Regresión logística.

Modelo lineal para predecir el comportamiento de una variable cualitativa con dos categorías de respuesta. Presentación de la asignatura. Páginas 68 a 81.

Fichero de datos para trabajar: kyphosis de la librería rpart.

kyphosis: Data on Children who have had Corrective Spinal Surgery

Description: The kyphosis data frame has 81 rows and 4 columns. representing data on children who have had corrective spinal surgery

Format: This data frame contains the following columns:

- Kyphosis: a factor with levels absent present indicating if a kyphosis (a type of deformation) was present after the operation.
- Age: in months
- Number: the number of vertebrae involved
- Start: the number of the first (topmost) vertebra operated on.

Objective:To predict the eficcacy of the surgery taking into acccount Age, Number, Start

Source: John M. Chambers and Trevor J. Hastie eds. (1992) Statistical Models in S, Wadsworth and Brooks/Cole, Pacific Grove, CA.

Para cargar los datos del ejemplo:
library(rpart)
head(kyphosis) ## muestra los primeros casos del fichero
summary(kyphosis) ## describe las variables del fichero
Etapas del análisis:

Describir las variables
Analizar las relaciones entre las variables Age, Number, Start and Kyphosis
Utilizar la regresión logística para predecir Kyphosis
Comentar los resultados obtenidos

Descripción de las variables

K <- kyphosis

```
> summary(K)
    Kyphosis       Age              Number           Start
 absent :64   Min.   :  1.00   Min.   : 2.000   Min.   : 1.00
 present:17   1st Qu.: 26.00   1st Qu.: 3.000   1st Qu.: 9.00
              Median : 87.00   Median : 4.000   Median :13.00
              Mean   : 83.65   Mean   : 4.049   Mean   :11.49
              3rd Qu.:130.00   3rd Qu.: 5.000   3rd Qu.:16.00
              Max.   :206.00   Max.   :10.000   Max.   :18.00
```

Relación de las variables con Kyphosis

numSummary(K[,2:4], groups=K$Kyphosis) las describe todas a la vez

Igualdad de distribuciones de Age, Number y Start según la variable Kyphosis

Se usará el test no paramétrico de Wilcoxon para dos muestras independientes

- **Age**

```
> numSummary(K$Age, groups=K$Kyphosis)

          Statistic
Group       mean     sd IQR 0% 25% 50% 75% 100%
  absent  79.891 61.861 113  1  18  79 131  206
  present 97.824 39.275  55 15  73 105 128  157
> wilcox.test(Age ~ Kyphosis, alternative="two.sided", data=K)
     Wilcoxon rank sum test with continuity correction
data:  Age by Kyphosis
W = 446.5, p-value = 0.2605
alternative hypothesis: true location shift is not equal to 0
```

No hay evidencias suficientes para considerar que Age esté relacionadas con la presencia o ausencia de la enfermedad.

- **Number**

```
> numSummary(K$Number, groups=K$Kyphosis)

          Statistic
Group      mean    sd IQR 0% 25% 50% 75% 100%
  absent  3.750 1.414   2  2   3   4   5    9
  present 5.176 1.879   2  3   4   5   6   10

> wilcox.test(Number ~ Kyphosis, alternative="two.sided", data=K)
     Wilcoxon rank sum test with continuity correction
data:  Number by Kyphosis
W = 289, p-value = 0.002504
alternative hypothesis: true location shift is not equal to 0
```

La variable Number parece asociada con Kyphosis ya que los niños en los que no se eliminó la enfermedad tienden a tener un número mayor de vértebras afectadas

- **Start**

```
> numSummary(K$Start, groups=K$Kyphosis)

          Statistic
Group       mean    sd IQR 0% 25% 50% 75% 100%
  absent  12.609 4.428   5  1  11  14  16   18
  present  7.294 4.283   7  1   5   6  12   14

> wilcox.test(Start ~ Kyphosis, alternative="two.sided", data=K)
     Wilcoxon rank sum test with continuity correction
data:  Start by Kyphosis
W = 896, p-value = 0.00004023
alternative hypothesis: true location shift is not equal to 0
```

La variable Start parece asociada con Kyphosis ya que los niños en los que se eliminó la enfermedad tienden a tener la primera vertebra afectada más baja que aquellos con la enfermedad presente.

Modelo de regresión logística

R commander: Estadísticos + Ajuste de modelos + Modelo lineal generalizado

```
> GLM.1<- glm(Kyphosis ~ Age+Number+Start,family=binomial(logit),data=K)

> summary(GLM.1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3124  -0.5484  -0.3632  -0.1659   2.1613

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.036934   1.449575  -1.405  0.15996
Age          0.010930   0.006446   1.696  0.08996 .
Number       0.410601   0.224861   1.826  0.06785 .
Start       -0.206510   0.067699  -3.050  0.00229 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234  on 80  degrees of freedom
Residual deviance: 61.380  on 77  degrees of freedom
AIC: 69.38

Number of Fisher Scoring iterations: 5

> exp(coef(GLM.1))  # Exponentiated coefficients ("odds ratios")
(Intercept)         Age      Number       Start
  0.1304281   1.0109904   1.5077239   0.8134181
```

La variable Start tiene un coeficiente claramente distinto de cero porque el p-valor 0.00229 es mucho más pequeño que los niveles de significación habituales (0,05 y 0.01). Además, como el coeficiente es negativo (el odds ratio es menor que 1) al aumentar el valor de esta variables se reduce la probabilidad de la categoría present.

Las otras dos variables tienen unos p-valores muy próximos a 0,05 y por tanto no se puede rechazar la hipótesis de que sean cero, pero parece razonable suponer que estén relacionados con el resultado del tratamiento.

Para decidir sobre esta cuestión se utilizara un procedimiento automático para ver que variables se incluyen en el modelo

```
Modelo + Selección de modelo paso a paso + adelante/atrás + AIC
> stepwise(GLM.1, direction='forward/backward', criterion='AIC')

Modelo recomendado: Incluye las tres variables
Coefficients:
(Intercept)        Start       Number          Age
   -2.03693     -0.20651      0.41060      0.01093

Modelo + Selección de modelo paso a paso + adelante/atrás + BIC
> stepwise(GLM.1, direction='forward/backward', criterion='BIC')
Modelo recomendado: Incluye una sola variable
Coefficients:
(Intercept)        Start
     0.8901      -0.2179
```

Estos resultados indican que no está nada claro que modelo elegir y que existen razones para explorar como funcionan ambos.

Modelo con las tres variables:

Vamos a tratar de cuantificar la calidad predictiva del modelo GLM.1 comparando las probabilidades estimadas en los grupos "present" y "absent". Estas probabilidades están entre los resultados almacenados en GLM.1 con el nombre de "fitted.values"

```
> names(GLM.1)
"coefficients" "residuals"    "fitted.values" "effects"           "R"
"rank"         "qr"           "family"        "linear.predictors" "deviance"
"aic"          "null.deviance" "iter"         "weights"          "prior.weights"
"df.residual"  "df.null"       "y"            "converged"         "boundary"
"model"        "call"          "formula"      "terms"             "data"
"offset"       "control"       "method"       "contrasts"         "xlevels"
```

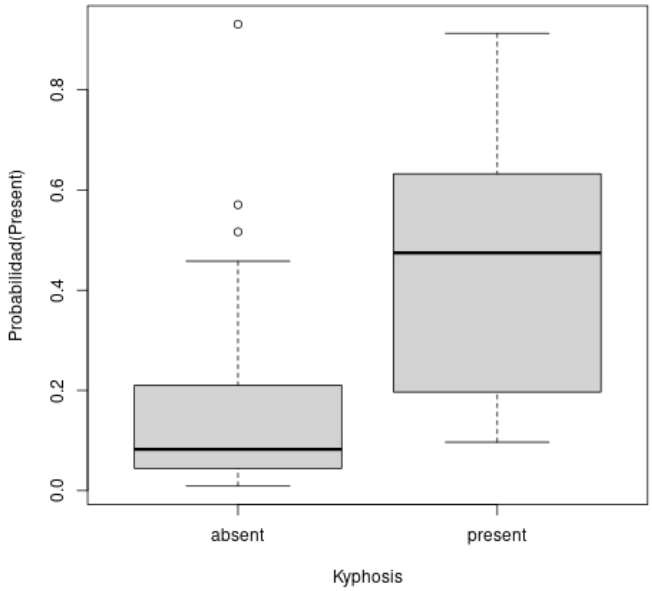
```
> numSummary(GLM.1$fitted.values, groups= K$Kyphosis)Statistic
Group       mean     sd    IQR     0%    25%    50%    75%   100%
  absent  0.1511 0.1698 0.1640 0.0090 0.0448 0.0823 0.2088 0.9310
  present 0.4313 0.2423 0.4355 0.0967 0.1968 0.4745 0.6322 0.9129
```

Estas probabilidades son, en general, bastante más pequeñas en el grupo "absent" que en "present". Esto se aprecia con claridad en la descripción de ambos grupos, ya que el valor 0.20 corresponde al percentil 75 de "absent" y sólo el 25 de "present".

A continuación se representan las probabilidades de "Present" en los dos grupos usando una gráfica de cajas y salvando la imagen con el nombre de "cajas.png"

```
> png(file = "/home/Bilbao_2022-2023/cajas.png", bg =
"transparent")
> boxplot(GLM.1$fitted.values ~ K$Kyphosis, xlab="Kyphosis",
ylab="Probabilidad(Present)" )
> dev.off()
```

A partir de esta gráfica se pueden cponsiderar, por ejemplo, dos puntos de referencia 0.20 y 0.60 para estudiar como se relacionan estas probabilidades con el resultado del tratamiento.



Ahora se calcula la matriz de confusión para describir la proporción de "Present" y "Absent" en cada uno de los tres grupos de probabilidades considerados.

```
> T<- table(K$Kyphosis,cut(GLM.1$fitted.values, c(0,0.20,0.60,1)))
> T
         (0,0.2] (0.2,0.6] (0.6,1]
  absent       47        16       1
  present       5         6       6

> round( 100 *prop.table(T,2),1)
         (0,0.2] (0.2,0.6] (0.6,1]
  absent     90.4      72.7    14.3
  present     9.6      27.3    85.7
```

Estos resultados permiten obtener algunas conclusiones interesantes:

- Cuando la probabilidad estimada es menor o igual a 0.20 se consigue eliminar el problema en el 90,4 % de los niños.
- Si la probabilidad esta entre (0.2,0.6] el porcentaje de éxito se reduce al 72,7%
- Para valores superiores a 0,6 más del 85% de los tratamientos resultaron ineficaces

Este procedimiento tiene el inconveniente de que los datos utilizados para estimar los parámetros del modelo se emplean también para evaluar su "calidad". Por ello es conveniente utilizar unos datos para calcular el modelo y otros independiente para estimar las probabilidades de acierto y error. Esto se suele hacer eligiendo al azar una proporción de s observaciones para "entrenar", por ejemplo el 75%, y el resto para "evaluar.

```
> n <- dim(K)[1]; n ## número de datos del fichero
[1] 81
> estima <- sample(1:n, round(n*0.75), replace=FALSE) ## muestra de
estimación
> valida <- setdiff(1:n, estima) ## quita de los n datos los que están en
"estima"
> length(valida) ## número de casos empleados para validar
[1] 20
```

Ahora se repite el procedimiento anterior teniendo en cuenta la división de la muestra

```
> GLM.2<- glm(Kyphosis ~
Age+Number+Start,family=binomial(logit),data=K, subset=estima)
> summary(GLM.2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.12968    1.58560  -1.343   0.1792
Age          0.01297    0.00791   1.639   0.1011
Number       0.36100    0.24096   1.498   0.1341
Start       -0.20398    0.08331  -2.448   0.0143 *

> T2 <- table(K$Kyphosis[valida], cut(GLM.2$fitted.values[valida],
c(0, 0.20, 0.60, 1)))
> T2
          (0,0.2] (0.2,0.6] (0.6,1]
  absent        5         5       1
  present       3         1       1

> round( 100 *prop.table(T2,2),1)

          (0,0.2] (0.2,0.6] (0.6,1]
  absent     62.5      83.3    50.0
  present    37.5      16.7    50.0
```

Se aprecia que la proporción de aciertos tiende a disminuir al emplear diferente conjunto de datos para entrenar y validar.

Como el conjunto de datos es pequeño, las estimaciones de las proporciones de aciertos y fallos pueden cambiar mucho debido a la elección aleatoria de ambas muestras y por tanto es conveniente repetir varias veces el mismo proceso y comparar las proporciones de aciertos y fallos de cada iteración.

Otro aspecto muy importante es analizar si las estimaciones de los parámetros son "estables". En el caso en que el modelo cambie mucho, de unas veces a otras ,se plantea una dificultad muy importante porque la interpretación también se modifica y eso afecta a la credibilidad de los

resultados. En esta situación se debería emplear otra técnica estadística para intentar evitar esa inestabilidad.

```
> ## nueva iteración
> estima <- sample(1:n, round(n*0.75), replace=FALSE)
> valida <- setdiff(1:n, estima)
> GLM.3<- glm(Kyphosis ~
Age+Number+Start,family=binomial(logit),data=K, subset=estima)
> summary(GLM.3)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.21024    2.61898  -1.989  0.04666 *
Age          0.02293    0.01043   2.198  0.02795 *
Number       0.92118    0.43167   2.134  0.03284 *
Start       -0.27504    0.10211  -2.694  0.00707 **

> T3 <- table(K$Kyphosis[valida], cut(GLM.2$fitted.values[valida],
c(0, 0.20, 0.60, 1)))
> T3
         (0,0.2] (0.2,0.6] (0.6,1]
  absent        8         1       0
  present       1         6       0

> round( 100 *prop.table(T3,2),1)
         (0,0.2] (0.2,0.6] (0.6,1]
  absent     88.9      14.3
  present    11.1      85.7

> ## nueva iteración
> estima <- sample(1:n, round(n*0.75), replace=FALSE)
> valida <- setdiff(1:n, estima)
> GLM.4<- glm(Kyphosis ~
Age+Number+Start,family=binomial(logit),data=K, subset=estima)
> summary(GLM.4)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.914953   1.738743  -1.101   0.2707
Age          0.009551   0.006912   1.382   0.1670
Number       0.385440   0.269467   1.430   0.1526
Start       -0.190208   0.074996  -2.536   0.0112 *

> T4 <- table(K$Kyphosis[valida], cut(GLM.2$fitted.values[valida],
c(0, 0.20, 0.60, 1)))
> T4
         (0,0.2] (0.2,0.6] (0.6,1]
  absent        7         5       1
  present       1         2       0
> round( 100 *prop.table(T4,2),1)
         (0,0.2] (0.2,0.6] (0.6,1]
  absent     87.5      71.4   100.0
  present    12.5      28.6     0.0
```

Comenta los resultados obtenidos y los puntos que consideres "débiles" del análisis.

Analiza lo que ocurre con el modelo en el que interviene solamente la variable "Start"