Regresión logística

Ejemplo: Una variable explicativa binaria

Datos:

Tabla de parámetetro desestabilizado * Exposición al agente

Recuento

		Exposición al agente		
		no	si	Total
parámetetro desestabilizado	no	74	12	86
	si	5	4	9
Total		79	16	95

p = probabilidad de que el trabajador presente cierto parámetro desestabilizado

Exposición al agente = variable explicativa binaria

Regresión logística

Ejemplo: Una variable explicativa binaria

Datos:

```
> a=read.table("ejemplo1.DAT",header=T)
> a
  respuest exposici frecue
1        1        1      4
2        1        0      5
3        0        1     12
4        0        0     74
```

Se estructura la respuesta en una matriz formada por la columna de éxitos y la de fracasos

```
> ma=matrix(a$frecue,ncol=2)# 2 columnas: exitos y fracasos
> ma
     [,1] [,2]
[1,]    4   12
[2,]    5   74
```

Declare categórica la variable binaria exposición

> expos=factor(c("si","no"))

Regresión logística

Función de ajuste del modelo

```
> glm(ma~expos,family = binomial)
Call:  glm(formula = ma ~ expos, family = binomial)
Coefficients:
(Intercept)      expossi
     -2.695        1.596
Degrees of Freedom: 1 Total (i.e. Null);  0 Residual
Null Deviance:      4.267
Residual Deviance: 1.021e-14    AIC: 10.4
```

Coeficientes del modelo

Contraste de hipótesis

Función de resumen del análisis

```
> logit=glm(ma~expos,family = binomial)
> summary(logit)
Call:
glm(formula = ma ~ expos, family = binomial)
Deviance Residuals:
[1]  0  0
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6946     0.4621  -5.832 5.49e-09 ***
expossi       1.5960     0.7395   2.158   0.0309 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 4.2674e+00  on 1  degrees of freedom
Residual deviance: 1.0214e-14  on 0  degrees of freedom
AIC: 10.397
Number of Fisher Scoring iterations: 3
```

P-valor asociado

Coeficientes estimados

P-valor=0.0309<0.05 Se rechaza la nulidad del coeficiente exposi

Otro modo de testar mediante test chi-cuadrado la exposición

```
anova(logit,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: ma
Terms added sequentially (first to last)
      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      1     4.2674
expos  1   4.2674          0  1.021e-14    0.0389
```

Tras la introducción de la variable binaria Exposici el estadístico –2log de la verisimititud se reduce 4.2674

El estadístico –2log de la verosimilitud es salvo cte la deviance para el modelo

La significatividad de la variable exposici viene dada por El estadístico con valor 4,267 que sigue un modelo chi-cuadrado con 1 g.l.

El valor **4.2674 con 1 g.l. es significativo (p.valor=0.0389)**
Se concluye que expos sirve para explicar el modelo

Modelo propuesto: $$\log it = \ln \frac{p}{1-p} = \beta_0 + \beta_1 \text{exposici}$$

Modelo ajustado: $$\log it = \ln \frac{p}{1-p} = -2{,}694 + 1{,}596 \text{exposici}$$

Interpretación de los coeficientes:

1,596 es el cambio esperado en el logit al pasar de un trabajador No expuesto al agente (0) a uno expuesto (1).

ln (RO exp/no exp) = logit(expuesto) – logit (no expuesto) =1,596

RO exp/no exp = exp(1,596) = 4,931

El riesgo de parámetro desestabilizado es casi 5 veces mayor en los trabajadores expuestos que en los no expuestos

Ejemplo2: Una variable explicativa continua

La tabla siguiente clasifica a un grupo de 298 trabajadores. Muestra el nivel o grado de exposición a un agente ambiental (medido por la concentración de determinado factor en la planta química en que desarrollan su trabajo) . Tras cierto periodo de tiempo se ha examinado a cada trabajador para establecer si presenta o no síntomas de alergia.

Datos:

Tabla de contingencia Alergia * Grado o nivel de exposición

Recuento

		Grado o nivel de exposición						
		1,08	1,16	1,21	1,26	1,31	1,35	Total
Alergia	no	35	25	24	26	21	20	151
	si	15	24	26	24	29	29	147
Total		50	49	50	50	50	49	298

p = probabilidad de presentar síntomas de alergia

Exposi=grado de exposición al algente ambiental

Modelo propuesto: $$\log it = \ln \frac{p}{1-p} = \beta_0 + \beta_1 \text{exposi}$$

Regresión logística binaria

Datos:

```
> a=read.table("ejemplo2.DAT",header=T,dec=",")
> a
   frec exposi resp
1    15   1.08    1
2    24   1.16    1
3    26   1.21    1
4    24   1.26    1
5    29   1.31    1
6    29   1.35    1
7    35   1.08    0
8    25   1.16    0
9    24   1.21    0
10   26   1.26    0
11   21   1.31    0
12   20   1.35    0
> ma=matrix(a$frec,ncol=2)# 2 columnas: exitos y fracasos
> ma
     [,1] [,2]
[1,]   15   35
[2,]   24   25
[3,]   26   24
[4,]   24   26
[5,]   29   21
[6,]   29   20
```

Se prepara la matriz de exitos y fracasos denominada ma

Resumen del análisis: Contrastes individuales

summary:

```
> summary(glm(ma~exposi,family = binomial))
Call:
glm(formula = ma ~ exposi, family = binomial)
Deviance Residuals:
      1        2        3        4        5        6
-0.7958   0.8839   0.6340  -0.6198   0.1146  -0.2544
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.810      1.621  -2.967  0.00301 **
exposi         3.893      1.315   2.960  0.00308 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 11.3030  on 5  degrees of freedom
Residual deviance:  2.2786  on 4  degrees of freedom
AIC: 32.245
Number of Fisher Scoring iterations: 3
```

Dado el bajo valor de la significatividad (0,003), se rechaza la hipótesis de que dicho parámetro sea cero. Exposición afecta a probabilidad del suceso, por tanto sirve para explicar .

La prueba que permite contrastar la bondad del ajuste, frente a la alternativa de que el modelo no se ajusta, puede aproximarse mediante el estadístico de valor **2.2786 con 4 g.l**

No significativo, por lo que se acepta la bondad del ajuste

Contrastes chi-cuadrado

anova:

```
> anova(glm(ma~exposi,family = binomial),test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: ma
Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       5    11.3030
exposi  1   9.0245         4     2.2786    0.0027
```

La variable exposi (exposición) permite reducir la deviance residual en 9.0245, con 1 grado de libertad, lo que supone una significativa reducción, con p-valor igual a 0.0027 < 0.05.

Se concluye que el término es importante en el modelo

La variable introducida, exposi, sirve para explicar la probabilidad de presentar alergia

Modelo ajustado:

$$\log it = \ln \frac{p}{1-p} = -4{,}810 + 3{,}893\text{exposi}$$

Interpretación del parámetro estimado:

B=3,893 indica el cambio esperado en el logit (logaritmo de la odds o ventaja) al incrementar una unidad el nivel de exposición.

El valor exp(b) = exp(3,893) = 49,051 es la RO que compara las odds aumentando una unidad en la exposici

El cambio esperado en el logit al aumentar 0,1 unidades el nivel de exposición viene dado por:

$$\text{logit}(\text{exposi} + 0{,}1) - \text{logit}(\text{exposi}) = 3{,}893 * 0{,}1 = 0{,}3893$$

La RO que compara las odds aumentando 0,1 unidades en la exposici es igual a exp(0,3893) = 1,48

El riesgo de tener síntomas de alergia es aproximadamente 1,5 veces mayor al incrementar 0,1 unidades el grado o nivel de exposición

Ejemplo3: Dos variables explicativas (una binaria y otra continua)

La tabla siguiente muestra la clasificación de un grupo de trabajadores sometidos a diferentes niveles de exposición de un factor ambiental, el tiempo bajo dicha exposición (16 meses y 24 meses) y los resultados obtenidos al final del estudio según presente o no síntomas de afección respiratoria.

DATOS

Tabla de contingencia Síntomas respiratorios * Nivel de exposición * TIEMPO

Recuento

TIEMPO			Nivel de exposición ,00	,45	,75
16 meses	Síntomas respiratorios	no	204	301	186
		si	1	3	7
24 meses	Síntomas respiratorios	no	742	790	469
		si	20	98	118

p = probabilidad de afección respiratoria

Tiempo = variable binaria (0=16 meses y 1= 24 meses)

Nivel de exposición = variable continua

Regresión logística

Lectura de datos

Datos

```
> #No olvide poner como separador decimal la coma
> a=read.table("ejemplo3.DAT",header=T,dec=",")
> a
   nivelc tiempo resp frec nivel2
1    0.00      0    1    1 0.0000
2    0.45      0    1    3 0.2025
3    0.75      0    1    7 0.5625
4    0.00      0    0  204 0.0000
5    0.45      0    0  301 0.2025
6    0.75      0    0  186 0.5625
7    0.00      1    1   20 0.0000
8    0.45      1    1   98 0.2025
9    0.75      1    1  118 0.5625
10   0.00      1    0  742 0.0000
11   0.45      1    0  790 0.2025
12   0.75      1    0  469 0.5625
```

La variable nivel2 contiene los valores de nivelc al cuadrado

Regresión logística

Preparación de los datos para el análisis

Datos

```
> #Generamos un data frame con los éxitos y las variables independientes
> ma=cbind(a1$frec,frac)# 2 columnas:total de exitos y fracasos
> ma
      frac
[1,]   1  204
[2,]   3  301
[3,]   7  186
[4,]  20  742
[5,]  98  790
[6,] 118  469
```

```
> #declaramos la variable tiempo como factor
> tiempo=factor(a1$tiempo,labels=c("16m","24m"))
```

Modelo propuesto: $$\log it = \ln \frac{p}{1-p} = \beta_0 + \beta_1 \text{nivelc} + \beta_2 \text{tiempo}$$

```
> glm(ma~a1$nivelc+tiempo,family = binomial)
Call:  glm(formula = ma ~ a1$nivelc + tiempo, family = binomial)
Coefficients:
(Intercept)    a1$nivelc    tiempo24m
     -5.568        2.857        2.099
Degrees of Freedom: 5 Total (i.e. Null);  3 Residual
Null Deviance:      198.5
Residual Deviance: 2.035        AIC: 34.29
```

Aunque en esta tabla no se muestra ningún test, ya puede deducirse la importancia de los términos; mediante los estadísticos Null Deviance y Residual Deviance, la gran reducción que conlleva la introducción de los dos términos en el modelo (de 198.5 a 2.035) refleja esto. La residual deviance, si se aproxima a un modelo chi-cuadrado con 3 g.l., muestra también indicios de buen ajuste. Al menos una de las variables introducidas en el modelo es significativa.

Reducción de la deviance: 198.5 – 2.035=**196.4** con 5-3= **2 g.l**. es altamente significativo valor para una chi-cuadrado con 2 g.l.

Nuevo modelo propuesto más complejo (mod2)

$$\log it == \beta_0 + \beta_1 \text{nivelc} + \beta_2 \text{tiempo} + \beta_3 \text{nivelc}^2 + \beta_4 \text{nivelc} * \text{tiempo}$$

```
> mod1=glm(ma~a1$nivelc+tiempo,family = binomial)
> mod2=glm(ma~a1$nivelc+I(a1$nivelc^2)+tiempo*a1$nivelc,family = binomial)
#El término I(a1$nivelc^2) permite evaluar previamente la variable nivelc al cuadrado
> mod2
Call:  glm(formula = ma ~ a1$nivelc + I(a1$nivelc^2) + tiempo * a1$nivelc,family = binomial)
Coefficients:
        (Intercept)            a1$nivelc       I(a1$nivelc^2)
            -5.9757               4.3230              -1.1714
          tiempo24m  a1$nivelc:tiempo24m
             2.3865              -0.4885
Degrees of Freedom: 5 Total (i.e. Null);  1 Residual
Null Deviance:      198.5
Residual Deviance: 0.8537       AIC: 37.11
```

Aunque en esta tabla no se muestra ningún test, ya puede deducirse la importancia de los términos; mediante los estadísticos Null Deviance y Residual Deviance, la gran reducción que conlleva la introducción de los términos en el modelo (de 198.5 a 0.8537) refleja esto. La residual deviance, si se aproxima a un modelo chi-cuadrado con 1 g.l., muestra también indicios de buen ajuste. Al menos una de las variables introducidas en el modelo es significativa.

summary(mod2) $\log it = \beta_0 + \beta_1 \text{nivelc} + \beta_2 \text{tiempo} + \beta_3 \text{nivelc}^2 + \beta_4 \text{nivelc} * \text{tiempo}$

Permite contrastar la hipótesis de significatividad de los términos en el modelo.

```
> summary(mod2)
Call:
glm(formula = ma ~ a1$nivelc + I(a1$nivelc^2) + tiempo * a1$nivelc,
    family = binomial)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.9757     0.9942  -6.011 1.85e-09 ***
a1$nivelc            4.3230     1.8752   2.305   0.0211 *
I(a1$nivelc^2)      -1.1714     1.1125  -1.053   0.2924
tiempo24m            2.3865     0.9944   2.400   0.0164 *
a1$nivelc:tiempo24m -0.4885     1.5831  -0.309   0.7576
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 198.5347  on 5  degrees of freedom
Residual deviance:   0.8537  on 1  degrees of freedom
AIC: 37.11
Number of Fisher Scoring iterations: 5
```

El modelo se ajusta bien a los datos pero es innecesariamente complejo

En esta tabla se muestra test z, de donde puede deducirse la importancia o no de los términos; Los términos nivelc y tiempo son significativos. No lo son nivel^2 ni la interacción entre tiempo y nivel

anova(mod1,mod2,test="Chisq")

Permite contrastar la hipótesis de significatividad de los términos añadidos en el modelo.
Permite comparar modelos anidados

```
> anova(mod1,mod2,test="Chisq")  #compara modelos mod1 y mod2
Analysis of Deviance Table
Model 1: ma ~ a1$nivelc + tiempo
Model 2: ma ~ a1$nivelc + I(a1$nivelc^2) + tiempo * a1$nivelc
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3    2.03544
2         1    0.85371  2  1.18173   0.55385
```

Los resultados anova muestran que ninguno de los términos añadidos al modelo mod1 es importante para explicar la variable dependiente. Sólo se obtiene una reducción de la deviance igual 1.18173 (2.03544 - 0.85371) con 2 grados de libertad, que no es significativa (p-valor=0.55385).

Modelo ajustado

$$\log it = \ln \frac{p}{1-p} = -5{,}568 + 2{,}857\,nivelc + 2{,}099\,\text{tiempo}$$

```
> summary(mod1)
Call:
glm(formula = ma ~ a1$nivelc + tiempo, family = binomial)
Deviance Residuals:
      1       2       3       4       5       6
 0.2397  -0.5931  0.3702  -0.6515  0.8912  -0.5202
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.5683     0.3484 -15.983  < 2e-16 ***
a1$nivelc     2.8569     0.2893   9.875  < 2e-16 ***
tiempo24m     2.0989     0.3133   6.700 2.08e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 198.5347  on 5  degrees of freedom
Residual deviance:   2.0354  on 3  degrees of freedom
AIC: 34.292
Number of Fisher Scoring iterations: 4
```

El estadístico de z muestra que todos los coeficientes son significativos con valor **p-valor = 0,000**. Lo que permite rechazar las hipótesis de nulidad de los mismos.

Interpretación de los parámetros estimados

2,857 es el incremento esperado en el logit al aumentar una unidad la variable continua Nivelc, supuestos estables el resto de las variables

2,099 es el incremento esperado en el logit al pasar del tiempo 16 meses (código 0) de la variable binaria tiempo, al periodo de 24 meses (código 1), supuestos estables el resto de las variables

Equivalentemente, 17,408 es la RO que compara las Odds de trabajadores que tienen una unidad más en el nivel de exposición. El riesgo es 17 veces mayor al incrementar una unidad el nivel de exposición.

Del mismo modo, 8,157, indica que el riesgo de padecer síntomas de afección respiratoria Es aproximadamentes 8 veces mayor al pasar de 16 a 24 meses de exposición