ESTADÍSTICA II
Tema 4: Clasificación y regresión logística

- Clasificación supervisada: Planteamiento del problema.
- La regla Bayes
- Regla lineal de Fisher
- Estimación del error de clasificación
- Regresión logística

El problema de clasificación supervisada

Este problema tiene diferentes nombres en la literatura en inglés: *supervised classification*, *statistical learning*, *discrimination*, *machine learning*, *pattern recognition*, etc.

El objetivo del *análisis discriminante* es encontrar las *características* (*features*) diferenciales de un vector aleatorio $\mathbf{X} \in \mathbb{R}^p$ que se observa en varias poblaciones o clases conocidas. Tratamos de encontrar *discriminantes*, es decir, funciones de las componentes de $\mathbf{X}$ cuyos valores numéricos están lo más separados posible en poblaciones diferentes.

El objetivo de la *clasificación supervisada* es asignar "óptimamente" una nueva observación $\mathbf{x}$ a una de las poblaciones mencionadas. El término "supervisada" hace referencia a la existencia de información dada por la *muestra de entrenamiento* (*training* o *learning sample*), cuyas observaciones están correctamente clasificadas por algún "experto".

La información disponible es la de la *muestra de entrenamiento* $\{(\mathbf{X}_i, Y_i), 1 \leq i \leq n\}$, donde $\mathbf{X}_i$, $i = 1, \ldots, n$, son realizaciones independientes de $\mathbf{X}$ medidas en n individuos elegidos al azar, e Y_i son los correspondientes valores de la clase ($Y_i = j$ siempre que el i-ésimo individuo pertenezca a la población P_j).

La distribución de probabilidad de $\mathbf{X}$ se supone diferente en poblaciones diferentes. Denotamos la distribución condicional $\mathbf{X}|Y = j$ por P_j, $\boldsymbol{\mu}_j = \mathbb{E}(\mathbf{X}|Y = j)$ y $\boldsymbol{\Sigma}_j = \mathbb{V}(\mathbf{X}|Y = j)$.

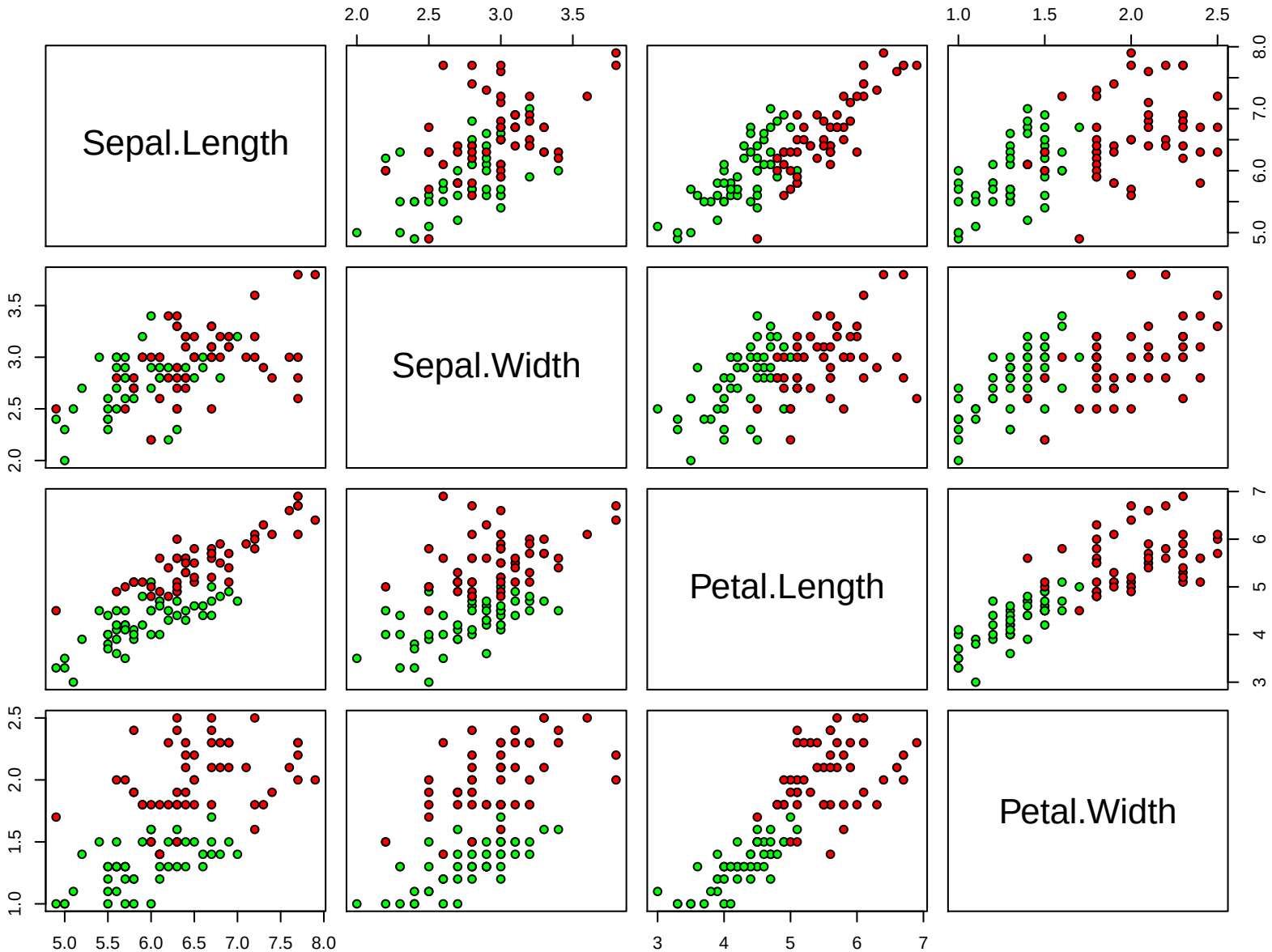
El objetivo último es clasificar una nueva observación $\mathbf{X}$ en alguna de las poblaciones P_j. Queremos predecir el correspondiente valor de Y utilizando la información de la muestra de entrenamiento.

Para simplificar el problema supondremos que sólo hay dos poblaciones, P_0 y P_1. Cuando hay más poblaciones, el problema se denomina *clasificación multiclase*.

Ejemplo (lirios): Se dispone de las medidas (en cm) del pétalo y del sépalo de 50 lirios de la especie *versicolor* y 50 de la especie *virginica*. Se trata de una parte de un conjunto de datos recogido por Anderson (1935) y analizado por Fisher (1936).

```
head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
Especie = iris$Species
levels(Especie) # Para ver las especies de lirios
[1] "setosa"     "versicolor" "virginica"
Setosa = (Especie == "setosa")
X = iris[!Setosa,(1:4)] # Datos para versicolor y virginica
# Diagrama de dispersion multiple
# verde = versicolor; rojo = virginica
colores <- c(rep("green",50),rep("red",50))
pairs(X, pch=21, bg=colores, oma=c(1.5,1.5,1.5,1.5))
```

Ejemplo (lirios): Diagrama de dispersión múltiple (en verde la especie `versicolor` y en rojo la especie `virginica`).



Es normal que una regla de clasificación cometa errores. Un buen procedimiento de clasificación debería realizar pocas clasificaciones erróneas: su probabilidad de asignación fallida debería ser pequeña.

El problema matemático es encontrar un *clasificador* $g : \mathbb{R}^p \to \{0, 1\}$, que represente nuestra predicción de Y dado $\mathbf{X}$. El clasificador falla en $\mathbf{x}$ si $g(\mathbf{x}) \neq y$. El *error de clasificación* o *riesgo* de un clasificador g es

$$L(g) = \mathbb{P}\{g(\mathbf{X}) \neq Y\}.$$

El clasificador óptimo (en general, desconocido) se denomina *clasificador Bayes* y es

$$g^* = \underset{g:\mathbb{R}^p \to \{0,1\}}{\arg\min}\ \mathbb{P}\{g(\mathbf{X}) \neq Y\}.$$

La menor probabilidad de error es el *error* o riesgo Bayes, $L^* = L(g^*)$.

Como la distribución de probabilidad de $(\mathbf{X}, Y)$ es desconocida, se construye un clasificador g_n utilizando la información de la muestra de entrenamiento y su comportamiento se evalúa mediante la probabilidad de error condicional

$$L_n = L(g_n) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y | \mathbf{X}_1, Y_1, \ldots, \mathbf{X}_n, Y_n\} \qquad (1)$$

De nuevo, el valor exacto de L_n es desconocido, pero se puede estimar mediante el *riesgo empírico*, la proporción de errores de clasificación en la muestra de entrenamiento

$$\hat{L}_n = \hat{L}_n(g_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{g_n(\mathbf{X}_i) \neq Y_i\}}.$$

Notación: n_j = número de observaciones de la población P_j.

$$n = n_0 + n_1$$

La regla Bayes

Sea $(\mathbf{X}, Y)$ un par de variables aleatorias con valores en $\mathbb{R}^p$ y $\{0, 1\}$ respectivamente. Describimos la distribución de probabilidad de $(\mathbf{X}, Y)$ mediante la distribución de probabilidad de $\mathbf{X}$

$$F(A) = \mathbb{P}\{\mathbf{X} \in A\}, \quad A \in \mathcal{B}_{\mathbb{R}^p}$$

y la *probabilidad a posteriori* de Y

$$\eta(\mathbf{x}) := \mathbb{P}\{Y = 1 | \mathbf{X} = \mathbf{x}\} = \mathbb{E}(Y | \mathbf{X} = \mathbf{x}).$$

Teorema: *La función de decisión*

$$g^*(\mathbf{x}) = \begin{cases} 1 & \text{si } \eta(\mathbf{x}) > 1/2 \\ 0 & \text{si no} \end{cases}$$

es el clasificador Bayes, es decir, para cualquier clasificador $g : \mathbb{R}^p \to \{0, 1\}$, *se cumple que*

$$L^* := \mathbb{P}\{g^*(\mathbf{X}) \neq Y\} \leq \mathbb{P}\{g(\mathbf{X}) \neq Y\}.$$

La regla Bayes equivale a clasificar $\mathbf{x}$ en P_1 si

$$\mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} > \mathbb{P}\{Y = 0|\mathbf{X} = \mathbf{x}\}$$

Si ...

- $\mathbf{X}$ tiene densidad f_0 en P_0 y densidad f_1 en P_1;
- las *probabilidades a priori* de las poblaciones son

$$\mathbb{P}(P_0) = \pi_0, \quad \mathbb{P}(P_1) = \pi_1 \quad (\pi_0 + \pi_1 = 1).$$

...entonces se tiene (fórmula de Bayes):

$$\mathbb{P}\{Y = 1|\mathbf{x}\} > \mathbb{P}\{Y = 0|\mathbf{x}\} \Leftrightarrow \pi_1 f_1(\mathbf{x}) > \pi_0 f_0(\mathbf{x}).$$

Regla Bayes bajo normalidad

Supongamos que f_0 y f_1 son normales: para $\mathbf{x} \in \mathbb{R}^p$,

$$f_i(\mathbf{x}) = \frac{1}{|\mathbf{\Sigma}_i|^{1/2}(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}, \quad i = 0, 1.$$

Entonces $\mathbf{x}$ se clasifica en P_0 si

$$d^2_{M_0}(\mathbf{x}, \boldsymbol{\mu}_0) < d^2_{M_1}(\mathbf{x}, \boldsymbol{\mu}_1) + 2\log\left(\frac{\pi_0|\mathbf{\Sigma}_1|^{1/2}}{\pi_1|\mathbf{\Sigma}_0|^{1/2}}\right)$$

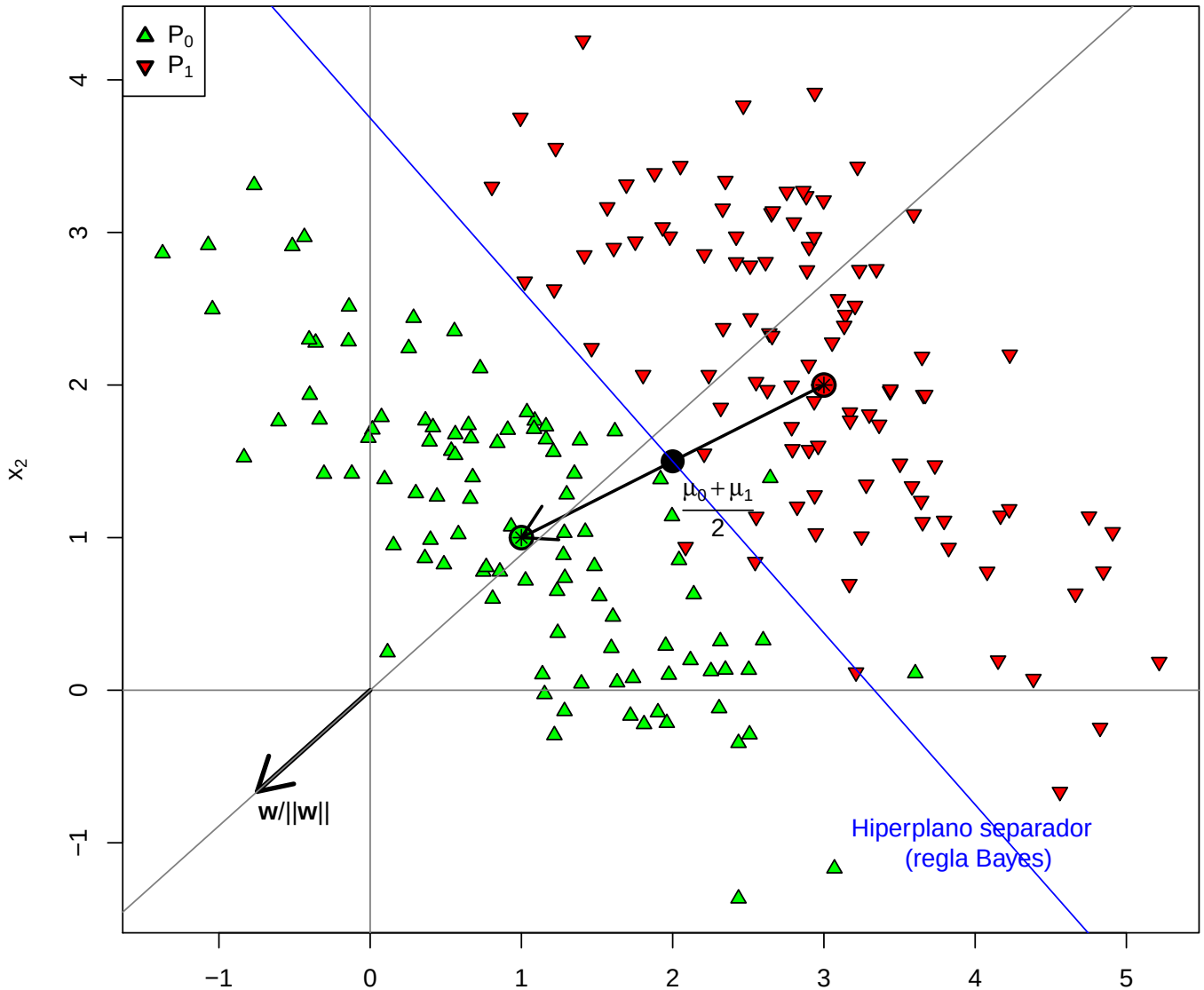
donde $d^2_{M_i}(\mathbf{x}, \boldsymbol{\mu}_i) = (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ es el cuadrado de la distancia de Mahalanobis entre $\mathbf{x}$ y $\boldsymbol{\mu}_i$ $(i = 0, 1)$.

Regla Bayes bajo normalidad y homocedasticidad ($\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$):

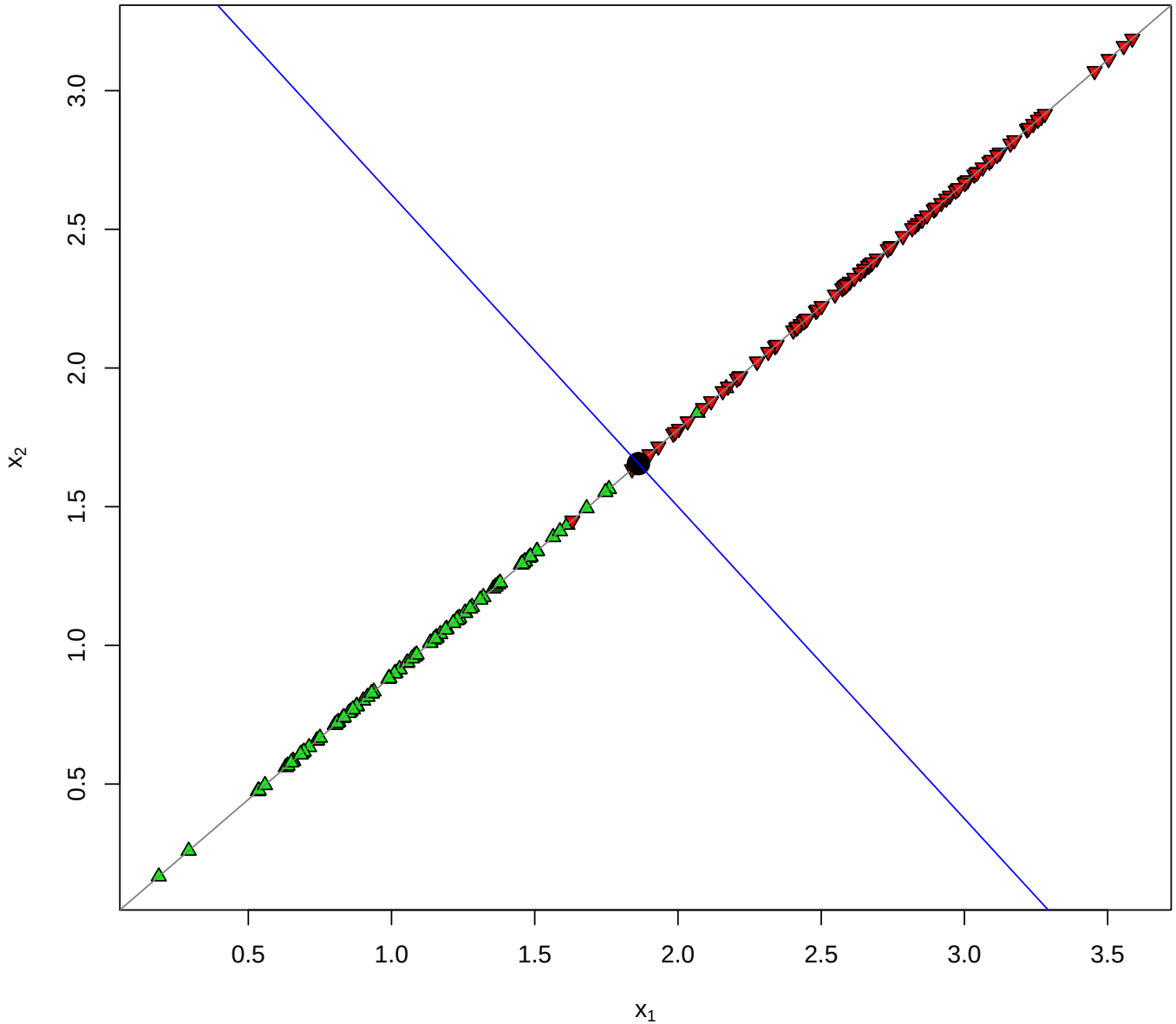
$\mathbf{x}$ se clasifica en P_0 si $\mathbf{w}'\mathbf{x} > \mathbf{w}'\left(\frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}\right) + \log\left(\frac{\pi_1}{\pi_0}\right)$, donde $\mathbf{w} := \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ y $\mathbf{\Sigma} = \mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$.

Ejemplo (normales homocedásticas): $\pi_0 = \pi_1 = 1/2$

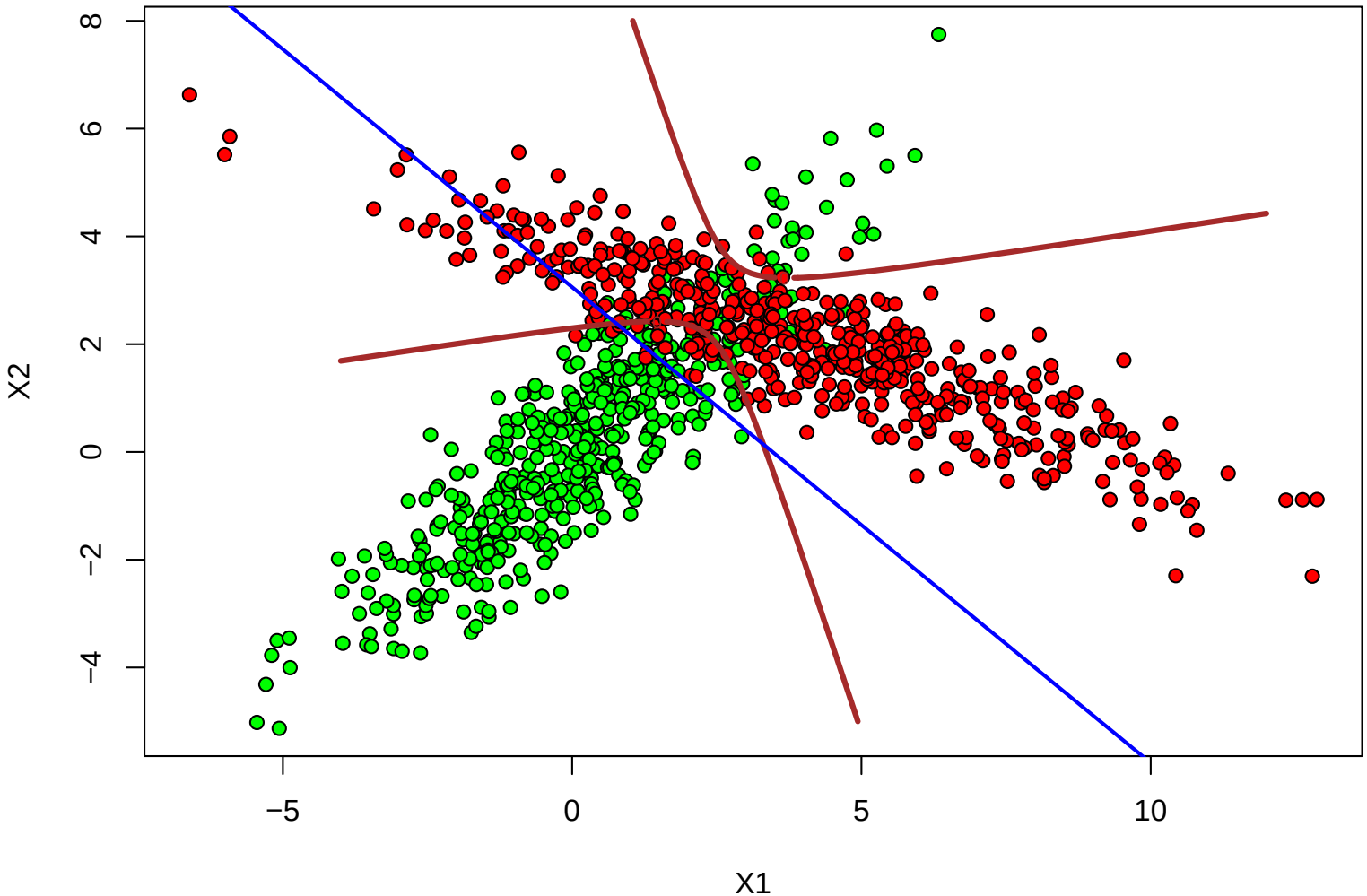
$$\boldsymbol{\mu}_0 = (1,1)', \quad \boldsymbol{\mu}_2 = (3,2)', \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.7 \\ 0.7 & 1 \end{pmatrix}$$



Ejemplo (normales homocedásticas): Proyección sobre la dirección **w**

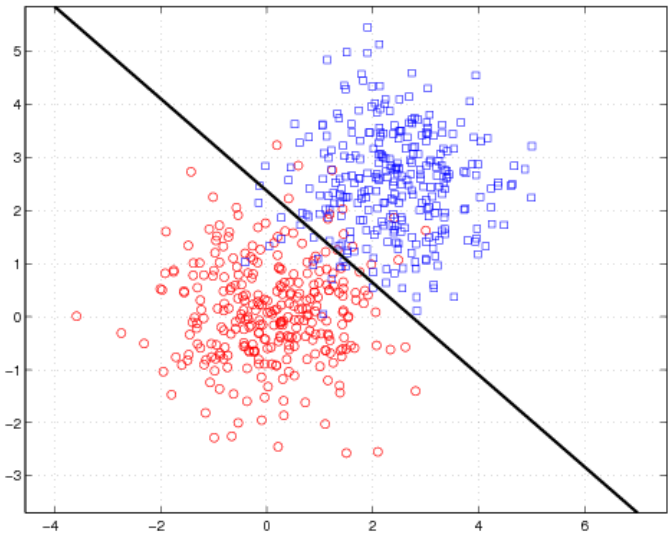


Ejemplo (normales heterocedásticas):



Regla lineal de Fisher

Una *regla de clasificación lineal* divide el espacio muestral ($\subseteq \mathbb{R}^p$) mediante un hiperplano afín y asigna una clase diferente (P_0 o P_1) a cada semiespacio.



Rosenblatt (1962) denominó *perceptrón* a este clasificador lineal.

Un hiperplano afín en $\mathbb{R}^p$ está determinado por una única ecuación lineal

$$a_1x_1 + a_2x_2 + \ldots + a_px_p = a_0 \Leftrightarrow \mathbf{a}'\mathbf{x} = a_0,$$

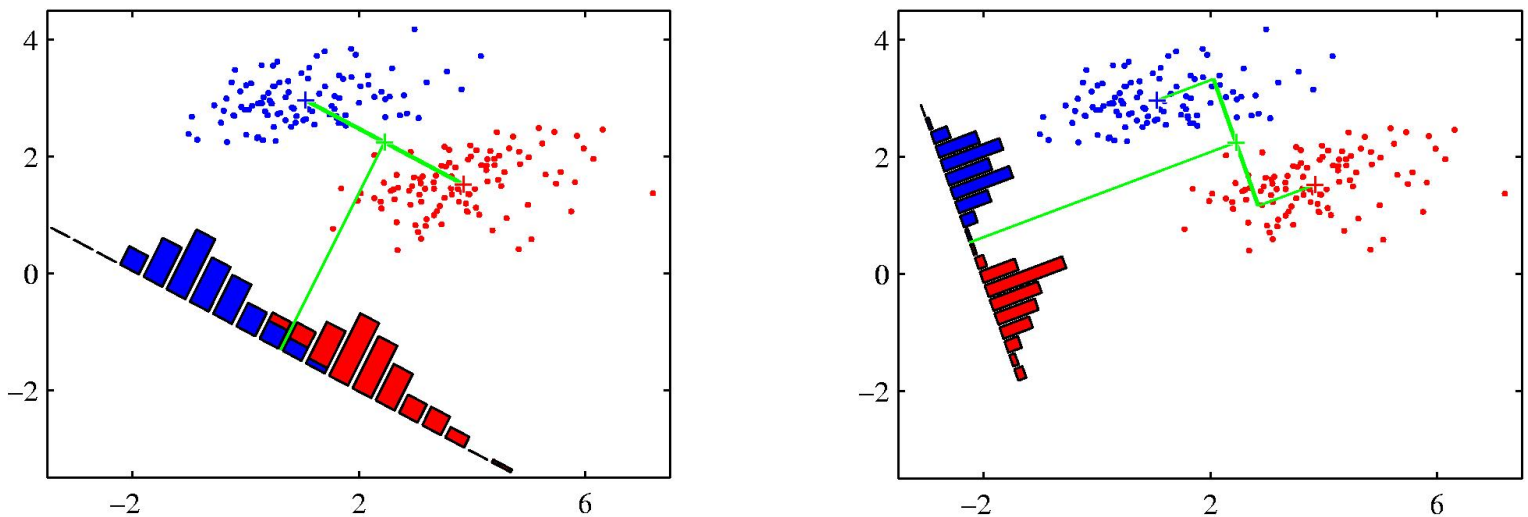
donde $\mathbf{a} = (a_1, \ldots, a_p) \neq \mathbf{0}$ y $\mathbf{x} = (x_1, \ldots, x_p)'$.

Las reglas lineales son sencillas de implementar e interpretar, pero sus errores de clasificación suelen ser muy superiores al riesgo Bayes. Sin embargo, los discriminantes lineales son la base de muchos procedimientos de clasificación satisfactorios (redes neuronales, máquinas de vector soporte, árboles,...).

La regla discriminante lineal con pesos $a_0, a_1, \ldots, a_p$ está dada por

$$g(\mathbf{x}) = \begin{cases} 1 & \text{si } \sum_{j=1}^{p} a_j x_j > a_0 \\ 0 & \text{si no} \end{cases} = \begin{cases} 1 & \text{si } \mathbf{a}'\mathbf{x} > a_0 \\ 0 & \text{si no.} \end{cases}$$

Se pueden usar diversos criterios para determinar direcciones de proyección $\mathbf{a}$ basados en la información muestral. La idea es seleccionar $\mathbf{a}$ de tal manera que, para la combinación lineal $Z = \mathbf{a}'\mathbf{X}$ de $\mathbf{X}$, los valores observados de Z en la población P_0 estén lo más separados posible de aquellos en la población P_1.



Una buena dirección debe separar bien los centros de los grupos, pero también la varianza de las proyecciones dentro de los grupos debe ser lo menor posible.

Clasificador lineal de Fisher (1936)

Su expresión se puede obtener a partir de la regla Bayes para poblaciones normales homocedásticas, con $\pi_0 = \pi_1$:

$\mathbf{x}$ se clasifica en P_0 si $\mathbf{w}' \left(\mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right) > 0$, con $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$.

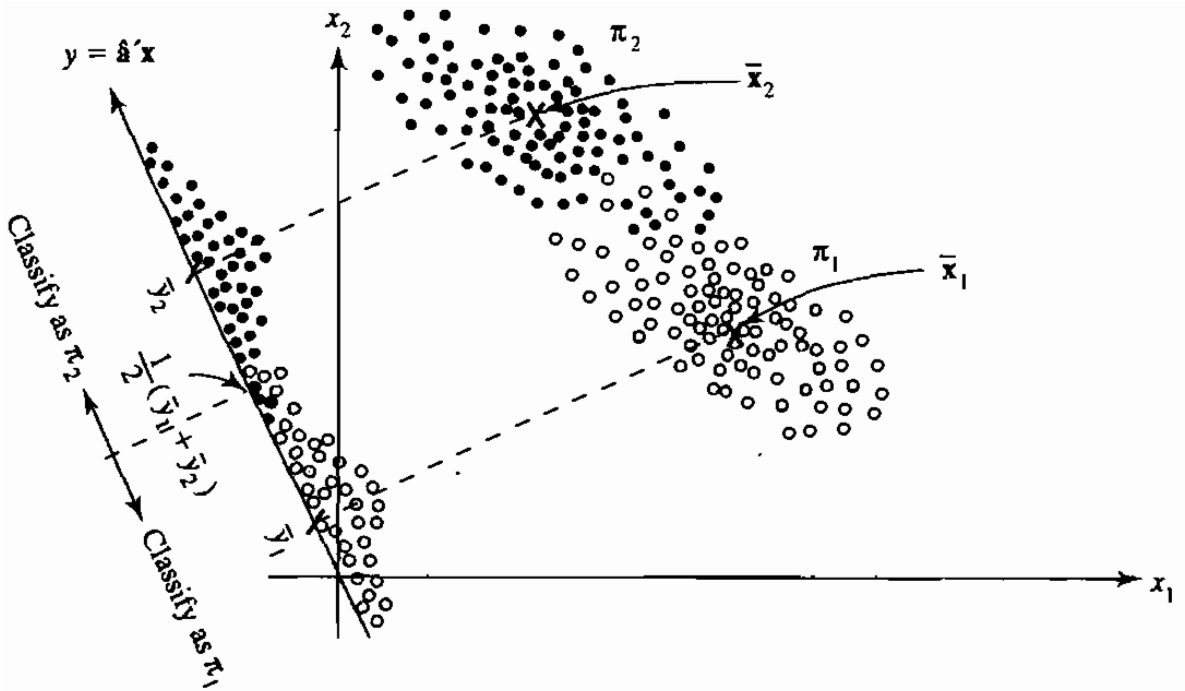
Estimamos $\boldsymbol{\mu}_j$ mediante la media muestral de $\mathbf{X}$ en P_j, $j = 0, 1$,

$$\bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}_{ji}.$$

Estimamos $\boldsymbol{\Sigma}$ mediante la matriz de covarianzas combinadas de $\mathbf{X}$

$$\mathbf{S}_{\mathbf{X}} = \frac{(n_0 - 1)\mathbf{S}_0 + (n_1 - 1)\mathbf{S}_1}{n_0 + n_1 - 2},$$

siendo $\mathbf{S}_j = \sum_{i=1}^{n_j} \frac{(\mathbf{x}_{ji} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ji} - \bar{\mathbf{x}}_j)'}{n_j - 1}$ la matriz de covarianzas de $\mathbf{X}$ en P_j.



Entonces la regla discriminante de Fisher tiene la expresión:

$$g_F(\mathbf{x}) = \begin{cases} 0 & \text{si } \mathbf{w}_F'\mathbf{x} > \mathbf{w}_F'\left(\frac{\bar{\mathbf{x}}_0+\bar{\mathbf{x}}_1}{2}\right) \\ 1 & \text{si no} \end{cases} \quad \text{con } \mathbf{w}_F = \mathbf{S}_{\mathbf{X}}^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$$

$$= \begin{cases} 0 & \text{si } (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)'\mathbf{S}_{\mathbf{X}}^{-1}\left(\mathbf{x} - \frac{\bar{\mathbf{x}}_0+\bar{\mathbf{x}}_1}{2}\right) > 0 \\ 1 & \text{si no} \end{cases}$$

En realidad Fisher (1936) obtuvo su clasificador lineal razonando de la manera siguiente:

Una buena dirección de proyección debe separar bien los centros de los grupos. La distancia entre las medias $(\mathbf{a}'\bar{\mathbf{x}}_0 - \mathbf{a}'\bar{\mathbf{x}}_1)^2 = \mathbf{a}'\mathbf{B}\mathbf{a}$, donde $\mathbf{B} = (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)'$, debe ser grande.

La varianza de las proyecciones dentro de los grupos $(\mathbf{a}'\mathbf{S}_\mathbf{X}\mathbf{a})$ debe ser lo menor posible.

Problema: Encontrar la dirección $\mathbf{a}$ que maximiza

$$\mathcal{S}^2(\mathbf{a}) = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{S}_\mathbf{X}\mathbf{a}} = \textit{cociente de Rayleigh} \text{ respecto a } \mathbf{a}.$$

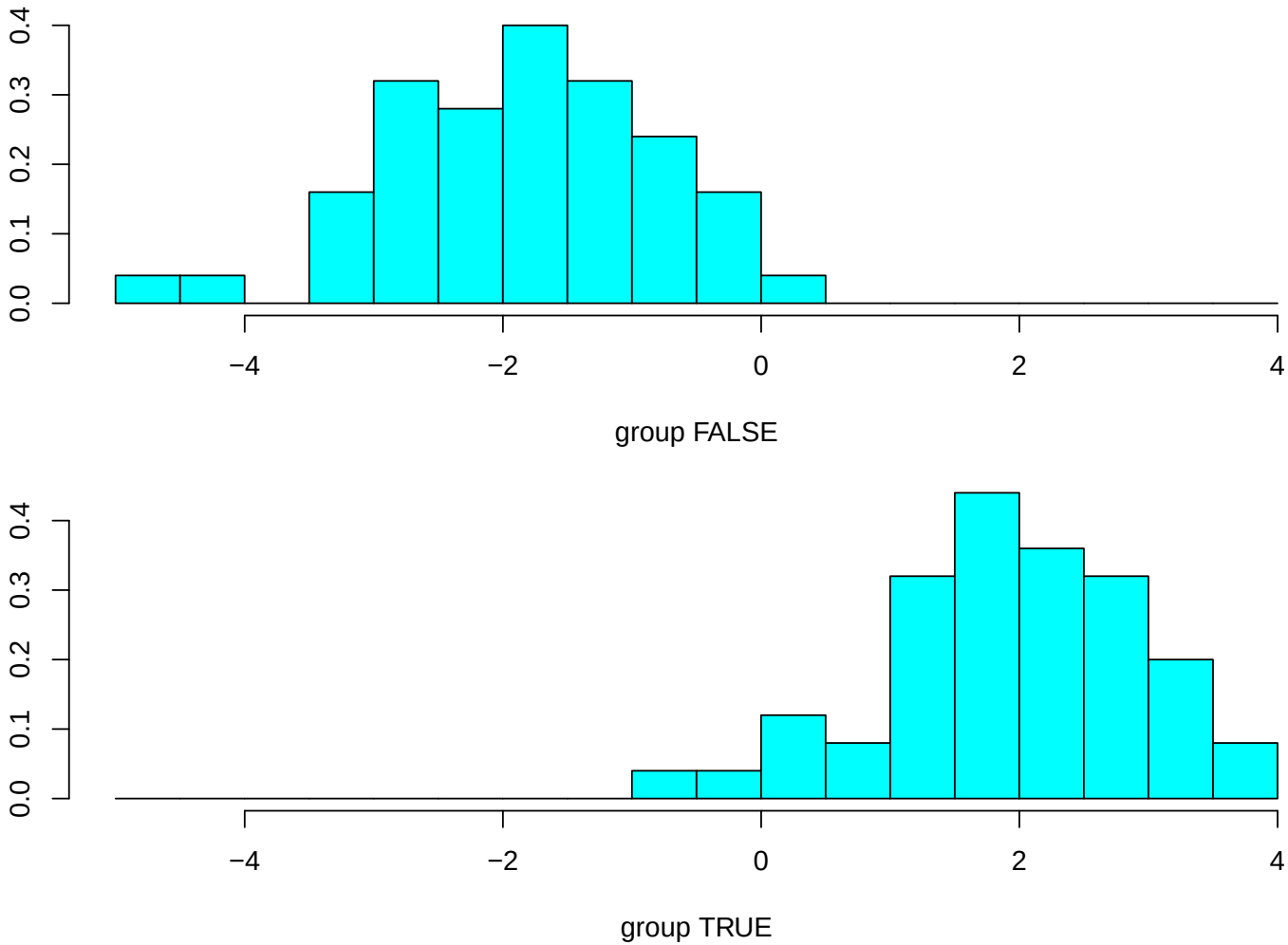
Teorema: *Cualquier vector* $\mathbf{a}$ *proporcional a* $\mathbf{w}_F = \mathbf{S}_\mathbf{X}^{-1}(\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$ *maximiza la separación* $\mathcal{S}$.

Ejemplo (lirios):

```
Especie = iris$Species
Setosa = (Especie == "setosa")
X = iris[!Setosa,(1:4)] # Datos de versicolor y virginica
Y = (Especie[!Setosa] == "versicolor") # T=versicolor F=virginica

library(MASS)
ClasFisher = lda(X,Y)

ClasFisher$scaling # Dirección de proyección
                    LD1
Sepal.Length  0.9431178
Sepal.Width   1.4794287
Petal.Length -1.8484510
Petal.Width  -3.2847304

ClasFisher$means # Medias por clases
      Sepal.Length Sepal.Width Petal.Length Petal.Width
FALSE        6.588       2.974        5.552       2.026
TRUE         5.936       2.770        4.260       1.326
```

Ejemplo (lirios):

```
plot(ClasFisher)
```



Estimación del error de clasificación

Es importante estimar la probabilidad de error de clasificación.

El riesgo empírico suele infraestimar el verdadero error ya que los datos se utilizan tanto para calcular el clasificador como para evaluarlo.

Existen diversos procedimientos para resolver este problema:

- Dividir la muestra en dos partes: *training data* y *test data*. Utilizar la primera parte para construir la regla de clasificación y estimar el error mediante la segunda.

- *Validación cruzada:* Hay muchos procedimientos de este tipo.

  k-fold cross-validation divide la muestra original en k submuestras del mismo tamaño. Se itera k veces el siguiente procedimiento: se toma una submuestra como *test data* y las $k-1$ submuestras restantes se juntan en una muestra de entrenamiento.

  Cuando $k = n$, el procedimiento de validación cruzada se denomina *leave-one-out*. En este caso, omitimos un dato de los n observados y generamos la regla de clasificación con los $n-1$ restantes. Clasificamos la observación apartada y repetimos el procedimiento para cada una de las observaciones.

$$\hat{L} = \frac{\text{Total de mal clasificados en la muestra por VC}}{n} 100\ \%.$$

Ejemplo (lirios):

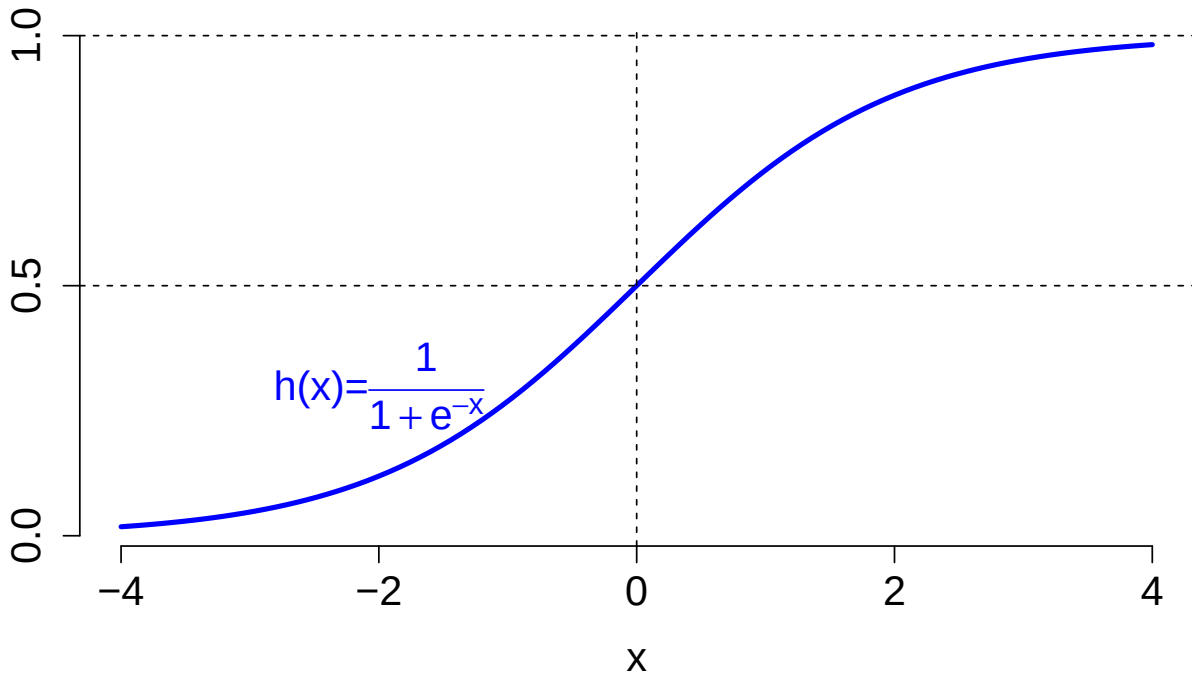
```
ClasFisher = lda(X,Y,CV=TRUE) # Con CV=T, obtenemos leave-one-out
n = nrow(X) # Tamaño muestral
sum(Y != ClasFisher$class)/n # TEVC
[1] 0.03
```

Regresión logística (o modelo logit)

La regresión logística plantea el siguiente modelo paramétrico para la función de regresión

$$\eta(\mathbf{x}) = \mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}} = h(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p),$$

donde $h(x) = 1/(1 + e^{-x}) = e^x/(1 + e^x)$ es la *función logística*.



Propiedades de la función logística

- $h(0) = 1/2$
- $h(-x) = 1 - h(x)$
- $h'(x) = h(x)(1 - h(x))$

La función logística no es la única que se ha utilizado para modelizar este tipo de datos (ver p. 257 de Devroye *et al.* 1997).

La regresión logística es un caso particular de *modelo lineal generalizado* en el que se supone que

$$\eta(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = H(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p),$$

siendo H una función de enlace entre el modelo lineal $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ y la función de regresión η. Por ejemplo, el modelo *probit* supone $\eta(\mathbf{x}) = \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$, donde Φ es la función de distribución N(0,1).

Interpretación de los parámetros del modelo logit

El modelo de regresión logística es equivalente a:

$$\text{logit}(\eta(\mathbf{x})) := \log\left(\frac{\eta(\mathbf{x})}{1-\eta(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Llamamos O a la *disparidad*, *momio* o *posibilidad* (*odds*):

$$O(\mathbf{x}) = \frac{\eta(\mathbf{x})}{1-\eta(\mathbf{x})} = \frac{\mathbb{P}\{Y=1|\mathbf{X}=\mathbf{x}\}}{\mathbb{P}\{Y=0|\mathbf{X}=\mathbf{x}\}}.$$

¿Cómo se interpreta el valor de O? ¿Qué significa, por ejemplo, $O(\mathbf{x}) = 2$?

Si se cumple el modelo de regresión logística, entonces

$$O(\mathbf{x}) = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}.$$

Razón de posibilidades si el regresor x_j aumenta una unidad (y los demás permanecen constantes):

$$\textit{Odds ratio} = \frac{O(\mathbf{x}+\Delta\mathbf{x})}{O(\mathbf{x})} = \frac{e^{\beta_0+\cdots+\beta_j(x_j+1)+\cdots+\beta_p x_p}}{e^{\beta_0+\cdots+\beta_j x_j+\cdots+\beta_p x_p}} = e^{\beta_j}.$$

Estimación de los parámetros de la regresión logística

Para estimar los parámetros se usa el método de máxima verosimilitud. Denotamos $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ y $\tilde{\mathbf{x}} = (1, x_1, \ldots, x_p)'$.

La distribución condicionada de $Y|\mathbf{X} = \mathbf{x}$ es una Bernoulli$(\eta(\mathbf{x}))$. La verosimilitud para la muestra $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ es

$$L(\boldsymbol{\beta}) = L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \eta(\mathbf{x}_i)^{y_i} (1 - \eta(\mathbf{x}_i))^{1-y_i}.$$

Log-verosimilitud:

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [y_i \log \eta(\mathbf{x}_i) + (1 - y_i) \log(1 - \eta(\mathbf{x}_i))] \\ &= \sum_{i=1}^{n} [y_i \boldsymbol{\beta}' \tilde{\mathbf{x}}_i - \log(1 + e^{\boldsymbol{\beta}' \tilde{\mathbf{x}}_i})] \end{aligned}$$

Ecuaciones de verosimilitud:

$$\mathbf{0} = \frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\left(y_i - \frac{1}{1 + e^{-\boldsymbol{\beta}' \tilde{\mathbf{x}}_i}} \right) \tilde{\mathbf{x}}_i \right] = \sum_{i=1}^{n} \left[(y_i - \eta(\mathbf{x}_i)) \, \tilde{\mathbf{x}}_i \right]$$

que es un sistema de $p+1$ ecuaciones no lineales en $\boldsymbol{\beta}$. Se resuelve utilizando el algoritmo de Newton-Raphson, basado en el desarrollo de Taylor de orden 1 del estadístico gradiente:

$$\mathbf{0} = \left. \frac{\partial \log L}{\partial \boldsymbol{\beta}} \right|_{\text{emv}(\boldsymbol{\beta})} \simeq \left. \frac{\partial \log L}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}} + \left. \frac{\partial^2 \log L}{\partial \boldsymbol{\beta}^2} \right|_{\boldsymbol{\beta}} (\text{emv}(\boldsymbol{\beta}) - \boldsymbol{\beta})$$

Por tanto, la iteración es

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \left(\left. \frac{\partial^2 \log L}{\partial \boldsymbol{\beta}^2} \right|_{\boldsymbol{\beta}^{(m)}} \right)^{-1} \left. \frac{\partial \log L}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}^{(m)}}.$$

En el caso particular de regresión logística:

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \mathbb{X}'(\mathbf{y} - \mathbf{p}) \qquad \text{y} \qquad \frac{\partial^2 \log L}{\partial \boldsymbol{\beta}^2} = -\mathbb{X}'\mathbf{W}\mathbb{X},$$

donde $\mathbf{p} = (\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n))'$ y
$\mathbf{W} = \text{diag}(\eta(\mathbf{x}_1)(1 - \eta(\mathbf{x}_1)), \ldots, \eta(\mathbf{x}_p)(1 - \eta(\mathbf{x}_p)))'$.

Por tanto,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbb{X}'\mathbf{W}^{(m)}\mathbb{X})^{-1}\mathbb{X}'(\mathbf{y} - \mathbf{p}^{(m)}).$$

Una vez obtenido $\text{emv}(\boldsymbol{\beta}) = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)'$, tenemos un estimador de la función de regresión

$$\hat{\eta}(\mathbf{x}) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_p x_p)}}$$

y el clasificador correspondiente, que es lineal aunque diferente en general del de Fisher,

$$g_n(\mathbf{x}) = \mathbb{1}_{\{\hat{\eta}(\mathbf{x}) > 1/2\}} = \mathbb{1}_{\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k > 0\}}.$$

Inferencia

Aplicando la teoría asintótica de los emv se demuestra que, si n es suficientemente grande,

$$\hat{\boldsymbol{\beta}} \overset{\text{aprox}}{\sim} \mathrm{N}_{p+1}\left(\boldsymbol{\beta}, (\mathbb{X}'\hat{\mathbf{W}}\mathbb{X})^{-1}\right),$$

donde $\hat{\mathbf{W}} = \mathrm{diag}(\hat{\eta}(\mathbf{x}_1)(1-\hat{\eta}(\mathbf{x}_1)), \ldots, \hat{\eta}(\mathbf{x}_n)(1-\hat{\eta}(\mathbf{x}_n)))$.

Esta aproximación es la base de los contrastes e intervalos para los parámetros del modelo.

Estadístico de Wald: Si $H_0 : \beta_j = 0$ es cierta, entonces

$$\frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)} \overset{\text{aprox}}{\sim} \mathrm{N}(0,1),$$

donde s.e.$(\hat{\beta}_j)$ es la raíz del j-ésimo elemento de la diagonal de $(\mathbb{X}'\hat{\mathbf{W}}\mathbb{X})^{-1}$.

Ejemplo con datos simulados:

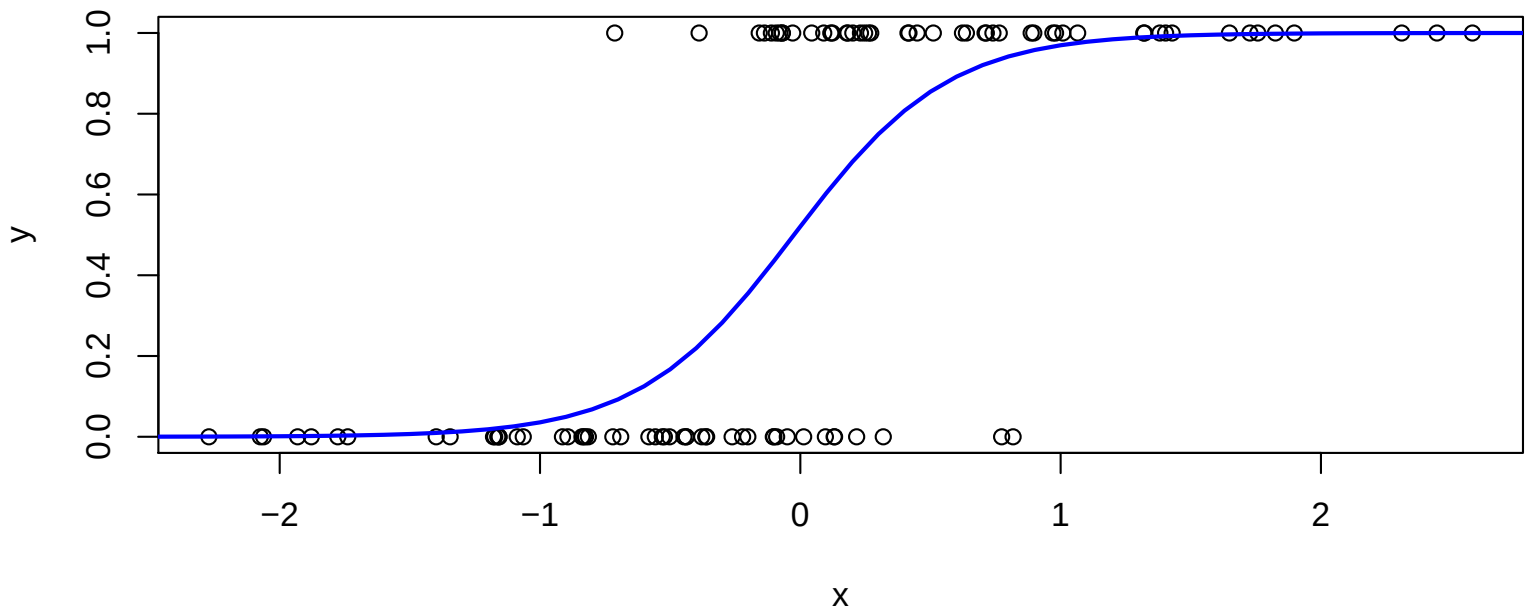
```
set.seed(100)
n <- 100
beta0 <- 0
beta1 <- 3
x <- rnorm(n)  # el modelo no asume normalidad de x
p = 1/(1+exp(-beta0-beta1*x))
y = rbinom(n, 1, p)

# Ajusta el modelo
reg = glm(y~x, family=binomial)
summary(reg)

Call:
glm(formula = y ~ x, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.40849  -0.53743  -0.00721   0.48375   2.19983
```

Ejemplo con datos simulados:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.08244    0.29764   0.277    0.782
x            3.37842    0.72712   4.646 3.38e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  70.219  on 98  degrees of freedom
AIC: 74.219

Number of Fisher Scoring iterations: 6
```

Ejemplo con datos simulados:

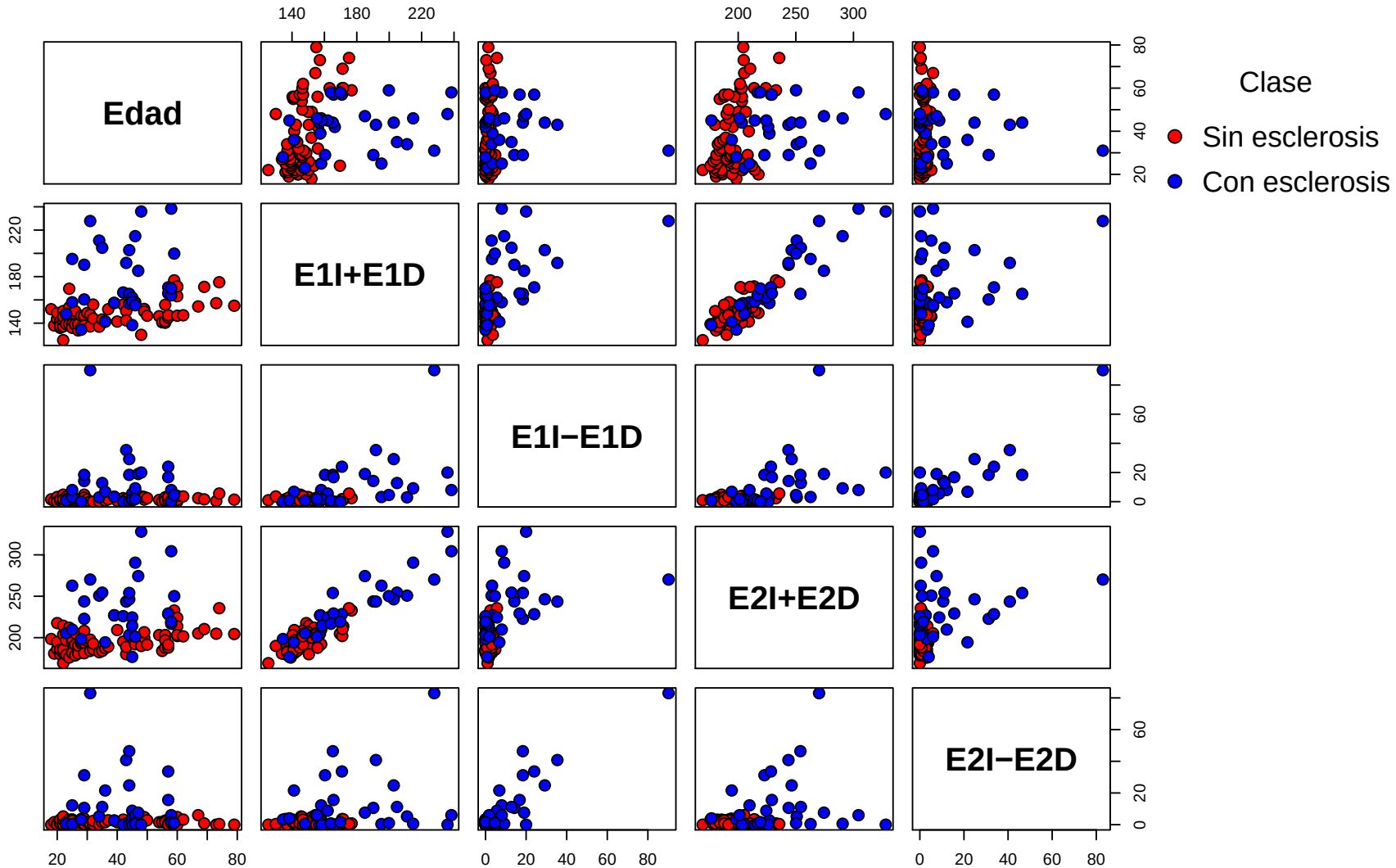
```
datos <- data.frame(x = seq(-4, 4, 0.1))
probabilidades <- predict(reg, datos, type = "response")
# por defecto calcula log p_i/(1-p_i), para calcular p_i
    usamos el argumento type
plot(x, y, pch = 21)
lines(datos$x, probabilidades, col = "blue", lwd = 2)
```



Ejemplo (esclerosis múltiple): El fichero `sclerosis.dat` contiene reacciones de sujetos de diferentes edades (columna 1), con y sin esclerosis múltiple (última columna: 0=sin, 1=con), a dos estímulos en ambos ojos. Las columnas 2 y 4 son la suma de las reacciones de ambos ojos a los dos estímulos por separado. Las columnas 3 y 5 son las diferencias entre las reacciones de los dos ojos a los estímulos 1 y 2 respectivamente.

```
Datos = read.table("sclerosis.dat",header=F,sep="")
Y = Datos$V6
X = Datos[,1:5]

pairs(X, labels=c("Edad","E1I+E1D","E1I-E1D","E2I+E2D","E2I-E2D"),
      cex = 1.5, pch = 21, bg=c("red","blue")[unclass(factor(Y))],
      cex.labels = 2, font.labels = 2,oma = c(2,2,2,18))
par(xpd=TRUE)
legend("topright", inset=c(0,0), c("Sin esclerosis","Con esclerosis"),
       pch = 21, pt.bg=c("red","blue"), cex=1.2, text.font=1,
       title="Clase",bty="n")
```

Ejemplo (esclerosis múltiple):



Ejemplo (esclerosis múltiple):

```
reglog = glm(V6~V1+V2+V3+V4+V5,data=Datos,family="binomial")
summary(reglog)

Call:
glm(formula = V6 ~ V1+V2+V3+V4+V5,family="binomial",data=Datos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.31250  -0.45006  -0.21017   0.01804   2.63976

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.84580    5.95999  -3.833 0.000126 ***
V1           -0.01728    0.02586  -0.668 0.503993
V2            0.02534    0.04683   0.541 0.588376
V3           -0.26880    0.20212  -1.330 0.183557
V4            0.08531    0.03856   2.212 0.026941 *
V5            0.44886    0.18532   2.422 0.015430 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 119.044  on 97  degrees of freedom
Residual deviance:  52.404  on 92  degrees of freedom
AIC: 64.404

Number of Fisher Scoring iterations: 7
```

Ejemplo (esclerosis múltiple):

- Escribe la fórmula estimada para la probabilidad de que un individuo padezca esclerosis múltiple en función de su edad (x_1) y de las variables x_2 = E1I+E1D, x_3 = E1I-E1D, x_4 = E2I+E2D, $x_5 = E2I - E2D$.

- Calcula un intervalo de confianza de nivel 95% para el coeficiente de la edad.

```
confint.default(reglog)
                      2.5 %      97.5 %
(Intercept) -34.527175537 -11.1644265
V1
V2           -0.066438005   0.1171236
V3           -0.664944976   0.1273521
V4            0.009732533   0.1608845
V5            0.085645495   0.8120798
```

- Lleva a cabo los contrastes de Wald para $H_0 : \beta_j = 0$.

Bondad de ajuste en regresión logística

En regresión logística hay diversas maneras de comparar los valores observados de Y con sus valores previstos por el modelo. Por ejemplo, definimos la desviación en el individuo i como

$$D_i^2 = -2[y_i \log \hat{\eta}(\mathbf{x}_i) + (1 - y_i) \log(1 - \hat{\eta}(\mathbf{x}_i))].$$

- Si $y_i = 1$, ¿cómo cambia D_i^2 cuando $\hat{\eta}(\mathbf{x}_i)$ decrece a 0?
- Si $y_i = 0$, ¿cómo cambia D_i^2 cuando $\hat{\eta}(\mathbf{x}_i)$ crece a 1?

En regresión logística, el análogo de la suma de cuadrados residual de la regresión lineal es la *desviación* (*deviance*)

$$D^2 = \sum_{i=1}^{n} D_i^2 = -2 \log L(\hat{\beta}_0, \ldots, \hat{\beta}_p).$$

El emv de $\boldsymbol{\beta}$ minimiza D^2.

Para valorar la bondad del ajuste del modelo logit a los datos se podría usar D^2.

Pero conviene tener en cuenta la complejidad del modelo. Una posibilidad es seleccionar el modelo que minimiza el *criterio de información de Akaike*:

$$\text{AIC} = -2 \log L(\hat{\beta}_0, \ldots, \hat{\beta}_p) + 2(p+1) = D^2 + 2(p+1).$$

término de penalización que mide la complejidad

Sobreajusta menos el *BIC* (*Bayesian Information Criterion*):

$$\text{BIC} = -2 \log L(\hat{\beta}_0, \ldots, \hat{\beta}_p) + p \log n = D^2 + p \log n.$$

término de penalización BIC

Ejemplo (esclerosis múltiple):

```
reglog = glm(V6~V1+V2+V3+V4+V5,data=Datos,family="binomial")
summary(reglog)

Call:
glm(formula = V6 ~ V1+V2+V3+V4+V5,family="binomial",data=Datos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.31250  -0.45006  -0.21017   0.01804   2.63976

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.84580    5.95999  -3.833 0.000126 ***
V1           -0.01728    0.02586  -0.668 0.503993
V2            0.02534    0.04683   0.541 0.588376
V3           -0.26880    0.20212  -1.330 0.183557
V4            0.08531    0.03856   2.212 0.026941 *
V5            0.44886    0.18532   2.422 0.015430 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 119.044  on 97  degrees of freedom
Residual deviance:  52.404  on 92  degrees of freedom
AIC: 64.404

Number of Fisher Scoring iterations: 7
```

Otra posibilidad es utilizar el *contraste de razón de verosimilitudes* para determinar si una variable o un grupo de variables, incluidas dentro del modelo logit, son significativas.

Sea V_0 un subespacio de $\mathbb{R}^{p+1}$ y $H_0 : \boldsymbol{\beta} \in V_0$.

La razón de verosimilitudes es:

$$\lambda_n = \frac{\sup_{\boldsymbol{\beta} \in V_0} L(\boldsymbol{\beta})}{\sup_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} L(\boldsymbol{\beta})} = \frac{L(\hat{\boldsymbol{\beta}}^{(0)})}{L(\hat{\boldsymbol{\beta}})}.$$

Se cumple que

$$-2 \log \lambda_n = -2 \log L(\hat{\boldsymbol{\beta}}^{(0)}) + 2 \log L(\hat{\boldsymbol{\beta}}) = D_0^2 - D^2.$$

Bajo $H_0 : \boldsymbol{\beta} \in V_0$, puede demostrarse que

$$-2 \log \lambda_n \xrightarrow[n \to \infty]{\mathrm{d}} \chi^2_k,$$

donde $k = p + 1 - \dim(V_0)$.
Región de rechazo de H_0: $R = \{-2 \log \lambda_n > \chi^2_{k,\alpha}\}$.

Ejemplo (esclerosis múltiple):

- Contrasta $H_0 : \beta_1 = \cdots = \beta_5 = 0$ mediante razón de verosimilitudes.

```
reglog = glm(V6~V1+V2+V3+V4+V5,data=Datos,family="binomial")
# Modelo sin regresores:
reglog_12345 <- glm(V6~1,data=Datos,family="binomial")

Test_12345 = anova(reglog_12345,reglog)
Test_12345
Analysis of Deviance Table

Model 1: V6 ~ 1
Model 2: V6 ~ V1 + V2 + V3 + V4 + V5
  Resid. Df Resid. Dev Df Deviance
1        97    119.044
2        92     52.404  5    66.64

pchisq(Test_12345$Deviance[2],Test_12345$Df[2],lower.tail=F)
5.118853e-13
```

Ejemplo (esclerosis múltiple):

- Contrasta $H_0 : \beta_1 = 0$ mediante razón de verosimilitudes.

```
reglog = glm(V6~V1+V2+V3+V4+V5,data=Datos,family="binomial")
reglog_1 = glm(V6~V2+V3+V4+V5,data=Datos,family="binomial")

Test_1 = anova(reglog_1,reglog)
Test_1
Analysis of Deviance Table

Model 1: V6 ~ V2 + V3 + V4 + V5
Model 2: V6 ~ V1 + V2 + V3 + V4 + V5
  Resid. Df Resid. Dev Df Deviance
1        93     52.866
2        92     52.404  1   0.4623

pchisq(Test_1$Deviance[2],Test_1$Df[2],lower.tail=F)
[1] 0.4965527
```

Los contrastes de Wald y de razón de verosimilitudes suelen dar p-valores parecidos pero no son equivalentes.

Referencias

Agresti, A. (2013). *Categorical Data Analysis*. 3rd ed. Wiley.
Secc. 2.2: "Comparing Two Proportions"

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59, 2–5.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.
Cap. 1: "Introduction"
Cap. 2: "The Bayes error"

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II, 179–188.

Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall.
Cap. 11: "Discrimination and classification"

Peña, D. (2002). *Regresión y diseño de experimentos*. Alianza.
Cap. 14: "Variables con respuesta cualitativa"

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.