3 Ejemplos simples sobre regresión logística

1. Ejemplo 1 (ejemplo1.DAT) tabla 2x2

Una variable explicativa binaria (Solución)

2. Ejemplo 2 (ejemplo2.DAT)

Una variable independiente continua (Solución)

3. Ejemplo 3 (ejemplo3.DAT)

Dos variables: una continua y otra dicotómica (Solución)

Ejemplos simples sobre regresión logística

En los 3 ejemplos que siguen, los datos están agrupados individuos con idénticos valores en sus variables independientes[1]:

Comenzamos con ajuste de un modelo con una sola variable independiente.

Ejemplo 1 (ejemplo1.DAT) tabla 2x2

Tras 85 semanas en estudio con un grupo de 95 trabajadores de los cuales 79 (controles) no estuvieron sometidos a un determinado agente, supuestamente nocivo, y 16 de ellos sí lo estuvieron, se realizó examen médico para detectar cuáles tenían cierto parámetro desestabilizado. Los resultados vienen en la tabla siguiente:

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

Ajuste de regresión logística para establecer si es significativo o no el factor exposición al agente.

Ejemplo 2 (ejemplo2.DAT) (variable independiente 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.

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

Análisis de regresión logística.

Ejemplo 3 (ejemplo3.DAT) (dos variables: una continua y otra dicotómica)
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.

[1] Nota: las observaciones Yi (total de éxitos en el grupo i-ésimo) son variables binomiales independientes B(ni, pi), donde las probabilidades son función de las variables independientes.

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

1. Se sospecha que el efecto del nivel de exposición es más alto cuanto más tiempo se lleva expuesto. Contraste la hipótesis de interacción tiempo*nivel de exposición para corroborar o no dicha sospecha.

Soluciones:

Ejemplo 1:

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

```
> 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
```

```
> ma=matrix(a$frecue,ncol=2)# 2 columnas: exitos y fracasos
> ma
     [,1] [,2]
[1,]    4   12
[2,]    5   74
> expos=factor(c("si","no"))
> 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
```

```
> 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
```

```
> anova(logit,test="Chisq")  Otro modo de testar mediante test chi-cuadrado exposición
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
```

La tabla muestra el mejoramiento del modelo al incluir la variable independiente. Al nivel de significación alfa del 5% es significativo(0,039 < 0,05). En este caso se reduce la deviance residual a cero (modelo saturado).

La tabla coeficientes muestra los parámetros del modelo estimado, así como su significación. El estadístico z nos permite contrastar la hipótesis de nulidad del parámetro correspondiente a la variable independiente exposición. Se observa que se rechaza la hipótesis al nivel alfa del 5%. Por lo que se admite que el factor influye en la variable dependiente.

Ecuación del modelo ajustado:

$\text{Log}^{2}(\text{odds}) = -2{,}694 + 1{,}596\ \text{Exposici}$

Interpretación del parámetro estimado:

B=1,596 indica el cambio esperado en el logit (logaritmo de la odds o ventaja) al pasar del grupo de no expuestos a la categoría de expuestos. El valor positivo del coeficiente indica un aumento del logit al pasar de la modalidad base (no expuestos) a la de expuestos.

O lo que es lo mismo, el valor $e^{1{,}596} = 4{,}931$ refleja la **razón de odds** que compara expuestos a no expuestos[3]. Lo que nos permite afirmar que el riesgo de sufrir desestabilización del parámetro estudiado es casi 5 veces mayor en los expuestos que en los no expuestos[4].

Ejemplo 2:

```
> #No olvide poner como separador decimal la coma, de lo contrario lee como factor, porque la varible exposi viene originalmente en el archivo ejemplo2.DAT, con el separador decimal coma.
> 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
```

[2] El logaritmo usado es el neperiano ln(Odds), aunque en R se note con log
[3] El valor exponenciado es más útil desde el punto de vista interpretativo.
[4] Observe que 4,93 es también la razón de odds calculada en la tabla 2x2: (4x74)/(5x12).

```
[6,]   29   20
> exposi=a[1:6,2]
> exposi
[1] 1.08 1.16 1.21 1.26 1.31 1.35

> glm(ma~exposi,family = binomial)

Call:  glm(formula = ma ~ exposi, family = binomial)

Coefficients:
(Intercept)       exposi
     -4.810        3.893

Degrees of Freedom: 5 Total (i.e. Null);  4 Residual
Null Deviance:      11.3
Residual Deviance: 2.279        AIC: 32.25



> 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
```

La deviance residual 2.2786 con 4 grados de libertad muestra indicios de un ajuste bueno. No obstante, la interpretación de estos estadísticos deberá hacerse con cautela. Cuando los datos no vienen agrupados, este resumen no es útil para interpretar la bondad de ajuste.

Otro modo de ver la significatividad del término añadido es mediante anova. En este caso usaremos el test chi-cuadrado:

```
> 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.

Obtenga la razón de Odds correspondiente a un aumento de una unidad en nivel de exposición.

```
> exp( 3.893)
[1] 49.05784
```

Obtenga la razón de Odds correspondiente a un aumento de una décima[5] en nivel de exposición.

```
> exp( 3.893*0.1)
[1] 1.475947
```

Ecuación del modelo ajustado

Log[6](odds) = -4,81 + 3,89 exposi

Ejemplo 3:

La variable tiempo es de tipo cualitativo en este ejemplo. Tiene dos modalidades 16 y 24 meses correspondientes a los niveles 0 y 1.
resp tiene niveles 0 y 1 correspondientes a no y si, respectivamente

```
> rm(list=ls(all=TRUE))
> #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
> #Generamos un data frame con los éxitos y las variables independientes
> a1=a[a$resp==1,] #exitos y variables independientes
> frac= a[a$resp==0,"frec"] #total fracasos sin variables independientes
> a1
  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
7   0.00      1    1   20 0.0000
8   0.45      1    1   98 0.2025
9   0.75      1    1  118 0.5625
> frac
[1] 204 301 186 742 790 469
> #formamos la respuesta a utilizar en glm
> 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$tiempo,labels=c("16m","24m"))
> glm(ma~a1$nivelc+tiempo,family = binomial)
```

[5] Valor más coherente con los incrementos de exposición (exposi)
[6] El logaritmo usado es el neperiano ln(Odds), aunque en R se note con log. Aquí lo notaremos de forma indistinta, aunque sea neperiano

```
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
```

> 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
```

> summary(mod2)

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

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

Deviance Residuals:
       1         2         3         4         5         6
 0.59184  -0.62207   0.28739  -0.10809   0.12917  -0.07415

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
```

Los resultados del resumen del mod2 sugieren que el modelo se ajusta bien a los datos (Residual deviance es 0.8537 con 1 g.l.) pero es innecesariamente complejo; no son significativos ni el término nivelc al cuadrado (I(a1$nivelc^2) con un p-valor 0.2924, mayor a 0.05), ni existe interacción entre tiempo y nivel (p-valor=0.7576).
El resultado anova que sigue corrobora las afirmaciones anteriores.

> 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 con 2 grados de libertad, que no es significativa (p-valor=0.55385).

En resumen, se observa que no existe interacción. Tampoco es significativa la variable que expresa el nivel al cuadrado (nivel2).

Los efectos son aditivos en la escala logarítmica. El cambio en el logit es una suma de efectos de nivelc y factor 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 modelo ajusta bien: 2.0354 con 3 gl. No es significativa la deviance residual.

Ecuación del modelo:

Ln(odds)= -5.5683 + 2.8569 nivelc + 2.0989 tiempo24m

Razón de odds por aumento de una unidad en nivel de concentración

Razón de odds que compara los expuestos 24 meses frente a los expuestos 16 meses