text{ y } \widehat{\beta}_2 = -0.3685788$$

de donde

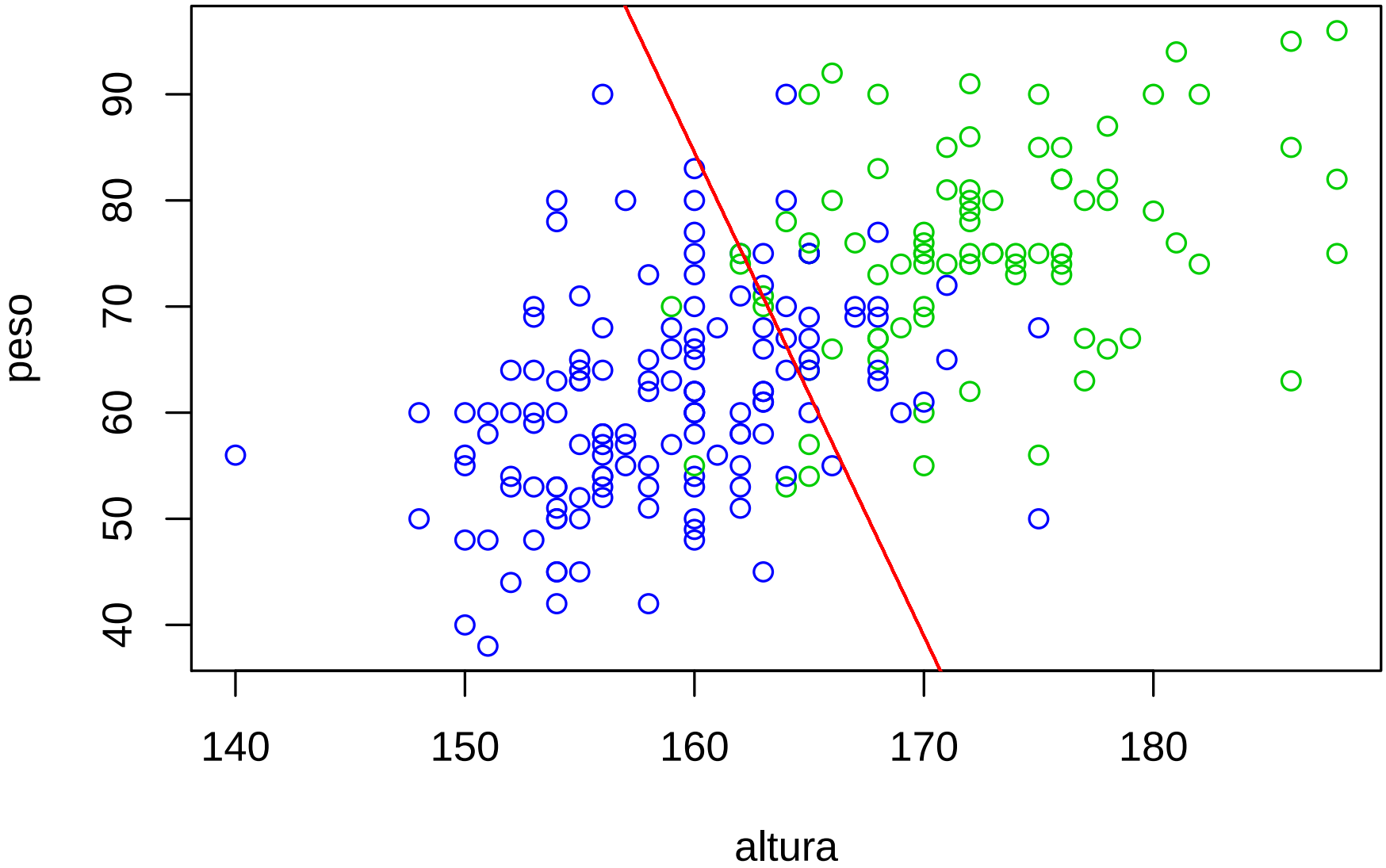
$$-\frac{\widehat{\beta}_1}{\widehat{\beta}_2}\,\frac{\texttt{sd(peso)}}{\texttt{sd(altura)}} = -4.5581236$$

y

$$\frac{\widehat{\beta}_1}{\widehat{\beta}_2}\,\frac{\texttt{sd(peso)}}{\texttt{sd(altura)}}\,\texttt{mean(altura)} + \texttt{mean(peso)} = 813.8330594$$

Entonces la ecuación discriminante se escribe como

$$\texttt{peso} = -4.5581236\,\texttt{altura} + 813.8330594.$$


peso
40
50
60
70
80
90
140
150
160
170
180
altura

Vamos a comprobar la eficacia del modelo empleando datos nuevos:

```
n_altura <- c(180, 177, 180, 180, 183, 184, 185, 184, 174,
180, 168,180, 183,181, 180, 190, 183, 167, 181, 179, 173,
170, 170, 183,179, 180, 188, 176,178, 185, 168, 157, 167,
168, 163, 167,166,164, 172, 165, 158, 161, 160,162, 165)
n_peso <- c(70, 57, 60, 66, 62, 68, 65, 72, 65, 72, 52, 75,
75, 68, 65,66, 78, 60, 67, 98, 75, 68, 59, 72, 73, 72, 70,
65, 72, 71, 52, 47,53,57, 65, 60, 68, 49, 57, 59, 62, 65,
61, 58, 58)
n_sexo <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0,0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
1,1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1, 1)
```

El modelo predice en los sexos en función de la altura y el peso (tipificados):

```
##  [1] 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 1 0 0 1 1 1 1 1 1 1 1
## [39] 1 1 1 1 1 1 1
## Levels: 0 1
```

Pasamos a comparar el sexo que predice el modelo y el verdadero sexo de la persona

```
##   sexe.true sexe.predicted difference
## 1         0              0          0
## 2         0              1         -1
## 3         0              0          0
## 4         0              0          0
## 5         0              0          0
## 6         0              0          0
```

Calculamos los aciertos y errores del modelo

```
verdaderos.H <- length(comparative$sexe.true[comparative$sexe.true == 0])
verdaderos.M <- length(comparative$sexe.true[comparative$sexe.true == 1])
falsos.H <- sum(abs(comparative$difference[comparative$sexe.true == 0]))
falsos.M <- sum(abs(comparative$difference[comparative$sexe.true == 1]))
verdaderos.H
```

```
## [1] 30
```

```
verdaderos.M
```

```
## [1] 15
```

```
falsos.H
```

```
## [1] 8
```

```
falsos.M
```

```
## [1] 0
```

```
porcentaje.correcto.H <- 100*
  ((verdaderos.H - falsos.H)/verdaderos.H)
porcentaje.correcto.M <- 100*
  ((verdaderos.M -falsos.M)/verdaderos.M)
porcentaje.falso.H <- 100*
  (falsos.H/verdaderos.H)
porcentaje.falso.M <- 100*
  (falsos.M /verdaderos.M)
porcentaje.correcto.H
```

```
## [1] 73.33333
```

```
porcentaje.correcto.M
```

```
## [1] 100
```

```
porcentaje.falso.H
```

```
## [1] 26.66667
```

```
porcentaje.falso.M
```

```
## [1] 0
```

Matriz de confusión del modelo predictivo

	Hombres	Mujeres
Hombres	73.3333333	26.6666667
Mujeres	0	100