Intervalos de Confianza y Tests de Hipótesis

Dos de las herramientas más usada de la inferencia estadística son los intervalos de confianza y los tests de hipótesis.

Por ejemplo, los tests de hipótesis son necesarios para comparar el ajuste de dos modelos ajustados a los datos.

Tanto para realizar tests como intervalos de confianza necesitamos las distribuciones muestrales de los estadísticos involucrados.

Distribución Asintótica

Fahrmeir y Kaufmann (1985, Annals of Statistics, 13, 342–368) deducen la consistencia y la distribución asintótica de los estimadores de máxima verosimilitud en el GLM bajo condiciones de regularidad allí establecidas.

Sea $\mathcal{I}_n = \mathcal{I}_n(\boldsymbol{\beta}_0) = D'V^{-1}D$ donde

$$\begin{aligned} D_{ij} &= \frac{\partial \mu_i}{\partial \beta_j} \\ V &= Diag(V(\mu_i)) \end{aligned}$$

evaluadas en $\boldsymbol{\beta}_0$

Fahrmeir y Kaufmann (1985) probaron que si

- (D) (Diveregencia) $\lambda_{min}(\mathcal{I}_n) \to \infty$
- (C) (Cota inferior) Para todo $\delta > 0$

$$\mathcal{I}_n(\beta) - c\mathcal{I}_n \text{ es semidefinida positiva}$$

para todo $\beta \in N_n(\delta)$ si $n \geq n_1(\delta)$, donde $N_n(\delta)$ es un entorno de β_0 y c es independiente de δ.

- (N) (Convergencia y Continuidad) Para todo $\delta > 0$

$$\max_{\beta \in N_n(\delta)} \|V_n(\beta) - I\| \to 0$$

donde

$$V_n(\boldsymbol{\beta}) = \mathcal{I}_n^{-1/2} \mathcal{I}_n(\boldsymbol{\beta}) \mathcal{I}_n^{-1/2}$$

es una matriz de información normalizada.

Existencia y Consistencia

Entonces, bajo (C) y (D) exite el EMV $\widehat{\boldsymbol{\beta}}$ y además

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$$

Distribución Asintótica

Entonces, bajo (D) y (N)

$$(\mathcal{I}_n)^{1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, I)$$

En la práctica, usaremos como matriz de covarianza asintótica a $\mathcal{I}_n(\widehat{\boldsymbol{\beta}}_n)$

Esto nos servirá para deducir intervalos de confianza para los parámetros y para deducir tests tipo Wald en tanto

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)' \mathcal{I}_n(\widehat{\boldsymbol{\beta}}_n) (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \overset{(a)}{\sim} \chi^2_p \,.$$

Por lo que ya vimos, entonces para n es suficientemente grande

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \overset{(a)}{\sim} N(\mathbf{O}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}) \,.$$

Para n sufcientemente grande, una aproximación razonable esperamos que sea

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \overset{(a)}{\sim} N(\mathbf{O}, \widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)) \,,$$

siendo

$$\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n) = (\mathbf{X}'\mathbf{W}(\widehat{\beta}_n)\mathbf{X})^{-1} \,.$$

Si queremos computar un intervalo de confianza de nivel asintótico $1-\alpha$ para β_j, éste será:

$$\widehat{\beta}_{nj} \pm z_\alpha \, \widehat{\sigma}(\widehat{\beta}_{nj}) \, ,$$

con

$$\widehat{\sigma}(\widehat{\beta}_j) = \left[\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})_{jj}\right]^{1/2} .$$

Inferencia acerca de una función de los coeficientes

Para una función lineal de los prámetros $\Psi = \mathbf{a}'\boldsymbol{\beta}_0$, una aproximación razonable para n suficientemente grande es

$$(\mathbf{a}'\widehat{\boldsymbol{\beta}}_n - \mathbf{a}'\boldsymbol{\beta}_0) \overset{(a)}{\sim} N(\mathbf{O}, \mathbf{a}'\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)\mathbf{a}) \, .$$

Para una función no lineal $\Psi = r(\boldsymbol{\beta}_0)$, para n grande tendremos

$$r(\widehat{\boldsymbol{\beta}}_n) \overset{(a)}{\sim} N(r(\boldsymbol{\beta}_0), \nabla r(\widehat{\boldsymbol{\beta}}_n)'\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)\nabla r(\widehat{\boldsymbol{\beta}}_n)) \, .$$

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 parsimoia 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, entonces $\widehat{\mu}_i = g^{-1}(\widehat{\beta_0})$.

Si llamamos $\widehat{\boldsymbol{\theta}}_s$ al valor estimado bajo el modelo saturado, el $L(\widehat{\boldsymbol{\theta}}_s, \mathbf{y})$, likelihood evaluado en dicho estimador, tomará el valor más grande posible para estas observaciones, asumiendo la misma distribución subyacente y la misma función de enlace.

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.

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

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

Un enfoque alternativo es el de comparar la bondad del ajuste de los dos 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	0.1667	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 S–plus 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
 -0.3122076 0.282141 -0.291303 0.5080521 0.1210355

Coefficients:
              Value Std. Error  t value
(Intercept) 9.586802   3.703679 2.588454
   logdosis 2.879164   1.101315 2.614296

(Dispersion Parameter for Binomial family taken to be 1 )
```

```
Null Deviance: 28.009 on 4 degrees of freedom

Residual Deviance: 0.5347011 on 3 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:
         (Intercept)
logdosis 0.9820848

salida$deviance
[1] 0.5347011

 pchisq(salida$deviance,3)
[1] 0.0887958

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

Resumiendo

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

Deviance Residuals:

1	2	3	4	5
-0.3122076	0.282141	-0.291303	0.5080521	0.1210355

Coefficients:

	Value Std.	Error	t value
(Intercept)	9.586802	3.703679	2.588454
logdosis	2.879164	1.101315	2.614296

Null Deviance: 28.009 on 4 degrees of freedom

Residual Deviance: 0.5347011 on 3 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients: (Intercept)

logdosis 0.9820848

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 S–plus 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

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

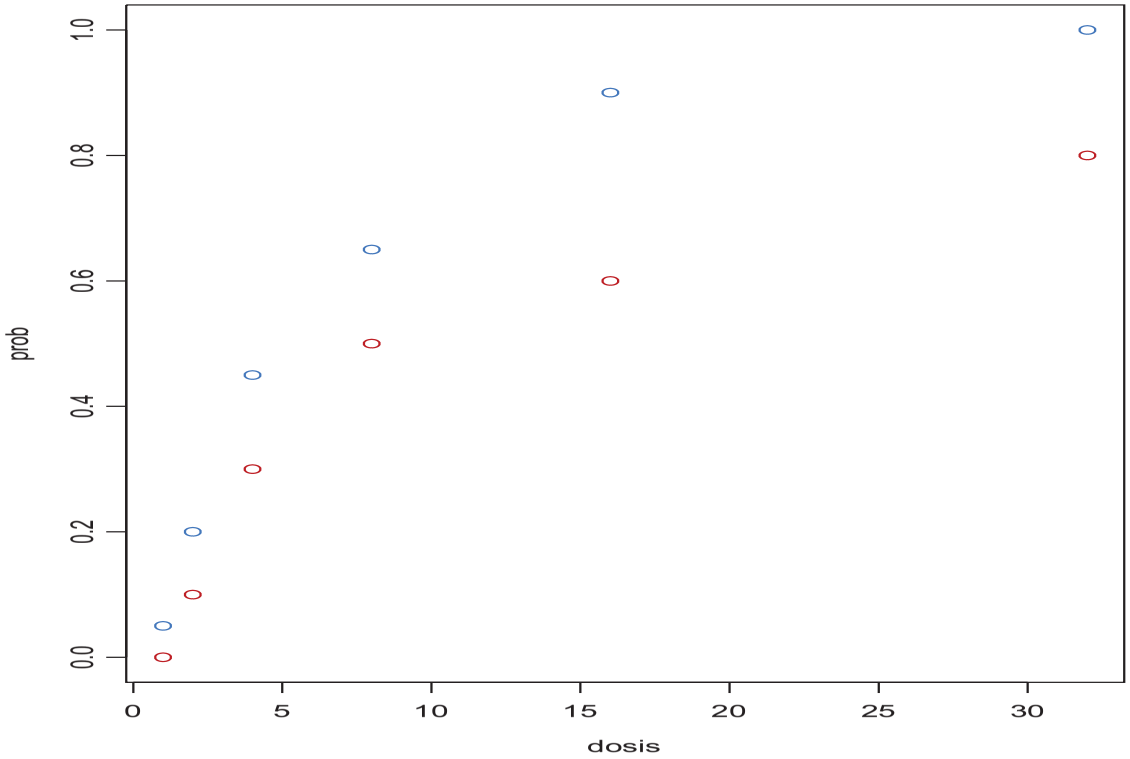

Figura 2: Gusanos del tabaco

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

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

de manera que cuando $sex = M$, para $ldose = 3$ tendríamos

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

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

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

Para ello hacemos

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

summary(salida.gusanos)
```

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

Coefficients:
                 Value Std. Error    t value
(Intercept) -2.9935414  0.5525295 -5.4178852
        sex  0.1749865  0.7781556  0.2248733
      ldose  0.9060363  0.1670577  5.4234939
  sex:ldose  0.3529131  0.2699444  1.3073547

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 124.8756 on 11 degrees of freedom

Residual Deviance: 4.993727 on 8 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

Aparentemente de la lectura de la tabla el efecto del sexo parece no signi-

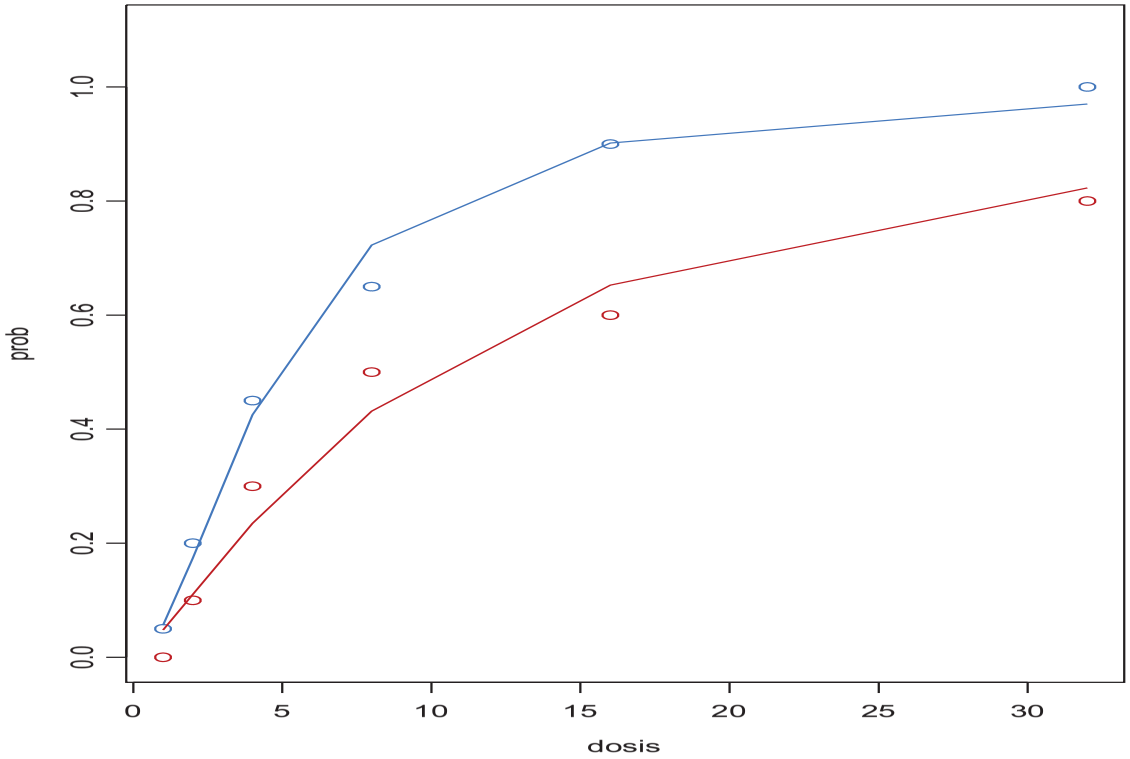

Figura 3: Gusanos del tabaco

ficativo, sin embargo debemos ser cuidadosos al interpretar esto. 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).

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

Coefficients:
                     Value Std. Error   t value
     (Intercept) -0.2754324  0.2304895 -1.194989
             sex  1.2337257  0.3769412  3.272992
    I(ldose - 3)  0.9060363  0.1670577  5.423494
sex:I(ldose - 3)  0.3529131  0.2699444  1.307355

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 124.8756 on 11 degrees of freedom

Residual Deviance: 4.993727 on 8 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

que muestra una diferencia significativa entre los dos sexos en la dosis 8.

Computamos el p–valor de la medida de ajuste global 1-pchisq(4.993727,8)= 0.7582464. Comparamos distintos modelos mediante la instrucción ANOVA

```
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 una pendiene para cada sexo:

```
salida3.gusanos<- glm(SF~sex+ldose-1, family=binomial)
summary(salida3.gusanos)

Coefficients:
          Value Std. Error   t value
 sexF -3.473154  0.4682939 -7.416612
 sexM -2.372411  0.3853875 -6.155911
ldose  1.064214  0.1310130  8.122959

    Null Deviance: 126.2269 on 12 degrees of freedom
Residual Deviance: 6.757064 on 9 degrees of freedom

Number of Fisher Scoring Iterations: 3
```

Otra forma sería

```
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
```

Matrices de cada uno:

```
 cbind(salida3.gusanos$x,salida4.gusanos$x)
   sexF sexM ldose (Intercept) sex ldose
 1    0    1     0           1   1     0
 2    0    1     1           1   1     1
 3    0    1     2           1   1     2
 4    0    1     3           1   1     3
 5    0    1     4           1   1     4
 6    0    1     5           1   1     5
 7    1    0     0           1   0     0
 8    1    0     1           1   0     1
 9    1    0     2           1   0     2
10    1    0     3           1   0     3
11    1    0     4           1   0     4
12    1    0     5           1   0     5
```

Interpretación de los coeficientes

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

$$\text{logit}(\pi) = \beta_0 + \beta_1 x$$

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
Presente	5	20	15	10	50
Ausente	20	10	10	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)
            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 $\theta(negro, blanco)$.

¿Qué podemos hacer cuando la variable es continua o discreta con muchos valores posibles?

El siguiente ejemplo corresponde al TP4 y se ha registrado la variable edad en forma discreta. Las variable independiente es **Age** y la dependiente **Low**. Primero consideraremos los cuartiles de la variable.

```
Analisis de cuartiles para Age:
> summary(age)
 Min. 1st Qu. Median  Mean    3rd Qu. Max.
 14      19     23   23.24      26   45

edad<- 1*(age<19)+2*(age>= 19 & age<23) +3*(age>= 23 & age<26)+ 4*(age>=26)
table(edad)
  1  2  3  4
 35 59 41 54
```