Clasificación y regresión logística

José R. Berrendero

Universidad Autónoma de Madrid

Contenidos

- Planteamiento del problema de clasificación supervisada
- Regla lineal de Fisher
- Regresión logística
- Optimalidad: la regla Bayes

El problema de clasificación supervisada

Disponemos de una muestra de k variables medidas en n unidades u objetos que pertenecen a dos grupos o poblaciones *(training data)*.

Cada observación $i = 1, \ldots, n$ consiste en un vector $(x_i', y_i)'$, donde $x_i \in \mathbb{R}^k$ son las k variables e $y \in \{0, 1\}$ indica el grupo al que pertenece la unidad en la que se han obtenido.

Objetivo: Asignar una nueva unidad con valores x (e y desconocida) a uno de los dos grupos (**obtener una regla de clasificación**).

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

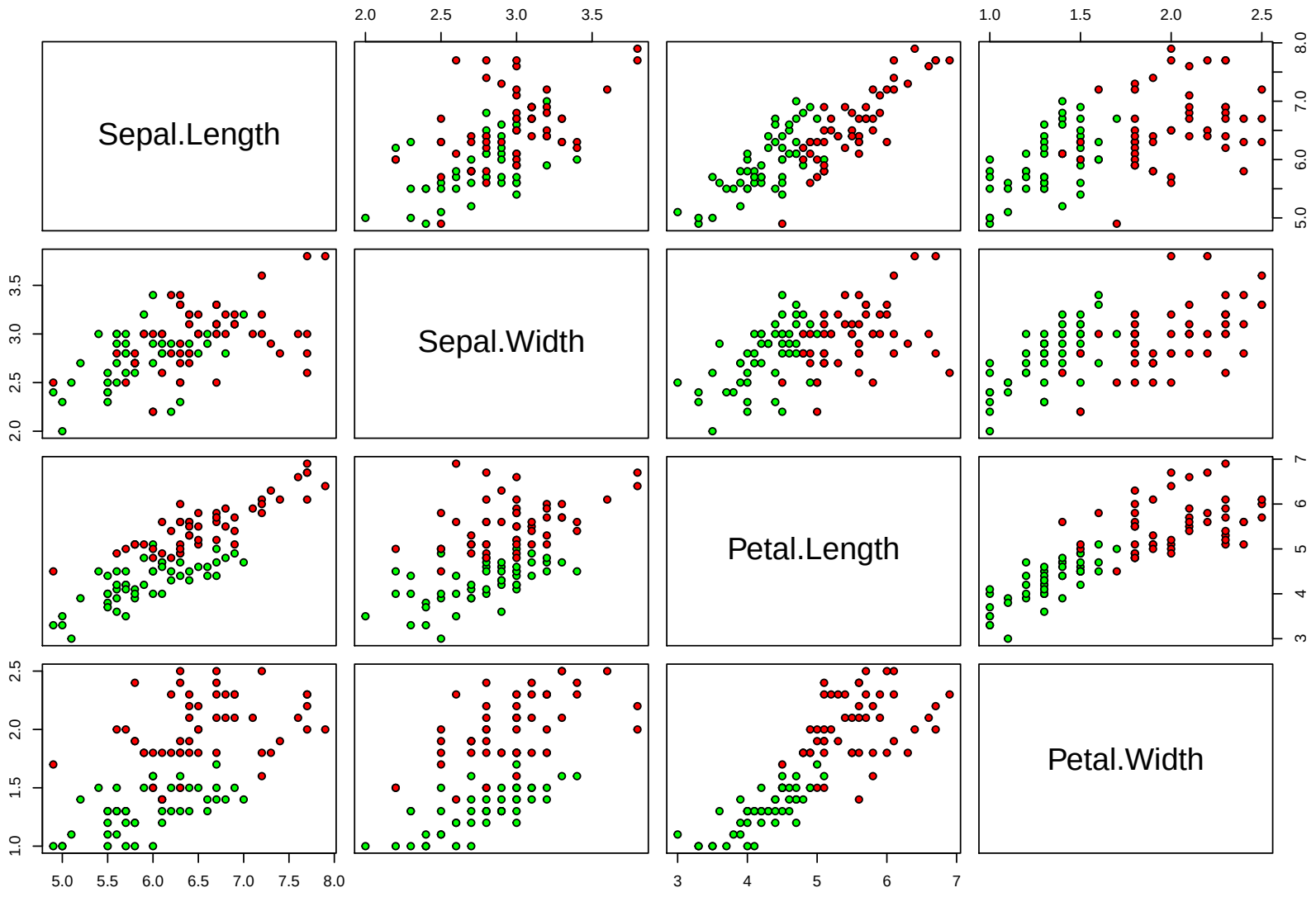
Ejemplo

Se dispone de las medidas del pétalo y del sépalo de 50 lirios de la especie *versicolor* y 50 de la especie *virginica*.

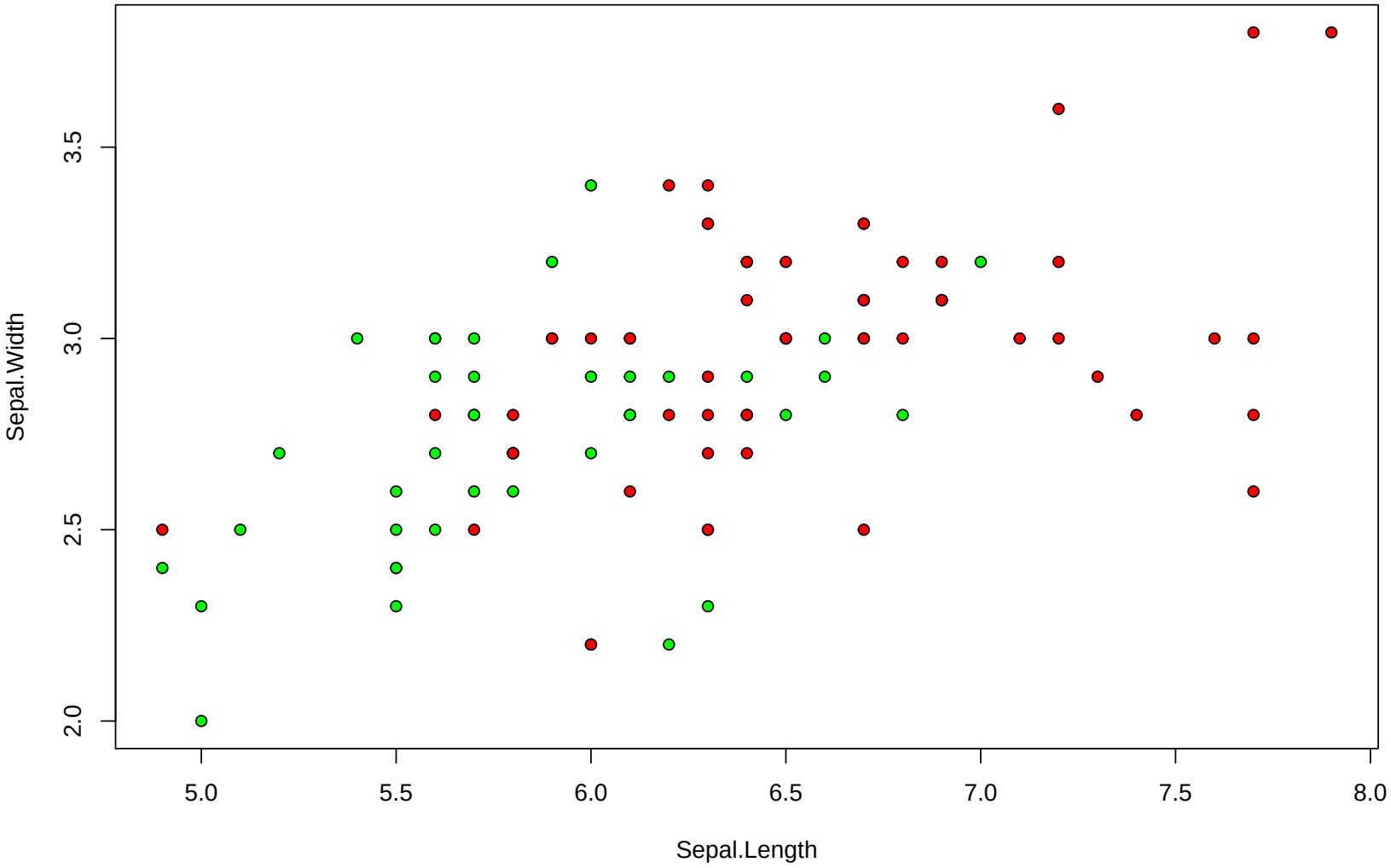
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1          7.0         3.2          4.7         1.4 versicolor
## 2          6.4         3.2          4.5         1.5 versicolor
## 3          6.9         3.1          4.9         1.5 versicolor
## 4          5.5         2.3          4.0         1.3 versicolor
## 5          6.5         2.8          4.6         1.5 versicolor
## 6          5.7         2.8          4.5         1.3 versicolor
```

Representamos en verde la especie *versicolor* y en rojo la especie *virginica*.

Ejemplo


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0
2.5
3.0
3.5
1.0
1.5
2.0
2.5
5.0
6.0
7.0
8.0
2.0
2.5
3.0
3.5
3
4
5
6
7
1.0
1.5
2.0
2.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

Ejemplo



Dos modelos ligeramente diferentes

Modelo 1: Fijamos n_0 y n_1 ($n_0 + n_1 = n$) y se observa

$$X_{0,1}, \ldots, X_{0,n_0} \text{ i.i.d. } P_0 \equiv X|Y = 0$$

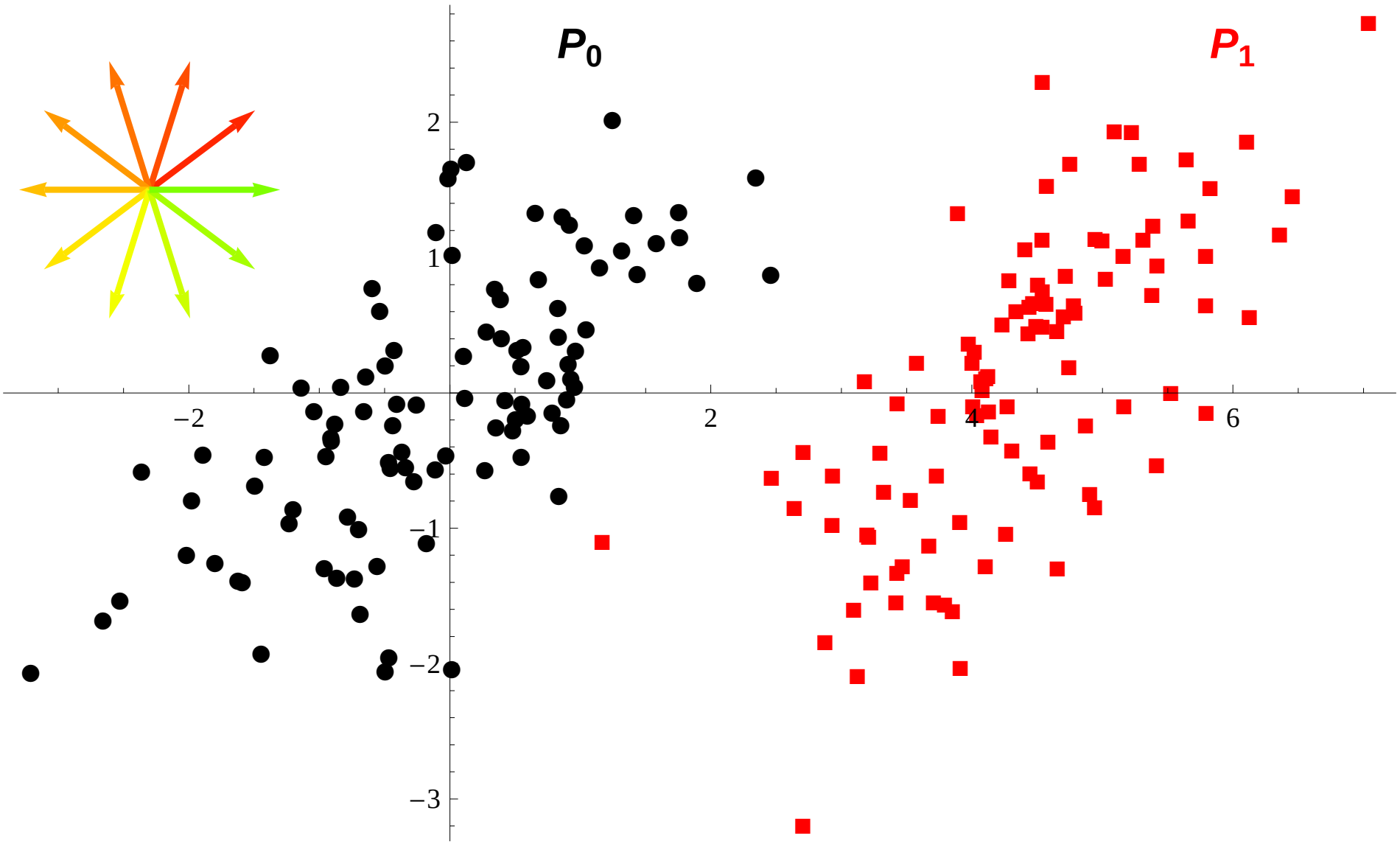
$$X_{1,1}, \ldots, X_{1,n_1} \text{ i.i.d. } P_1 \equiv X|Y = 1$$

Modelo 2: Las variables $Y_1, \ldots, Y_n$ son independientes con distribución de Bernoulli, $Y_i \equiv \text{Binom}(1, \eta(X_i))$ para cierta función $\eta(\cdot)$.

Denotaremos $\mu_i = \mathbb{E}(X|Y = i)$, $\Sigma_i = \text{Cov}(X|Y = i)$, para $i = 0, 1$.

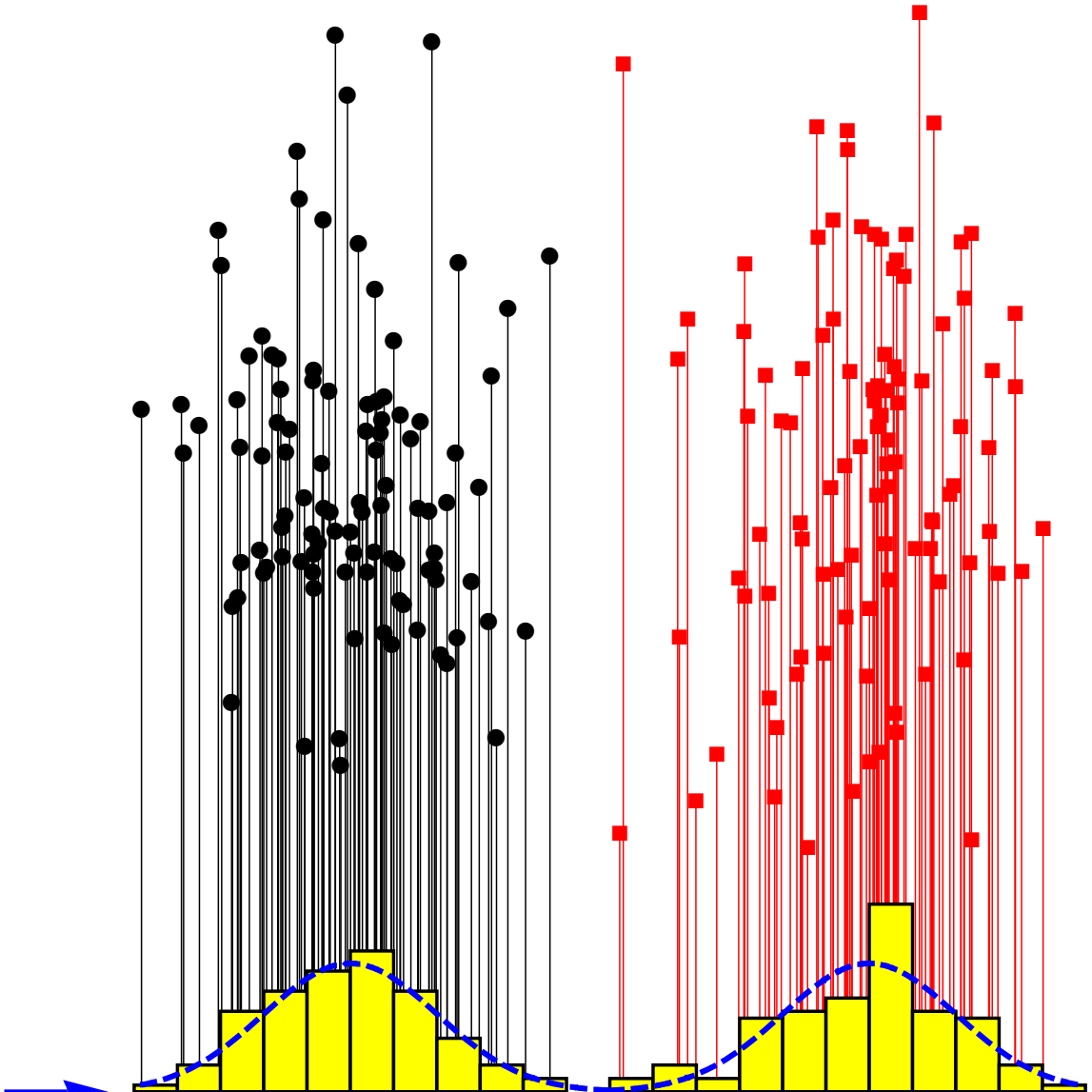
En el modelo 2, los valores n_0 y n_1 son aleatorios.

Regla lineal de Fisher



Regla lineal de Fisher

El enfoque de Fisher: Proyectar los datos en la dirección a más conveniente y utilizar las proyecciones $a'x_i$ para discriminar.



Regla lineal de Fisher

Suponemos $\Sigma_0 = \Sigma_1 = \Sigma$.

Una buena dirección debe separar bien los centros de los grupos. La distancia entre las medias $(a'\mu_0 - a'\mu_1)^2 = a'Ba$, donde $B = (\mu_0 - \mu_1)(\mu_0 - \mu_1)'$, debe ser grande.

La varianza de las proyecciones dentro de los grupos $(a'\Sigma a)$ debe ser lo menor posible.

Problema: Encontrar la dirección a que maximiza

$$f(a) = \frac{a'Ba}{a'\Sigma a} \qquad (\textbf{cociente de Rayleigh}).$$

Para cualquier $\lambda \neq 0$, $f(\lambda a) = f(a)$, por lo que es necesario normalizar. En **R** se impone $a'\Sigma a = 1$.

La solución es proporcional al vector $w = \Sigma^{-1}(\mu_1 - \mu_0)$.

Regla lineal de Fisher

- Proyectamos en la dirección w el punto x que queremos clasificar y los vectores de medias de los dos grupos.
- Clasificamos x en P_1 si su proyección está más cerca de la proyección de la media del grupo 1, que de la del grupo 0.

Regla de Fisher: Clasificar x en el grupo 1 (i.e. $Y = 1$) si y solo si

$$w' \left(x - \frac{\mu_0 + \mu_1}{2} \right) > 0,$$

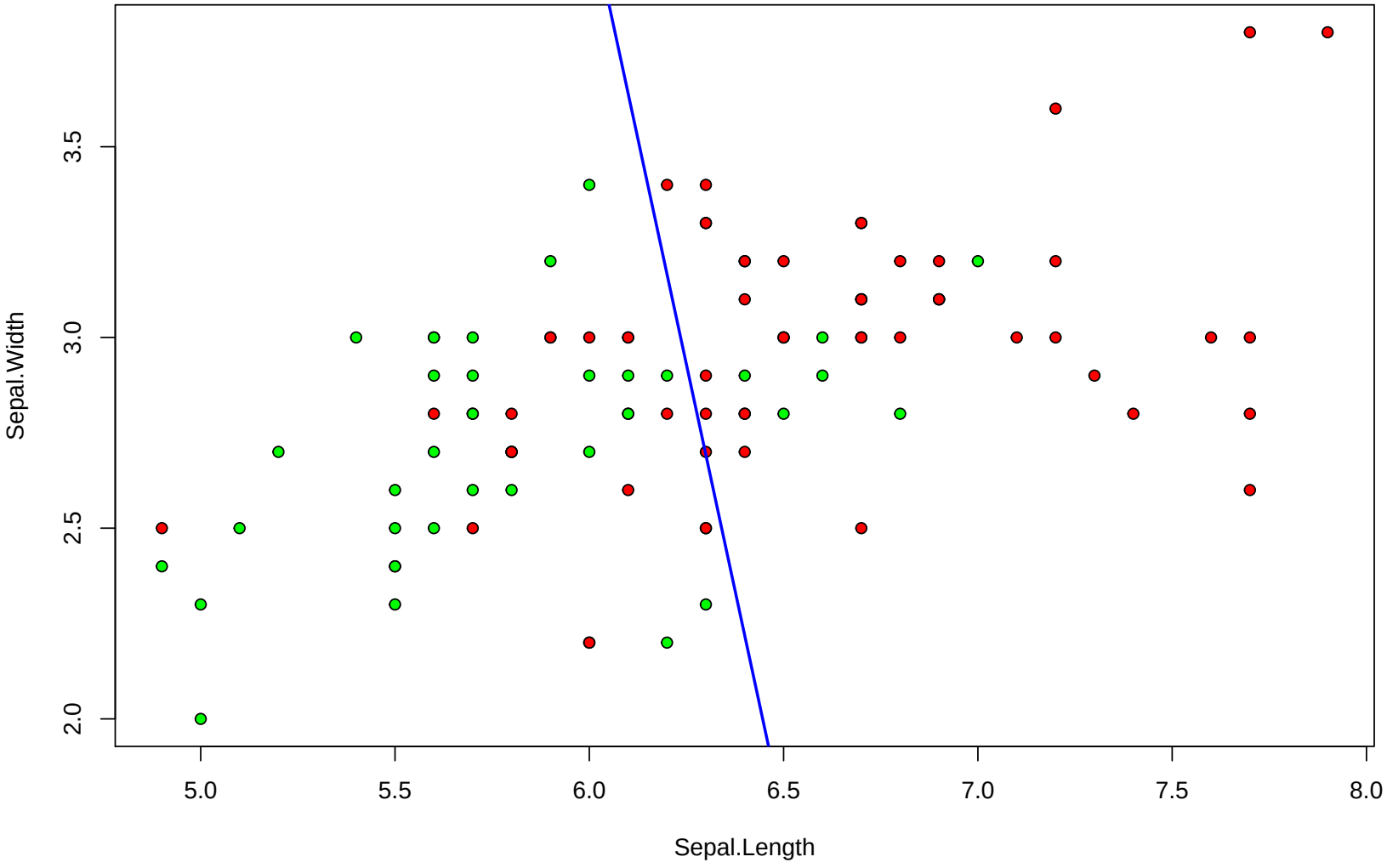
donde $w = \Sigma^{-1}(\mu_1 - \mu_0)$.

- En la práctica se usan los vectores de medias muestrales y la matriz de covarianzas estimada combinada siguiente:

$$\hat{\Sigma} = \frac{n_0 - 1}{n_0 + n_1 - 2} S_0 + \frac{n_1 - 1}{n_0 + n_1 - 2} S_1,$$

donde S_i es la matriz de covarianzas muestral del grupo i, $i = 0, 1$.

Ejemplo



Código para la figura anterior

```
library(MASS)
resultadoSep <- lda(Species~., data=lirios2)
w <- resultadoSep$scaling
medias <- resultadoSep$means
w0 <- sum(colMeans(medias)*w)

plot(lirios2[,1:2], pch=21, bg=colores)
abline(w0/w[2], -w[1]/w[2], lwd=2, col='blue')
```

Estimación del error de clasificación

Es importante estimar la probabilidad de error de clasificación.

La **tasa de error aparente** (TEA) es:

$$\text{TEA} := \frac{\text{Total de mal clasificados en la muestra}}{n} 100\%.$$

La TEA tiende a infraestimar el verdadero error ya que los datos se utilizan tanto para calcular la regla de clasificación como para evaluarla.

Estimación del error de clasificación

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:** 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.

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

Estimación del error de clasificación

```r
# Tasa de error aparente (sépalo)
n <- sum(resultadoSep$counts)
sum(lirios2$Species != predict(resultadoSep)$class) / n
```

```
## [1] 0.25
```

```r
# Tasa de error por VC (sépalo)
resultadoSepVC <- lda(Species~., data=lirios2, CV=TRUE)
sum(lirios2$Species != resultadoSepVC$class) / n
```

```
## [1] 0.27
```

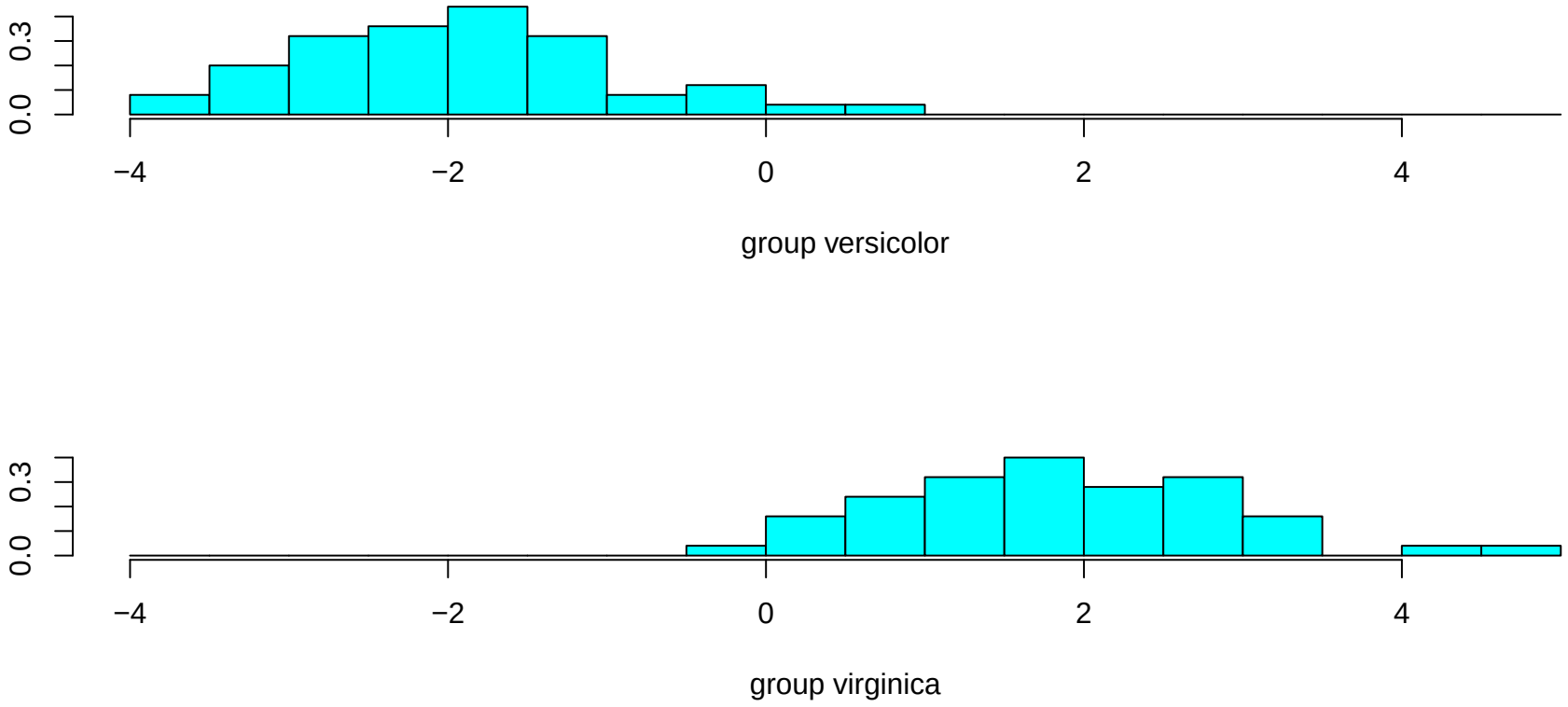
La regla de Fisher para las cuatro variables

```
resultado <- lda(Species~., data=lirios)
resultado$scaling
```

```
##                    LD1
## Sepal.Length -0.9431178
## Sepal.Width  -1.4794287
## Petal.Length  1.8484510
## Petal.Width   3.2847304
```

La regla de Fisher para las cuatro variables

```
plot(resultado)
```



Estimación del error de clasificación

```r
# Tasa de error aparente (pétalo y sépalo)
resultado <- lda(Species~., data=lirios)
sum(lirios$Species != predict(resultado)$class) / n
```

```
## [1] 0.03
```

```r
# Tasa de error por VC (pétalo y sépalo)
resultadoVC <- lda(Species~., data=lirios, CV=TRUE)
sum(lirios$Species != resultadoVC$class) / n
```

```
## [1] 0.03
```

Regresión logística

Disponemos de n observaciones. Cada observación $(x_{i1}, \ldots, x_{ik}, y_i)'$ está formada por un vector de variables regresoras $x_i = (1, x_{i1}, \ldots, x_{ik})'$ y el valor de la variable respuesta y_i.

Las variables $Y_1, \ldots, Y_n$ son independientes y tienen distribución de Bernoulli.

La probabilidad de "éxito" depende de las variables regresoras. Denotamos $p_i = \eta(x_i) = \mathbb{P}(Y_i = 1 \mid x_i)$.

Una relación lineal $p_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ no es adecuada.

Regresión logística

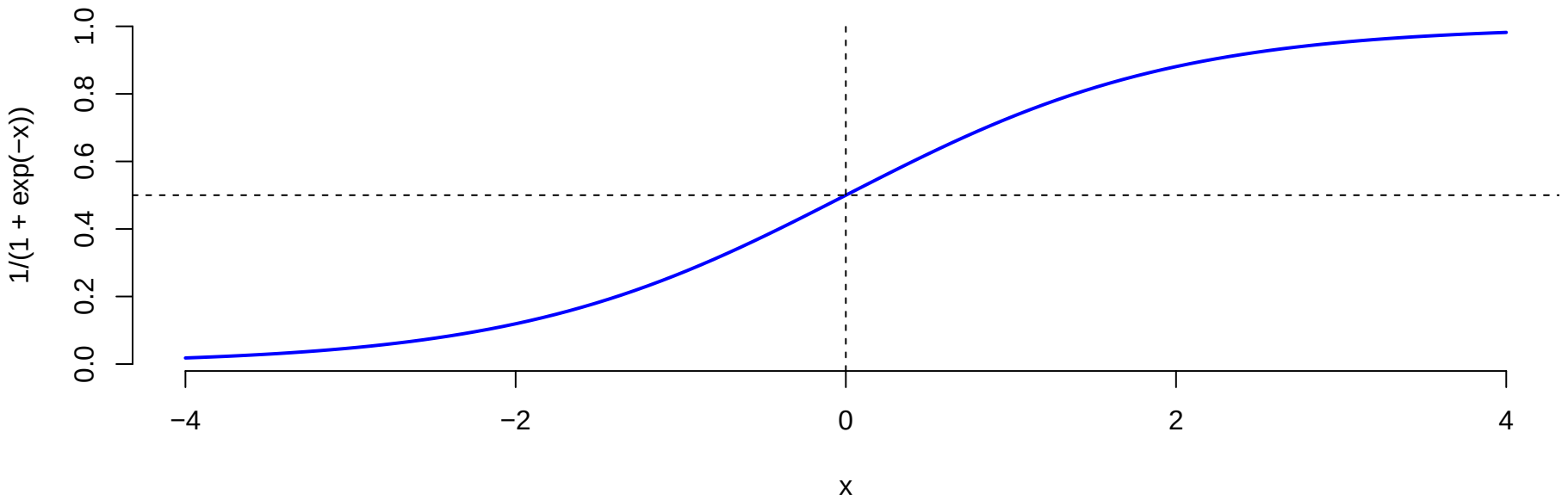
Suponemos que la relación entre p_i y x_i viene dada por

$$p_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}}},$$

es decir,

$$p_i = f(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}),$$

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



Algunas propiedades de la función logística

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

La función logística no es la única que se ha utilizado para modelizar este tipo de datos.

El modelo **probit** consiste en suponer $p_i = \Phi(x_i)$, donde Φ es la función de distribución normal estándar.

Interpretación de los parámetros del modelo

Llamamos O_i a la **razón de probabilidades** para la observación i:

$$O_i = \frac{p_i}{1 - p_i}$$

¿Cómo se interpreta el valor de O_i? ¿Qué significa, por ejemplo, $O_i = 2$?

Si se cumple el modelo de regresión logística, entonces

$$O_i = e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}$$

¿Cómo varía la razón de probabilidades si la variable regresora x_{ij} se incrementa una unidad?

$$\frac{O_i'}{O_i} = \frac{e^{\beta_0 + \cdots + \beta_j (x+1) + \cdots + \beta_k x_{ik}}}{e^{\beta_0 + \cdots + \beta_j x + \cdots + \beta_k x_{ik}}} = e^{\beta_j}.$$

Por tanto e^{β_j} es la variación de la razón de probabilidades cuando la variable regresora j se incrementa en una unidad y el resto de variables permanece constante.

Estimación

Para estimar los parámetros se usa el método de máxima verosimilitud.

Por ejemplo, si observamos los datos $(2,0),(1,1),(3,1)$, entonces $\hat{\beta}_0$ y $\hat{\beta}_1$ son los valores que maximizan la función de verosimilitud

$$L(\beta_0,\beta_1) = P(Y=0 \mid x=2)P(Y=1 \mid x=1)P(Y=1 \mid x=3)$$

$$L(\beta_0,\beta_1) = \left(1 - \frac{1}{1+e^{-\beta_0-2\beta_1}}\right)\left(\frac{1}{1+e^{-\beta_0-\beta_1}}\right)\left(\frac{1}{1+e^{-\beta_0-3\beta_1}}\right)$$

Esta función es cóncava. Se pueden aplicar algoritmos estándar de optimización para maximizarla.

Estimación

Verosimilitud:

$$L(\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}.$$

Log. de la verosimilitud:

$$\ell(\beta) = \log L(\beta) = \sum_{i=1}^{n}[Y_i \log p_i + (1-Y_i)\log(1-p_i)]$$

El EMV es el valor para el que se anula el gradiente:

$$\nabla(\hat{\beta}) = \sum_{i=1}^{n}\left[Y_i x_i - \frac{1}{1+e^{-x_i'\hat{\beta}}}x_i\right] = 0.$$

Estas ecuaciones son análogas a las ecuaciones normales en regresión lineal:

$$\sum_{i=1}^{n}(Y_i - \hat{p}_i)x_i = 0 \Leftrightarrow X'Y = X'\hat{p}.$$

Desviaciones

Las desviaciones (*deviances*) se definen:

$$D_i^2 = -2[Y_i \log \hat{p}_i + (1 - Y_i) \log(1 - \hat{p}_i)]$$

- Si $Y_i = 1$, ¿cómo cambia D_i^2 cuando $\hat{p}_i$ decrece a 0?
- Si $Y_i = 0$, ¿cómo cambia D_i^2 cuando $\hat{p}_i$ crece a 1?

Los valores D_i^2 hacen el papel de los residuos en regresión lineal.

El análogo de SCE es $\sum_{i=1}^n D_i^2$. Se cumple $D^2 = \sum_{i=1}^n D_i^2 = -2\ell(\hat{\beta})$.

Desviaciones

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

Resulta conveniente tener en cuenta la complejidad del modelo. Una posibilidad es usar el **criterio de información de Akaike**:

$$\text{AIC} = -2\ell(\hat{\beta}) + 2(k+1) = D^2 + 2(k+1).$$

Inferencia

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

$$\hat{\beta} \cong \mathrm{N}_{k+1}(\beta, (X'\hat{W}X)^{-1}),$$

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

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

Estadístico de Wald: si $\beta_j = 0$,

$$\frac{\hat{\beta}_j}{\mathrm{e.t.}(\hat{\beta})} \cong \mathrm{N}(0,1),$$

donde e.t.$(\hat{\beta})$ es la raíz del elemento correspondiente de la diagonal de $(X'\hat{W}X)^{-1}$.

Un ejemplo con datos simulados

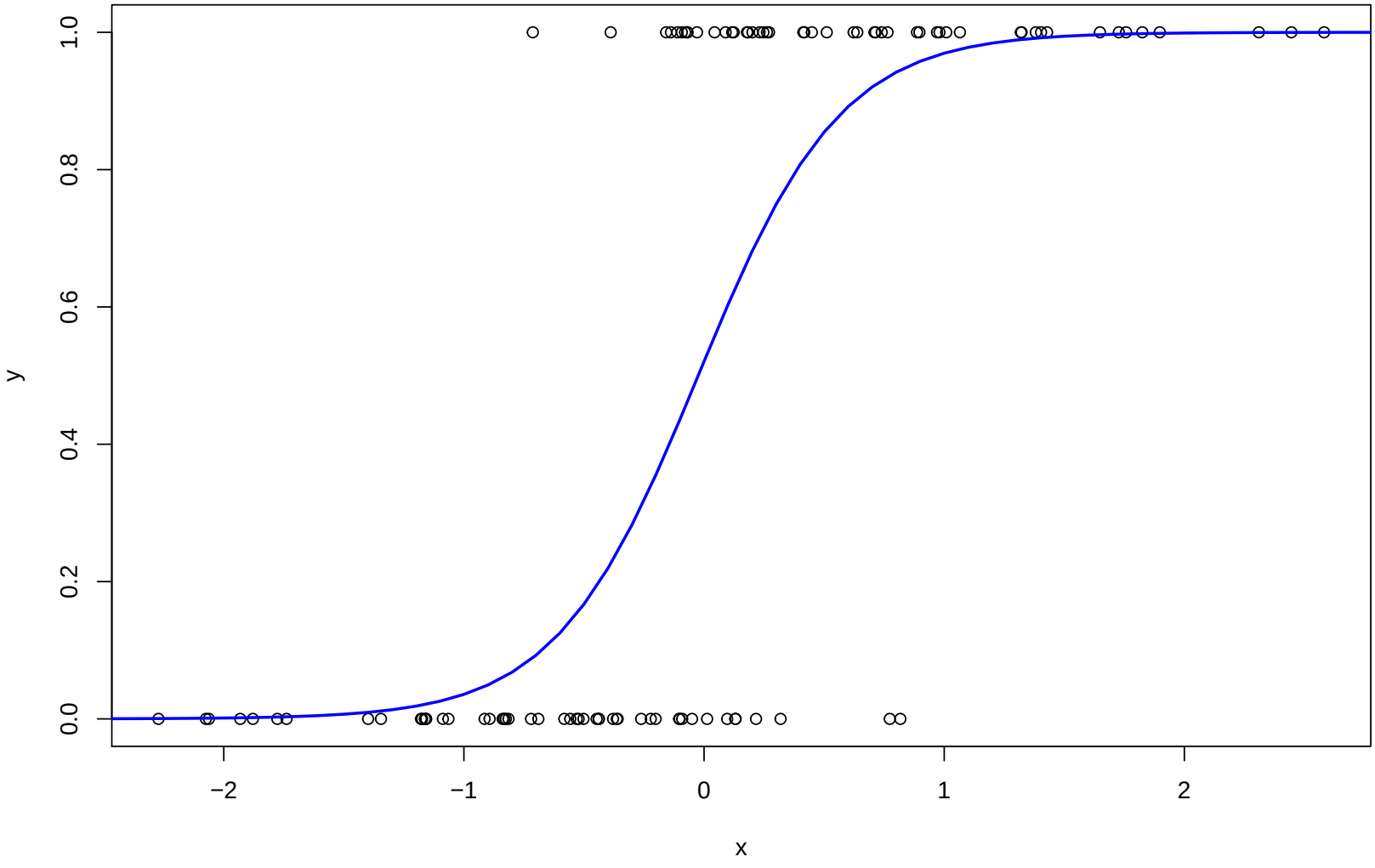
```r
set.seed(100)
n <- 100
beta0 <- 0
beta1 <- 3
x <- rnorm(n)  # el modelo no asume normalidad de x
p = 1/(1+exp(-beta0-beta1*x))
y = rbinom(n, 1, p)

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

Un ejemplo con datos simulados

```
##
## Call:
## glm(formula = y ~ x, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.40849  -0.53743  -0.00721   0.48375   2.19983
##
## 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
```

Probabilidades estimadas



Contraste de razón de verosimilitudes

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

La razón de verosimilitudes es:

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

Se verifica:

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

Puede demostrarse que, bajo H_0,

$$-2 \log \lambda_n \to_d \chi^2_p,$$

donde $p = k + 1 - \dim(V_0)$.

Se rechaza H_0 en $R = \{-2 \log \lambda_n > \chi^2_{p,\alpha}\}$.

Ejemplo: datos de lirios

Codificamos: $Y = 0$ (versicolor) $Y = 1$ (virginica).

```
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -42.638     25.707  -1.659    0.097
## lirios$Sepal.Length    -2.465      2.394  -1.030    0.303
## lirios$Sepal.Width     -6.681      4.480  -1.491    0.136
## lirios$Petal.Length     9.429      4.737   1.991    0.047
## lirios$Petal.Width     18.286      9.743   1.877    0.061
```

Null deviance:

```
## [1] 138.6294
```

Deviance:

```
## [1] 11.89855
```

Cuestiones

- Escribe la fórmula estimada para la probabilidad de que un lirio pertenezca a la especie virginica en función de las medidas de su pétalo y su sépalo.
- Calcula un intervalo de confianza de nivel 95% para el coeficiente de la longitud del sépalo. (`confint` no da el IC de Wald)

```
##                                2.5 %     97.5 %
## (Intercept)              -118.866840 -9.8781379
## lirios$Sepal.Length        -9.099914  1.4787955
## lirios$Sepal.Width        -18.787029  0.1886706
## lirios$Petal.Length         3.332356 25.7555533
## lirios$Petal.Width          5.463641 45.7719798
```

- Lleva a cabo los contrastes de Wald para $H_0 : \beta_j = 0$.
- Contrasta $H_0 : \beta_1 = \cdots = \beta_4 = 0$ mediante razón de verosimilitudes.
- Contrasta $H_0 : \beta_1 = 0$ mediante razón de verosimilitudes.

Cuestiones

```
anova(reg0, reg)
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ lirios$Sepal.Width + lirios$Petal.Length + lirios$Pe
## Model 2: y ~ lirios$Sepal.Length + lirios$Sepal.Width + lirios$Pe
##     lirios$Petal.Width
##   Resid. Df Resid. Dev Df Deviance
## 1        96     13.266
## 2        95     11.899  1   1.3673
```

```
1-pchisq(1.3673, 1)
```

```
## [1] 0.2422763
```

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

Regla de clasificación logística

Se clasifica x en el grupo 1 (i.e. $Y = 1$) si y solo si

$$\mathbb{P}(\widehat{Y = 1}|x) > \mathbb{P}(\widehat{Y = 0}|x)$$

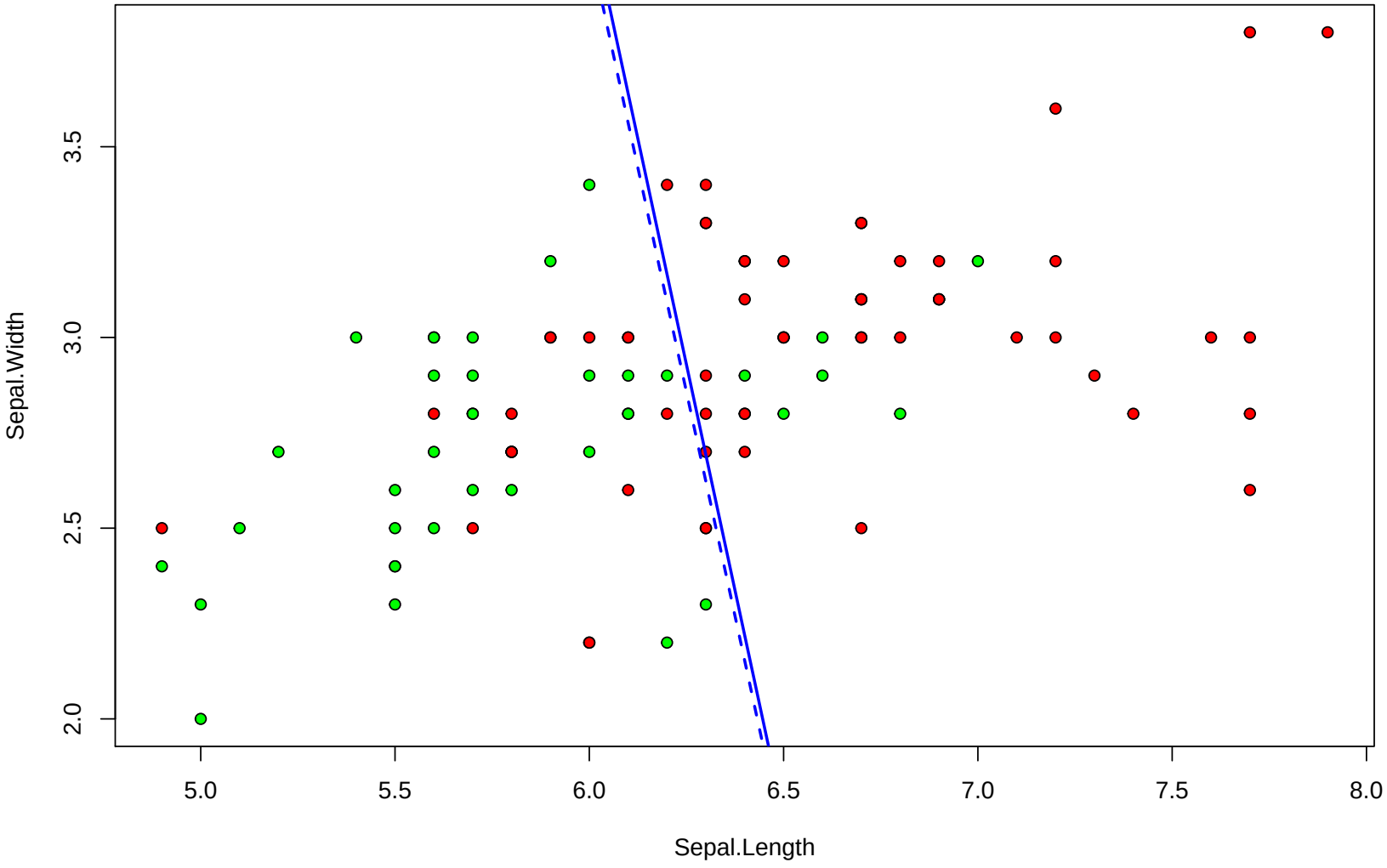
Se obtiene una regla lineal (diferente en general a la de Fisher): se clasifica x en el grupo 1 (i.e. $Y = 1$) si y solo si

$$\hat{\beta}_0 + \hat{\beta}_k x_1 + \ldots + \hat{\beta}_k x_k > 0.$$

En el **ejemplo**, clasificamos a un lirio como virginica si y solo si

$-42.65 - 2.46 \cdot \texttt{long.sep.} - 6.68 \cdot \texttt{anch.sep} + 9.43 \cdot \texttt{long.pet.} + 18.29 \cdot \texttt{anch.pet.} > 0.$

Ejemplo: medidas del sépalo (Fisher y regla logística)



Optimalidad: la regla Bayes

Regla Bayes: x se clasifica en P_1 si y solo si

$$\mathbb{P}(Y = 1|x) > \mathbb{P}(Y = 0|x)$$

En el caso en que

- P_0 tiene densidad f_0 y P_1 tiene densidad f_1,
- las **probabilidades a priori** de las poblaciones son

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

se tiene (fórmula de Bayes):

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

La regla Bayes es óptima (su error de clasificación es el mínimo posible). A este error se le llama **error Bayes**.

Regla Bayes bajo normalidad

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

$$f_i(x) = |\Sigma_i|^{-1/2}(2\pi)^{-k/2} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\}, \quad i = 0, 1.$$

Regla Bayes bajo normalidad

x se clasifica en P_0 si

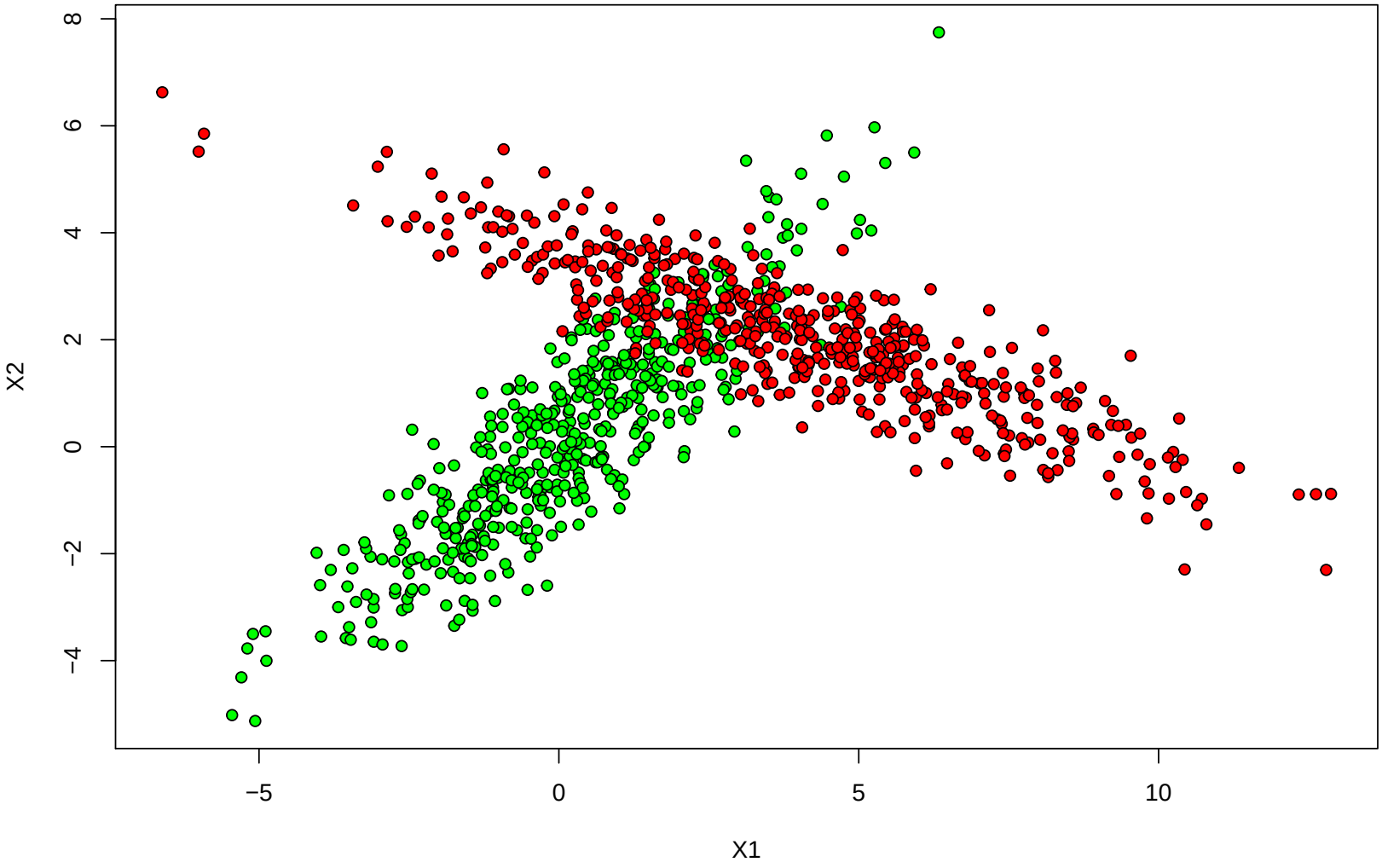
$$d^2_{M_0}(x, \mu_0) < d^2_{M_1}(x, \mu_1) + 2\log\left(\frac{\pi_0|\Sigma_1|^{1/2}}{\pi_1|\Sigma_0|^{1/2}}\right)$$

donde $d^2_{M_i}(x, \mu_i) = (x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)$ es el cuadrado de la distancia de Mahalanobis entre x y μ_i $(i = 0, 1)$.

Regla Bayes bajo normalidad y homocedasticidad: $(\Sigma_0 = \Sigma_1)$

x se clasifica en P_0 si $w'x < w'\left(\frac{\mu_0+\mu_1}{2}\right) + \log\left(\frac{\pi_0}{\pi_1}\right)$

Ejemplo: datos simulados



Ejemplo: datos simulados