Ejemplo simple de regresión logística

CONTENIDO:
Regresión logística con variables cuantitativas y cualitativas

OBJETIVO:

1. Mostrar los modos en que se puede introducir la información para el análisis
2. Familiarizar con los términos y funciones usuales del análisis de regresión logística
3. Interpretar los resultados del análisis

Regresión logística

Regresión logística

1.Datos: Formas en que se pueden presentar

Los datos pueden darse de varios modos, según se presente la información relativa a los exitos y fracasos de la variable dependiente.

1) Un vector de valores que representan proporciones de éxitos. Nº de éxitos ***yi*** entre el total (***ni=éxitos + fracasos***). En este caso los totales deben introducirse como el argumento weights.
2) Un vector de 0's y 1's (fracasos y éxitos, respectivamente). En este caso no hay que especificar el argumento weights.
3) Un vector con valores que representan más de dos niveles. En este caso se trata como en el caso 2), anterior, asumiendo que el nivel más bajo representa el cero o fracaso y los otros el 1(éxito).
4) Una matriz formada por dos columnas que representan los éxitos y fracasos. En este caso se asume que la primera columna contiene los éxitos (***yi***) y la segunda los fracasos (***ni-yi***). Tampoco es necesario el argumento weights.

Ejemplo:

Datos:

Varios grupos de trabajadores (hombres), han estado expuestos a diversas dosis de un compuesto químico. Los datos se muestran en la tabla siguiente:

Introduzca estos datos

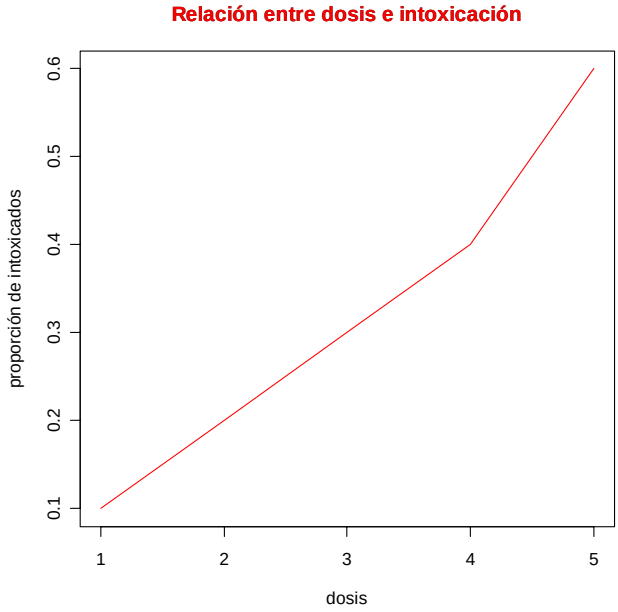
```
> datos=edit(data.frame())
> datos
  dosis numexpos numintoxi
1     1       20         2
2     2       30         6
3     3       30         9
4     4       25        10
5     5       20        12
```

```
d=datos
d$frec=d$numintoxi/d$numexpos   #proporción de intoxicados (éxitos)
d$frac= d$numexpos-d$numintoxi   #número de no intoxicados (fracasos)
```

```
> d
  dosis numexpos numintoxi frec frac
1     1       20         2  0.1   18
2     2       30         6  0.2   24
3     3       30         9  0.3   21
4     4       25        10  0.4   15
5     5       20        12  0.6    8
```

Representación gráfica:

```
plot(d$dosis, d$frec,type="l",col=2,ylab="proporción de intoxicados",xlab="dosis")
title("Relación entre dosis e intoxicación",col.main="red")
```



1.1.Datos en modo frecuencias relativas

> Mod1=glm(d$frec~d$dosis, weights=d$numexpos, family=binomial)
>Mod1=glm(frec~dosis, weights=numexpos,family=binomial,data=d)#equivalente anterior
> Mod1

```
Call:  glm(formula = frec ~ dosis, family = binomial, data = d, weights = d$numexpos)

Coefficients:
(Intercept)        dosis
    -2.6814       0.6013

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
Null Deviance:      14.92
Residual Deviance: 0.2037       AIC: 20.93
```

Nota: observe que no es necesario calcular previamente las variables. Se puede modelar directamente operando con ellas. Por ejemplo, es equivalente ejecutar la orden:

> glm((numintoxi/numexpos)~dosis, weights=numexpos,family=binomial,data=d)

```
Call:  glm(formula = (numintoxi/numexpos) ~ dosis, family = binomial,       data = d,
weights = numexpos)

Coefficients:
(Intercept)        dosis
    -2.6814       0.6013

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
Null Deviance:      14.92
Residual Deviance: 0.2037       AIC: 20.93
```

1.2.Datos en modo matriz de éxitos y fracasos

> md=cbind(d$numintoxi, d$frac)
> Mod2=glm(md~ dosis, family = binomial,data=d)
> Mod2

```
Call:  glm(formula = md ~ dosis, family = binomial, data = d)

Coefficients:
(Intercept)        dosis
    -2.6814       0.6013

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
```

```
Null Deviance:      14.92
Residual Deviance: 0.2037       AIC: 20.93
```

1.3.Datos en modo vector de éxitos y fracasos

Preparación de los datos en formato vector de éxitos y fracasos

```
>y=c(rep(1,sum(d$numintoxi)) , rep(0,sum(d$frac)) )
>dosis2=c(rep(d$dosis,d$numintoxi),rep(d$dosis,d$frac))
```

```
> y
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[112] 0 0 0 0 0 0 0 0 0 0 0 0 0
> dosis2
  [1] 1 1 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5
 [38] 5 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4
[112] 4 4 4 4 4 4 5 5 5 5 5 5 5 5
```

```
>glm(y~dosis2,family=binomial)
```

```
Call:  glm(formula = y ~ dosis2, family = binomial)

Coefficients:
(Intercept)       dosis2
    -2.6814       0.6013

Degrees of Freedom: 124 Total (i.e. Null);  123 Residual
Null Deviance:      155.2
Residual Deviance: 140.5        AIC: 144.5
```

Observe que sí han cambiado, no obstante, los grados de libertad, el valor de la deviance y el valor del estadístico AIC. Antes los datos se estructuraban en una tabla de 5 filas. Ahora en una de 125.

b) Suponga que a iguales niveles de dosis y para iguales tamaños de expuestos (numexpos) se han obtenido los siguientes intoxicadas para subgrupos de mujeres.

```
  dosis numexpos numintoxi
1     1       20         4
2     2       30        10
3     3       30        16
4     4       25        20
5     5       20        19
```

Genere un nuevo archivo que contenga la información completa diferenciando por sexo.

```
> #Data frame con nuevos datos para mujeres añadidos
> d2=rbind(datos,datos)
> d2
   dosis numexpos numintoxi
1      1       20         2
2      2       30         6
3      3       30         9
4      4       25        10
5      5       20        12
11     1       20         2
21     2       30         6
31     3       30         9
41     4       25        10
51     5       20        12
```

```
> d2$numintoxi=c(datos$numintoxi,4,10,16,20,19)  #sustituye los valores para mujeres
> d2$sexo=c(rep('h',5),rep('m',5))
> d2
   dosis numexpos numintoxi sexo
1      1       20         2    h
2      2       30         6    h
3      3       30         9    h
```

```
4      4       25        10    h
5      5       20        12    h
11     1       20         4    m
21     2       30        10    m
31     3       30        16    m
41     4       25        20    m
51     5       20        19    m
```

>d2$sexo=factor(d2$sexo)
> glm(d2$numintoxi/d2$numexpos~d2$dosis+d2$sexo, weights=d2$numexpos, family=binomial)
>#o la orden equivalente
> glm(numintoxi/ numexpos~ dosis+ sexo, weights= numexpos, family=binomial,data=d2)

```
Call:  glm(formula = numintoxi/numexpos ~ dosis + sexo, family = binomial,       data =
d2, weights = numexpos)

Coefficients:
(Intercept)        dosis        sexom
    -3.3385       0.7972       1.2467

Degrees of Freedom: 9 Total (i.e. Null);  7 Residual
Null Deviance:      69.06
Residual Deviance: 3.803        AIC: 42.4
```

> summary(glm(numintoxi/ numexpos~ dosis+ sexo+dosis*sexo, weights= numexpos, family =binomial, data=d2))

```
Call:
glm(formula = numintoxi/numexpos ~ dosis + sexo + dosis * sexo,
    family = binomial, data = d2, weights = numexpos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.56947  -0.13528   0.06966   0.19441   0.61606

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.68138    0.59066  -4.540 5.63e-06 ***
dosis        0.60134    0.16726   3.595 0.000324 ***
sexom       -0.01132    0.82765  -0.014 0.989090
dosis:sexom  0.41042    0.25600   1.603 0.108885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.0618  on 9  degrees of freedom
Residual deviance:  1.1771  on 6  degrees of freedom
AIC: 41.777

Number of Fisher Scoring iterations: 4
```

Nota: el mismo resultado da si se sustituye la fórmula anterior (más extensa) por la más simple: numintoxi/numexpos~dosis*sexo, dado que se asume que los efectos de orden inferior al de interacción se incluyen por defecto (modelos jerárquicos).

En la tabla anterior, eliminamos el término menos significativo (p-valor más grande) que es **dosis:sexom** (recuerde que aunque **sexom** presenta un p-valor mayor, no puede salir del modelo antes que otro término de orden superior del que forma parte (principio de jerarquía). La eliminación del término de interacción da lugar a un nuevo contexto, dentro del cual vemos que la variable sexo recupera importancia.

> summary(glm(numintoxi/ numexpos~ dosis+ sexo, weights= numexpos, family=binomial,data=d2))

```
Call:
glm(formula = numintoxi/numexpos ~ dosis + sexo, family = binomial,
    data = d2, weights = numexpos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-0.63047  -0.49580   0.04172   0.55702   1.20603

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3385     0.4770  -7.000 2.57e-12 ***
dosis         0.7972     0.1255   6.352 2.12e-10 ***
sexom         1.2467     0.3023   4.123 3.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.0618  on 9  degrees of freedom
Residual deviance:  3.8027  on 7  degrees of freedom
AIC: 42.403

Number of Fisher Scoring iterations: 4
```

En la tabla coeficientes, tanto dosis como sexom son altamente significativas, con p-valores bajísimos, lo que demuestra que ambas variables son importantes para explicar la variable dependiente (probabilidad de éxito).

La deviance residual con un valor de 3.8027 y 7 grados de libertad, puede tomarse como indicio de buen ajuste del modelo. Aunque la interpretación de estos resultados deben realizarse con cautela, en este caso, dado que los datos se han presentado agrupados, la aproximación no resulta arriesgada. Cuando el ajuste se realiza con datos sin agrupar, no debe efectuarse.

2.Presentación de la ecuación del modelo en términos de logits, de odds y de probabilidad

```
Call:  glm(formula = numintoxi/numexpos ~ dosis + sexo, family = binomial,      data =
d2, weights = numexpos)

Coefficients:
(Intercept)        dosis        sexom
    -3.3385       0.7972       1.2467

Degrees of Freedom: 9 Total (i.e. Null);  7 Residual
Null Deviance:      69.06
Residual Deviance: 3.803        AIC: 42.4
```

$$\ln\left[\frac{P(\acute{e}xito)}{P(\,fracaso)}\right] = \log it = -3.3385 + 0.7972\,dosis + sexom\,1.2467$$

$$Odds = \frac{P(\acute{e}xito)}{P(\,fracaso)} = e^{\log it} = e^{-3.3385+0.7972\,dosis+sexom\,1.2467}$$

$$P(\acute{e}xito) = \frac{e^{\log it}}{1+e^{\log it}} = \frac{e^{-3.3385+0.7972\,dosis+sexom\,1.2467}}{1+e^{-3.3385+0.7972\,dosis+sexom\,1.2467}}$$

3.Interpretación de los coeficientes del modelo

```
Call:  glm(formula = numintoxi/numexpos ~ dosis + sexo, family = binomial,      data =
d2, weights = numexpos)

Coefficients:
(Intercept)        dosis        sexom
    -3.3385       0.7972       1.2467
```

```
Degrees of Freedom: 9 Total (i.e. Null);  7 Residual
Null Deviance:      69.06
Residual Deviance: 3.803        AIC: 42.4
```

Coeficiente de dosis:

1. 0,7972 es el cambio esperado en el logit al aumentar una unidad la dosis, supuestas estables el resto de las variables en el modelo.

2. exp(0,7972) es la razón de Odds al aumentar una unidad la dosis, supuestas estables el resto de las variables en el modelo.

```
> exp(0.7972)= 2.219318
```

Es decir, al aumentar una unidad la dosis, la Odds o ventaja de intoxicación se duplica; es 2.22 (podríamos decir que el riesgo de intoxicación frente a no intoxicación se multiplica por 2.22)

Coeficiente de sexo:

1.2467 es el cambio esperado en el logit al pasar de un hombre a una mujer.

La razón de odds que compara mujeres con hombres es igual a exp(1.2467)= 3.478844. Es decir, es 3,5 veces superior la odds o ventaja de intoxicación en la mujer que en el hombre.

4. Otros resultados del análisis del modelo ajustado sin interacción

```
> logit1=glm(numintoxi/numexpos~dosis+sexo, weights=numexpos, family=binomial,data=d2)
> residuals(logit1)
         1          2          3          4          5         11         21
 0.4411415  0.7559416  0.2487533 -0.6304705 -0.5248322 -0.1653203 -0.5101743
        31         41         51
-0.4526643  0.5956509  1.2060331

> fitted.values(logit1)
         1          2          3          4          5          6          7
0.07300754 0.14877628 0.27947247 0.46258795 0.65638441 0.21505610 0.37811381
         8          9         10
0.57434218 0.74964897 0.86919871

> coef(logit1)
(Intercept)    d2$dosis    d2$sexom
 -3.3385342   0.7971514   1.2466694

> deviance(logit1)
[1] 3.802703

> logit1$null.deviance
[1] 69.06178

> logit1$y
        1         2         3         4         5         6         7         8
0.1000000 0.2000000 0.3000000 0.4000000 0.6000000 0.2000000 0.3333333 0.5333333
        9        10
0.8000000 0.9500000

> logit1$xlevels
$`d2$sexo`
[1] "h" "m"

> logit1$residuals
          1           2           3           4           5           6
 0.39883992  0.40447691  0.10194060 -0.25176133 -0.24999304 -0.08919121
```

```
          7           8           9          10
-0.19043870 -0.16774370  0.26828786  0.71070144
```

5. Comparación de modelos mediante anova

La función anova permite comparar modelos anidados. Cuando se usa un solo modelo se determina la significatividad de cada término añadido.

```
> anova(logit1,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: numintoxi/numexpos

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     9     69.062
dosis  1   46.997         8     22.065 7.109e-12
sexo   1   18.262         7      3.803 1.925e-05
```

Comprobaremos de otro modo que el término interacción no es significativo:

```
> logit2=update(logit1,.~.+sexo*dosis)
> logit2

Call:  glm(formula = numintoxi/numexpos ~ dosis + sexo + dosis:sexo,        family =
binomial, data = d2, weights = numexpos)

Coefficients:
(Intercept)        dosis        sexom  dosis:sexom
   -2.68138      0.60134     -0.01132      0.41042

Degrees of Freedom: 9 Total (i.e. Null);  6 Residual
Null Deviance:      69.06
Residual Deviance: 1.177        AIC: 41.78


> anova(logit2,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: numintoxi/numexpos

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          9     69.062
dosis       1   46.997         8     22.065 7.109e-12
sexo        1   18.262         7      3.803 1.925e-05
dosis:sexo  1    2.626         6      1.177     0.105
```

La última, fila de la tabla anterior, correspondiente al término de interacción presenta un p-valor mayor que 0.05. No se puede rechazar la hipótesis de nulidad o no importancia del término de interacción para explicar la probabilidad de éxito (intoxicación)
La reducción de la deviance en 2.626 con 1 g.l. no es importante (p-valor=0.105 > 0.05)

Otro modo de comparar los modelos es mediante

```
> anova(logit1,logit2,test = "Chisq")

Analysis of Deviance Table

Model 1: numintoxi/numexpos ~ dosis + sexo
Model 2: numintoxi/numexpos ~ dosis + sexo + dosis:sexo
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
```

```
1         7     3.8027
2         6     1.1771  1   2.6256    0.1052
```

6.Valores predichos del modelo sin interacción

>predict(logit1) #predice los valores de los logits (valores de los logits ajustados)

```
> predict(logit1,type="response") #predice las probabilidades de intoxicación
         1          2          3          4          5          6          7
0.07300754 0.14877628 0.27947247 0.46258795 0.65638441 0.21505610 0.37811381
         8          9         10
0.57434218 0.74964897 0.86919871
```

Inserción de los valores predichos en el data.frame

```
d2$probabilidad= predict(logit1,type="response")
d2
   dosis numexpos numintoxi sexo probabilidad
1      1       20         2    h   0.07300754
2      2       30         6    h   0.14877628
3      3       30         9    h   0.27947247
4      4       25        10    h   0.46258795
5      5       20        12    h   0.65638441
11     1       20         4    m   0.21505610
21     2       30        10    m   0.37811381
31     3       30        16    m   0.57434218
41     4       25        20    m   0.74964897
51     5       20        19    m   0.86919871
```

7.Predicción para valores nuevos con el modelo ajustado

Notas importantes:
Tenga en cuenta que si ha usado en la **fórmula** nombres de variables que implican el data.frame que las contiene, el nuevo data.frame debe llamarse igual y con los mismos nombres de variables. Es decir, si la fórmula es a$y~a$x, para predecir debe usar el mismo nombre, **a**, para el data.frame, y por supuesto los mismos nombres de variables, aunque los valores sean distintos.
Si usa sólo el nombre de las variables en la fórmula debe construir el data.frame nuevo con las columnas cuyos nombres representen a dichas variables.

```
#predicción de valores nuevos
dosis=7
sexo="h"
ndatos1=data.frame(dosis,sexo)
ndatos1
predict(logit1, newdata = ndatos1,type = "response")
```

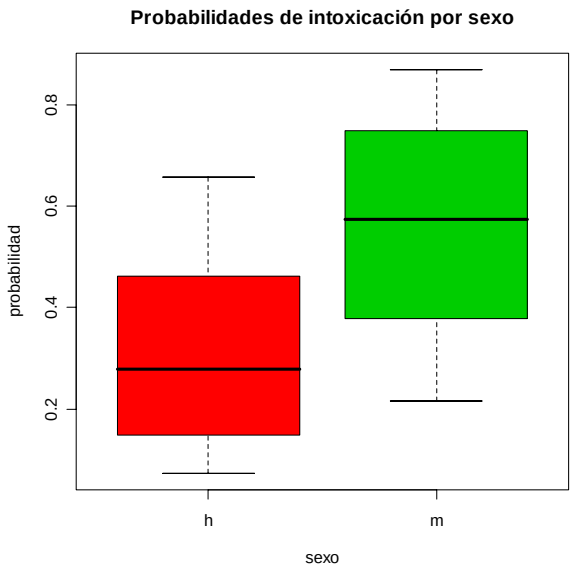
```
[1] 0.05021325
```

```
#predicción de valores nuevos
dosis=7
sexo="h"
ndatos1=data.frame(dosis,sexo)  #datos con valores en las variables independientes
predict(logit1, newdata = ndatos1,type = "response")
```

```
[1] 0.903917
```

8. Representación gráfica del modelo

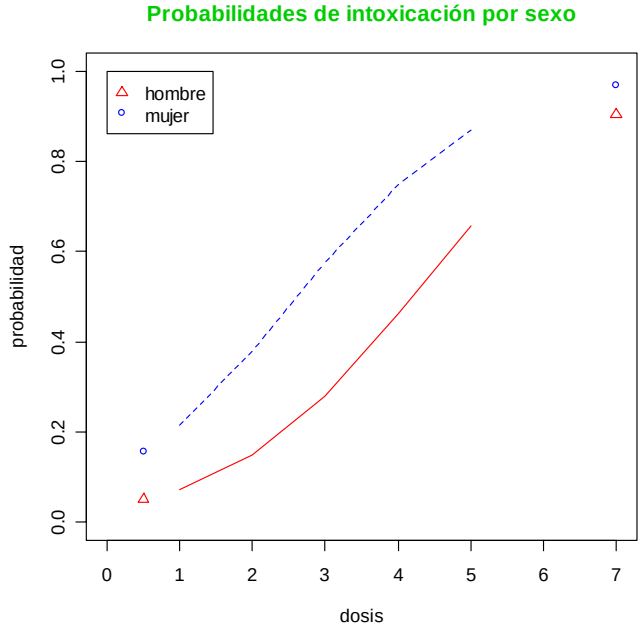
```
plot(probabilidad~sexo, data=d2,col=2:3,main="Probabilidades de intoxicación por sexo")
```




```
#predicción de valores nuevos para hombres
ndosis=c(0.5,7)
sexo=c("h","h")

ndatosh=data.frame(ndosis,sexo)  #datos con valores en las variables independientes
ndatosh
ph=predict(logit1, newdata = ndatosh,type = "response")

#predicción de valores nuevos para mujeres
ndosis=c(0.5,7)
sexo=c("m","m")
ndatosm=data.frame(ndosis,sexo)
ndatosm
pm=predict(logit1, newdata = ndatosm,type = "response")
```

Representación gráfica de la curva

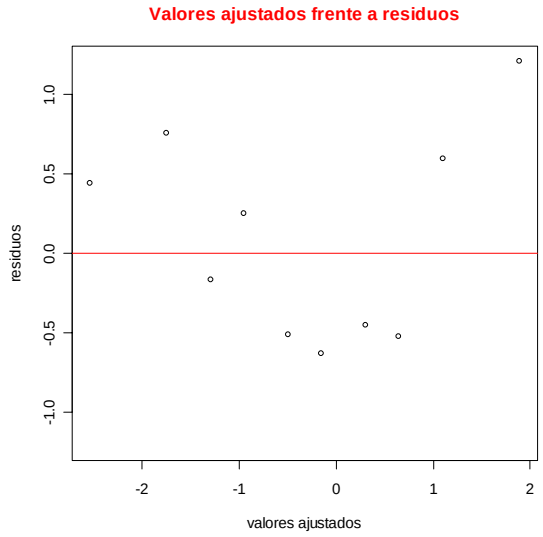
```
>#Valores ajustados
>plot(probabilidad~dosis, data=d2,subset=sexo=="h",type="l",lty=1,col=2,ylim=c(0,1),xlim=c(0,7))
>lines(probabilidad~dosis, data=d2,subset=sexo=="m",type="l",lty=2,col=4,ylim=c(0,1),xlim=c(0,7))

>#Valores nuevos ajustados
>points(ndosis,pm,col=4,pch=1)
>points(ndosis,ph,col=2,pch=2)

title("Probabilidades de intoxicación por sexo",col.main=3)
legend(0,1,legend=c("hombre","mujer"),col=c(2,4),pch=2:1)
```



9. Diagnóstico de residuos

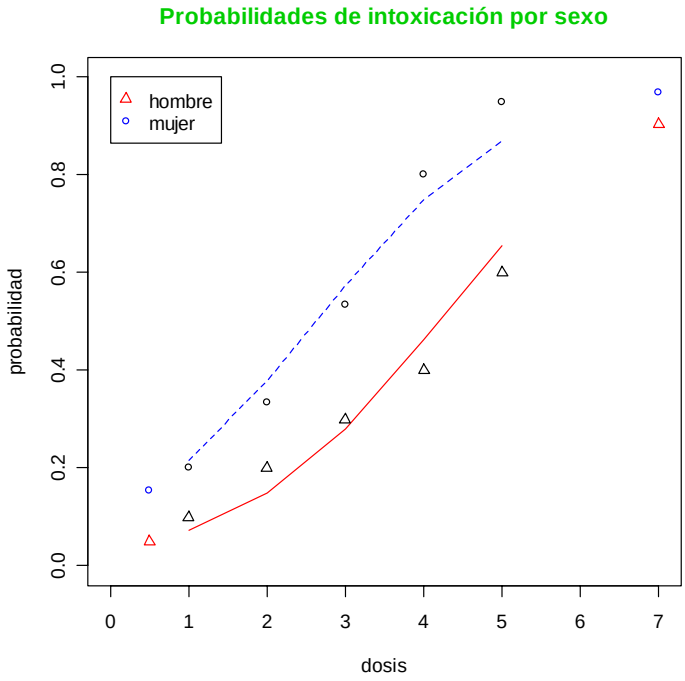
La representación gráfica de los valores ajustados y los residuos tipo “deviance”, permite ver si existen o no anomalías en el ajuste.

```
p=predict(logit1)#valores de los logits ajustados
r=residuals(logit1)
plot(p,r,xlab="valores ajustados",ylab="residuos",ylim=c(-1,1)*max(abs(r)),pch=levels(d2$sexo))
abline(0,0,col="red")
title("Valores ajustados frente a residuos",col.main="red")
```



Los residuos no caen fuera de la banda (-2, 2). Prácticamente están en la banda (-1,1), lo que es buen indicio de ajuste; sin embargo, parece existir cierta pauta de descenso

cuando crecen los valores ajustados hasta un valor medio, a partir del cual tienden a crecer.
El gráfico de probabilidades ajustadas en función de la dosis y sexo, muestra que en el grupo de mujeres, a valores bajos de la dosis, el modelo tiende a sobreestimar la probabilidad y para valores altos, hace lo contrario. En el grupo de los hombres el modelo subestima cuando las dosis son bajas y sobrestima la probabilidad cuando los valores de la dosis son altos.



10. Resumen de las funciones usadas

En orden aproximado de ejecución

```
edit(data.frame())

plot(d$dosis, d$frec,type="l",col=2,ylab="proporción de intoxicados",xlab="dosis")
title("Relación entre dosis e intoxicación",col.main="red")

glm((numintoxi/numexpos)~dosis, weights=numexpos,family=binomial,data=d)

md=cbind(d$numintoxi, d$frac)
Mod2=glm(md~ dosis, family = binomial,data=d)

y=c(rep(1,sum(d$numintoxi)) , rep(0,sum(d$frac)) )
dosis2=c(rep(d$dosis,d$numintoxi),rep(d$dosis,d$frac))
glm(y~dosis2,family=binomial)

d2$sexo=factor(d2$sexo)
glm(numintoxi/ numexpos~ dosis+ sexo, weights= numexpos, family=binomial,data=d2)
summary( glm(numintoxi/ numexpos~ dosis+ sexo+dosis*sexo, weights= numexpos, family =binomial, data=d2))
```

```
logit1=glm(numintoxi/numexpos~dosis+sexo, weights=numexpos, family=binomial,data=d2)
residuals(logit1)
fitted.values(logit1)
coef(logit1)
deviance(logit1)

logit1$null.deviance
logit1$y
logit1$residuals


anova(logit1,test="Chisq")
logit2=update(logit1,.~.+sexo*dosis)

predict(logit1)  #predice los valores de los logits (valores de los logits ajustados)
predict(logit1,type="response") #predice las probabilidades de intoxicación
predict(logit1, newdata = ndatos1,type = "response")


plot(probabilidad~dosis, data=d2,subset=sexo=="h",type="l",lty=1,col=2,ylim=c(0,1),xlim=c(0,7))
lines(probabilidad~dosis, data=d2,subset=sexo=="m",type="l",lty=2,col=4,ylim=c(0,1),xlim=c(0,7))
points(ndosis,pm,col=4,pch=1)
```

Prácticas:

1) Con la opción copiar y pegar seleccione y copie en el *editor bloc de notas* los datos del data.frame d en un archivo de datos denominado tóxico. Léalo desde R.
2) Con la función de R:(write.table) guarde los datos del data data.frame d en un archivo denominado toxico2. Léalo luego desde R.
3) Efectúe una predicción para una mujer expuesta a una dosis de 9
4) Razón de odds que compara hombres con mujeres
5) Incremento esperado en el logit al aumentar 2 unidades la dosis de exposición
6) Probabilidad de intoxicación de una mujer expuesta a una dosis de 15