Ejemplo

Supongamos un problema de dosis–respuesta en el que un grupo de animales son expuestos a una sustancia peligrosa en distintas concentraciones. Sea n_i el número de animales que recibe la dosis i , Y_i el número de animales que muere y por lo tanto $p_i = Y_i/n_i$ la proporción de muertos en el i–ésimo grupo.

Llamemos π_i a la probabilidad de muerte y modelemos a π_i en términos de $z_i = \log_{10}(\text{concentración})$.

Proponemos el modelo:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 z_i \, .$$

Un parámetro de interés en estos problemas suele ser el valor de z para el cual se obtiene el 50 % de muertes. Llamemos a dicho valor M_{50}.

Como $\text{logit}(1/2) = 0$, tenemos que $M_{50} = -\dfrac{\beta_0}{\beta_1}$. Por lo tanto,

$$\frac{\partial M_{50}}{\partial \beta_0} = \frac{-1}{\beta_1}$$
$$\frac{\partial M_{50}}{\partial \beta_1} = \frac{\beta_0}{{\beta_1}^2}$$

La varianza estimada de $-\frac{\widehat{\beta}_0}{\widehat{\beta}_1}$ es

$$\left[\frac{-1}{\widehat{\beta}_1}, \frac{\widehat{\beta}_0}{\widehat{\beta}_1^2}\right] \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1} \begin{bmatrix} \frac{-1}{\widehat{\beta}_1} \\ \\ \frac{\widehat{\beta}_0}{\widehat{\beta}_1^2} \end{bmatrix},$$

donde $\widehat{\mathbf{W}} = diag(n_i\widehat{\pi}_i(1-\widehat{\pi}_i))$.

Tests de Hipótesis

En el contexto de GLM abordaremos el problema de comparar dos modelos cuando tienen la misma distribución subyacente y la misma función link.

Consideraremos la comparación de dos modelos anidados, es decir la diferencia entre los dos modelos será que la componente lineal de un modelo tendrá más parámetros que el otro.

El modelo más simple, que corresponderá a H_o, será un caso especial de un modelo más general.

Si el modelo más simple ajusta a los datos tan bien como el más general, entonces, en virtud del principio de parsimonia no rechazaremos H_o.

Si el modelo más general ajusta significativamente mejor, rechazaremos H_o en favor de H_1, que corresponde al modelo más complejo. Para realizar estas comparaciones deberemos usar medidas de *bondad de ajuste*. .

Las medidas de bondad de ajuste pueden basarse en el máximo valor de la función de verosimilitud, en el máximo valor del log de la función de verosimilitud,

en el mínimo valor de la suma de cuadrados o en un estadístico combinado basado en los residuos.

El proceso de comparación será como sigue: $M_o \subset M_1$

1. Especificamos un modelo M_o correspondiente a H_o y un modelo más general, M_1, que corresponde a H_1.
2. Ajustamos M_o y calculamos el estadístico de bondad de ajuste G_o. Idem con M_1 y su correspondiente G_1.
3. Computamos la *mejoría* computando una medida de la discrepancia entre $G1$ y G_o.
4. A partir de la distribución de esta medida de discrepancia, testeamos H_o vs. la alternativa H_1, es decir M_o vs. M_1.
5. Si la hipótesis H_o no es rechazada, preferimos el modelo M_o. Si rechazamos H_o, elegiremos M_1.

Estadístico de Cociente de Verosimilitud

El modelo con el máximo número de parámetros que pueden ser estimados se conoce como **modelo saturado**. Es un GLM con la misma distribución subyacente y la misma función de enlace que el modelo de interés, que podría tener tantos parámetros como observaciones. Llamemos m al máximo número de parámertos que puede especificarse. (Si hay observaciones que tienen las mismas covariables (replicaciones), el modelo saturado podría determinarse con menos de n parámetros.)

En el modelo saturado los μ justan exactamente a los datos: $\widehat{\mu}_i = Y_i$.

Por lo tanto, en el modelo saturado se asigna toda la variación a la componente sistemática y ninguna a la componente aleatoria. Este modelo no se usa ya que no resume la información presente en los datos, pero provee una base para medir la discrepancia para un modelo intermedio entre el modelo saturado y el **modelo nulo**, en el que hay un único parámetro para todas las observaciones.

Sean $\widehat{\boldsymbol{\theta}}_s$ el valor estimado bajo el modelo saturado y $L(\widehat{\boldsymbol{\theta}}_s, \mathbf{y})$, el likelihood evaluado en dicho estimador

Sea $L(\widehat{\boldsymbol{\theta}}, \mathbf{y})$ el máximo valor del likelihood para el modelo de interés. El cociente de verosimilitud será

$$\lambda = \frac{L(\widehat{\boldsymbol{\theta}}_s, \mathbf{y})}{L(\widehat{\boldsymbol{\theta}}, \mathbf{y})},$$

que nos da una idea de cuán bueno es el ajuste del modelo, respecto del modelo saturado.

En la práctica se usa el logaritmo de este cociente

$$\log(\lambda) = \ell(\widehat{\boldsymbol{\theta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\theta}}, \mathbf{y}).$$

Grandes valores de $log(\lambda)$ sugieren un pobre ajuste del modelo respecto al modelo saturado.

Un estadístico cercano y muy usado en el contexto de GLM es la **deviance**,

introducida por Nelder y Wedderburn (1972).

Asumamos que $a_i(\phi) = \phi/w_i$, entonces consideremos

$$D^* = 2\left[\ell(\widehat{\boldsymbol{\theta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\theta}}, \mathbf{y})\right] = 2\sum_{i=1}^{n} w_i\{y_i(\widehat{\boldsymbol{\theta}}_{si} - \widehat{\boldsymbol{\theta}}_i) - b(\widehat{\boldsymbol{\theta}}_{si}) + b(\widehat{\boldsymbol{\theta}}_i)\}/\phi = D/\phi$$

D es conocida como la deviance y D^* es la deviance escalada.

Nota: A veces es conveniente expresar el log likelihood en términos de las medias μ's más que de $\boldsymbol{\beta}$ o $\boldsymbol{\theta}$. En ese caso llamaríamos $\ell(\widehat{\mu}, \mathbf{y})$ al likelihood maximizado sobre $\boldsymbol{\beta}$, mientras que el máximo alcanzado en el modelo saturado sería $\ell(\mathbf{y}, \mathbf{y})$.

Ejemplos

Caso Normal

Recordemos que $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\Phi = \sigma^2$ $(w_i = 1)$.

Entonces,

$$D = 2 \sum_{i=1}^{n} (y_i(y_i - \mu_i) - \frac{1}{2}y_i^2 + \frac{1}{2}\mu_i^2) = \sum_{i=1}^{n} (y_i - \mu_i)^2 .$$

Caso Binomial

Recordemos que $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, es decir $\pi = \frac{e^\theta}{1+e^\theta}$,

$b(\theta) = \log(1 + e^\theta) = -\log(1 - \pi)$, entonces

$$\begin{aligned} D &= 2 \sum_{i=1}^{n} n_i \{\frac{y_i}{n_i}(\widehat{\boldsymbol{\theta}}_{si} - \widehat{\boldsymbol{\theta}}_i) - b((\widehat{\boldsymbol{\theta}}_{si}) + b(\widehat{\boldsymbol{\theta}}_i)\} \\ &= 2 \sum_{i=1}^{n} n_i \left[\frac{y_i}{n_i}\left(\log \frac{y_i/n_i}{1 - y_i/n_i} - \log \frac{\widehat{\pi}_i}{1 - \widehat{\pi}_i}\right) + \right. \end{aligned}$$

$$
\begin{aligned}
& \log(1-\frac{y_i}{n_i}) - \log(1-\widehat{\pi}_i)\Big] \\
= \; & 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i/n_i}{\widehat{\pi}_i} + y_i \log \frac{1-\widehat{\pi}_i}{1-y_i/n_i} + n_i \log \frac{1-y_i/n_i}{1-\widehat{\pi}_i} \right] \\
= \; & 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i/n_i}{\widehat{\pi}_i} + (n_i - y_i) \log \frac{1-y_i/n_i}{1-\widehat{\pi}_i} \right] \\
= \; & 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{\widehat{\mu}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \widehat{\mu}_i} \right]
\end{aligned}
$$

Para realizar los tests de bondad de ajuste debemos conocer la distribución de D.

Heurísticamente podríamos deducir la la distribución de D. Si hacemos un desarrrollo de Taylor de segundo orden alrededor de un punto dado $\mathbf{b}$, tenemos que:

$$\ell(\boldsymbol{\beta}) \simeq \ell(\mathbf{b}) + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{U}(\mathbf{b}) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathcal{I}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})\,.$$

donde $\mathbf{U} = (U_1, \ldots, U_p)'$

$$\begin{aligned} U_j = \frac{\partial \ell(\boldsymbol{\beta}, \mathbf{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta}, y_i)}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_j} x_{ij} \quad j = 1, \ldots, p\,. \end{aligned}$$

$$E(\mathbf{U}) = 0 \qquad E(\mathbf{U}\mathbf{U}') = \mathcal{I}\,,$$

siendo $\mathcal{I}$ la matriz de información de Fisher.

Si $\mathbf{b}$ es el punto donde ℓ alcanza su máximo, entonces

$$\ell(\boldsymbol{\beta}) - \ell(\mathbf{b}) \simeq -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})'\mathcal{I}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})\,.$$

Por lo tanto

$$2(\ell(\mathbf{b}) - \ell(\boldsymbol{\beta})) \simeq (\boldsymbol{\beta} - \mathbf{b})'\mathcal{I}(\mathbf{b})\boldsymbol{\beta} - \mathbf{b})\,.$$

y en consecuencia, para n suficientemente grande

$$2(\ell(\mathbf{b}) - \ell(\boldsymbol{\beta})) \overset{(a)}{\sim} \chi^2_p\,.$$

de este resultado, obtenemos

$$\begin{aligned} D/\phi &= 2\left[\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}, \mathbf{y})\right] \\ &= 2\left[\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\boldsymbol{\beta}_s, \mathbf{y})\right] \\ &\quad -2\left[\ell(\widehat{\boldsymbol{\beta}}, \mathbf{y}) - \ell(\boldsymbol{\beta}, \mathbf{y})\right] + 2\left[\ell(\boldsymbol{\beta}_s, \mathbf{y}) - \ell(\boldsymbol{\beta}, \mathbf{y})\right] \end{aligned}$$

Luego,

$$D/\phi \overset{(a)}{\sim} \chi^2_{m-p,\nu}\,,$$

siendo

$$\nu = 2\left[\ell(\boldsymbol{\beta}_s, \mathbf{y}) - \ell(\boldsymbol{\beta}, \mathbf{y})\right]\,,$$

donde ν es una constante positiva cercana a 0 si el modelo propuesto ajusta a los datos *tan bien* como el modelo saturado.

Aplicaciones a Test de Hipótesis

Las hipótesis relativas al parámetro $\boldsymbol{\beta}$ de longitud p pueden testearse usando el estadístico de Wald y su distribución asintótica

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \, \mathcal{I}_n(\widehat{\boldsymbol{\beta}}) \, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{(a)}{\sim} \chi^2_p \, .$$

Si queremos decidir entre dos modelos anidados, un enfoque alternativo es comparar la bondad del ajuste de los modelos involucrados. Consideremos la hipótesis nula:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0q})'$$

correspondiente al Modelo M_0 y una hipótesis más general

$$H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = (\beta_{01}, \ldots, \beta_{0p})'$$

correspondiente al Modelo M_1 con $q < p < n$.

Si testeamos H_0 vs. H_1 usando la diferencia de los estadísticos de cociente del

logaritmo de la verosimilitud tenemos

$$\begin{aligned}\Delta D/\phi &= \frac{D_0 - D_1}{\phi} \\ &= 2\left[\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}_0, \mathbf{y})\right] - 2\left[\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}_1, \mathbf{y})\right] \\ &= 2\left[\ell(\widehat{\boldsymbol{\beta}}_1, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}_0, \mathbf{y})\right] .\end{aligned}$$

Compararíamos a ΔD con una $\phi\chi^2_{p-q}$ ya que bajo H_0 tendríamos que $\Delta D \overset{(a)}{\sim} \phi\chi^2_{p-q}$.

Si el valor observado de $\Delta D/\phi$ fuera mayor que el percentil $\chi^2_{p-q,\alpha}$ rechazaríamos a H_0 en favor de H_1, bajo el supuesto de que H_1 da una mejor descripción de los datos.

Ejemplo:

Los siguientes datos corresponden a un experimento de dosis–respuesta en el que 5 grupos de 6 animales fueron expuestos a una sustancia peligrosa (Schafer, 2000). Y_i denota al número de animales que murieron al ser expuestos a la $i-$ésima dosis.

obs.	$x_i = \log_{10}(conc.)$	y_i	$n_i - y_i$	y_i/n_i	$\widehat{\pi}_i$
1	-5	0	6	0.000	0.0080899
2	-4	1	5	0.167	0.1267669
3	-3	4	2	0.667	0.7209767
4	-2	6	0	1.000	0.9787199
5	-1	6	0	1.000	0.9987799

El comando R que usamos es:

```
 yy<- c(0,1,4,6,6)
 sf<- cbind(yy,6-yy)
 logdosis<- -c(5:1)

 salida<- glm(sf~logdosis,family=binomial)

 summary(salida)

Call:
glm(formula = sf ~ logdosis, family = binomial)

Deviance Residuals:
      1        2        3        4        5
-0.3122   0.2821  -0.2913   0.5081   0.1210

Coefficients:
            Estimate   Std. Error z value Pr(>|z|)
(Intercept)    9.587      3.707   2.586   0.0097 **
logdosis       2.879      1.102   2.612   0.0090 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 28.0090  on 4  degrees of freedom
Residual deviance:  0.5347  on 3  degrees of freedom
AIC: 8.58

Number of Fisher Scoring iterations: 6

> salida$devian
[1] 0.5347011

> pchisq(salida$deviance,3)          1-pchisq(salida$deviance,3)
[1] 0.0887958                        0.9112042
```

Análisis de la deviance

El análisis de la deviance es una generalización del análisis de la varianza para los GLM obtenido para una secuencia de modelos anidados (cada uno incluyendo más términos que los anteriores). Suponemos en todos ellos la misma distribución y la misma función link.

Dada una secuencia de modelos anidados usamos la deviance como una medida de discrepancia y podemos formar una tabla de diferencias de deviances.

Sean $M_{p_1}, M_{p_2}, \ldots, M_{p_r}$ una sucesión de modelos anidados de dimensión $p_1 < p_2 < \ldots < p_r$ y matrices de diseño $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}, \ldots, \mathbf{X}_{p_r}$ y deviances $D_{p_1} > D_{p_2} > \ldots > D_{p_r}$.

La diferencia $D_{p_i} - D_{p_j}$, $p_j > p_i$, es interpretada como una medida de la variación de los datos explicada por los términos que están en M_{p_j} y no están en M_{p_i}, incluidos los efectos de los términos de que están en M_{p_i} e ignorando los efectos de cualquier término que no está en M_{p_j}.

De esta manera, si $D_{p_i} - D_{p_j} > \chi^2_{p_j - p_i, \alpha}$ habría que incorporar al modelo los términos que están en M_{p_j} y no están en M_{p_i}.

Cada secuencia de modelos corresponde a una tabla de análisis de la varianza diferente. La secuencia de los modelos estará determinada por el interés del investigador.

Veamos otro ejemplo:

Collett (1991) reporta los datos de un experimento sobre toxicidad en gusanos de la planta de tabaco dosis de *pyrethroid trans–cypermethrin* al que los gusanos empezaron a mostrar resistencia. Grupos de 20 gusanos de cada sexo fueron expuestos a por 3 días al *pyrethroid* y se registró el número de gusanos muertos o knockeados en cada grupo.

Los resultados se muestran en la siguiente tabla.

	dosis (μg)					
sexo	1	2	4	8	16	32
Machos	1	4	9	13	18	20
Hembras	0	2	6	10	12	16

Cuadro 1: Gusanos del tabaco

Ajustamos un modelo de regresión logística usando $\log_2(dosis)$, dado que las dosis son potencias de 2.

Para procesar con R usamos los comandos

```
options(contrasts=c("contr.treatment", "contr.poly"))
ldose<- rep(0:5,2)
numdead<- c(1,4,9,13,18,20,0,2,6,10,12,16)
sex<- factor(rep(c("M","F"),c(6,6)))
SF<- cbind(numdead,numalive=20-numdead)

contrasts(sex)
     M
   F 0
   M 1
```

Comenzaremos por un gráfico

```
probas=numdead/20
plot(2^ldose, probas,type="n",xlab="dosis",ylab="prob")
lines(2^ldose[sex=="M"],type="p", probas[sex=="M"],col=6)
lines(2^ldose[sex=="F"], probas[sex=="F"],type="p",col=8)
```

Queremos investigar la posibilidad de que haya diferentes rectas para los dos sexos. Para ello plantearemos y ajustaremos el modelo

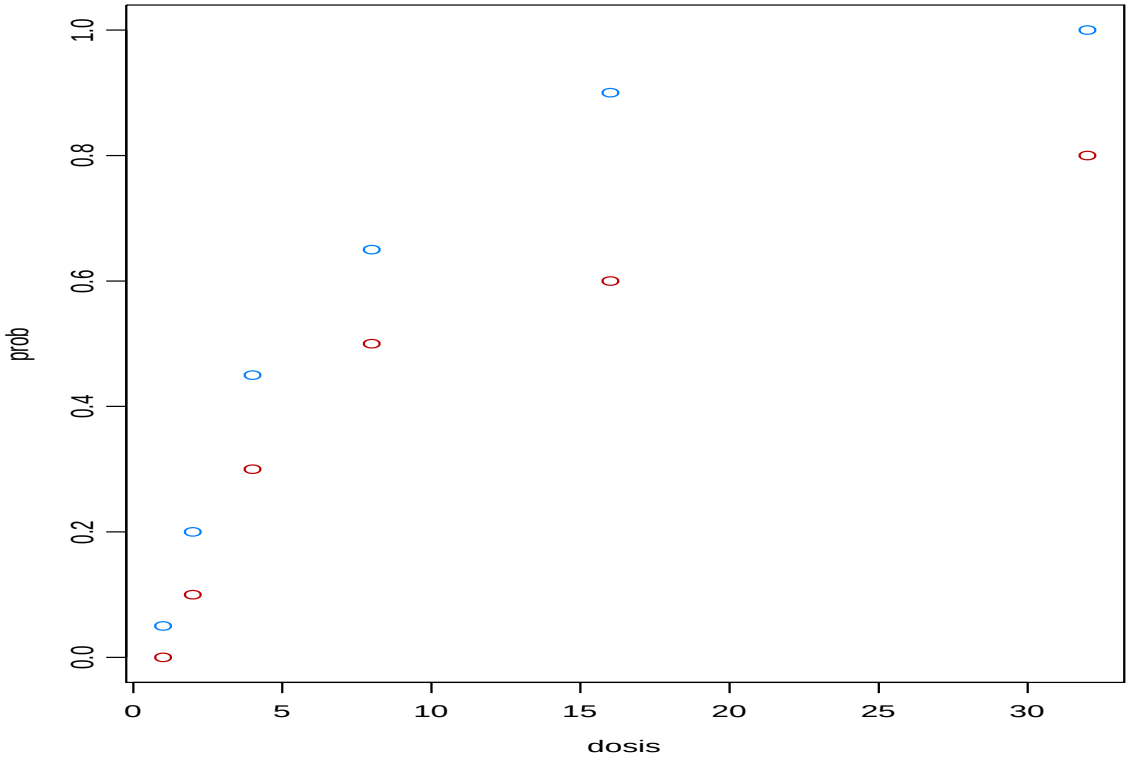

Figura 2: Gusanos del tabaco

$$logit(\pi_i) = \beta_0 + \beta_1 sex + \beta_2 ldose + \beta_3\, sex{:}ldose$$

Por ejemplo, si $sex = M$, para $ldose = 3$ tendríamos

$$logit(\pi_{3,i}) = \beta_0 + \beta_1 + 3\ (\beta_2 + \beta_3)$$

en cambio si $sex = F$, para $ldose = 3$

$$logit(\pi_{3,i}) = \beta_0 + 3\ \beta_2$$

Para ello hacemos:

```
salida.gusanos<- glm(SF~sex*ldose, family=binomial)
summary(salida.gusanos)

Call:
glm(formula = SF ~ sex * ldose, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.39849  -0.32094  -0.07592   0.38220   1.10375

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
```

```
AIC: 43.104
Number of Fisher Scoring iterations: 4

1-pchisq(4.993727,8)= 0.7582464

> salida.gusanos$x
   (Intercept) sexM ldose sexM:ldose
1            1    1     0          0
2            1    1     1          1
3            1    1     2          2
4            1    1     3          3
5            1    1     4          4
6            1    1     5          5
7            1    0     0          0
8            1    0     1          0
9            1    0     2          0
10           1    0     3          0
11           1    0     4          0
12           1    0     5          0
```

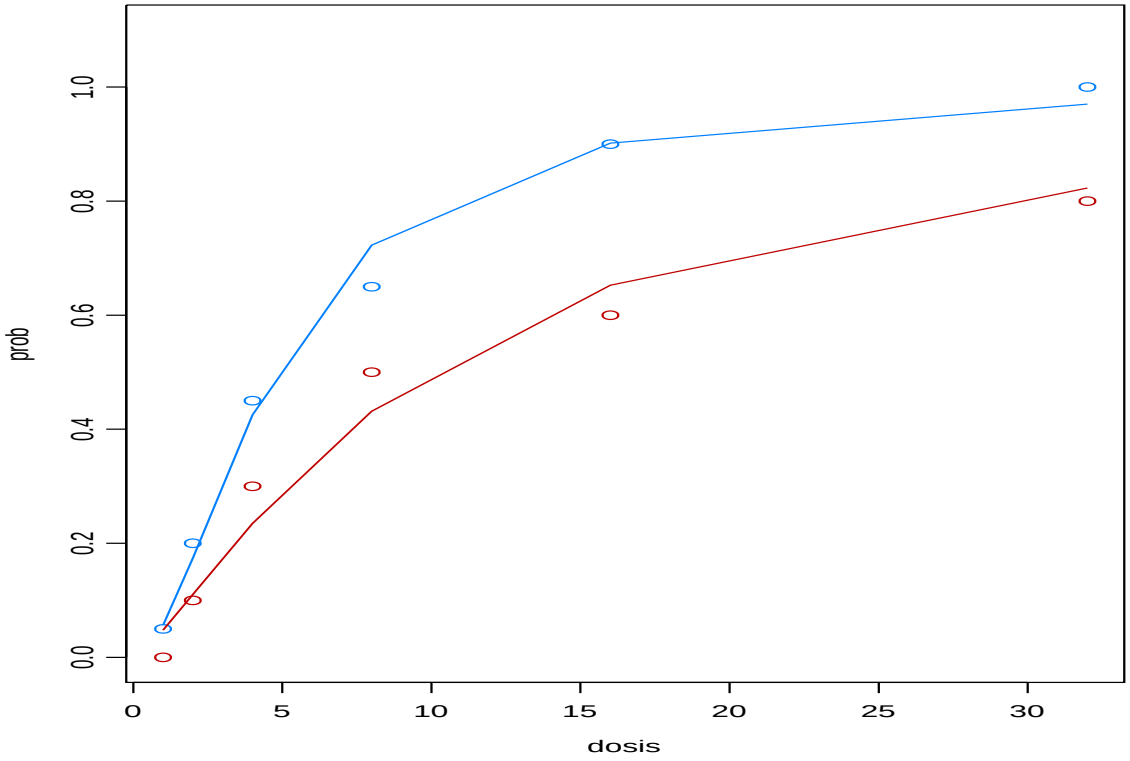

Figura 3: Gusanos del tabaco

Aparentemente de la lectura de la tabla el efecto del sexo parece no significativo. Dado que estamos ajustando distintas pendientes para cada sexo, el test individual para este parámetro testea la hipótesis de que las curvas no difieren cuando la log dosis es 0. Vamos a reparametrizar de manera de incluir la intercept en una dosis central (8), que muestra una diferencia significativa entre los dos sexos en la dosis 8. Computamos el p–valor de la medida de ajuste global

Comparamos distintos modelos mediante la instrucción ANOVA

```
> salida2<-glm(SF~sex*I(ldose-3), family=binomial)
> summary(salida2)
Call:
glm(formula = SF ~ sex * I(ldose - 3), family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.39849  -0.32094  -0.07592   0.38220   1.10375

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.2754     0.2305  -1.195  0.23215
sexM               1.2337     0.3770   3.273  0.00107 **
I(ldose - 3)       0.9060     0.1671   5.422 5.89e-08 ***
sexM:I(ldose - 3)  0.3529     0.2700   1.307  0.19117
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom
AIC: 43.104
Number of Fisher Scoring iterations: 4
```

```
#######################################################

anova(salida.gusanos, test="Chisq")
Analysis of Deviance Table

Binomial model

Response: SF

Terms added sequentially (first to last)
          Df Deviance Resid. Df Resid. Dev   Pr(Chi)
     NULL                    11   124.8756
      sex  1   6.0770        10   118.7986 0.0136955 = 1-pchisq(6.0770,1)
    ldose  1 112.0415         9     6.7571 0.0000000 = 1-pchisq(112.0415,1)
sex:ldose  1   1.7633         8     4.9937 0.1842088 = 1-pchisq(1.7633,1)
```

Ahora ajustamos sin la interacción

```
salida4.gusanos<- glm(SF~sex+ldose, family=binomial)
summary(salida4.gusanos)

Call: glm(formula = SF ~ sex + ldose, family = binomial)

Coefficients:
                        Value         Std. Error             t value
(Intercept) -3.473154184226693 0.4682938902230899 -7.416612210277016
        sex  1.100742853982481 0.3557226321395416  3.094385216262218
      ldose  1.064213642005792 0.1310130474986223  8.122959219134131

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 124.8755926044078 on 11 degrees of freedom

Residual Deviance: 6.757064232235749 on 9 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

La matriz de diseño sería:

```
 cbind(salida4.gusanos$x)
   (Intercept) sex ldose
 1           1   1     0
 2           1   1     1
 3           1   1     2
 4           1   1     3
 5           1   1     4
 6           1   1     5
 7           1   0     0
 8           1   0     1
 9           1   0     2
10           1   0     3
11           1   0     4
12           1   0     5
```

Interpretación de los coeficientes

Supongamos que tenemos una variable independiente que también es dicotómica. Nuestro modelo será

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x \Longrightarrow \pi = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

donde $X = 0$ ó $X = 1$.

Los valores de nuestro modelo son

	$Y = 1$	$Y = 0$
$X = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$
$X = 0$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$	$1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$

Cuadro 2: Variables dicotómicas

El odds ratio es

$$\theta = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}$$

que resulta

$$\theta = e^{\beta_1}$$

por lo tanto el logaritmo del *odds ratio* es

$$\log \theta = \beta_1$$

y un intervalo de confianza de nivel aproximado $1-\alpha$ para θ será

$$\exp(\widehat{\beta}_1 \pm z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\beta}_1)})$$

Consideremos el caso de una variable cualtitativa que toma varios valores, como en la siguiente situación

	blanco	negro	hispanico	otros	Total
Ausente	20	10	10	10	50
Presente	5	20	15	10	50
Total	25	30	25	20	100
θ	1	8	6	4	

Cuadro 3: Ejemplo hipótetico

```
options(contrasts=c("contr.treatment", "contr.poly"))
yy<- c(5,20,15,10)
nn<- c(25,30,25,20)
color<- factor(rep(c("blanco","negro","hipanico","otros"),c(1,1,1,1)))
SF<- cbind(yy,nyy=nn-yy)
```

```
contrasts(color) d            Variables de Diseno
               D1     D2    D3
           hipanico negro otros
  blanco         0     0     0
hipanico         1     0     0
   negro         0     1     0
   otros         0     0     1

Call: glm(formula = SF ~ color, family = binomial)
Coefficients:
                  Value Std. Error   t value
  (Intercept) -1.386294  0.4999999 -2.772589
colorhipanico  1.791759  0.6454971  2.775782
   colornegro  2.079442  0.6324554  3.287886
   colorotros  1.386294  0.6708203  2.066566

    Null Deviance: 14.04199 on 3 degrees of freedom
Residual Deviance: 0 on 0 degrees of freedom
```

Veamos que

$$\begin{aligned}\exp(1.791759) &= 5.999997\\ \exp(2.079442) &= 8.000004\\ \exp(1.386294) &= 3.999999\end{aligned}$$

Observemos además que como

$$\text{logit}(\pi) = \beta_0 + \beta_{11}D_1 + \beta_{12}D_2 + \beta_{13}D_3$$

$$\begin{aligned}\log\widehat{\theta}(negro, blanco) &= \\ &= \beta_0 + \beta_{11}(D_1 = 0) + \beta_{12}(D_2 = 1) + \beta_{13}(D_3 = 0)\\ &\quad - [\beta_0 + \beta_{11}(D_1 = 0) + \beta_{12}(D_2 = 0) + \beta_{13}(D_3 = 0)]\\ &= \beta_{12}\end{aligned}$$

y en base a la distribución asintótica de los parámetros podemos obtener un intervalo de confianza para $\theta(negro, blanco)$.

Algunas herramientas de diagnóstico

- Como en regresión lineal al graficar los residuos vs. el predictor lineal $\widehat{\eta}$ esperamos encontrar una banda horizontal, más o menos paralela al eje de las abscisas alrededor del 0.

- Podríamos encontrar una curvatura o un ancho de la banda variable. Una curvatura podría sugerir:

  - elección incorrecta de la función de enlace
  - omisión de algún término no lineal de una covariable

- El ancho de banda variable puede sugerir que la función de varianza es incorrecta.

- También estos gráficos pueden ayudar a detectar residuos muy grandes, es decir mayores que 2.5 ó 3.

- Otra posibilidad es graficar los residuos vs. cada covariable por separado, tal como lo hacíamos en Modelo Lineal. Una curvatura en este gráfico nuevamente puede sugerir que la covariable debería transformarse, como x^2, $\sqrt{x}$ o $\log x$.

Residuos

En general se definen los siguientes residuos:

- Residuos de Respuesta: $Y_i - \widehat{\mu}_i$
- Residuos de Pearson: $\frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}$

  Tienen media 0 y varianza constante si la función de varianza es la correcta.
- Residuos de Trabajo: $(Y_i - \widehat{\mu}_i)\frac{\partial \eta_i}{\partial \mu_i}$
- Residuos de la Deviance: $sgn(Y_i - \widehat{\mu}_i)\sqrt{d_i}$

  Generalmente, más normales que los de Pearson, es decir menos asimétricos. Utiles para visualizar outliers.

Problemas con la función de varianza

Como en el modelo lineal el gráfico del valor absoluto de los residuos vs. $\widehat{\mu}$ puede ser útil para detectar problemas en la función de varianza.

Un gráfico sin ninguna tendencia indicaría una función de varianza correcta. En cambio, por ejemplo, una tendencia positiva sugeriría utilizar una función de varianza que aumente más rapidamente. Debemos tener en cuenta que dentro de una familia particular de distribuciones no es posible cambiar la función de varianza, sino que ésta está fijada por el modelo.

En el GLM la situación es muy parecida a la del Modelo Lineal: si la función de varianza no es la correcta el estimador de $\boldsymbol{\beta}$ será asintóticamente insesgado y normal, pero no eficiente. Así mismo, no estaremos estimando consistentemente la matriz de covarianza de $\widehat{\boldsymbol{\beta}}$.

Algunas estrategias para construir un modelo en regresión logística

Hosmer y Lemeshow (1989) sugieren algunas estrategias a la hora de ajustar un modelo de regresión logística. Enumeramos algunas de ellas:

- **Recomiendan comenzar por un análisis cuidadoso de cada variable a través de un ajuste univariado**. Para variables nominales, ordinales y continuas con muy pocos valores sugieren hacerlo a través de una tabla de contingencia para la respuesta ($y = 0, 1$) y los k valores de la variable independiente. Además de realizar un test de ajuste global (cociente de verosimilitud), para aquellas variables que exhiben un moderado nivel de asociación, proponen estimar los odds ratios usando uno de los niveles como referencia.
- **En este punto sugieren tener cuidado con aquellas tablas de contingencia que tienen alguna casilla con 0**. Una estrategia sencilla para evitar esto puede ser colapsar algunas categorías de la variable independiente de alguna manera razonable o eliminar la categoría completamente.

- **Cuando la variable es continua** puede hacerse un gráfico suavizado, dividiendo a la variable independiente en clases o intervalos. Podrían usarse los cuartos o cuartiles para dividir la variable continua.

- **Una vez realizado el análisis univariado seleccionan las variables para un análsis multivariado**. Recomiendan como candidato para la regresión multivariada a toda variable que en el test univariado tenga un p–valor < 0.25, así como a toda variable que se sepa es importante desde el punto de vista biológico (o del problema).

  Una vez que todas estas variables han sido identificadas, comienzan con un modelo multivariado que las contiene a todas.

  Este punto de corte 0.25 fue sugerido por Mickey and Greenland (1989). El uso de un punto tan grande (el usual es 0.05) tiene la desventaja de que pueden introducirse variables de dudosa importancia.

  Un problema de la aproximación por los modelos univariados es que variables que están en forma individual débilmente asociadas con la respuesta pueden

ser predictores importantes cuando se consideran en forma conjunta.

Por este motivo, debe revisarse la incorporación de todas las variables antes de arribar a un modelo final.

- La importancia de cada variable en el modelo multivariado puede ser evaluada a través del estadístico de Wald de cada una y una comparación del coeficiente estimado del modelo multivariado con el coeficiente estimado en el modelo univariado que sólo contiene esa variable.

  Hosmer y Lemeshow sugieren eliminar las variables que no contribuyen al modelo cuando nos basamos en estos criterios y ajustar un nuevo modelo. Proponen comparar los coeficientes estimados de las variables que quedan en el nuevo modelo con los estimados en el viejo modelo. En particular, sugieren preocuparnos por aquellas variables cuyos coeficientes cambian mucho en magnitud. Esto podría indicar que algunas de las variables eliminadas son importantes en el efecto de las variables restantes en el ajuste.

  Este procedimiento de eliminación, reajuste y verificación continúa hasta

que parezca que las variables importantes han sido incluidas y las excluidas son las biológica o estadísticamente sin importancia.

- En general, la decisión de comenzar con todas las variables posibles depende de la cantidad de observaciones. Cuando los datos no son adecuados para soportar este análisis, podría llegarse a resultados inestables: los estadísticos de Wald no serían adecuados para la selección de las variables. En este caso habría que refinar los resultados del análsis univariado y ver que es lo relevante desde el punto de vista científico.
- Un análisis alternativo puede ser utilizar un *método stepwise* en el que las variables son incluidas o excluidas secuencialmente de manera de poder identificar un modelo *full* y luego proceder como hemos descripto.
- Para las variable continuas deberemos chequear el supuesto de linealidad. Box–Tidwell (1962) sugieren incorporar un término de la forma $xln(x)$ y ver si su coeficiente es significativo o no. Un coeficiente significativo daría evidencias de no linealidad. Sin embargo, advierten sobre la falta de potencia del método para detectar pequeños apartamientos de la linealidad.

- Una vez que obtenemos un modelo que creemos que contiene las variables esenciales deberemos considerarar la necesidad de incorporar interacciones entre ellas. Sugieren incorporar la interacción y evaluar su significación en términos del cociente de verosimilitud. Ellos recomiendan no incorporar interacciones cuyo único efecto es agrandar los errores standard sin cambiar el valor estimado.

Observaciones Agrupadas en el caso Binomial

Como hemos visto cuando las variables son discretas puede haber replicaciones. Podemos encontrar que algunas de nuestras n observaciones toman el mismo valor en x_i. Si llamamos $x_1^*, \ldots, x_m^*$ a los valores distintos de las covariables (sin tener en cuenta las repeticiones), $m \leq n$, podemos comprimir los valores de las respuesta en

$$y_i^* = \sum_{j:x_j=x_i^*} y_j \quad n_i^* = \sum_{j:x_j=x_i^*} n_j \, .$$

Si los n_i^* son grandes podremos tener estadísticos de bondad de ajuste X^2 o

G^2 bien aproximados. Como ya observamos, estos estadísticos tendrán $m - p$ grados de libertad en lugar de $n - p$.

Si el modelo es cierto, al colapsar los valores con igual x_i no hay pérdida de información al sumar las Y_i's correspondiente. Sin embargo, si el modelo no es cierto, las π_i's de observaciones con igual x_i's no serán necesariamente idénticas y en ese caso no será necesariamente fácil detectar apartamientos al modelo.

El hecho de agrupar observaciones también puede limitar la posibilidad de detectar sobredispersión, que ocurre cuando las variables Y_i's tienen varianza mayor que $n_i\pi_i(1 - \pi_i)$.

La falta de ajuste del modelo se pude deber a:

- covariables omitidas
- función link incorrecta
- presencia de outliers
- sobredispersión