Regresión Logística

Introducción

El problema de clasificación en dos grupos puede abordarse introduciendo una variable ficticia binaria para representar la pertenencia de una observación a uno de los dos grupos. Por ejemplo, si se desea discriminar entre créditos que se devuelven o que presentan problemas para su cobro, puede añadirse a la base de datos una nueva variable y que tome el valor 0, cuando el crédito se devuelve sin problemas y valor 1 en otro caso. El problema de discriminación es equivalente a la previsión del valor de la variable ficticia y. Si el valor previsto está más próximo a 0 que a 1, clasificaremos al elemento en la primera población. En otro caso, lo haremos en la segunda.

Se construye un modelo que permita prever el valor de la variable ficticia binaria de un elemento de una población, en función de ciertas características medibles $\mathbf{x}$. Supongamos que se dispone de una muestra de n elementos del tipo $(y_i, \mathbf{x}_i)$, donde y_i es igual a 0 cuando el elemento pertenece a la primera población P_1 y 1 cuando pertenece a la segunda P_2. A su vez, $\mathbf{x}_i$ es un vector de variables explicativas.

El primer enfoque es formular el siguiente modelo de regresión:

$$y = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x} + \mathbf{u} \tag{1}$$

y estimar los parámetros por mínimos cuadrados de la forma habitual. Este método es equivalente a la función lineal discriminante de Fisher. Como ya se vio, este procedimiento es óptimo para clasificar si la distribución conjunta de las variables explicativas es normal multivariante, con la misma matriz de covarianzas. Sin embargo, la discriminación

lineal puede funcionar mal en otros contextos, cuando las covarianzas sean distintas o las distribuciones muy alejadas de la normal. Además, si un objetivo importante del estudio es identificar qué variables son mejores para clasificar entre las dos poblaciones, la función lineal se encuentra con problemas de interpretación, tanto del modelo como de sus coeficientes estimados.

En concreto, tomando esperanzas en (1) para $\mathbf{x} = \mathbf{x}_i$

$$E\left[y|\mathbf{x}_i\right] = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_i$$

Llamamos p_i a la probabilidad de que y tome el valor 1 cuando $\mathbf{x} = \mathbf{x}_i$

$$p_i = P\left(y = 1|\mathbf{x}_i\right)$$

y la esperanza de y es:

$$E\left[y|\mathbf{x}_i\right] = P\left(y = 1|\mathbf{x}_i\right) \cdot 1 + P\left(y = 0|\mathbf{x}_i\right) \cdot 0 = p_i$$

por tanto,

$$p_i = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_i$$

que es una expresión equivalente del modelo. En consecuencia, la predicción $\widehat{y}_i$ estima la probabilidad de que un individuo con características definidas por $\mathbf{x} = \mathbf{x}_i$ pertenezca a la población correspondiente a $y = 1$.

El inconveniente principal de esta formulación es que p_i debe estar entre cero y uno, y no hay ninguna garantía de que la predicción, $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1'\mathbf{x}_i$ verifique esta restricción, ya que el modelo puede prever probabilidades mayores que la unidad. Esto no es un problema insalvable para clasificar, pero lo es si queremos interpretar el resultado de la regla de clasificación como una probabilidad de pertenencia a cada población.

A pesar de estos inconvenientes, este modelo simple conduce a una buena regla de clasificación, ya que según la interpretación de Fisher, maximiza la separación entre los grupos, sea cual sea la distribución de los datos. Sin embargo, cuando los datos no son normales, o no tienen la misma matriz de covarianzas, la clasificación mediante una ecuación

de relación lineal no es necesariamente óptima, y el modelo logístico puede conducir a mejores resultados.

El modelo logístico (*Logit*)

Si queremos que el modelo proporcione directamente la probabilidad de pertenecer a cada uno de los grupos, debemos transformar la variable respuesta de algún modo para garantizar que la respuesta prevista esté entre cero y uno. Si tomamos,

$$p_i = F\left(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i\right),$$

garantizaremos que p_i esté entre cero y uno si exigimos que F tenga esa propiedad.

La clase de funciones no decrecientes, acotadas entre cero y uno, es la clase de las funciones de distribución, por lo que el problema se resuelve tomando como F cualquier función de distribución.

Habitualmente se toma como F la función de distribución logística, dada por:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i)}}.$$

Esta función tiene la ventaja de ser continua. Además, como,

$$1 - p_i = \frac{e^{-(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i)}}{1 + e^{-(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i)}} = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i)}}$$

resulta que

$$g_i = \log \frac{p_i}{1 - p_i} = \log \left(\frac{\frac{1}{1+e^{-\left(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i\right)}}}{\frac{e^{-\left(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i\right)}}{1+e^{-\left(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i\right)}}} \right) = \log \left(\frac{1}{e^{-(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i)}} \right) = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i. \qquad (2)$$

de modo que, al hacer la transformación, se tiene un modelo lineal que se denomina *logit*.

La variable g representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones y, al ser una función lineal de las variables explicativas, facilita la estimación y la interpretación del modelo.

Una ventaja adicional del modelo logit es que si las variables son normales verifican el modelo logit y, además, también es cierto para una amplia gama de situaciones distintas a la normal.

En efecto, si las variables son normales multivariantes

$$g_i = \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)' \mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)' \mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$$

simplificando,

$$g_i = \frac{1}{2}\left(\boldsymbol{\mu}_2 \mathbf{V}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \mathbf{V}^{-1} \boldsymbol{\mu}_1\right) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{V}^{-1} \mathbf{x}.$$

Por tanto, g_i es una función lineal de las variables $\mathbf{x}$. Comparando con (2) la ordenada en el origen, β_0, es igual a

$$\beta_0 = -\frac{1}{2}\boldsymbol{\omega}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

donde $\boldsymbol{\omega} = \mathbf{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, y el vector de pendientes es $\boldsymbol{\beta}_1 = \boldsymbol{\omega}$.

Aunque se puede demostrar que la estimación de $\widehat{\boldsymbol{\omega}}$ mediante el modelo logístico no es eficiente en el caso normal, dicho modelo puede ser más eficaz cuando los poblaciones no tienen la misma matriz de covarianzas, o son claramente no normales.

Interpretación del Modelo Logístico

Los parámetros del modelo son: β_0, la ordenada en el origen, y $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_k)$. A veces, se utilizan también como parámetros $\exp(\beta_0)$ y $\exp(\beta_i)$, que se denominan los *odds ratios* o *ratios de probabilidades*. Estos valores indican cuánto se modifican las probabilidades por unidad de cambio en las variables x. En efecto, de (2) se deduce que

$$O_i = \frac{p_i}{1-p_i} = \exp(\beta_0).\prod_{j=1}^{k} \exp(\beta_j)^{x_j}.$$

Supongamos que consideramos dos elementos que tienen valores iguales en todas las variables menos en una. Sean $(x_{i1}, ..., x_{ih}, ..., x_{ik})$ los valores de las variables para el primer elemento y $(x_{j1}, ..., x_{jh}, ..., x_{jk})$ para el segundo, y todas las variables son las mismas en

ambos elementos menos en la variable h donde $x_{ih} = x_{jh} + 1$. Entonces, el odds ratio para estas dos observaciones es:

$$\frac{O_i}{O_j} = e^{\beta_h}$$

e indica cuánto se modifica el ratio de probabilidades cuando la variable x_j aumenta en una unidad.

Si consideramos $p_i = 0{,}5$ en el modelo logit, entonces

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} = 0$$

es decir,

$$x_{1i} = -\frac{\beta_0}{\beta_1} - \sum_{j=2}^{k} \frac{\beta_j x_{ji}}{\beta_1}$$

y x_{1i} representa el valor de x_1 que hace igualmente probable que un elemento cuyas restantes variables son $x_{2i}, ..., x_{ki}$, pertenezca a la primera o la segunda población.

Diagnosis

Los residuos del modelo (que a veces se denominan residuos de Pearson) se definen como,

$$e_i = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1 - \widehat{p}_i)}}$$

y, si el modelo es correcto, serán variables de media cero y varianza unidad que pueden servir para hacer la diagnosis de dicho modelo. El estadístico $\chi_0^2 = \sum_i e_i^2$ permite realizar un contraste global de la bondad del ajuste. Se distribuye asintóticamente como una χ^2 con $(n - k - 1)$ grados de libertad, donde $k + 1$ es el número de parámetros en el modelo.

En lugar de los residuos de Pearson se pueden utilizar, también, las desviaciones o *pseudoresiduos* definidos por

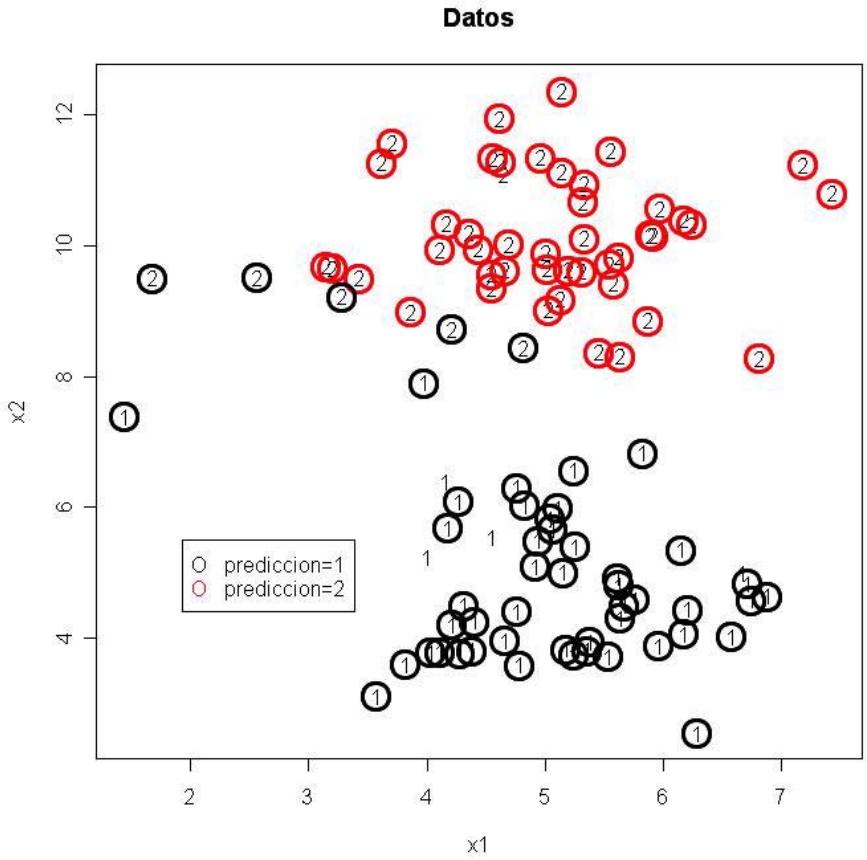
$$d_i = -2(y_i \log \widehat{p}_i + (1 - y_i) \log(1 - \widehat{p}_i))$$

Comparación entre el modelo logístico y el análisis discriminante lineal

Efron (1975) demostró que cuando los datos son normales multivariantes y se estiman los parámetros en la muestra, la función de discriminante lineal de Fisher funciona mejor que la regresión logística. Por ejemplo, en el campo de la concesión de créditos existen numerosos estudios comparando ambos métodos. La conclusión general es que ninguno de los dos métodos supera al otro de manera uniforme y que depende de la base de datos utilizada.

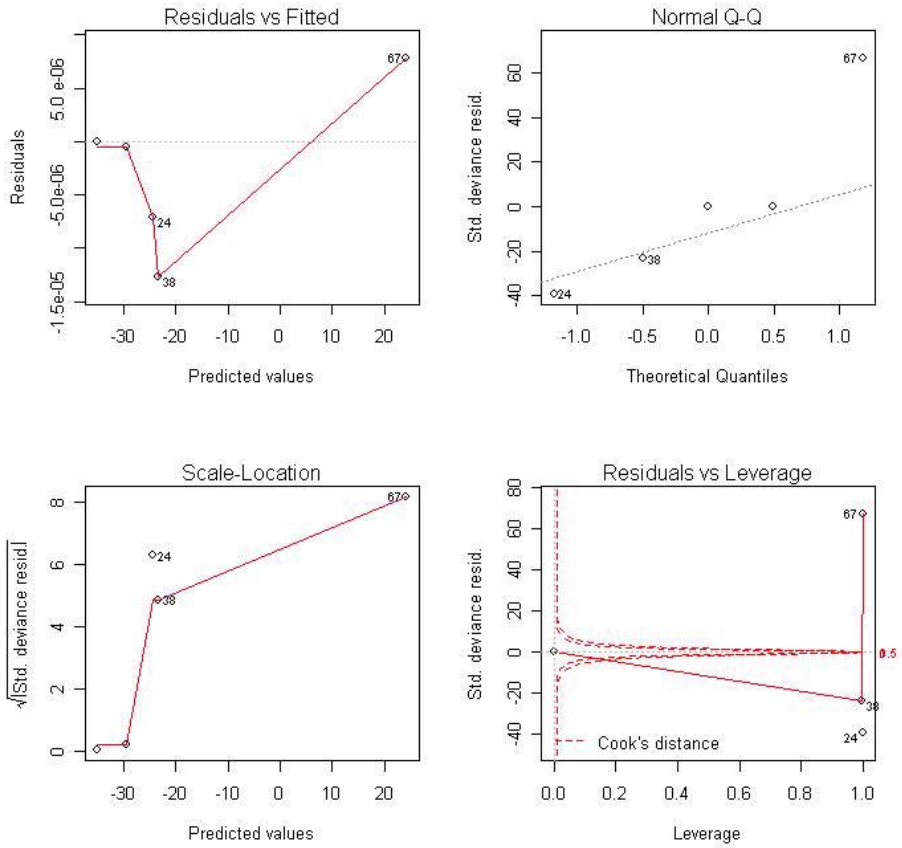
Ejemplo

Aplicamos la regresión logística a los datos usados en el ejemplo de Análisis Discriminante suponiendo que se han observado solamente las poblaciones 1 y 2. Seguimos utilizando un entrenamiento con 5 observaciones. El resultado de la clasificación esta resumido en la siguiente figura.



Se puede ver que sólo la variable x_2 tiene poder discriminante (p-*valor* del test $\chi^2_1 =$ 0,02) mientras que la variable x_1 no tiene ningún poder discriminante y puede ser eliminada del conjunto de variables explicativas.

Para chequear el modelo hay que mirar los residuos en la siguiente figura.



Aparece un problema ya que con sólo 5 observaciones en la muestra de entrenamiento es muy difícil chequear el modelo.A continuación se muestra código R utilizado en este ejemplo.

```
# Ejemplo de Regresion Logistica


rm(list=ls(all=TRUE))
# Inicializo el generador de numeros aleatorios para obtener
# siempre los mismos datos y la misma muestra de entrenamiento
set.seed(666)


# Matriz de datos
```

```
x1 <- c(rnorm(100,mean=5),rnorm(50,mean=10))
x2 <- c(rnorm(50,mean=5),rnorm(50,mean=10),rnorm(50,mean=10))

# Genero 50 repeticiones de los valores 1, 2 y 3
Ig <- gl(3,50)

# Selecciono los valores de x1 y x2 distintos del grupo 3
x1 <- x1[Ig != "3"]
x2 <- x2[Ig != "3"]
Ig <- as.factor(as.numeric(Ig[Ig!="3"]))

# Dibujo los datos
plot(x1, x2, pch=as.character(Ig), main="Datos")
losdatos <- data.frame(x1,x2,Ig)

# Elijo la muestra de entrenamiento
train <- sample(1:100,5)
# Cuento el numero de elementos de cada clase
table(losdatos$Ig[train])

# Calculo los 3 parametros de la funcion de regresion logistica:
# (intercept, b1 para x1 y b2 para x2)
z <- glm(Ig ~ ., data=losdatos, subset=train,
family=binomial(link="logit"))
summary(z)

# Esta funcion elimina cada variable (x1, x2) a la vez y hace el test Chi2
```

```
# Mira help(drop1) para mas detalles
drop1(z, test="Chisq")

# Hago predicciones
# Por defecto, al redondear a un solo valor se tiene 0 y 1, por lo que
# se suma 1 a todos para tener los grupos 1 y 2.
prediccion <- 1+round(predict(z, newdata=losdatos[-train,],
type="response"), 0)

# Dibujo de las predicciones
points(x=losdatos[-train,1], y=losdatos[-train,2], pch="O", cex=2,
col=as.numeric(prediccion))
a <- locator(1)
legend(a, c("prediccion=1", "prediccion=2"), col=1:2, pch="O")

# Chequeo el modelo analizando los residuos
windows()
par(mfrow=c(2,2))
plot(z)
```

Ejemplo

Consideramos los datos del fichero `womensrole` con dos variables explicativas: `sex` y `education`. Se aplica una regresión logística:

```
data("womensrole", package="HSAUR")
womensrole
fm1 <- cbind(agree,disagree) ~ sex+education
glm_1 <- glm(fm1, data=womensrole, family = binomial())
```

```
summary(glm_1)
```

Se obtiene que `education` tiene importancia y `sex` no la tiene.

Se considera una función para dibujar la variable `education` frente a las predicciones

```
predis1 <- predict(glm_1, type="response")


elgrafico <- function(predis) {
f <- womensrole$sex == "Female"
plot(womensrole$education, predis, type = "n",
ylab = "Probabilidad de coincidencia",
xlab = "Educaci\'{o}n", ylim = c(0,1))
lines(womensrole$education[!f], predis[!f], lty=1, col="red")
lines(womensrole$education[f], predis[f], lty=2, col="blue")
lgtxt <- c("Predicci\'{o}n (Hombres)", "Predicci\'{o}n (Mujeres)")
legend("topright", lgtxt, lty = 1:2, bty = "n",col=c('red','blue'))
y <- womensrole$agree / (womensrole$agree +
womensrole$disagree)
text(womensrole$education, y, ifelse(f, "\\VE", "\\MA"),
family = "HersheySerif", cex = 1.25,col=c('red','blue'))
}


elgrafico(predis1)
```

Ambas curvas, para mujeres y hombres equivalen aunque sugieren la existencia de interacción entre `sex` y `education`.

Se considera otro modelo:

```
fm2 <- cbind(agree,disagree) ~ sex*education
```

```
glm_2 <- glm(fm2, data = womensrole,
family = binomial())
summary(glm_2)
```

Alternativa: Se puede usar también la librería `RWeka`:

```
library(RWeka)
# Usamos los datos de un ejemplo: infert
# Infertility after Spontaneous and Induced Abortion
help(infert)

status <- factor(infert$case, labels=c("control", "caso"))
m <- Logistic(status ~ spontaneous + induced, data=infert)
m

# Evaluacion del clasificador:
e <- evaluate_Weka_classifier(m)
e
e$details

# Comparo el resultado con la orden glm de R:
glm(status ~ spontaneous + induced, data=infert, family=binomial())
```

Otra manera más general de hacerlo:

```
library(RWeka)

list_Weka_interfaces()
NB <- make_Weka_classifier("weka/classifiers/functions/Logistic")
```

```
NB
WOW(NB)


status <- factor(infert$case, labels = c("control", "case"))
modelo <- NB(status ~ spontaneous + induced, data = infert)
print(modelo)
```

Ampliaciones y Complementos

Para trabajar con regresión logística en `Tanagra`, basta seguir este tutorial:

`http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Variable_Selection_Binary_Logistic_Regression.pdf`

Para generalizar la regresión logística considerando más de 2 clases, se puede seguir los tutoriales:

`http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Multinomial_Logistic_Regression.pdf`

`http://en.wikipedia.org/wiki/Multinomial_logit`

y

`http://www.stat.psu.edu/~jglenn/stat504/08_multilog/01_multilog_intro.htm`